Uncertainty in Distributed Information Retrieval

Ilya Markov

University of Lugano

January 31, 2014
Outline

1. Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2. Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3. Summary and Future Work
   - Summary
   - Future Work
Information Management Systems

- Collect
- Store
- **Provide access**
- Distribute
ARMISTICE SIGNED, END OF THE WAR!
BERLIN SEIZED BY REVOLUTIONISTS;
NEW CHANCELLOR BEGS FOR ORDER;
OUSTED KAISER FLEES TO HOLLAND

SON FLEES WITH EX-KAISER
Hindenburg Also Believed to be Among Those in His Party.

ALL ARE HEAVILY ARMED
Automobiles Brisk with Rifles as Fugitives Arrive at Dutch Frontier.

ON THEIR WAY TO DE STEEG
Belgian Yell to There, "Are You On Your Way to Paris?"

LONDON, Nov. 10.—Both the former German Emperor and his eldest son, Frederick Wilhelm, crossed the Dutch frontier Sunday morning, according to reports from The Hague. His reported destination is De Steeg, near Utrecht.

The former German Emperor, which is believed to include Field Marshal von Hindenburg, arrived in Emden, [midway between Emden and Hanover] on the Dutch frontier.

Kaiser Fought Hindenburg's Call for Abdication; Failed to Get Army's Support in Keeping Throne

By GEORGE KIRWIN.

AMSTERDAM, Nov. 12.—1800 troops crossed the Rhine, Vermont at 10 a.m. on the line, 700 engineers and 2000 officers and men of the German army, according to reports from The Hague.

The European News, which was founded in 1828, is a daily newspaper published in the Netherlands. It is an independent newspaper, and is not affiliated with any political party or organization.

BERLIN SOLDIERS JOIN REVOLT

A Red Shell Building in Which Officers Vainly Resist.

THINGS DEMAND REPUBLIC

Revolutionary Flag on Royal Palace—Crowd Princes' Palace Also Seized.

GENERAL STRIKE IS BEGAN

Burgomaster and Police Submit—War Office Now Under Socialistic Control.

LONDON, Nov. 10. — The greater part of Berlin is in control of revolutionaries, the former Kaiser having fled to Holland, and Friedrich Ebert, the new Socialist Chancellor, has taken command of the situation. The revolt is spreading throughout the country with great rapidity.

Disputes resolved in London today announce these startling developments. The Workmen's and Soldiers' Council is now administering the municipal government of the German capital. The Kaiser and his government have been driven out. The coup was planned in the military circles, and the military leaders, under the direction of General von Hindenburg, have been instrumental in its success.

Socialist Chancellor Appeals to All Germans To Help Him Save Fatherland from Anarchy

BEIJING, Nov. 10. (Christian Post) — In an address to the people, the new German Chancellor, Friedrich Ebert, said:

"The Socialists have always been in favor of a peaceful settlement of all problems. We believe that the best way to achieve this is by negotiation. We are ready to discuss anything, so long as it is done in good faith."

"We are not afraid of the future. We are looking forward to a better world for all of us."

WAR ENDS AT 6 O'CLOCK THIS MORNING

The State Department in Washington made the announcement at 2:45 o'clock.

ARMISTICE WAS SIGNED IN FRANCE AT MIDNIGHT

Terms Include Withdrawal from Alsace-Lorraine, Disarming and Demobilization of Army and Navy, and Occupation of Strategic Naval and Military Points.

By The Associated Press.

WASHINGTON, Monday, Nov. 11, 1918, 4 A. M.—The armistice between Germany, on the one hand, and the Allied Governments and the United States, on the other, has been signed.

The State Department announced at 2:45 o'clock this morning that the armistice has been signed.

The department's announcement simply stated that the armistice was signed.

The armistice was signed by the German government and the Allied powers, including the United States, on November 11, 1918, bringing an end to World War I.
Information Access

- Information Retrieval
- Text Mining
- Machine Translation
- Text Summarization
- etc.
Outline

1. Introduction
   - Information Retrieval
     - Distributed Information Retrieval
     - Thesis Goals

2. Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3. Summary and Future Work
   - Summary
   - Future Work
Given a query and a set of docs

1. Estimate doc’s relevance
2. Rank docs by relevance
Evaluation in IR

- **Dataset**
  - TREC GOV2: crawl of .gov in 2004, 25 million web pages
  - 150 queries
  - Relevance judgements

- **Measures**
  - Precision at 10 (P@10)
  - Average Precision (AP)
  - Mean AP (MAP)
Outline

1 Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2 Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3 Summary and Future Work
   - Summary
   - Future Work
Distributed Information Retrieval (DIR) is concerned with aggregating multiple searchable sources of information within a single interface.
Broker-based DIR

1. Resource Description
2. Resource Selection
3. Score Normalization
4. Results Merging
Broker-based DIR

- Resource description
- Broker-based document ranking
- Resource selection
- Source-specific document ranking
- Score normalization

Resource description
Resource selection
Score normalization, Results merging
Resource descriptions are approximate.

Uncertainty can be measured by divergence between a description and a source.

Existing studies on improving descriptions implicitly reduced uncertainty.
Uncertainty in Document Ranking

- Core IR problem
- Approximation of a relevance-based ranking
Uncertainty in Resource Selection

- Uncertain input (descriptions, document ranking) $\implies$ uncertain ranking of sources
- Measure, analyze, reduce
Uncertainty in Score Normalization

- Uncertain input (ranking of sources, document rankings)  
  \[\Rightarrow\]  uncertain normalized rankings
- Measure, analyze, reduce
Uncertainty in DIR

- Resource description
- Broker-based document ranking
- Resource selection
- Source-specific document ranking
- Score normalization

Resource description
Resource selection
Score normalization, Results merging
Outline

1 Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2 Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3 Summary and Future Work
   - Summary
   - Future Work
1. Analyze and evaluate state-of-the-art resource selection and score normalization methods in a unified framework
2. Improve resource selection and score normalization methods
3. Analyze, measure and reduce uncertainty in resource selection and score normalization
# Thesis Structure

<table>
<thead>
<tr>
<th></th>
<th>Analysis, Evaluation</th>
<th>Improvement</th>
<th>Uncertainty Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Selection</td>
<td></td>
<td>Chapter 3</td>
<td></td>
</tr>
<tr>
<td>Score Normalization</td>
<td></td>
<td>Chapter 4</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>Results Merging</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2. Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3. Summary and Future Work
   - Summary
   - Future Work
Experimental Setup

- **Testbeds**
  - gov2.250
  - gov2.30
- **Resource Selection**
  - Method: CRCS(e)
  - Parameter: $k$
- **Score Normalization**
  - Method: CORI
  - Parameter: $\lambda$
Small-Document Resource Selection

QUERY -> RESOURCE SELECTION -> RETRIEVAL SYSTEM -> ...
CORI Score Normalization

\[
CORI = f(s(R|q)) \cdot \frac{s(d|q) - s_{\text{min}}(q)}{s_{\text{max}}(q) - s_{\text{min}}(q)}
\]
Outline

1 Introduction
   • Information Retrieval
   • Distributed Information Retrieval
   • Thesis Goals

2 Reducing Uncertainty
   • General Idea
   • Perturbation Methods
   • Resource Selection
   • Score Normalization

3 Summary and Future Work
   • Summary
   • Future Work
Small-Document Resource Selection

QUERY → RESOURCE SELECTION → RETRIEVAL SYSTEM → ...
Exploring the Space of Source Rankings
Using Multiple Document Rankings
Obtaining Multiple Documents Rankings
Components to Perturb

sample of documents

retrieval system

QUERY

user’s query

document ranking

 queried document
Outline

1. Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2. Reducing Uncertainty
   - General Idea
   - Perturbation Methods
     - Resource Selection
     - Score Normalization

3. Summary and Future Work
   - Summary
   - Future Work
Query Perturbation

- The weight of each query term is decreased one term at a time.
- This results in $L + 1$ perturbations, where $L$ is the query length.
Ten different retrieval functions are used.
The parameters of each function *could* also be changed in order to obtain more perturbations.
Document Ranking Perturbation

- Sample with replacement $n$ documents from the original ranking.
- The sampling is weighed by document relevance scores.
- The process is repeated 100 times (converges after about 50 iterations).
Perturbation Methods

1. Query perturbation
2. Retrieval system perturbation
3. Document ranking perturbation
Reducing Uncertainty in Resource Selection

Diagram showing the process of resource selection with uncertainty reduction.
Outline

1 Introduction
   • Information Retrieval
   • Distributed Information Retrieval
   • Thesis Goals

2 Reducing Uncertainty
   • General Idea
   • Perturbation Methods
   • Resource Selection
   • Score Normalization

3 Summary and Future Work
   • Summary
   • Future Work
Figure: Gov2.30
Figure: Gov2.30
Resource Selection Summary

- *Retrieval system perturbation* improves performance while having the same or lower uncertainty, or reduces uncertainty for the same level of performance.
- *Query perturbation* usually performs worse than the baseline while having the same uncertainty, or achieves the same performance but increases uncertainty.
- *Ranking perturbation* improves performance/decreases uncertainty of CRCS(e) on two testbeds.
Outline

1. Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2. Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3. Summary and Future Work
   - Summary
   - Future Work
CORI

Figure: Top-10, gov2.250
CORI

Figure: Top-10, gov2.250
Score Normalization Summary

- *Retrieval system* and *ranking perturbations* usually improve performance of the baseline, decreasing uncertainty in some cases.
- *Query perturbation* almost always performs worse than the baseline and has the same or higher uncertainty.
Outline

1 Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2 Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3 Summary and Future Work
   - Summary
   - Future Work
Outline

1 Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2 Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3 Summary and Future Work
   - Summary
   - Future Work
Summary

1. Analyze and evaluate state-of-the-art resource selection and score normalization methods in a unified framework
2. Improve resource selection and score normalization methods
3. Analyze, measure and reduce uncertainty in resource selection and score normalization
Contributions: Goal 1

- **Resource Selection**
  - A unified framework for theoretical, qualitative and quantitative analyses of small-document (SD) approaches to resource selection
  - Factors affecting SD resource selection
  - A classification of SD approaches
  - A standard implementation of SD techniques

- **Score Normalization**
  - Theoretical and empirical analysis of unsupervised linear score normalization within a unified framework
  - Underlying assumptions of each method
  - A classification of score normalization methods
  - A standard implementation of score normalization techniques
Contributions: Goal 2

- Resource Selection
  - A method for dampening source sizes in SD resource selection
  - Unsupervised methods for setting parameters of SD techniques
  - A combination of SD techniques with different properties

- Score Normalization
  - Resource selection to relax assumptions of linear score normalization
  - Weighted linear score normalization, two methods for estimating weights
  - Parameterized CORI
Contributions: Goal 3

- The statement of the problem of uncertainty in DIR
- A method for measuring this uncertainty
- Analysis of uncertainty of SD resource selection and CORI score normalization
- A method for reducing the uncertainty in resource selection and score normalization, several implementations of this method
Outline

1. Introduction
   - Information Retrieval
   - Distributed Information Retrieval
   - Thesis Goals

2. Reducing Uncertainty
   - General Idea
   - Perturbation Methods
   - Resource Selection
   - Score Normalization

3. Summary and Future Work
   - Summary
   - Future Work
Results Presentation in Aggregated Search

Take into account user behavior when presenting federated results.


I. Markov and F. Crestani.
Theoretical, qualitative and quantitative analyses of small-document approaches to resource selection.
Accepted with minor revision at ACM Transactions on Information Systems.

I. Markov, L. Azzopardi, and F. Crestani.
Reducing the uncertainty in resource selection.

I. Markov, A. Arampatzis, and F. Crestani.
On CORI results merging.

F. Crestani and I. Markov.
Distributed information retrieval and applications.

I. Markov, A. Arampatzis, and F. Crestani.
Unsupervised linear score normalization revisited.

I. Markov.
Modeling document scores for distributed information retrieval.