Machine Learning Approaches to Information Retrieval

Hang Li
Microsoft Research Asia

Joint Work with Jun Xu, Yunbo Cao, Guoping Hu
Talk Outline

- Introduction to Information Retrieval
- Search by Type
- Factoid Search
- New Learning Algorithm for Ranking
- Information Desk
- Summary
Introduction to Information Retrieval
What Is Information Retrieval?

I want to access information X
Why Important?
Current Approach

Information Retrieval = Ranking
General Model for Ranking

documents (information)

query (or question)

relevance scores for ranking

\[
d_1 \sim f(q, d_1) \\
d_2 \sim f(q, d_2) \\
\vdots \\
d_n \sim f(q, d_n)
\]
Relevance: Matching between Query and Document

\[ f(q,d) \]
Probabilistic Model

Let $d_1, d_2, \ldots, d_n$ be documents and $q$ be a query (or question). The relevant scores for ranking are

\[ d_1 \sim P(r \mid q, d_1) \]
\[ d_2 \sim P(r \mid q, d_2) \]
\[ \vdots \]
\[ d_n \sim P(r \mid q, d_n) \]

where $R \in \{r, \bar{r}\}$. The probabilistic model is

\[ P(R \mid Q, D) \]

with $R \in \{r, \bar{r}\}$.
Okapi or BM25
(Robertson and Walker 1994)

\[ \sum_{w \in d \cap q} \frac{(k + 1) tf(w)}{(1 - b)k + b \frac{dl}{avgdl} + tf(w)} \]
Language Mode
(Ponte and Croft 1998; Lafferty and Zhai, 2001)

document = bag of words

\[ d_1 = w_{11} w_{12} \cdots w_{1l_1} \]
\[ d_2 = w_{21} w_{22} \cdots w_{2l_2} \]
\[ \vdots \]
\[ d_n = w_{n1} w_{n2} \cdots w_{nl_n} \]

\[ q = w_{q1} w_{q2} \cdots w_{ql_q} \]

relevance scores for ranking

\[ d_1 \sim P(q \mid d_1) \]
\[ d_2 \sim P(q \mid d_2) \]
\[ \vdots \]
\[ d_n \sim P(q \mid d_n) \]
Classification Model

query (or question) $q$

documents $d_1, d_2, ..., d_n$

relevance scores for ranking $d_1 \sim P(r \mid q, d_1), d_2 \sim P(r \mid q, d_2), ..., d_n \sim P(r \mid q, d_n)$

$P(R \mid Q, D)$

$R \in \{r, \bar{r}\}$

Logistic Regression or SVM
Ordinal Regression Model
(Herbrich et al., 2000; Joachims, 2002)

documents

query (or question)

\( d_1 \)

\( d_2 \)

\( \vdots \)

\( d_n \)

\( R = f(Q, D) \)

\( R \in \{1, 2, \ldots k\} \)

relevance scores for ranking

\( d_1 \sim f(q, d_1) \)

\( d_2 \sim f(q, d_2) \)

\( \vdots \)

\( d_n \sim f(q, d_n) \)

Ranking SVM
General Model for Ranking (2)

query (or question) $q$

documents $d_1, d_2, \ldots, d_n$

scores for ranking

$$g(d_1) + f(q,d_1)$$

$$g(d_2) + f(q,d_2)$$

$$\vdots$$

$$g(d_n) + f(q,d_n)$$

$f(Q,D)$

$g(D)$
Page Rank
(Brin and Page, 1998)

\[ P(d_i) = \alpha \frac{1}{n} + (1 - \alpha) \sum_{d_j \in M(d_i)} \frac{P(d_j)}{L(d_j)} \]
Challenges

• Proximity
• Synonym and polysemy
• Quality and freshness of document (webpage)
• Spamming
• Information granularity
• Search need understanding
• Evaluation
• Personalization
• Training data collection (relevance feedback, click-through)
Search by Document Type
Search by Document Type

I want to find documents relevant to X in type Y
Probabilistic Model

query with type $q$

documents $d_1, d_2, \ldots, d_n$

ranking function

$P(R,T | Q,D) \approx P(R | Q,D)P(T | D)$

$P(R,T | Q,D)$

$R \in \{r, \bar{r}\}$

$T \in \{t, \bar{t}\}$
Example – Manuals Search

- **Query**
  “create an external link”

- **Document1**
  **External Links**
  External links are to Web sites external to your Web site. You can create external links by using buttons, highlighting text, or creating hot spots on images.

  For example creating a linkage to the Far Eastern Economic Review [FEER] one would go to the FEER through the Internet. You might want to right click on the FEER logo or recent issue cover and save the image to your computer [must be saved as a graphical interface file - GIF]. To make either the button or the text or the picture (or all three) interactive you highlight what you want to connect and then you click the hyperlink tool, or the Edit-Hyperlink. Press the World Wide Web tab in the Create Hyperlink dialogue box [see box below]. Whatever address you have in the running Net browser will show up in the box, or you can browse. In this case, the FEER address is there, one only need to press OK and the external hyperlink is created. All three examples to the right will link you to the FEER site.

- **Document2**
  **Creating an External Link**
  To create an anchor that is a link to another document:

  Select to select (by click and drag or by keyboard) the text for the link you are creating.

  Click the Link button (first case) or select the entry "Create or change link" in the Links menu (second case).

  **In the first case,** the cursor changes from an arrow to a hand to let you click the target document.
  - If the target document is displayed in another Amaya window, click anywhere within that window to create the link.
  - If the target document is not displayed in another Amaya window, press the F2 or Delete key, or click a part of the document which cannot be a valid target. A dialog prompts you for the location of the target document. Type the URI of the target document and then Confirm to create the link.

  **In the second case,** a dialog prompts you for the location of the target document.
  - If the target document is displayed in another Amaya window and you want to select it by click, click the Click button then click anywhere within that window to create the link.
  - If the target document is not displayed in another Amaya window, type the URI of the target document and then Confirm to create the link.
Manuals Search -- Experiment

• MS intranet data
  – 50 queries from log of Microsoft Web
  – 1.22 answers per query (from 5000 documents)

• Evaluation Measure

\[ RR_i = \frac{1}{\text{Rank}_i} \]

\[ MRR = \frac{\sum_{i=1}^{Q} RR_i}{Q} \]

• Results

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type Only</td>
<td>0.3651</td>
</tr>
<tr>
<td>Relevance Only</td>
<td>0.5688</td>
</tr>
<tr>
<td>Combined Model (our approach)</td>
<td><strong>0.7278</strong></td>
</tr>
</tbody>
</table>
Factoid Search
People Search
(Cao et al., 2005)

I want to find people related to X
Probabilistic Model

query with type $q$

documents

entities

$q_{deP} \sum_{e} P(e|q) \propto \sum_{d} P(d)P(q|d)P(e|d,q)$

ranking function

$P(E|Q,D)$
Example – People Search

• Query
  “Who knows about digital ink”

• Document
  Jian Wang, Ph.D.
  Senior Researcher & Research Manager
  Multimodal User Interface Group, Microsoft Research Asia
  Dr. Jian Wang is Research Manager of the Multimodal User Interface Group at Microsoft Research Asia (MSR Asia). Dr. Jian Wang's research specializations are ink and pen computing, usability, multimodal user interface, virtual reality and human cognition.
  The Multimodal UI Group's current research projects include: advanced digital ink parser, digital ink annotation and representation of digital ink for Tablet PC. The group previously invented an inline input and correction user interface for Asian languages called Modeless Input User Interface, which allows Chinese users of Office XP to smoothly enter English and Chinese text without constantly switching between input language modes.

• Answer
  Jian Wang
People Search -- Experiment

• MSR Corpus
  – 32 queries searching in 3109 documents
  – 810 Candidates

• Evaluation Measure

\[
\text{Top-5 Precision} = \frac{\# \{ \text{persons which appear in both Top-5 ranked candidates and ground truth} \}}{5}
\]

• Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Top-5 Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile-based Model</td>
<td>0.428</td>
</tr>
<tr>
<td>Our Model</td>
<td>0.563 (+31.5%)</td>
</tr>
</tbody>
</table>
I want to find time related to X
Probabilistic Model

passages

$\begin{align*}
    t_1 & \rightarrow e_1 \\
    t_2 & \rightarrow e_2 \\
    \vdots & \\
    t_n & \rightarrow e_m \\
\end{align*}$

entities

query with type

$q$

$P(E|Q,T)$

ranking function

$e \sim f(q,t) = \sum_i \lambda_i f_i(q,t)$
Time Search -- Example

• Query:
  Nixon visit China

• Document:
  Nixon's most significant achievement in foreign affairs may have been the
  establishment of direct relations with the People's Republic of China after a
  21-year estrangement. Following a series of low-level diplomatic contacts in
  1970 and the lifting of U.S. trade and travel restrictions the following year,
  the Chinese indicated that they would welcome high-level discussions, and
  Nixon sent his national security adviser, Henry Kissinger, to China for secret
  talks. The thaw in relations became apparent with the “ping-pong
  diplomacy” conducted by American and Chinese table-tennis teams in
  reciprocal visits in 1971–72. Nixon's visit to China in February–March 1972,
  the first by an American president while in office, concluded with the
  Shanghai Communiqué, in which the United States formally recognized the
  “one-China” principle—that there is only one China, and that Taiwan is a
  part of China.

• Answer
  February–March 1972
Time Search -- Experiment

• MS intranet data
  – 100 queries from log of http://msweb, containing ‘when’, ‘schedule’, ‘day’, and ‘time’
  – 8.73 answers per query (from 10000 documents)

• Evaluation Measure

\[ RR_i = \frac{1}{\text{Rank}_i} \]

\[ MRR = \frac{\sum_{i=1}^{Q} RR_i}{Q} \]

• Results

<table>
<thead>
<tr>
<th>Method</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Baseline</td>
<td>0.5033</td>
</tr>
<tr>
<td>Learning</td>
<td>0.5809 (+15%)</td>
</tr>
</tbody>
</table>
New Learning Algorithm for Ranking
Ranking Learning

- Given:
  - \( S = \{(\tilde{x}_i, y_i)\}_{i=1}^m \subset X \times Y \), where \( Y = \{r_1 \prec \cdots \prec r_q\} \)
  - \( H = \{h : X \mapsto Y\} \) (hypothesis space)
  - \( L : Y \times Y \mapsto R \) (loss function)

- Question: based on \( S \) find \( h^* = \arg\min_{h} \mathbb{E}(L(h(\tilde{x}), y)) \)

- ERM: choose \( h_{ERM}^* = \arg\min_{h} \sum_{i=1}^{m} L(h(\tilde{x}_i), y_i) \)

- SRM: choose \( h_{SRM}^* = \arg\min_{h} \sum_{i=1}^{m} L(h(\tilde{x}_i), y_i) + \lambda Q(h) \)
Viewing Ranking as Classification

• Formulizing ranking problem as classification of example pairs

• Measuring the loss of $h$ by inversions

$$L(y_1, y_2, \hat{y}_1, \hat{y}_2) = \begin{cases} 
1 & (y_1 < y_2) \land \neg (\hat{y}_1 < \hat{y}_2) \\
1 & (y_1 > y_2) \land \neg (\hat{y}_1 > \hat{y}_2) \\
0 & \text{otherwise}
\end{cases}$$
Viewing Ranking as Classification (cont’)

• Given $S = \{(\tilde{x}_i, y_i)\}_{i=1}^m$, find $h_{ERM}$ that minimizes

$$\sum_{i=1}^m \sum_{j=1}^m L(y_i, y_j, h(\tilde{x}_i), h(\tilde{x}_j))$$

• Equivalently, find a $f_h$ that minimizes

$$\sum_{((\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}), z_i) \in S'} L_{0-1}(f_h(\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}), z_i),$$

where $S' = \left\{ (\tilde{x}_i, \tilde{x}_j), z = \begin{cases} +1 & y_i \succ y_j \\ -1 & y_i \prec y_j \end{cases} : (\tilde{x}_i, y_i), (\tilde{x}_j, y_j) \in S \right\}$
Ranking Model

- Model: instances are ranked by $U(x)$
Simulation Experiment 1

Error rate of pairs

NDCG curve

Low penalty
High penalty
High penalty

Our approach
Ranking SVM
Information Desk
(Li et al., 2005)
Could you please take five minutes to fill in the survey form? The form>>>:

**What is...**
Definition of a technical term, group name, product name or code name. Expansion of an acronym.
- What is Longhorn?
- What is ATM?

**Who is...**
Aliases, title, department, oil and associated documents of a person.
- Who is Bill Gates?
- Who is Steve?

**Where is the homepage of...**
Homepage of a group or product.
- Where is the homepage of MSDN?
- Where is the homepage of Office?

**Who knows about...**
People who know about technical terms, group names or product names.
- Who knows about database?
- Who knows about Trustworthy Computing?

Looking for InfoDesk maintained by the sales group? Find it here.

Microsoft Confidential
Features -- ‘what is’

Definition of **Longhorn**

**Longhorn** is the codename for the next release of the Windows platform, which will further Microsoft’s long term vision...
Features – ‘who is’

Bill Gates
US-Executive-Chairman
+1 (425) XXXXXXX XXXXXX

Documents of Bill Gates(118)
- My advice to students: Education counts
  [File Link]
- Evento NET Reviewers – Seattle – 7/8 Novembro
  [File Link]
- A Vision for Life Long Learning – Year 2020
  [File Link]
- Bill Gates answers most frequently asked questions.
  [File Link]

Top 10 terms appearing in documents of Bill Gates

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>984.4443</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>816.4247</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>595.0771</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>578.5604</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>565.7299</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>435.5366</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>412.4467</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>385.446</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>346.5993</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>345.3285</td>
</tr>
</tbody>
</table>

>>more
Features – ‘where is homepage of ’

Homepages of Office

Office Portal Site
This is the internal site for Office
Features - ‘who knows about’

Data Mining

What is  Who is  Where is homepage of

What is  Who is  Where is homepage of

People Associated with **Data mining**
Jamie MacLennan

US-SQL Data Warehouse
+1 (425) XXXXXXX XXXXXXX

Associated documents(4):
Summary
Summary

• Information retrieval = helping people access information
• Currently search = ranking
• Matching between query and document
• Our work
  – Search by Document Type, Factoid Search
  – New learning algorithm for ranking
• Many issues to study
References


Hang Li, Yunbo Cao, Jun Xu, Yuhua Hu, Shenjie Li, and Dmitriy Meyerzon, A New Approach to Intranet Search Based on Information Extraction. Proc. of ACM-CIKM'05 industry track.


Yunbo Cao, Jingjing Liu, Shenghua Bao, and Hang Li (2005), Microsoft Research Asia (MSRA) at Enterprise Track of TREC 2005: Expert Search. In Proceedings of TREC.