Statistical Learning Approach to Matching of Query and Document in Search

Hang Li
Microsoft Research Asia
Talk Outline

• Background
  1. Key Problem in Search = Matching between Query and Document
  2. IR Models (BM25) = Matching Functions
  3. Mismatch Problem

• Problems and Our Solutions
  1. Generic IR Model as Asymmetric Kernel
  2. Extension of IR Models as Asymmetric Kernels
  3. Learning of Robust IR Models as Asymmetric Kernels
  4. Learning of Matching Function
How Search Works
Search Architecture

User Interface → Ranker → Index

Query understanding → Retrieval, Matching, Ranking

Search result presentation

Document understanding

Crawler → Indexer

Web
Three Important Processes

• Retrieval
  – Finding documents from inverted index

• Matching
  – Calculating relevance score between query and document pair

• Ranking
  – Ranking documents based on not only relevance scores but also importance scores, etc
Key Factor for Search: Matching between Query and Document

$f(q, d)$

$q$

$q'$

$q$

$d$
Matching between Heterogeneous Data is Everywhere

- Matching between languages (translation)
- Matching between text and image (image annotation)
- Matching between people (dating)
- Matching between user and item (collaborative filtering)
IR Models
Vector Space Model

query (or question) $q$

relevant scores for ranking

$$d_1 \sim \cos(q, d_1)$$
$$d_2 \sim \cos(q, d_2)$$
$$\vdots$$
$$d_n \sim \cos(q, d_n)$$

documents

$\begin{align*}
d_1 \\
d_2 \\
\vdots \\
d_n
\end{align*}$
Probabilistic Model

documents

query (or question)

relevant scores for ranking

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

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

\[ \vdots \]

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

\[ r \in \{1,0\} \]
Okapi or BM25
(Robertson and Walker 1994)

query (or question) $q$

$\sum_{w \in d \cap q} \frac{(k + 1)tf(w)}{(1 - b)k + b \frac{dl}{avgdl}} + tf(w)$

documents $d_1, d_2, \ldots, d_n$
Language Mode
(Ponte and Croft 1998)

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

\[ d_1 \sim P(q \mid d_1) \]
\[ d_2 \sim P(q \mid d_2) \]
\[ \vdots \]
\[ d_n \sim P(q \mid d_n) \]
Term Mismatch = Main Challenge in Search
Examples of Term Mismatch

- Query $\leftrightarrow$ document
- pool schedule $\leftrightarrow$ swimming pool schedule
- seattle best hotel $\leftrightarrow$ seattle best hotels
- natural logarithm transformation $\leftrightarrow$ logarithm tranformation
- china kong $\leftrightarrow$ china hong kong
- why are windows so expensive $\leftrightarrow$ why are macs so expensive
Different Queries Can Represent Same Intent

“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 between earth sun
- distance between sun and earth
- distance between the earth and sun
- 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 sun to earth
- distance from sun to earth
- distance from the earth to the sun
- distance from the earth to the sun
- distance from the earth to earth
- distance from sun to the earth
- distance from the sun to earth
- distance from the sun to 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 sun from earth
- distance of sun from the earth
- distance of sun to earth
- distance of sun to the earth
- distance of the earth from earth
- distance of the earth from the earth
- distance of the earth to earth
- distance of the earth to the earth
- 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 sun and earth
- distance sun earth
- distance to earth from sun
- distance to earth from the sun
- distance to sun from earth
- distance to sun from 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 the sun
- how far is earth from earth
- how far is earth from earth
- how far is earth from the sun
- how far is earth to earth
- how far is earth to sun
- how far is earth to the sun
- how far is it from earth to earth
- how far is it from earth to the earth
- how far is earth from earth
- how far is earth from the sun
- how far is earth to earth
- how far is earth to the sun
- how far is it from earth to earth
- how far is it from earth to the earth
- how far is earth from earth
- how far is earth from the sun
- how far is earth to earth
- how far is earth to the sun
- how far is the earth from earth
- how far is the earth from earth
- how far is the earth from the sun
- how far is the earth to earth
- how far is the earth to sun
- how far is the earth to the sun
- how far is the earth away from the earth
- how far is the earth away from the earth
- how far is the earth from earth
- how far is the earth from earth
- how far is the earth from the sun
- how far is the earth to earth
- how far is the earth to sun
- how far is the earth to the sun
- how far is the sun from 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 earth
- how far is the sun to the earth
- how far is the sun away from earth
- 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
- how far is the sun to earth
Different Queries Can Represent Same Intent
“Youtube”

<table>
<thead>
<tr>
<th>Query</th>
<th>Query</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>yutube</td>
<td>yuotube</td>
<td>yuo tube</td>
</tr>
<tr>
<td>ytube</td>
<td>youtubr</td>
<td>yu tube</td>
</tr>
<tr>
<td>youtubo</td>
<td>youtuber</td>
<td>youtubecom</td>
</tr>
<tr>
<td>youtube om</td>
<td>youtube music videos</td>
<td>youtube videos</td>
</tr>
<tr>
<td>youtube</td>
<td>youtube com</td>
<td>youtube co</td>
</tr>
<tr>
<td>youtub com</td>
<td>you tube music videos</td>
<td>youtube</td>
</tr>
<tr>
<td>youtub</td>
<td>you tube com yourtube</td>
<td>your tube</td>
</tr>
<tr>
<td>you tube</td>
<td>you tub</td>
<td>you tube video clips</td>
</tr>
<tr>
<td>you tube videos</td>
<td>www you tube com</td>
<td>wwww youtube com</td>
</tr>
<tr>
<td>www youtube</td>
<td>www youtube com</td>
<td>www youtube co</td>
</tr>
<tr>
<td>yotube</td>
<td>www you tube</td>
<td>www utube com</td>
</tr>
<tr>
<td>ww youtube com</td>
<td>www utube</td>
<td>www u tube</td>
</tr>
<tr>
<td>utube videos</td>
<td>utube com</td>
<td>utube</td>
</tr>
<tr>
<td>u tube com</td>
<td>utub</td>
<td>u tube videos</td>
</tr>
<tr>
<td>u tube</td>
<td>my tube</td>
<td>toutube</td>
</tr>
<tr>
<td>outube</td>
<td>our tube</td>
<td>toutube</td>
</tr>
</tbody>
</table>
Matching between Two Worlds
Problems to be Addressed
Problems to be Addressed

1. Is there unified and general framework for IR models (matching models)?
2. How to make extensions of IR models
3. How to make the IR models robust (deal with mismatch) by learning?
4. Is it possible to directly learn a matching function given training data?
Similarity Learning for Information Retrieval

Joint work with Wei Wu and Jun Xu
1. Generic IR Model as Asymmetric Kernel
Asymmetric Kernel

- Kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
  - Definition: $k(x, x') = \langle \phi(x), \phi(x') \rangle$, where $\phi : \mathcal{X} \rightarrow \mathcal{H}$
  - Given $k_1$ and $k_2$ are kernels, create new kernels:

\[
\alpha k_1 + k_2, \quad k_1 \cdot k_2;
\]

- Kernel $k : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$
  - Definition: $k(x, y) = \langle \phi(x), \phi'(y) \rangle$, where $\phi : \mathcal{X} \rightarrow \mathcal{H}$
  - $\mathcal{H}$

$k_2$ are asymmetric kernels, create new kernels:

$\alpha k$, where $\alpha \in \mathbb{R}; k_1 + k_2; k_1 \cdot k_2$
Kernel vs Asymmetric Kernel

\[ k(x, x') = \langle \phi(x), \phi(x') \rangle \]

\[ k(x, y) = \langle \phi(x), \phi'(y) \rangle \]

Hilbert Space \( H \)

Space \( X \)

Space \( Y \)
IR Models are Asymmetric Kernels

- **VSM**
  - $BM25(q, d) = \langle \phi_{Q}^{VSM}(q), \phi_{D}^{VSM}(d) \rangle$, for all $w \in V$
    \[
    \phi_{Q}^{VSM}(q)_w = tfidf(w, q) \quad \text{and} \quad \phi_{D}^{VSM}(d)_w = tfidf(w, d)
    \]

- **BM25**
  - $BM25(q, d) = \langle \phi_{Q}^{BM25}(q), \phi_{D}^{BM25}(d) \rangle$, for all $w \in V$
    \[
    \phi_{Q}^{BM25}(q)_w = \frac{(k_3+1) \times tf(w,q)}{k_3+tf(w,q)}
    \]
    \[
    \phi_{D}^{BM25}(d)_w = \text{IDF}(w) \cdot \frac{(k_1+1) \times tf(w,d)}{k_1(1-b+b \cdot \frac{\text{len}(d)}{\text{avgDocLen}})+tf(w,d)}
    \]

- **LMIR**
  - $LMIR(q, d) = \langle \phi_{Q}^{LMIR}(q), \phi_{D}^{LMIR}(d) \rangle + \text{len}(q) \cdot \log \frac{\mu}{\text{len}(d)+\mu}$, for all $w \in V$
    \[
    \phi_{Q}^{LMIR}(q)_w = tf(w, q)
    \]
    \[
    \phi_{D}^{LMIR}(d)_w = \log \left(1 + \frac{tf(w,d)}{\mu \cdot P(w)}\right), \text{where } P(w) \text{ plays similar role as IDF in BM25}
    \]
IR Models as Asymmetric Kernels

Query Space

Document Space

New Space
Open Questions on Asymmetric Kernels

• Sufficient and necessary condition
• Counterpart of Mercer’s theorem
• What function class should work for matching in search
2. Extension of IR Models as Asymmetric Kernels
Relevance beyond Unigram

- Machine learning book
  - Unigram
    - Bigram
      - 2-dependent-terms
        - Co-occur (machine, learning)
          - Co-occur (machine, book)
        - Co-occur (learning, book)
      - Co-occur (learning, book)
    - Machine learning
      - Learning book
        - Co-occur (machine, learning)
          - Co-occur (machine, book)
Extension of IR models

• **BM25**
  - \( BM25(q, d) = \langle \phi_Q^{BM25}(q), \phi_D^{BM25}(d) \rangle \), and for all \( w \in V \)
    \[
    \phi_Q^{BM25}(q)_w = \frac{(k_3+1) \times tf(w, q)}{k_3 + tf(w, q)} \\
    \phi_D^{BM25}(d)_w = IDF(t) \cdot \frac{(k_1+1) \times tf(w, d)}{k_1 \left(1-b+b \frac{len(d)}{avgDocLen} \right) + tf(w, d)}
    \]

• **BM25_Kernel**
  - \( BM25\_Kernel(q, d) = \sum_t BM25\_Kernel_t(q, d) \) where \( t \) is dependence type
  - \( BM25\_Kernel_t(q, d) = \langle \phi_Q^{BM25}(q), \phi_D^{BM25}(d) \rangle \), and for all \( x \in V_t \)
    \[
    \phi_{Q,t}^{BM25}(q)_x = \frac{(k_3+1) \times f_t(x, q)}{k_3 + f_t(x, q)} \\
    \phi_{D,t}^{BM25}(d)_x = IDF_t(x) \cdot \frac{(k_1+1) \times f_t(x, d)}{k_1 \left(1-b+b \frac{f_t(d)}{avgDocLen_t} \right) + f_t(x, d)}
    \]
Experimental Results

Ranking accuracies on web search data

Kernel model with different term dependences
Experimental Results

Ranking accuracies on OHSUMED

- MAP
- NDCG@1
- NDCG@3
- NDCG@5

**Evaluation Metrics**
- MAP
- NDCG@1
- NDCG@3
- NDCG@5

**Models**
- LMIR
- BM25
- MRF
- LMIR Kernel
- BM25 Kernel

Ranking accuracies on AP

**Evaluation Metrics**
- MAP
- NDCG@1
- NDCG@3
- NDCG@5

**Models**
- LMIR
- BM25
- MRF
- LMIR Kernel
- BM25 Kernel
3. Learning of Robust IR Models as Asymmetric Kernels
Matching = Subset to Subset Matching

Query Space
Semantically related queries are close to each other

Document Space
Semantically related documents are close to each other

Matching between two semantic spaces

Microsoft Stock Price
MSFT Price
MSFT Quote

http://moneycentral.msn.com/
http://finance.yahoo.com/
http://money.cnn.com/
Asymmetric Kernel Learning

- Asymmetric Kernel: $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$
- Input
  - Training data $S = \{(x_i, y_i), t_i\}_{1 \leq i \leq N}$
- Output
  - Asymmetric kernel function
- Optimization

$$\min_{k \in \mathcal{K} \subseteq \mathcal{A}} \frac{1}{N} \sum_{i=1}^{N} l(k(x_i, y_i), t_i) + \Omega(k)$$
Asymmetric Kernel Learning Using Kernel Methods

- **Assumption**
  - Space of asymmetric kernels is RKHS generated by positive-definite kernel $\bar{k}: (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$
  - Hyper Asymmetric Kernel (HAK)

- **Optimization**

$$
\min_{k \in \mathcal{K}} \frac{1}{N} \sum_{i=1}^{N} l(k(x_i, y_i), t_i) + \frac{\lambda}{2} \|k\|_{\mathcal{K}}^2
$$

- **Solution**
  - By representer theorem
    $$
k^*(x, y) = \sum_{i=1}^{N} \alpha_i \bar{k}((x_i, y_i), (x, y))
$$

- **Hyper Asymmetric Kernel**

$$
\bar{k}((x, y), (x', y')) = g(x, y) k_X(x, x') k_Y(y, y') g(x', y')
$$
Learning Robust BM25

- BM25 = asymmetric kernel $k_{BM25}(q, d)$
- HAK
  \[ \bar{k}((q, d), (q', d')) = k_{BM25}(q, d)k_Q(q, q')k_D(d, d')k_{BM25}(q', d') \]
- Solution (called Robust BM25)
  \[ k_{RBM25}(q, d) = k_{BM25}(q, d) \cdot \sum_{i=1}^{N} \alpha_i k_Q(q, q_i)k_D(d, d_i)k_{BM25}(q_i, d_i) \]
- Deal with term mismatch
Mapping to Space of Query Document Pairs - KNN in New Space

Query-document pair space

Query space

Hilbert space $k_Q(q, q')$

Document space

Hilbert space $k_D(d, d')$

Asymmetric kernels

Matching

Hilbert space $k_{IR}(q, d)$

Hilbert space $k_{IR}(q', d')$

Microsoft Confidential
## Experimental Results

Table 5: Ranking accuracies on web search and enterprise search data.

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>NDCG@1</th>
<th>NDCG@3</th>
<th>NDCG@5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web search</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robust BM25</td>
<td>0.1192</td>
<td>0.2480</td>
<td>0.2587</td>
<td>0.2716</td>
</tr>
<tr>
<td>Pairwise Kernel</td>
<td>0.1123</td>
<td>0.2241</td>
<td>0.2418</td>
<td>0.2560</td>
</tr>
<tr>
<td>Query Expansion</td>
<td>0.0963</td>
<td>0.1797</td>
<td>0.2061</td>
<td>0.2237</td>
</tr>
<tr>
<td>BM25</td>
<td>0.0908</td>
<td>0.1728</td>
<td>0.2019</td>
<td>0.2180</td>
</tr>
<tr>
<td><strong>Enterprise search</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robust BM25</td>
<td>0.3122</td>
<td>0.4780</td>
<td>0.5065</td>
<td>0.5295</td>
</tr>
<tr>
<td>Pairwise Kernel</td>
<td>0.2766</td>
<td>0.4465</td>
<td>0.4769</td>
<td>0.4971</td>
</tr>
<tr>
<td>Query Expansion</td>
<td>0.2755</td>
<td>0.4076</td>
<td>0.4712</td>
<td>0.4958</td>
</tr>
<tr>
<td>BM25</td>
<td>0.2745</td>
<td>0.4246</td>
<td>0.4531</td>
<td>0.4741</td>
</tr>
</tbody>
</table>
Theorem 4 (Power Series Construction) Given two Mercer kernels \( k_X : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) and \( k_Y : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \), for any asymmetric kernel \( g(x, y) \) and \( \{c_i\}_{i=1}^{n} \subset \mathbb{R}^+ \), \( \bar{k}_p \) defined below is a hyper asymmetric kernel.

\[
\bar{k}_p((x, y), (x', y')) = \sum_{i=0}^{\infty} c_i \cdot g(x, y) (k_X(x, x')k_Y(y, y'))^i g(x', y'), \quad (7)
\]

where the convergence radius of \( \sum_{i=0}^{\infty} c_i \xi^i \) is \( R \), \( |k_X(x, x')| < \sqrt{R} \), \( |k_Y(y, y')| < \sqrt{R} \), for any \( x, x', y, y' \).

Theorem 5 (Multiple Kernel Construction) Given two finite sets of Mercer kernels \( K_X = \left\{ k_X^i(x, x') \right\}_{i=1}^{n} \) and \( K_Y = \left\{ k_Y^i(y, y') \right\}_{i=1}^{n} \). For any asymmetric kernel \( g(x, y) \) and \( \{c_i\}_{i=1}^{n} \subset \mathbb{R}^+ \), \( \bar{k}_M \) defined below is a hyper asymmetric kernel.

\[
\bar{k}_M((x, y), (x', y')) = \sum_{i=1}^{n} c_i \cdot g(x, y)k_X^i(x, x')k_Y^i(y, y')g(x', y'). \quad (8)
\]
4. Learning of Matching Function
Learning Similarity Function between Objects in Two Spaces

Query Space

Document Space

New Space

$q_1, q_2, \ldots, q_m$

$d_1, d_2, \ldots, d_n$

$L_1, L_2$
Keywords and Images Represented in Same Space
**Problem Formulation**

- **Setting**
  - Two spaces: \( \mathcal{X} \subset \mathbb{R}^m \) and \( \mathcal{Y} \subset \mathbb{R}^n \).

- **Input**
  - Training data: \( \{(x_i, y_i, t_i)\}_{1 \leq i \leq N} \)

- **Output**
  - Similarity function \( f(x, y) \)

- **Assumption**
  - Two linear (and orthonormal) transformations \( L_x \) and \( L_y \)
  - Dot product as similarity function \( \langle L_x^T x, L_y^T y \rangle = x^T L_x \ L_y^T y \)

- **Optimization**

\[
\arg\max_{L_X, L_Y} \sum_{r_i = +1} x_i^T L_X \ L_y^T \ y_i - \sum_{r_i = -1} x_i^T L_X \ L_y^T \ y_i
\]

subject to \( L_x^T L_X = I_{k \times k}, L_y^T L_y = I_{k \times k} \)
Our Solution

• Non-convex optimization
• Can prove that global optimal solution exists
• Global optimal can be found by solving SVD (Singular Value Decomposition)
• SVD of Matrix $M_S - M_D = U \Sigma V^T$

• Algorithm

1: Input: training data $\{(x_i, y_i, r_i)\}_{i=1}^N$, parameter $k \leq \min(n, m)$.
2: Calculate $M_S$ and $M_D$ through $\sum_{i=1}^N y_i x_i^T$ and $\sum_{i=-1}^{r_i} y_i x_i^T$, respectively.
3: Calculate SVD of $M_S - M_D$.
4: Choose left and right singular vectors $(u_1, \cdots, u_k)$ and $(v_1, \cdots, v_k)$ w.r.t the top $k$ singular values.
5: Output: $L_y = (u_1, \cdots, u_k)$ and $L_x = (v_1, \cdots, v_k)$. 
Talk Outline

• Background
  1. Key Problem in Search = Matching between Query and Document
  2. IR Models (BM25) = Matching Functions
  3. Mismatch Problem

• Problems and Our Solutions
  1. Generic IR Model as Asymmetric Kernel
  2. Extension of IR Models as Asymmetric Kernels
  3. Learning of Robust IR Models as Asymmetric Kernels
  4. Learning of Matching Function
Thank You!