Axiomatic Result Re-Ranking

Matthias Hagen  Michael Völske  Steve Göring  Benno Stein

Bauhaus-University Weimar

www.webis.de
Axiomatic Result Re-Ranking

Abstract

- Axiomatic IR has identified a diverse set of constraints that retrieval models should fulfill, but so far these have been limited to theoretical analysis.
- We incorporate axioms into the retrieval process via re-ranking.
- Large-scale study on Clueweb corpora to show feasibility.
A Brief Tour of Axiomatic IR
A Brief Tour of Axiomatic IR

Observations

- Common strong baseline retrieval models perform similarly well, although derived very differently (BM25, PL2, Query Likelihood...)

- Even minor variations tend to fail in some way or another; why?

Axiomatic IR Answer: these models share beneficial properties, independently of how they are derived

Research Goal: identify and formalize these properties as axioms.
A Brief Tour of Axiomatic IR

Axioms

Successful retrieval functions share similar heuristics:

Example:

\[
BM25(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}
\]
A Brief Tour of Axiomatic IR

Axioms

Successful retrieval functions share similar heuristics:

- TF weighting

Example:

\[
BM25(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}
\]
A Brief Tour of Axiomatic IR

Axioms

Successful retrieval functions share similar heuristics:

- TF weighting
- IDF weighting

Example:

\[ BM25(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \]
A Brief Tour of Axiomatic IR

Axioms

Successful retrieval functions share similar heuristics:

- TF weighting
- IDF weighting
- Length normalization

Example:

\[ BM25(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \left( 1 - b + b \cdot \frac{|D|}{\text{avgdl}} \right)} \]
A Brief Tour of Axiomatic IR

Axioms

Successful retrieval functions share similar heuristics:

- TF weighting
- IDF weighting
- Length normalization

Example:

\[
BM_{25}(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}
\]

Axioms formally capture these heuristics, and how they should be used.
# A Brief Tour of Axiomatic IR

## Axioms

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Acronyms</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term frequency</td>
<td>TFC1–TFC3</td>
<td>[Fang, Tao, Zhai; SIGIR’04]</td>
</tr>
<tr>
<td></td>
<td>TDC</td>
<td>[Fang, Tao, Zhai; SIGIR’04]</td>
</tr>
<tr>
<td>Document length</td>
<td>LNC1 + LNC2</td>
<td>[Fang, Tao, Zhai; SIGIR’04]</td>
</tr>
<tr>
<td></td>
<td>TF-LNC</td>
<td>[Fang, Tao, Zhai; SIGIR’04]</td>
</tr>
<tr>
<td></td>
<td>QLNC</td>
<td>[Cummins, O’Riordan; CIKM’12]</td>
</tr>
<tr>
<td>Lower bound</td>
<td>LB1 + LB2</td>
<td>[Lv, Zhai; CIKM’11]</td>
</tr>
<tr>
<td>Query aspects</td>
<td>REG</td>
<td>[Zheng, Fang; ECIR’10]</td>
</tr>
<tr>
<td></td>
<td>DIV</td>
<td>[Gollapurdi, Sharma; WWW’09]</td>
</tr>
<tr>
<td>Semantic similarity</td>
<td>STMC1 + STMC2</td>
<td>[Fang, Zhai; SIGIR’06]</td>
</tr>
<tr>
<td></td>
<td>STMC3</td>
<td>[Fang, Zhai; SIGIR’06]</td>
</tr>
<tr>
<td></td>
<td>TSSC1 + TSSC2</td>
<td>[Fang, Zhai; SIGIR’06]</td>
</tr>
<tr>
<td>Term proximity</td>
<td>PHC + CCC</td>
<td>[Tao, Zhai; SIGIR’07]</td>
</tr>
</tbody>
</table>
A Brief Tour of Axiomatic IR

Term Frequency Constraints

TFC1  Give a higher score to a document with more occurrences of a query term.

TFC2  The amount of increase in the score due to adding a query term must decrease as we add more terms.

TFC3  Favor a document with more distinct query terms.

Length Normalization Constraints

LNC1  Penalize long documents.

LNC2  Avoid over-penalizing long documents.

TF-LNC Regularize the interaction of TF and document length.

Lower-bounding Term Frequency Constraints

LB1  The presence-absence gap shouldn’t be closed due to length normalization.

LB2  Repeated occurrence isn’t as important as first occurrence.
A Brief Tour of Axiomatic IR

Term Frequency Constraints

TFC1 Give a higher score to a document with more occurrences of a query term.
TFC2 The amount of increase in the score due to adding a query term must decrease as we add more terms.
TFC3 Favor a document with more distinct query terms.

Length Normalization Constraints

LNC1 Penalize long documents.
LNC2 Avoid over-penalizing long documents.
TF-LNC Regularize the interaction of TF and document length.

Lower-bounding Term Frequency Constraints

LB1 The presence-absence gap shouldn’t be closed due to length normalization.
LB2 Repeated occurrence isn’t as important as first occurrence.
A Brief Tour of Axiomatic IR

Axiom Examples: TFC1

TFC1  Give a higher score to a document with more occurrences of a query term.

Given:

- Single-term query $Q = \{q\}$
- Documents $D_1, D_2$ with $|D_1| = |D_2|$

$$\text{IF TF}(q, D_1) > \text{TF}(q, D_2) \text{ THEN Score}(Q, D_1) > \text{Score}(Q, D_2)$$
A Brief Tour of Axiomatic IR

Axiom Examples: LB2

LB2 Repeated occurrence isn’t as important as first occurrence.

Given:

- Two-term query $Q = \{q_1, q_2\}$
A Brief Tour of Axiomatic IR

Axiom Examples: LB2

LB2  Repeated occurrence isn’t as important as first occurrence.

Given:

- Two-term query \( Q = \{q_1, q_2\} \)
- Documents \( D_1, D_2 \) with \( TF(q_1, D_1) > 0 \) and \( TF(q_1, D_2) > 0 \) and
  \[ TF(q_2, D_1) = TF(q_2, D_2) = 0 \]
A Brief Tour of Axiomatic IR
Axiom Examples: LB2

LB2  Repeated occurrence isn’t as important as first occurrence.

Given:

- Two-term query $Q = \{q_1, q_2\}$
- Documents $D_1, D_2$ with $TF(q_1, D_1) > 0$ and $TF(q_1, D_2) > 0$ and $TF(q_2, D_1) = TF(q_2, D_2) = 0$
- Document $D'_1 = D_1 \cup \{q_1\} \setminus \{t_1\}$ for any $t_1 \in D_1$, $t_1 \notin Q$
A Brief Tour of Axiomatic IR

Axiom Examples: LB2

LB2  Repeated occurrence isn’t as important as first occurrence.

Given:

- Two-term query $Q = \{q_1, q_2\}$
- Documents $D_1, D_2$ with $TF(q_1, D_1) > 0$ and $TF(q_1, D_2) > 0$ and $TF(q_2, D_1) = TF(q_2, D_2) = 0$
- Document $D'_1 = D_1 \cup \{q_1\} \setminus \{t_1\}$ for any $t_1 \in D_1, t_1 \notin Q$
- Document $D'_2 = D_2 \cup \{q_2\} \setminus \{t_2\}$ for any $t_2 \in D_2, t_2 \notin Q$
A Brief Tour of Axiomatic IR

Axiom Examples: LB2

LB2 Repeated occurrence isn’t as important as first occurrence.

Given:

- Two-term query $Q = \{q_1, q_2\}$
- Documents $D_1, D_2$ with $TF(q_1, D_1) > 0$ and $TF(q_1, D_2) > 0$ and $TF(q_2, D_1) = TF(q_2, D_2) = 0$
- Document $D'_1 = D_1 \cup \{q_1\} \setminus \{t_1\}$ for any $t_1 \in D_1, t_1 \notin Q$
- Document $D'_2 = D_2 \cup \{q_2\} \setminus \{t_2\}$ for any $t_2 \in D_2, t_2 \notin Q$

IF $\text{Score}(Q, D_1) = \text{Score}(Q, D_2)$ THEN $\text{Score}(Q, D'_1) < \text{Score}(Q, D'_2)$
A Brief Tour of Axiomatic IR

Axiomatic Analysis

- BM25 (no matter the parameter setting) violates the LB2 constraint
- A minor modification corrects this, for consistently better performance
  
  [Lv and Zhai, CIKM’11]

\[
BM25\ (Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}
\]
A Brief Tour of Axiomatic IR

Axiomatic Analysis

- BM25 (no matter the parameter setting) violates the LB2 constraint
- A minor modification corrects this, for consistently better performance

[Lv and Zhai, CIKM’11]

\[ BM25^+(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \left( \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} + \delta \right) \]
A Brief Tour of Axiomatic IR

Axiomatic Analysis

- BM25 (no matter the parameter setting) violates the LB2 constraint
- A minor modification corrects this, for consistently better performance

[Lv and Zhai, CIKM’11]

\[
BM25^+(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \left( \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} + \delta \right)
\]

Our research question:
How can we automate the “axiomatization” of retrieval models?
Axiomatic Result Re-ranking
Axiomatic Result Re-ranking

Axiomatic Re-ranking Pipeline

1. Retrieve an initial top-$k$ result set.
2. (a) Compute re-ranking preferences of various axioms.
   (b) Aggregate re-ranking preferences.
3. Re-rank the initial result set.

Axiom 1: TFC1

Axiom 2: ORIG

Meta learning of axiom impact

KwikSort algorithm

$1' \rightarrow \rightarrow \cdots \rightarrow k'$
Axiomatic Result Re-ranking

Axiomatic Re-ranking Pipeline

1. Retrieve an initial top-\(k\) result set.
Axiomatic Result Re-ranking
Axiomatic Re-ranking Pipeline

1. Retrieve an initial top-\(k\) result set.

2. (a) Compute re-ranking preferences of various axioms.
Axiomatic Result Re-ranking
Axiomatic Re-ranking Pipeline

1. Retrieve an initial top-\(k\) result set.

2. (a) Compute re-ranking preferences of various axioms.
   (b) Aggregate re-ranking preferences.
Axiomatic Result Re-ranking

Axiomatic Re-ranking Pipeline

1. Retrieve an initial top-$k$ result set.

2. (a) Compute re-ranking preferences of various axioms.
   (b) Aggregate re-ranking preferences.

3. Re-rank the initial result set.
Axiomatic Result Re-ranking
Axiomatic Re-ranking Pipeline

1. Retrieve an initial top-$k$ result set.

2. (a) Compute re-ranking preferences of various axioms.
   (b) Aggregate re-ranking preferences.

3. Re-rank the initial result set.
Axiomatic Result Re-ranking

Requirements on Axioms

Re-state each axiom as a triple:

\[ A = (\text{precondition}, \text{filter}, \text{conclusion}) \]
Re-state each axiom as a triple:

\[ A = (\text{precondition}, \text{filter}, \text{conclusion}) \]

Given axiom \( A \) and document pair \( D_1, D_2 \):

- The \textit{precondition} evaluates whether or not \( A \) can be applied to \( D_1, D_2 \)
- If the \textit{filter} condition is satisfied...
- The \textit{conclusion} is a ranking preference of the form \( D_1 >_A D_2 \)
Axiomatic Result Re-ranking

Adapting Existing Axioms

1. Convert to our triple formulation
2. Relax equality constraints and tighten inequality constraints
3. Modify conclusion to express a ranking preference
Axiomatic Result Re-ranking

Adapting Existing Axioms

1. Convert to our triple formulation
2. Relax equality constraints and tighten inequality constraints
3. Modify conclusion to express a ranking preference

Example: TFC1

Given:

- Single-term query $Q = \{q\}$
- Documents $D_1, D_2$ with $|D_1| = |D_2|$

IF $TF(q, D_1) > TF(q, D_2)$ THEN $Score(Q, D_1) > Score(Q, D_2)$
Axiomatic Result Re-ranking
Adapting Existing Axioms

1. Convert to our triple formulation
2. Relax equality constraints and tighten inequality constraints
3. Modify conclusion to express a ranking preference

Example: TFC1

Given:
- Single-term query \( Q = \{ q \} \)
- Documents \( D_1, D_2 \) with \( |D_1| = |D_2| \)

\[
\text{IF} \quad TF(q, D_1) > TF(q, D_2) \quad \text{THEN} \quad \text{Score}(Q, D_1) > \text{Score}(Q, D_2)
\]

Precondition \( := |D_1| \approx_{10\%} |D_2| \)

Filter \( := TF(q, D_1) >_{10\%} TF(q, D_2) \)

Conclusion \( := D_1 >_{TFC1} D_2 \)
### Axiomatic Result Re-ranking

#### Adapting Existing Axioms

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Acronyms</th>
<th>Adapted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term frequency</td>
<td>TFC1–TFC3</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>TDC</td>
<td>✓</td>
</tr>
<tr>
<td>Document length</td>
<td>LNC1 + LNC2</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>TF-LNC</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>QLNC</td>
<td>✗</td>
</tr>
<tr>
<td>Lower bound</td>
<td>LB1 + LB2</td>
<td>✓</td>
</tr>
<tr>
<td>Query aspects</td>
<td>REG</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>DIV</td>
<td>✓</td>
</tr>
<tr>
<td>Semantic similarity</td>
<td>STMC1 + STMC2</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>STMC3</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>TSSC1 + TSSC2</td>
<td>✗</td>
</tr>
<tr>
<td>Term proximity</td>
<td>PHC + CCC</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>QPHRA</td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>PROX1–5</td>
<td>New</td>
</tr>
<tr>
<td>Other</td>
<td>ORIG</td>
<td>New</td>
</tr>
</tbody>
</table>
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents $D_1$, $D_2$ and multi-term query $Q = \{q_1, q_2, \ldots, q_n\}$

Precondition: both documents contain all query terms.

Give preference to the document where:

PROX1  Query term pairs are closer together on average.

PROX2  Query terms first occur earlier in the document.

PROX3  The whole query as a phrase occurs earlier in the document.

PROX4  The number of non-query terms in the closest grouping of all query terms is smaller.

PROX5  The average shortest text span containing all query terms is smaller.
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents $D_1$, $D_2$ and multi-term query $Q = \{q_1, q_2, \ldots, q_n\}$

**PROX1** Query term pairs are closer together on average.
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents $D_1, D_2$ and multi-term query $Q = \{q_1, q_2, \ldots, q_n\}$

PROX1 Query term pairs are closer together on average.

Consider query term pairs $P = \{(i, j) \mid q_i, q_j \in Q, i < j\}$
Axiomatic Result Re-ranking
New Term Proximity Axioms

Given two documents $D_1$, $D_2$ and multi-term query \( Q = \{q_1, q_2, \ldots, q_n\} \)

PROX1 Query term pairs are closer together on average.

Consider query term pairs \( P = \{(i, j) \mid q_i, q_j \in Q, i < j\} \)

Precondition := both documents contain all query terms

\[ Q = \{1, 2, 3\} \quad P = \{(1,2), (1,3), (2,3)\} \]
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents $D_1$, $D_2$ and multi-term query $Q = \{q_1, q_2, \ldots, q_n\}$

PROX1 Query term pairs are closer together on average.

Consider query term pairs $P = \{(i, j) \mid q_i, q_j \in Q, i < j\}$ and count $\delta(D, i, j)$ average number of words between $q_i, q_j$ in $D$

Precondition $\implies$ both documents contain all query terms

Filter $\implies \pi(Q, D_1) < \pi(Q, D_2)$ where $\pi(Q, D) = \frac{1}{|P|} \sum_{(i,j) \in P} \delta(D, i, j)$

<table>
<thead>
<tr>
<th>Q</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P = {(1,2), (1,3), (2,3)}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\pi(Q, D_1)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| D_1 | | | | | |
| D_2 | | | | | |

$1/3 (1 + 3 + 1) = 5/3$
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents $D_1$, $D_2$ and multi-term query $Q = \{q_1, q_2, \ldots, q_n\}$

PROX1  Query term pairs are closer together on average.

Consider query term pairs $P = \{(i, j) \mid q_i, q_j \in Q, i < j\}$ and count $\delta(D, i, j)$ average number of words between $q_i, q_j$ in $D$

Precondition  :=  both documents contain all query terms

Filter  :=  $\pi(Q, D_1) < \pi(Q, D_2)$ where $\pi(Q, D) = \frac{1}{|P|} \sum_{(i,j) \in P} \delta(D, i, j)$

\[
\begin{array}{c|c|c|c|c}
Q & 1 & 2 & 3 \\
\hline
D_1 & Y & G & B & W & W & W \\
\hline
D_2 & Y & Y & G & B & W & W \\
\end{array}
\]

$P = \{(1,2), (1,3), (2,3)\}$  \hspace{1cm} $\pi(Q,D_i)$

$\frac{1}{3} \left( 1 + 3 + 1 \right) = \frac{5}{3}$
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents \( D_1, D_2 \) and multi-term query \( Q = \{q_1, q_2, \ldots, q_n\} \)

PROX1 Query term pairs are closer together on average.

Consider query term pairs \( P = \{(i, j) \mid q_i, q_j \in Q, i < j\} \) and count \( \delta(D, i, j) \)
average number of words between \( q_i, q_j \) in \( D \)

Precondition \( := \) both documents contain all query terms

Filter \( := \pi(Q, D_1) < \pi(Q, D_2) \) where \( \pi(Q, D) = \frac{1}{|P|} \sum_{(i,j) \in P} \delta(D, i, j) \)

\[
\begin{align*}
Q & \quad 1 \quad 2 \quad 3 \\
D_1 & \quad \text{[ ] } \quad \text{[ ] } \quad \text{[ ] } \quad \text{[ ] } \quad \text{[ ] } \\
D_2 & \quad \text{[ ] } \quad \text{[ ] } \quad \text{[ ] } \quad \text{[ ] } \quad \text{[ ] } \\
\end{align*}
\]

\( P = \{ (1,2), (1,3), (2,3) \} \)

\( \pi(Q,D_i) \)

\[
\begin{align*}
1/3 \ ( 1 + 3 + 1 ) &= 5/3 \\
1/3 \ ((3+0)/2 + (4+1)/2 + 0 ) &= 4/3 \\
\end{align*}
\]
Axiomatic Result Re-ranking

New Term Proximity Axioms

Given two documents $D_1$, $D_2$ and multi-term query $Q = \{q_1, q_2, \ldots, q_n\}$

PROX1 Query term pairs are closer together on average.

Consider query term pairs $P = \{(i,j) \mid q_i, q_j \in Q, i < j\}$ and count $\delta(D, i, j)$ average number of words between $q_i, q_j$ in $D$

Precondition := both documents contain all query terms

Filter := $\pi(Q, D_1) < \pi(Q, D_2)$ where $\pi(Q, D) = \frac{1}{|P|} \sum_{(i,j) \in P} \delta(D, i, j)$

Conclusion := $D_1 >_{PROX1} D_2$

$$\pi(Q, D_1) = \frac{1}{3} \left( 1 + 3 + 1 \right) = \frac{5}{3}$$

$$\pi(Q, D_2) = \frac{1}{3} \left( \frac{3+0}{2} + \frac{4+1}{2} + 0 \right) = \frac{4}{3}$$

$D_1 >_{PROX1} D_2 = 0$

$D_2 >_{PROX1} D_1 = 1$
An initial result set \( \{D_1, \ldots, D_k\} \) and axiom \( A \) yield a \( k \)-by-\( k \) preference matrix

\[
M_A[i, j] = \begin{cases} 
1 & \text{if } D_i \succ_A D_j, \\
0 & \text{otherwise.}
\end{cases}
\]
Axiomatic Result Re-ranking
Axiom Preference Aggregation

An initial result set \( \{D_1, \ldots, D_k\} \) and axiom \( A \) yield a \( k \)-by-\( k \) preference matrix

\[
M_A[i, j] = \begin{cases} 
1 & \text{if } D_i >_A D_j, \\
0 & \text{otherwise}.
\end{cases}
\]

**Axiom 1: TFC1**

\[
M_1 = \begin{bmatrix} 
1 & 2 & 3 & \ldots & k \\
1 & 0 & 1 & 1 \\
2 & 1 & 0 \\
3 & 1 \\
\vdots & \vdots \\
k & 
\end{bmatrix}
\]

Meta learning of axiom impact
KwikSort algorithm

\( f_R(M_1', \ldots, M_{23}') \)
Axiomatic Result Re-ranking
Axiom Preference Aggregation

An initial result set \( \{D_1, \ldots, D_k\} \) and axiom \( A \) yield a \( k \)-by-\( k \) preference matrix

\[
M_A[i,j] = \begin{cases} 
1 & \text{if } D_i >_A D_j, \\
0 & \text{otherwise}.
\end{cases}
\]

\( A_1 : TFC1 \)

\[
M_1 = \begin{bmatrix}
1 & 0 & 1 & 1 \\
2 & 1 & 0 & 1 \\
3 & . & 1 & . \\
. & . & . & . \\
k & . & . & .
\end{bmatrix}
\]

\( A_{23} : ORIG \)

\[
M_{23} = \begin{bmatrix}
1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 \\
3 & . & 1 & . \\
. & . & . & . \\
k & . & . & .
\end{bmatrix}
\]
For a set of $m$ axioms $\{A_1, \ldots A_m\}$, aggregate the individual preference matrices and use the aggregate preference matrix for re-ranking.
Axiomatic Result Re-ranking
Axiom Preference Aggregation

For a set of $m$ axioms $\{A_1, \ldots, A_m\}$, aggregate the individual preference matrices and use the aggregate preference matrix for re-ranking.

Hypothesis: Different basis retrieval models will deviate from the axiomatic constraints in different ways.
Axiomatic Result Re-ranking
Axiom Preference Aggregation

For a set of $m$ axioms $\{A_1, \ldots A_m\}$, aggregate the individual preference matrices and use the aggregate preference matrix for re-ranking.

Hypothesis: Different basis retrieval models will deviate from the axiomatic constraints in different ways.

Approach: Given a set of queries with known relevance judgments, learn a retrieval-model-specific aggregation function that optimizes the average retrieval performance of the re-ranking.
Axiomatic Result Re-ranking

Frame as Classification Problem:

- Individual axiom preferences as predictors
Axiomatic Result Re-ranking

Axiom Preference Aggregation

Frame as Classification Problem:

- Individual axiom preferences as predictors
- Relative document relevance as response
Axiomatic Result Re-ranking
Axiom Preference Aggregation

Frame as Classification Problem:
- Individual axiom preferences as predictors
- Relative document relevance as response
- One training example per document pair
The aggregated preference matrix may contain contradictions

E.g. $M[i, j] = M[j, i]$

Need to solve rank-aggregation problem at this step

We use a Kemeny rank-aggregation scheme [Kemeny; 1959]

Solve using the KwikSort approximation algorithm [Ailon, Charikar, Newman; 2008]
Experimental Evaluation
Experimental Evaluation
Impact of Axiomatic Reranking on Web Track Performance

- Consider 16 basis retrieval models implemented in the Terrier\(^1\) framework
- Index the ClueWeb09 corpus using each basis retrieval model
- Retrieve top 50 results and re-rank
- 120 queries from TREC Web tracks 2009–2014 as training set
- 60 queries as test set
- Measure difference in nNDCG@10 using
  - Axiomatic reranking (AX)
  - Markov Random Field term dependency (MRF)
  - Both (MRF+AX)

\(^1\)http://terrier.org
## Experimental Evaluation

**Average nDCG@10 on Test Set (n=60) Using Top-50 Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPH</td>
<td>0.273</td>
</tr>
<tr>
<td>DFRee</td>
<td>0.205</td>
</tr>
<tr>
<td>In_expC2</td>
<td>0.205</td>
</tr>
<tr>
<td>TF_IDF</td>
<td>0.202</td>
</tr>
<tr>
<td>In_expB2</td>
<td>0.201</td>
</tr>
<tr>
<td>DFReeKLIM</td>
<td>0.199</td>
</tr>
<tr>
<td>BM25</td>
<td>0.198</td>
</tr>
<tr>
<td>InL2</td>
<td>0.197</td>
</tr>
<tr>
<td>BB2</td>
<td>0.195</td>
</tr>
<tr>
<td>DFR_BM25</td>
<td>0.194</td>
</tr>
<tr>
<td>LemurTF_IDF</td>
<td>0.187</td>
</tr>
<tr>
<td>DLH13</td>
<td>0.164</td>
</tr>
<tr>
<td>PL2</td>
<td>0.160</td>
</tr>
<tr>
<td>DLH</td>
<td>0.153</td>
</tr>
<tr>
<td>DirichletLM</td>
<td>0.139</td>
</tr>
<tr>
<td>Hiemstra_LM</td>
<td>0.107</td>
</tr>
</tbody>
</table>
# Experimental Evaluation

Average nDCG@10 on Test Set (n=60) Using Top-50 Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Basis</th>
<th>AX</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPH</td>
<td>0.273</td>
<td><strong>0.291</strong></td>
</tr>
<tr>
<td>DFRee</td>
<td>0.205</td>
<td><strong>0.236</strong></td>
</tr>
<tr>
<td>In_expC2</td>
<td>0.205</td>
<td><strong>0.214</strong></td>
</tr>
<tr>
<td>TF_IDF</td>
<td>0.202</td>
<td><strong>0.228</strong></td>
</tr>
<tr>
<td>In_expB2</td>
<td>0.201</td>
<td>0.202</td>
</tr>
<tr>
<td>DFReeKLIM</td>
<td>0.199</td>
<td><strong>0.213</strong></td>
</tr>
<tr>
<td>BM25</td>
<td>0.198</td>
<td>0.188</td>
</tr>
<tr>
<td>InL2</td>
<td>0.197</td>
<td>0.197</td>
</tr>
<tr>
<td>BB2</td>
<td>0.195</td>
<td>0.197</td>
</tr>
<tr>
<td>DFR_BM25</td>
<td>0.194</td>
<td><strong>0.206</strong></td>
</tr>
<tr>
<td>LemurTF_IDF</td>
<td>0.187</td>
<td><strong>0.224</strong></td>
</tr>
<tr>
<td>DLH13</td>
<td>0.164</td>
<td><strong>0.187</strong></td>
</tr>
<tr>
<td>PL2</td>
<td>0.160</td>
<td><strong>0.213</strong></td>
</tr>
<tr>
<td>DLH</td>
<td>0.153</td>
<td>0.187</td>
</tr>
<tr>
<td>DirichletLM</td>
<td>0.139</td>
<td><strong>0.242</strong></td>
</tr>
<tr>
<td>Hiemstra_LM</td>
<td>0.107</td>
<td>0.167</td>
</tr>
<tr>
<td>(# higher)</td>
<td></td>
<td>14</td>
</tr>
</tbody>
</table>
## Experimental Evaluation

Average nDCG@10 on Test Set (n=60) Using Top-50 Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Basis</th>
<th>AX</th>
<th>MRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPH</td>
<td>0.273</td>
<td>0.291</td>
<td>0.307</td>
</tr>
<tr>
<td>DFRee</td>
<td>0.205</td>
<td>0.236</td>
<td>0.230</td>
</tr>
<tr>
<td>In_expC2</td>
<td>0.205</td>
<td>0.214</td>
<td>0.229</td>
</tr>
<tr>
<td>TF_IDF</td>
<td>0.202</td>
<td>0.228</td>
<td>0.239</td>
</tr>
<tr>
<td>In_expB2</td>
<td>0.201</td>
<td>0.202</td>
<td>0.234</td>
</tr>
<tr>
<td>DFReeKLIM</td>
<td>0.199</td>
<td>0.213</td>
<td>0.224</td>
</tr>
<tr>
<td>BM25</td>
<td>0.198</td>
<td>0.188</td>
<td>0.229</td>
</tr>
<tr>
<td>InL2</td>
<td>0.197</td>
<td>0.197</td>
<td>0.235</td>
</tr>
<tr>
<td>BB2</td>
<td>0.195</td>
<td>0.197</td>
<td>0.236</td>
</tr>
<tr>
<td>DFR_BM25</td>
<td>0.194</td>
<td>0.206</td>
<td>0.236</td>
</tr>
<tr>
<td>LemurTF_IDF</td>
<td>0.187</td>
<td>0.224</td>
<td>0.221</td>
</tr>
<tr>
<td>DLH13</td>
<td>0.164</td>
<td>0.187</td>
<td>0.184</td>
</tr>
<tr>
<td>PL2</td>
<td>0.160</td>
<td>0.213</td>
<td>0.190</td>
</tr>
<tr>
<td>DLH</td>
<td>0.153</td>
<td>0.187</td>
<td>0.181</td>
</tr>
<tr>
<td>DirichletLM</td>
<td>0.139</td>
<td>0.242</td>
<td>0.192</td>
</tr>
<tr>
<td>Hiemstra_LM</td>
<td>0.107</td>
<td>0.167</td>
<td>0.161</td>
</tr>
</tbody>
</table>

(# higher) 14 9 (16)
### Experimental Evaluation

Average nDCG@10 on Test Set (n=60) Using Top-50 Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Basis</th>
<th>AX</th>
<th>MRF</th>
<th>MRF+AX</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPH</td>
<td>0.273</td>
<td>0.291</td>
<td>0.307</td>
<td>0.314</td>
</tr>
<tr>
<td>DFRee</td>
<td>0.205</td>
<td>0.236</td>
<td>0.230</td>
<td>0.245</td>
</tr>
<tr>
<td>In_expC2</td>
<td>0.205</td>
<td>0.214</td>
<td>0.229</td>
<td>0.238</td>
</tr>
<tr>
<td>TF_IDF</td>
<td>0.202</td>
<td>0.228</td>
<td>0.239</td>
<td>0.200</td>
</tr>
<tr>
<td>In_expB2</td>
<td>0.201</td>
<td>0.202</td>
<td>0.234</td>
<td>0.237</td>
</tr>
<tr>
<td>DFReeKLIM</td>
<td>0.199</td>
<td>0.213</td>
<td>0.224</td>
<td>0.224</td>
</tr>
<tr>
<td>BM25</td>
<td>0.198</td>
<td>0.188</td>
<td>0.229</td>
<td>0.216</td>
</tr>
<tr>
<td>InL2</td>
<td>0.197</td>
<td>0.197</td>
<td>0.235</td>
<td>0.212</td>
</tr>
<tr>
<td>BB2</td>
<td>0.195</td>
<td>0.197</td>
<td>0.236</td>
<td>0.234</td>
</tr>
<tr>
<td>DFR_BM25</td>
<td>0.194</td>
<td>0.206</td>
<td>0.236</td>
<td>0.220</td>
</tr>
<tr>
<td>LemurTF_IDF</td>
<td>0.187</td>
<td>0.224</td>
<td>0.221</td>
<td>0.237</td>
</tr>
<tr>
<td>DLH13</td>
<td>0.164</td>
<td>0.187</td>
<td>0.184</td>
<td>0.201</td>
</tr>
<tr>
<td>PL2</td>
<td>0.160</td>
<td>0.213</td>
<td>0.190</td>
<td>0.211</td>
</tr>
<tr>
<td>DLH</td>
<td>0.153</td>
<td>0.187</td>
<td>0.181</td>
<td>0.197</td>
</tr>
<tr>
<td>DirichletLM</td>
<td>0.139</td>
<td>0.242</td>
<td>0.192</td>
<td>0.253</td>
</tr>
<tr>
<td>Hiemstra_LM</td>
<td>0.107</td>
<td>0.167</td>
<td>0.161</td>
<td>0.163</td>
</tr>
<tr>
<td>(# higher)</td>
<td>14</td>
<td>9 (16)</td>
<td></td>
<td>10 (15)</td>
</tr>
</tbody>
</table>
Experimental Evaluation

Average nDCG@10 on Test Set (n=60) Using Top-50 Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Basis</th>
<th>AX</th>
<th>MRF</th>
<th>MRF+AX</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPH</td>
<td>0.273</td>
<td>0.291</td>
<td>0.307</td>
<td>0.314</td>
<td>0.642</td>
</tr>
<tr>
<td>DFRee</td>
<td>0.205</td>
<td>0.236</td>
<td>0.230</td>
<td>0.245</td>
<td>0.599</td>
</tr>
<tr>
<td>In_expC2</td>
<td>0.205</td>
<td>0.214</td>
<td>0.229</td>
<td>0.238</td>
<td>0.591</td>
</tr>
<tr>
<td>TF_IDF</td>
<td>0.202</td>
<td>0.228</td>
<td>0.239</td>
<td>0.200</td>
<td>0.589</td>
</tr>
<tr>
<td>In_expB2</td>
<td>0.201</td>
<td>0.202</td>
<td>0.234</td>
<td>0.237</td>
<td>0.592</td>
</tr>
<tr>
<td>DFReeKLIM</td>
<td>0.199</td>
<td>0.213</td>
<td>0.224</td>
<td>0.224</td>
<td>0.591</td>
</tr>
<tr>
<td>BM25</td>
<td>0.198</td>
<td>0.188</td>
<td>0.229</td>
<td>0.216</td>
<td>0.587</td>
</tr>
<tr>
<td>InL2</td>
<td>0.197</td>
<td>0.197</td>
<td>0.235</td>
<td>0.212</td>
<td>0.593</td>
</tr>
<tr>
<td>BB2</td>
<td>0.195</td>
<td>0.197</td>
<td>0.236</td>
<td>0.234</td>
<td>0.587</td>
</tr>
<tr>
<td>DFR_BM25</td>
<td>0.194</td>
<td>0.206</td>
<td>0.236</td>
<td>0.220</td>
<td>0.591</td>
</tr>
<tr>
<td>LemurTF_IDF</td>
<td>0.187</td>
<td>0.224</td>
<td>0.221</td>
<td>0.237</td>
<td>0.576</td>
</tr>
<tr>
<td>DLH13</td>
<td>0.164</td>
<td>0.187</td>
<td>0.184</td>
<td>0.201</td>
<td>0.499</td>
</tr>
<tr>
<td>PL2</td>
<td>0.160</td>
<td>0.213</td>
<td>0.190</td>
<td>0.211</td>
<td>0.550</td>
</tr>
<tr>
<td>DLH</td>
<td>0.153</td>
<td>0.187</td>
<td>0.181</td>
<td>0.197</td>
<td>0.470</td>
</tr>
<tr>
<td>DirichletLM</td>
<td>0.139</td>
<td>0.242</td>
<td>0.192</td>
<td>0.253</td>
<td>0.564</td>
</tr>
<tr>
<td>Hiemstra_LM</td>
<td>0.107</td>
<td>0.167</td>
<td>0.161</td>
<td>0.163</td>
<td>0.397</td>
</tr>
</tbody>
</table>

(# higher) 14 9 (16) 10 (15) 16
Conclusions
Conclusions

Summary

- Axiom-based framework for re-ranking
- Incorporating axiomatic ideas into the retrieval process directly
- New axioms for modeling term proximity preferences
Conclusions

Summary

- Axiom-based framework for re-ranking
- Incorporating axiomatic ideas into the retrieval process directly
- New axioms for modeling term proximity preferences

Future Work

- Formulate further axiomatic ideas in a way suitable for re-ranking (e.g. for near-duplicates)
- Comparison with learning-to-rank algorithms
- Improve efficiency of the re-ranking process
Conclusions

Summary

- Axiom-based framework for re-ranking
- Incorporating axiomatic ideas into the retrieval process directly
- New axioms for modeling term proximity preferences

Future Work

- Formulate further axiomatic ideas in a way suitable for re-ranking (e.g. for near-duplicates)
- Comparison with learning-to-rank algorithms
- Improve efficiency of the re-ranking process

Thank you!