Problems and Methods in Distributed Information Retrieval

Ilya Markov
ilya.markov@usi.ch

University of Lugano, Lugano, Switzerland

July 5, 2012
Supervisor - Prof. Fabio Crestani
IR@USI
- 6 PhD students
- Blog search, patent search, mobile IR

Distributed Information Retrieval
aka Federated Search
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
4. Score Normalization/Results Merging
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Information Retrieval in Practice

University of Lugano - Università della Svizzera italiana
www.usi.ch/en
⇒ Graduate studies - the Master's ... - Faculty of Informatics - Job opportunities at USI

USI - Faculty of Informatics
www.inf.usi.ch/
⇒ Show map of Via Giuseppe Buffi 13, 6900 Lugano

USI - Università della Svizzera italiana
www.usi.ch/ - Translate this page
Università di Lugano. University of Lugano. Scegli la lingua. Choose the language. 1. Italian. Italian. 2. Inglese. English. Benvenuto sul sito dell'Università ...
⇒ Show map of Largo Bernasconi 2, 6850 Mendrisio
⇒ Risorse e servizi online - Concorsi e offerte di lavoro ... - Università di Lugano

Details
Bus: Lugano, Ponte Madonneta

More reviews
google.com (3)
Given a query and a set of docs

1. Estimate doc’s relevance

2. Rank docs by relevance

3. ...
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Aggregated Search
DIR Problems

1. Resource Selection
2. Score Normalization/Results Merging
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
1 Information Retrieval

2 Distributed Retrieval Scenarios

3 Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches

4 Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Objective of Resource Selection

Given a query and a set of document sources select sources that are likely to contain relevant documents.
Information Retrieval

1. Given a query and a set of docs
2. Estimate doc’s relevance
3. Rank docs by relevance

Ilya Markov ilya.markov@usi.ch

Problems and Methods in DIR
Given a query and a set of sources

1. Estimate number of relevant docs

2. Rank sources by this number

3. ...
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
First Generation (CORI)

1. Sources as documents
2. Retrieval on documents
3. Documents are sources
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Second Generation (ReDDE)

1. All documents from all sources are merged in a centralized index.
2. Documents are ranked with regards to a query.
3. Top $n$ documents are supposed to be relevant to a query.
4. The number of relevant documents in a source $R$ is calculated as a proportion of its documents in the top $n$. 

```
1 2 3 4 5 ... top n ... k-1 k
```

Ilya Markov ilya.markov@usi.ch

Problems and Methods in DIR
Second Generation (ReDDE)

\[ \hat{R}_{red} = \frac{3}{8}, \quad \hat{R}_{green} = \frac{2}{8}, \quad \hat{R}_{yellow} = \frac{3}{8}, \quad \hat{R}_{blue} = 0 \]
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Objective of Score Normalization

Given multiple lists of document scores returned by different sources make these scores comparable across sources.
### Score Normalization

#### Introduction

**Linear Score Normalization**

**Results Merging**

---

#### Score Normalization/Results Merging

<table>
<thead>
<tr>
<th>Score</th>
<th>Merged Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>87</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>0.87</td>
</tr>
<tr>
<td>15</td>
<td>0.41</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Linear score normalization is applied to each score before merging.
- Results are merged using a simple linear approach.

---

#### Example

- Source scores: 100, 87, 41, 15
- Merged scores: 1.0, 1.0, 0.87, 0.41

---

Ilya Markov ilya.markov@usi.ch

Problems and Methods in DIR 26
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Linear Score Normalization

- MinMax
  
  \[ s_{\text{norm}} = \frac{s - \text{min}}{\text{max} - \text{min}} \]

- Z-Score
  
  \[ s_{\text{norm}} = \frac{s - \mu}{\sigma} \]

- Sum
  
  \[ s' = s - \text{min}, \quad s_{\text{norm}} = \frac{s'}{\sum_i s'_i} \]

MinMax has the weakest assumptions. It is the most robust and best performing method.
### MinMax

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Score 1</th>
<th>Source 2</th>
<th>Score 2</th>
<th>Source 3</th>
<th>Score 3</th>
<th>Merged Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.1</td>
<td>1250</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>87</td>
<td>0.03</td>
<td>815</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>41</td>
<td>0.015</td>
<td>800</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>15</td>
<td>0.001</td>
<td>234</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Each source contains at least 1 relevant document.
- This document is most likely to be ranked 1st.
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging
Results Merging (CORI)

\[ s_{\text{norm}} = (1 + 0.4 \cdot w_{\text{source}}) \frac{s - \min}{\max - \min} \]
Alternative ways to calculate source weights?
Outline

1. Information Retrieval
2. Distributed Retrieval Scenarios
3. Resource Selection
   - Introduction
   - First Generation Approaches
   - Second Generation Approaches
4. Score Normalization/Results Merging
   - Introduction
   - Linear Score Normalization
   - Results Merging