Pseudo-relevance Feedback & Query Models

Kuan-Yu Chen (陳冠宇)

2020/11/27 @ TR-313, NTUST
<table>
<thead>
<tr>
<th>#</th>
<th>Team Name</th>
<th>Score</th>
<th>Entries</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M10915010_盧克涵</td>
<td>0.63447</td>
<td>46</td>
<td>6d</td>
</tr>
<tr>
<td>2</td>
<td>M10915045_施信宏</td>
<td>0.62661</td>
<td>26</td>
<td>2d</td>
</tr>
<tr>
<td>3</td>
<td>M10915095_薛宇翔</td>
<td>0.61474</td>
<td>49</td>
<td>10h</td>
</tr>
<tr>
<td>4</td>
<td>M10907505_游照臨</td>
<td>0.61284</td>
<td>13</td>
<td>9h</td>
</tr>
<tr>
<td>5</td>
<td>M10915027_石成峰</td>
<td>0.60911</td>
<td>31</td>
<td>9h</td>
</tr>
<tr>
<td>6</td>
<td>M10915028_陳柏勳</td>
<td>0.60886</td>
<td>29</td>
<td>12h</td>
</tr>
<tr>
<td>7</td>
<td>M10915100_郭智威</td>
<td>0.60248</td>
<td>68</td>
<td>9h</td>
</tr>
<tr>
<td>8</td>
<td>M10915201_陳牧凡</td>
<td>0.59516</td>
<td>70</td>
<td>3d</td>
</tr>
<tr>
<td>9</td>
<td>D10907005_陳昱宏</td>
<td>0.59213</td>
<td>36</td>
<td>14h</td>
</tr>
<tr>
<td>10</td>
<td>M10815048_張嘉銘</td>
<td>0.58963</td>
<td>63</td>
<td>3d</td>
</tr>
<tr>
<td>11</td>
<td>B10615043_何嘉峻</td>
<td>0.58052</td>
<td>70</td>
<td>13h</td>
</tr>
<tr>
<td>12</td>
<td>B10615034_黃柏翰</td>
<td>0.57919</td>
<td>128</td>
<td>2d</td>
</tr>
<tr>
<td>13</td>
<td>B10615036_黃承銘</td>
<td>0.57787</td>
<td>11</td>
<td>1d</td>
</tr>
<tr>
<td>14</td>
<td>B10615026_溫承勳</td>
<td>0.57495</td>
<td>48</td>
<td>2d</td>
</tr>
<tr>
<td>15</td>
<td>M10915006_黃穎宏</td>
<td>0.57439</td>
<td>19</td>
<td>1d</td>
</tr>
<tr>
<td>16</td>
<td>M10915080_羅建程</td>
<td>0.57409</td>
<td>11</td>
<td>12h</td>
</tr>
<tr>
<td>17</td>
<td>TEST</td>
<td>0.57201</td>
<td>1</td>
<td>15h</td>
</tr>
<tr>
<td>18</td>
<td>M10815036_王仁德</td>
<td>0.57185</td>
<td>2</td>
<td>12h</td>
</tr>
<tr>
<td>19</td>
<td>B10632026_吳苡瑋</td>
<td>0.57150</td>
<td>17</td>
<td>17h</td>
</tr>
</tbody>
</table>

**FYI: with 32 topics**

<table>
<thead>
<tr>
<th>#</th>
<th>Team Name</th>
<th>Score</th>
<th>Entries</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>M10915010_盧克涵</td>
<td>0.52665</td>
<td>46</td>
</tr>
<tr>
<td>2</td>
<td>▲ 5</td>
<td>M109150100_郭智威</td>
<td>0.50803</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>▲ 14</td>
<td>TEST</td>
<td>0.50714</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>—</td>
<td>M10907505_游照臨</td>
<td>0.50159</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>▼ 3</td>
<td>M10915045_施信宏</td>
<td>0.50113</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>▲ 3</td>
<td>D10907005_陳昱宏</td>
<td>0.49624</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>▲ 5</td>
<td>B10615034_黃柏翰</td>
<td>0.49327</td>
<td>128</td>
</tr>
<tr>
<td>8</td>
<td>▼ 2</td>
<td>M10915028_陳柏勳</td>
<td>0.49014</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>▼ 1</td>
<td>M10915201_陳牧凡</td>
<td>0.48948</td>
<td>70</td>
</tr>
<tr>
<td>10</td>
<td>▼ 5</td>
<td>M10915027_石成峰</td>
<td>0.48665</td>
<td>31</td>
</tr>
<tr>
<td>11</td>
<td>—</td>
<td>B10615043_何嘉峻</td>
<td>0.48597</td>
<td>70</td>
</tr>
<tr>
<td>12</td>
<td>▼ 9</td>
<td>M10915095_薛宇翔</td>
<td>0.48540</td>
<td>49</td>
</tr>
<tr>
<td>13</td>
<td>▲ 6</td>
<td>B10632026_吳苡瑋</td>
<td>0.48165</td>
<td>17</td>
</tr>
<tr>
<td>14</td>
<td>▼ 4</td>
<td>M10815048_張嘉銘</td>
<td>0.48067</td>
<td>63</td>
</tr>
<tr>
<td>15</td>
<td>▲ 7</td>
<td>B10615033_王豊禎</td>
<td>0.48044</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>▲ 7</td>
<td>M10915012_黃偉愷</td>
<td>0.47906</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>▲ 13</td>
<td>M10815013_陳思妮</td>
<td>0.47847</td>
<td>31</td>
</tr>
<tr>
<td>18</td>
<td>▲ 2</td>
<td>B10615024_李韋宗</td>
<td>0.47838</td>
<td>51</td>
</tr>
<tr>
<td>19</td>
<td>▼ 1</td>
<td>M10815036_王仁德</td>
<td>0.47793</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>▼ 7</td>
<td>B10615036_黃泰銘</td>
<td>0.47630</td>
<td>11</td>
</tr>
</tbody>
</table>
About Final Project

- Group your team!
  - 2~4 team members
  - Choose a paper

<table>
<thead>
<tr>
<th>Date</th>
<th>Syllabus</th>
<th>Homework</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/18</td>
<td>Course Overview</td>
<td></td>
</tr>
<tr>
<td>9/25</td>
<td>Break for Roelting2020</td>
<td></td>
</tr>
<tr>
<td>10/2</td>
<td>Holiday for Moon Festival</td>
<td></td>
</tr>
<tr>
<td>10/9</td>
<td>Holiday for National Day</td>
<td></td>
</tr>
<tr>
<td>10/16</td>
<td>Classic Models</td>
<td>Homework-1 (deadline: 10/29 23:59)</td>
</tr>
<tr>
<td>10/23</td>
<td>Extended Probabilistic Models</td>
<td>Homework-2 (deadline: 11/5 23:59)</td>
</tr>
<tr>
<td>10/30</td>
<td>Evaluation &amp; Benchmark Collections</td>
<td>Homework-3 (deadline: 11/12 23:59)</td>
</tr>
<tr>
<td>11/6</td>
<td>Latent Semantic Analysis</td>
<td></td>
</tr>
<tr>
<td>11/13</td>
<td>Statistical Topic Models</td>
<td>Homework-4 (deadline: 11/26 23:59)</td>
</tr>
<tr>
<td>11/20</td>
<td>Search Results Diversification</td>
<td></td>
</tr>
<tr>
<td>12/4</td>
<td>Talk</td>
<td>Submit Your Member List!</td>
</tr>
<tr>
<td>12/11</td>
<td>Representation Learning for Information Retrieval</td>
<td></td>
</tr>
<tr>
<td>12/18</td>
<td>Supervised Retrieval Models &amp; Information Retrieval in Practice</td>
<td>Homework-6 (deadline: 12/31 23:59) &amp; Submit Your Paper Title!</td>
</tr>
<tr>
<td>12/25</td>
<td>Break for Your Final Project</td>
<td></td>
</tr>
<tr>
<td>1/1</td>
<td>Holiday for Founding Anniversary</td>
<td></td>
</tr>
<tr>
<td>1/8</td>
<td>Presentation-1</td>
<td></td>
</tr>
<tr>
<td>1/15</td>
<td>Presentation-2</td>
<td></td>
</tr>
</tbody>
</table>
Review

• These methods mainly differ in **diversity modeling**
  – **Implicitly**: The diversity is implicitly modeled through document similarities
    • MMR
    • SMM
  – **Explicitly**: It can be explicitly modeled through the coverage of query subtopics, and document dependency
    • xMMR
    • WUME
    • xQuAD

[Diagram of relevance, coverage, novelty, and diversity metrics]
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 document 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.

- 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**.
The Rocchio Algorithm – 2

• The idea can be fulfilled by using the vector space model with pseudo relevant and non-relevant documents

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

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

• A simplified variant is to consider the positive feedback documents only

\[ \tilde{q}' = \alpha \cdot \tilde{q} + \beta \cdot \frac{1}{|R_q|} \cdot \left( \sum_{d_j \in R_q} \overrightarrow{d_j} \right) \]
The full process will become

1. Perform VSM
2. Select a set of top-ranked documents
3. Reformulate the query vector
4. Perform VSM with the new query vector
KL-Divergence Measure

- Query likelihood measure is a classic way to employ LM to IR

\[ P(d_j|q) = \frac{P(q|d_j)P(d_j)}{P(q)} \propto P(q|d_j)P(d_j) \]

\[ \approx P(q|d_j) \approx \prod_{i=1}^{|q|} P(w_i|d_j) \]

- 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 \textbf{probabilistic model} rather than simply an \textbf{observation}

- KL-divergence supports us to achieve a better result by considering \textbf{both} query and document models
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 \textit{co-occur} with the query in the concept will have higher probabilities

\[
P_{RM}(w) \equiv \frac{P(w, q | R)}{\sum_{w' \in V} P(w', q | R)} \approx \frac{\sum_{d_j \in R_q} P(d_j)P(w, q | d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j)P(w', q | d'_j)}
= \frac{\sum_{d_j \in R_q} P(d_j)P(w | d_j)P(q | d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j)P(w' | d'_j)P(q | d_j)}
= \frac{\sum_{d_j \in R_q} P(d_j)P(w | d_j) \prod_{i=1}^{\|q\|} P(w_i | d_j)}{\sum_{w' \in V} \sum_{d'_j \in R_q} P(d'_j)P(w' | d'_j) \prod_{i'=1}^{\|q\|} P(w_{i'} | d'_j)}
\]
Consequently, for a given pair of query and document, the relevance degree can be determined by using the new query language model.

In order to incorporate the general information, the background model can also be integrated.

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

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

\[
= - \sum_{w \in V} [\alpha \cdot P_{ULM}(w) + \beta \cdot P_{RM}(w) + (1 - \alpha - \beta) \cdot P_{BG}(w)] \log P(w|d_j)
\]
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 $P_{SMM}(w)$ is estimated by maximizing the log-likelihood of the set of top-ranked documents $R_q$ expressed as follows:

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

- Estimate the parameters
  - E-step

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

- M-step

\[
P_{SMM}(w) = \frac{\sum_{d_j \in R_q} c(w, d_j) P(T_{SMM}|w)}{\sum_{w' \in V} \sum_{d_j' \in R_q} c(w', d_j') P(T_{SMM}|w')}
\]

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

\[
= \prod_{d_j \in R_q} \prod_{w \in V} (P_{SMM}(w|T_{SMM})P(T_{SMM}) + P(w|BG)P(BG))^{c(w,d_j)}
\]
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

  \[
  P(T_{TMM}|w, d_j) = \frac{(1 - \alpha - \beta) \cdot P_{TMM}(w)]}{(1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w|d_j) + \beta \cdot P(w|BG)}
  \]

  \[
  P(T_{d_j}|w, d_j) = \frac{\alpha \cdot P(w|d_j)}{(1 - \alpha - \beta) \cdot P_{TMM}(w) + \alpha \cdot P(w|d_j) + \beta \cdot P(w|BG)}
  \]

  - M-step

  \[
  P_{TMM}(w) = \frac{\sum_{d_j \in R_q} c(w, d_j)P(T_{TMM}|w, d_j)}{\sum_{w' \in V} \sum_{d_j' \in R_q} c(w', d_j')P(T_{TMM}|w', d_j')}
  \]

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

• 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

\[ \mathcal{L} = \prod_{d_j \in R_q} \prod_{w \in V} \left( (1 - \alpha) \cdot P_{\text{SMM}}(w) + \alpha \cdot P(w|BG) \right)^{c(w,d_j)} \]
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 $P(w|d_{w_i})$

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

$$P_{RM}(w) \approx \frac{\sum_{j \in R_q} P(d_j)P(w|d_j)P(q|d_j)}{\sum_{w' \in V} \sum_{j \in R_q} P(d'_j)P(w'|d'_j)P(q|d'_j)}$$
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
    \[
    Rel(d) \equiv KL(q||d) = \sum_{w \in V} P(w|q) \log \frac{P(w|q)}{P(w|d)}
    \]
  - Density
    - Jeffreys divergence
      \[
      Density(d) \equiv -\frac{1}{|D|} \sum_{d_j \in D} (KL(d_j||d) + KL(d||d_j))
      \]
  - Diversity
    \[
    Diversity(d) \equiv \min_{\tilde{d} \in \tilde{D}} (KL(\tilde{d}||d) + KL(d||\tilde{d}))
    \]
  - Active-RDD
    \[
    d^* = \arg\max_{d \in \{D-\tilde{D}\}} \alpha \cdot Rel(d) + \beta \cdot Density(d) + (1 - \alpha - \beta) \cdot Diversity(d)
    \]
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
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

J. Rocchio

C.X. Zhai
Illinois University

V. Lavrenko
Edinburgh

Illinois University
Homework 5 – Description

• In this project, you will have
  – 150 Queries
    • 60% Public Queries & 40% Private Queries
  – 30,000 Documents

• Our goal is to implement a PRF algorithm for retrieval
  – In addition to the PRF model, you can combine any models/strategies to achieve a good performance

• Please submit a report and your source codes to the Moodle system, otherwise you will get 0 point
  – The report will be judged by TA, and the score is either 1 or 2
Homework 5 – Scoring

- Please login our competition page at Kaggle
  - [https://www.kaggle.com/t/46f5a4ea8bed4a59bd1ba632226adea0](https://www.kaggle.com/t/46f5a4ea8bed4a59bd1ba632226adea0)
  - Your team name is ID_Name
    - M123456_陳冠宇
  - The evaluation measure is **MAP@5000**
  - The maximum number of daily submissions is 20
  - The **hard** deadline is 12/10 23:59am
    - You point is depended on your performance on the *private* leaderboard!
  - Your Score = \[
  \frac{YourMAP - BaselineMAP}{HighestMAP - BaselineMAP} \times 13\%
  \]
Homework 5 – Warning!!

- Please follow our rules
  - Don’t cheat yourself, your friends, and me!
  - Don’t create multiple accounts!
  - Implement the IR system by YOUSELF!
    - Enjoy the Information Retrieval Methods
Questions?

kychen@mail.ntust.edu.tw