Pseudo-relevance Feedback & Query Models

Kuan-Yu Chen (陳冠宇)

2018/11/09 @ TR-514, NTUST
<table>
<thead>
<tr>
<th>#</th>
<th>Δpub</th>
<th>Team Name</th>
<th>Kernel</th>
<th>Team Members</th>
<th>Score</th>
<th>Entries</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>M10715004</td>
<td></td>
<td></td>
<td>0.50982</td>
<td>42</td>
<td>2d</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>M10715052</td>
<td></td>
<td></td>
<td>0.48138</td>
<td>9</td>
<td>5h</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>M10715062</td>
<td></td>
<td></td>
<td>0.47324</td>
<td>42</td>
<td>1d</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>B10415018_沈政一</td>
<td></td>
<td></td>
<td>0.47312</td>
<td>50</td>
<td>21h</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>B10415004_楊晉復</td>
<td></td>
<td></td>
<td>0.45749</td>
<td>48</td>
<td>13h</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>B10415045_施弘仰</td>
<td></td>
<td></td>
<td>0.44680</td>
<td>14</td>
<td>16h</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>B10332015_羅子原</td>
<td></td>
<td></td>
<td>0.44400</td>
<td>8</td>
<td>2h</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>M10615110</td>
<td></td>
<td></td>
<td>0.43939</td>
<td>13</td>
<td>2h</td>
</tr>
<tr>
<td>9</td>
<td>↑1</td>
<td>M10715025_廖傑明</td>
<td></td>
<td></td>
<td>0.42423</td>
<td>27</td>
<td>1d</td>
</tr>
<tr>
<td>10</td>
<td>↑1</td>
<td>M10715010</td>
<td></td>
<td></td>
<td>0.42025</td>
<td>58</td>
<td>2d</td>
</tr>
</tbody>
</table>
# Progress

<table>
<thead>
<tr>
<th>Date</th>
<th>Syllabus</th>
<th>Homework</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/14</td>
<td>Course Overview</td>
<td></td>
</tr>
<tr>
<td>9/21</td>
<td>Classic Models</td>
<td>Homework-1</td>
</tr>
<tr>
<td>9/28</td>
<td>Extended Probabilistic Models</td>
<td></td>
</tr>
<tr>
<td>10/5</td>
<td>Break @ Rocling 2018</td>
<td></td>
</tr>
<tr>
<td>10/12</td>
<td>Evaluation &amp; Benchmark Collections</td>
<td>Homework-2 &amp; HW2 Description</td>
</tr>
<tr>
<td>10/19</td>
<td>Latent Semantic Analysis and Topic Models</td>
<td>Homework-3</td>
</tr>
<tr>
<td>10/26</td>
<td>Search Results Diversification</td>
<td></td>
</tr>
<tr>
<td>11/2</td>
<td>Midterm Exam</td>
<td></td>
</tr>
<tr>
<td>11/9</td>
<td>Pseudo-Relevance Feedback &amp; Query Models</td>
<td>Homework-4</td>
</tr>
<tr>
<td>11/16</td>
<td>Invited Talk</td>
<td></td>
</tr>
<tr>
<td>11/23</td>
<td>Introduction to Deep Learning</td>
<td>Submit Your Member List and Paper Title!</td>
</tr>
<tr>
<td>11/30</td>
<td>Representation Learning for Information Retrieval</td>
<td>Homework-5</td>
</tr>
<tr>
<td>12/7</td>
<td>Supervised Retrieval Models</td>
<td></td>
</tr>
<tr>
<td>12/14</td>
<td>Presentations</td>
<td></td>
</tr>
<tr>
<td>12/21</td>
<td>Presentations</td>
<td></td>
</tr>
<tr>
<td>12/28</td>
<td>Presentations</td>
<td></td>
</tr>
<tr>
<td>1/4</td>
<td>Competition</td>
<td></td>
</tr>
</tbody>
</table>
Review

• Topic Models
  – PLSA
  – LDA

• Search Results Diversification
  – MMR
  – SMM
  – xMMR
  – WUME
  – xQuAD

• Clarity
Introduction

- An information need can be defined as **the reason** for which the user turns to a search engine

- Each query usually consists of **only a few words**, the corresponding representation might not be appropriately estimated
  - Several effective formulations to enhance the query representation by **pseudo-relevance feedback** process

A General Flowchart of PRF

- “Pseudo” means that we assume top-ranked documents are relevant documents
Research Issues

- The main issues in pseudo-relevance feedback
  - How to select relevant documents from the top-retrieved documents
  - How to select expansion terms
The Rocchio Algorithm – 1

- Rocchio’s relevance feedback model is a classic query expansion method and it has been shown to be effective in boosting information retrieval performance
  - It is a way of incorporating pseudo relevance feedback information into the vector space model

\[ \tilde{q}' = \alpha \cdot \tilde{q} + \beta \cdot \frac{1}{|R_q|} \cdot \left( \sum_{d_j \in R_q} \tilde{d}_j \right) - \gamma \cdot \frac{1}{|\bar{R}_q|} \cdot \left( \sum_{d_{j'} \in \bar{R}_q} \tilde{d}_{j'} \right) \]

- \( R_q \) be the set of relevant documents to a given query \( q \)
- \( \bar{R}_q \) be the set of non-relevant documents to query \( q \)
- Each word is represented by the TFIDF score
The Rocchio Algorithm – 2

- Starting from the original query $\tilde{q}$, the new query moves you some distance toward the centroid of the relevant documents and some distance away from the centroid of the non-relevant documents.

- A simplified variant is to consider the positive feedback documents only:

$$\tilde{q}' = \alpha \cdot \tilde{q} + \beta \cdot \frac{1}{|R_q|} \cdot \left( \sum_{d_j \in R_q} \vec{d}_j \right)$$
KL-Divergence Measure

- Another basic formulation of LM for IR is the Kullback-Leibler (KL)-Divergence measure

\[
KL(q||d_j) = \sum_{w \in V} P(w|q) \log \frac{P(w|q)}{P(w|d_j)} \propto -\sum_{w \in V} P(w|q) \log P(w|d_j)
\]

- A query is treated as a **probabilistic model** rather than simply an **observation**
- KL-divergence supports us to achieve a better result by considering both query and document models
Relevance Model

- The relevance modeling (RM) is a well-practiced approach
  - Each query is assumed to be associated with a concept $R$ (or relevance class/information need)
    - Both the query and relevant documents are drawn from the concept $R$
  - The RM model assumes that words $w$ that co-occur with the query in the concept will have higher probabilities

\[
P_{RM}(w) \equiv \frac{P(w, q|R)}{\sum_{w' \in V} P(w', q|R)} \approx \frac{\sum_{d_j \in R_q} P(d_j) P(w, q|d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} P(d'_j) P(w', q|d'_j)}
\]

\[
= \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) \prod_{i=1}^{q} P(w_i|d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} P(d'_j) P(w'|d'_j) \prod_{i=1}^{q} P(w_i|d'_j)}
\]
Simple Mixture Model – 1

• An alternative formulation to extract relevance cues is simple mixture model (SMM)
  – It assumes that words in the set of pseudo-relevance feedback documents are drawn from two-component mixture model:
    • One component is the query model
    • The other is a background model

• The SMM model $P_{SMR}(w)$ is estimated by maximizing the log-likelihood of the set of top-ranked documents $R_q$ expressed as follows:

$$\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left( (1 - \alpha) \cdot P_{SMR}(w) + \alpha \cdot P(w|BG) \right)^{c(w,d_j)}$$
Simple Mixture Model – 2

- Estimate the parameters
  - E-step

\[
P(T_{SM} \mid w) = \frac{(1 - \alpha) \cdot P_{SM}(w)}{(1 - \alpha) \cdot P_{SM}(w) + \alpha \cdot P(w \mid BG)}
\]

- M-step

\[
P_{SM}(w) = \frac{\sum_{d_j \in R_q} c(w, d_j) P(T_{SM} \mid w)}{\sum_{w' \in V} \sum_{d_{j'} \in R_q} c(w', d_{j'}) P(T_{SM} \mid w')}
\]

\[
\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} ((1 - \alpha) \cdot P_{SM}(w) + \alpha \cdot P(w \mid BG))^{c(w, d_j)}
\]

\[
= \prod_{d_j \in R_q} \prod_{w \in V} (P_{SM}(w \mid T_{SM}) P(T_{SM}) + P(w \mid BG) P(BG))^{c(w, d_j)}
\]
The TriMM model $P_{TM_M}(w)$ is estimated by maximizing the log-likelihood of the set of top-ranked documents.

- It assumes that words in the set of pseudo-relevance feedback documents are drawn from three-component mixture model:
  - One component is the query model
  - Another component is the document-specific model
  - The other is a background model

$$\mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left( (1 - \alpha - \beta) \cdot P_{TM_M}(w) + \alpha \cdot P(w|d_j) + \beta \cdot P(w|BG) \right)^{c(w,d_j)}$$
Tri-Mixture Model – 2

- Estimate the parameters
  - E-step
    \[
    P(T_{TMM} \mid w, d_j) = \frac{(1 - \alpha - \beta) \cdot P_{TMM}(w)}{(1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w \mid d_j) + \beta \cdot P(w \mid BG)}
    \]
    \[
    P(T_{d_j} \mid w, d_j) = \frac{\alpha \cdot P(w \mid d_j)}{(1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w \mid d_j) + \beta \cdot P(w \mid BG)}
    \]
  - M-step
    \[
    P_{TMM}(w) = \frac{\sum_{d_j \in R_q} c(w, d_j)P(T_{TMM} \mid w, d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} c(w', d_j')P(T_{TMM} \mid w', d_j')} \]
    \[
    P(w \mid d_j) = \frac{c(w, d_j)P(T_{d_j} \mid w, d_j)}{\sum_{w' \in V} c(w', d_j)P(T_{d_j} \mid w', d_j)}
    \]
• It is obvious that the major difference among the representative models mentioned above is how to capitalize on the set of documents and the original query.

• A principled framework can be obtained to unify all of these models (and their extensions) by using a generalized objective likelihood function:

\[
\mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left( \sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)}
\]
A Unified Framework – 2

\[ \mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left( \sum_{m \in M} P(w|m) P(m) \right)^{c(w,e)} \]

- **Relevance modeling (RM):** when \( E \) only consists of the user query, \( M \) consists of a set of document models corresponding to the top-ranked (pseudo-relevant) documents, and we assume the document models are known, then it can be deduced to the RM model

\[
P_{RM}(w) \approx \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) \prod_{i=1}^{q} P(w_i|d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} P(d_j') P(w'|d_j') \prod_{i=1}^{q} P(w_i|d_j')} \]

\[
= \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) P(q|d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} P(d_j') P(w'|d_j') P(q|d_j')} \]

\[
= \sum_{d_j \in R_q} P(w|d_j) \frac{P(d_j) P(q|d_j)}{\sum_{d_j' \in R_q} P(d_j') P(q|d_j')} \]

\[ \sum_{w' \in V} P(w'|d_j') = 1 \]
A Unified Framework – 3

\[ \mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left( \sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)} \]

- **Simple mixture modeling (SMM):** if we hypothesize that \( M \) consists of two components: one component is a generic background model and the other is an unknown query-specific topic model, the weight of each component is presumably fixed in advance, and the observations are those top-ranked documents

\[ \mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left( (1 - \alpha) \cdot P_{SMM}(w) + \alpha \cdot P(w|BG) \right)^{c(w,d_j)} \]
A Unified Framework – 4

\[ \mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left( \sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)} \]

- **Tri-Mixture modeling (TMM):** if we hypothesize that \( M \) consists of three components: the first component is a generic background model, the second model is a document-specific model, and the last one is an unknown query-specific topic model, the weight of each component is presumably fixed in advance, and the observations are those top-ranked documents

\[ \mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left( (1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w|d_j) + \beta \cdot P(w|BG) \right)^{c(w,d_j)} \]
A Unified Framework – 5

\[ \mathcal{L} = \prod_{e \in E} \prod_{w \in V} \left( \sum_{m \in M} P(w|m)P(m) \right)^{c(w,e)} \]

- **Others**: without loss of generality, some other state-of-the-art language models also can be deduced from the proposed general objective function, such as the **positional relevance model**, the **cluster-based methods**, the **topic models**, and among others

\[
\mathcal{L} = \prod_{w_i \in V} \prod_{d_j \in D} P(w_i, d_j)^{c(w_i,d_j)} = \prod_{d_j \in D} \prod_{i=1}^{|d_j|} P(w_i, d_j) \\
= \prod_{d_j \in D} \prod_{i=1}^{|d_j|} \left( P(d_j) \sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j) \right)
\]
Topic-based Relevance Modeling

- TRM assumes that the additional cues of how words are distributed across a set of latent topics can carry useful global topic structure for relevance modeling
  - The pseudo-relevant documents are assumed to share a set of pre-defined latent topic variables \{T_1, \ldots, T_k, \ldots, T_K\}

\[
P_{TRM}(w) \approx \frac{\sum_{d_j \in R_q} \sum_{k=1}^{K} P(d_j) P(T_k|d_j) P(w|T_k) P(q|T_k)}{\sum_{w' \in V} \sum_{d_j' \in R_q} \sum_{k'=1}^{K} P(d_j') P(T_{k'}|d_j') P(w|T_{k'}) P(q|T_{k'})}
\]

- As with PLSA and LDA, the probabilities \(P(w|T_k)\) and \(P(T_k|d_j)\) can be estimated using inference algorithms like EM or VB-EM algorithms on the whole document collection

\[
P_{RM}(w) \approx \frac{\sum_{d_j \in R_q} P(d_j) P(w|d_j) P(q|d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} P(d_j') P(w'|d_j') P(q|d_j')}
\]
Word-based Relevance Modeling

- The most challenging aspect facing RM is how to efficiently infer the relevance class
  - The relevance class of a given query is commonly approximated by the top-ranked documents returned by an IR system

- The WRM model of each word in the language can be trained by concatenating those words occurring within a context window to form a relevant observation sequence for estimating $P(w|d_{w_i})$

$$P_{WRM}(w) \approx \frac{\sum_{w_i \in q} P(d_{w_i})P(w|d_{w_i})P(q|d_{w_i})}{\sum_{w'_i \in V} \sum_{w_i \in q} P\left(d'_{w_i}\right)P\left(w'|d'_{w_i}\right)P\left(q|d'_{w_i}\right)}$$

$$P_{RM}(w) \approx \frac{\sum_{d_j \in R_q} P(d_j)P(w|d_j)P(q|d_j)}{\sum_{w'_j \in V} \sum_{d'_j \in R_q} P(d'_j)P(w'|d'_j)P(q|d'_j)}$$
Research Issues

• The main issues in pseudo-relevance feedback
  – How to select relevant documents from the top-retrieved documents
  – How to select expansion terms
In order to select a set of pseudo-relevant documents, which can cover most of the possible aspects of the query, a few selecting methods have been proposed

- **Gapped Top $K$**
  - partition the documents into $K$ clusters based solely on the relevance scores
  - select documents with the highest relevance score in each cluster to form the feedback document set

- **Cluster Centroid**
  - partition top-ranked documents into $K$ clusters
  - select the most representative document from each cluster
Active Relevance, Density, & Diversity

- Active-RDD algorithm extends the MMR algorithm by adding an extra term, which reflects the document density
  - Relevance
    
    \[ \text{Rel}(d) \equiv KL(q||d) = \sum_{w \in V} P(w|q) \log \frac{P(w|q)}{P(w|d)} \]
  
  - Density
    - Jeffreys divergence
    
    \[ \text{Density}(d) \equiv -\frac{1}{|D|} \sum_{d_j \in D} \left( KL(d_j||d) + KL(d||d_j) \right) \]
  
  - Diversity
    
    \[ \text{Diversity}(d) \equiv \min_{\tilde{d} \in \tilde{D}} (KL(\tilde{d}||d) + KL(d||\tilde{d})) \]
  
  - Active-RDD
    
    \[ d^* = \arg\max_{d \in \{D \setminus \tilde{D}\}} \alpha \cdot \text{Rel}(d) + \beta \cdot \text{Density}(d) + (1 - \alpha - \beta) \cdot \text{Diversity}(d) \]
Resampling Method

- The essential idea is that a document that appears in multiple highly-ranked clusters will contribute more to the query terms than other documents
  - The **dominate documents** in the sampled clusters are used for feedback with **redundancy**
  - The overlapping cluster method is used to identify **dominant documents** for the query to emphasize good representative terms in dominant documents
Conclusions

• The methods for tackling the fundamental problem can be classified into **global** methods and **local** methods
  – Global methods are techniques for expanding or reformulating query terms independent of the query and initial search results
    • Thesaurus or WordNet
    • automatic thesaurus generation
    • spelling correction
  – Local methods adjust a query relative to the documents that initially appear to match the query
    • Relevance feedback
    • Pseudo relevance feedback (Blind relevance feedback)
    • (Global) indirect relevance feedback
Homework 4 – Description

• In this project, you will have
  – 800 Queries
  – 2265 Documents

• Our goal is to implement the Rocchio algorithm (or LM-based methods) for retrieval
Homework 4 – Kaggle

- Please login our competition page at Kaggle
  - https://www.kaggle.com/t/aed5e4d90570477d8fb1745552cd904e

- Your team name is ID_Name
  - M123456_陳冠宇
Homework 4 – Submission Format
Homework 4 – Scoring

• The evaluation measure is \textbf{MAP@50}

• The maximum number of daily submissions is 20

• The \textbf{hard} deadline is 11/23 11:00am

  \[ \text{YourScore} = 4 + \frac{\text{YourMAP} - \text{BaselineMAP}}{\text{HighestMAP} - \text{BaselineMAP}} \times 6\% \]

• You should submit source codes and a mini report onto the Moodle system
  – TA will ask you to demo your program
  – In this HW, you can \textbf{ONLY} leverage PRF models to do retrieval
The Evolution

1957 Term Frequency
1965 Rocchio Algorithm
1972 Inverse Document Frequency
1973 Boolean Model
1975 Vector Space Model
1976 Probabilistic Model
1978 Language Modeling Approaches
1994 Best Match Models (Okapi Systems)
1998 Language Modeling Approaches
1999 Probabilistic Latent Semantic Analysis
2001 Relevance-based LM & Simple Mixture Model
2003 Latent Dirichlet Allocation

David M. Blei
Columbia University, USA

Thomas Hofmann
ETH Zurich, Switzerland

Scott Deerwester

V. Lavrenko
Edinburgh

C.X. Zhai
Illinois University

J. Rocchio
Gossiping

Yoshua Bengio @UMONTREAL

Michael Jordan @Berkeley

Andrew Ng @Stanford

Thomas Hofmann @ETH Zurich, Switzerland

W. Bruce Croft @UMass

David M. Blei @Columbia University

John Lafferty @Yale

V. Lavrenko @Edinburgh

C.X. Zhai @Illinois University
Questions?

kychen@mail.ntust.edu.tw