Models for Metasearch

Javed Aslam
The Metasearch Problem

Search for: chili peppers
Search Engines

- Provide a ranked list of documents.
- May provide relevance scores.
- May have performance information.
Search Engine: Alta Vista
Search Engine: Ultraseek
Search Engine: inq102 TREC3

Queryid (Num): 50
Total number of documents over all queries
Retrieved: 50000
Relevant: 9805
Rel_ret: 7305

Interpolated Recall - Precision Averages:
at 0.00  0.8992
at 0.10  0.7514
at 0.20  0.6584
at 0.30  0.5724
at 0.40  0.4982
at 0.50  0.4272
at 0.60  0.3521
at 0.70  0.2915
at 0.80  0.2173
at 0.90  0.1336
at 1.00  0.0115

Average precision (non-interpolated)
for all rel docs (averaged over queries)
0.4226

Precision:
At 5 docs: 0.7440
At 10 docs: 0.7220
At 15 docs: 0.6867
At 20 docs: 0.6740
At 30 docs: 0.6267
At 100 docs: 0.4902
At 200 docs: 0.3848
At 500 docs: 0.2401
At 1000 docs: 0.1461

R-Precision (precision after R
(= num_rel for a query) docs retrieved):
Exact: 0.4524
External Metasearch

Metasearch Engine

Search Engine A

Database A

Search Engine B

Database B

Search Engine C

Database C
Internal Metasearch

Search Engine

- Text Module
- URL Module
- Image Module

Metasearch core

HTML Database

Image Database
Outline

- Introduce problem
- Characterize problem
- Survey current techniques
- Describe new approaches
  - decision theory, social choice theory
  - experiments with TREC data
- Upper bounds for metasearch
- Future work
# Classes of Metasearch Problems

<table>
<thead>
<tr>
<th>Relevance Scores</th>
<th>No Training Data</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranks only</td>
<td>Borda, Condorcet, rCombMNZ</td>
<td>Bayes</td>
</tr>
<tr>
<td>CombMNZ</td>
<td>LC model</td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Introduce problem
- Characterize problem
  - Survey current techniques
  - Describe new approaches
    - decision theory, social choice theory
    - experiments with TREC data
  - Upper bounds for metasearch
- Future work
Classes of Metasearch Problems

- Borda, Condorcet, rCombMNZ
- Bayes
- CombMNZ
- LC model
CombsUM [Fox, Shaw, Lee, et al.]

- Normalize scores: [0,1].
- For each doc:
  - sum relevance scores given to it by each system (use 0 if unretrieved).
- Rank documents by score.
- Variants: MIN, MAX, MED, ANZ, MNZ
CombMNZ [Fox, Shaw, Lee, et al.]

- Normalize scores: [0,1].
- For each doc:
  - sum relevance scores given to it by each system (use 0 if unretrieved), and
  - multiply by number of systems that retrieved it (MNZ).
- Rank documents by score.
How well do they perform?

- Need *performance metric*.
- Need *benchmark data*. 
Metric: Average Precision

\[
\begin{align*}
R & \quad 1/1 \\
N & \quad 2/3 \\
R & \quad 3/5 \\
N & \quad 4/8 \\
R & \quad 0.6917
\end{align*}
\]
Benchmark Data: TREC

- Annual *Text Retrieval Conference*.
- Millions of documents (AP, NYT, etc.)
- 50 queries.
- Dozens of retrieval engines.
- Output lists available.
- Relevance judgments available.
## Data Sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number systems</th>
<th>Number queries</th>
<th>Number of docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC3</td>
<td>40</td>
<td>50</td>
<td>1000</td>
</tr>
<tr>
<td>TREC5</td>
<td>61</td>
<td>50</td>
<td>1000</td>
</tr>
<tr>
<td>Vogt</td>
<td>10</td>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>TREC9</td>
<td>105</td>
<td>50</td>
<td>1000</td>
</tr>
</tbody>
</table>
CombX on TREC5 Data

TREC 5: Combining the top i systems in order.
Experiments

- Randomly choose \( n \) input systems.
- For each query:
  - combine, trim, calculate avg precision.
- Calculate mean avg precision.
- Note best input system.
- Repeat (statistical significance).
CombMNZ on TREC5

TREC 5: avg precision over 200 random sets of systems.
Outline

- Introduce problem
- Characterize problem
- Survey current techniques
  - Describe new approaches
    - decision theory, social choice theory
    - experiments with TREC data
  - Upper bounds for metasearch
  - Future work
New Approaches [Aslam, Montague]

- Analog to *decision theory*.
  - Requires only rank information.
  - Training required.
- Analog to *election strategies*.
  - Requires only rank information.
  - No training required.
## Classes of Metasearch Problems

<table>
<thead>
<tr>
<th>Relevance Scores</th>
<th>No Training Data</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranks Only</td>
<td>Borda, Condorcet, rCombMNZ</td>
<td>Bayes</td>
</tr>
<tr>
<td></td>
<td>CombMNZ</td>
<td>LC model</td>
</tr>
</tbody>
</table>
Decision Theory

- Consider two alternative explanations for some observed data.
  - Medical example:
    - Perform a set of blood tests.
    - Does patient have disease or not?

- Optimal method for choosing among the explanations: *likelihood ratio test.*
  [Neyman-Pearson Lemma]
Metasearch via Decision Theory

- Metasearch analogy:
  - *Observed data* – document rank info over all systems.
  - *Hypotheses* – document is relevant or not.

- Ratio test:
  \[ O_{rel} = \frac{\Pr[rel \mid r_1, r_2, \ldots, r_n]}{\Pr[irr \mid r_1, r_2, \ldots, r_n]} \]
Bayesian Analysis

\[ P_{rel} = \Pr[rel \mid r_1, r_2, \ldots, r_n] \]
\[ P_{rel} = \frac{\Pr[r_1, r_2, \ldots, r_n \mid rel] \cdot \Pr[rel]}{\Pr[r_1, r_2, \ldots, r_n]} \]
\[ O_{rel} = \frac{\Pr[r_1, r_2, \ldots, r_n \mid rel] \cdot \Pr[rel]}{\Pr[r_1, r_2, \ldots, r_n \mid irr] \cdot \Pr[irr]} \]
\[ O_{rel} \cong \frac{\Pr[rel] \cdot \prod_i \Pr[r_i \mid rel]}{\Pr[irr] \cdot \prod_i \Pr[r_i \mid irr]} \]
\[ LO_{rel} \sim \sum_i \log \frac{\Pr[r_i \mid rel]}{\Pr[r_i \mid irr]} \]
Bayes on TREC3

TREC 3: avg precision over 200 random sets of systems.

- Bayes-fuse
- CombMNZ
- The best input system

Avg precision

Number of randomly chosen input systems
Bayes on TREC5

TREC 5: avg precision over 200 random sets of systems.

![Graph showing avg precision over 200 random sets of systems.](image)
Bayes on TREC9

TREC 9: avg precision over 200 random sets of systems.

- Bayes-fuse
- CombMNZ
- The best input system

Number of randomly chosen input systems

Avg precision
Beautiful theory, but...

*In theory, there is no difference between theory and practice; in practice, there is.*

—variably: Chuck Reid, Yogi Berra

**Issue:** *independence assumption*...
Naïve-Bayes Assumption

\[ O_{rel} = \frac{\Pr[r_1, r_2, \ldots, r_n \mid rel] \cdot \Pr[rel]}{\Pr[r_1, r_2, \ldots, r_n \mid irr] \cdot \Pr[irr]} \]

\[ O_{rel} \approx \frac{\Pr[rel] \cdot \prod_i \Pr[r_i \mid rel]}{\Pr[irr] \cdot \prod_i \Pr[r_i \mid irr]} \]
Bayes on Vogt Data

TREC 5 subset: avg precision over between 1 and 200 random sets of systems.
New Approaches [Aslam, Montague]

- Analog to decision theory.
  - Requires only rank information.
  - Training required.

- Analog to election strategies.
  - Requires only rank information.
  - No training required.
Classes of Metasearch Problems

<table>
<thead>
<tr>
<th></th>
<th>no training data</th>
<th>training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranks only</td>
<td>Borda, Condorcet, rCombMNZ</td>
<td>Bayes</td>
</tr>
<tr>
<td>relevance scores</td>
<td>CombMNZ</td>
<td>LC model</td>
</tr>
</tbody>
</table>

CombMNZ LC model
Election Strategies

- Plurality vote.
- Approval vote.
- Run-off.
- Preferential rankings:
  - instant run-off,
  - Borda count (positional),
  - Condorcet method (head-to-head).
Metasearch Analogy

- Documents are *candidates*.
- Systems are *voters* expressing preferential rankings among candidates.
Condorcet Voting

- Each ballot ranks all candidates.
- Simulate head-to-head run-off between each pair of candidates.
- Condorcet winner: candidate that beats all other candidates, head-to-head.
Condorcet Paradox

- Voter 1: A, B, C
- Voter 2: B, C, A
- Voter 3: C, A, B
- Cyclic preferences: cycle in Condorcet graph.
- Condorcet consistent path: Hamiltonian.
- For metasearch: any CC path will do.
Condorcet Consistent Path
Hamiltonian Path Proof

Base Case:

Inductive Step:
Condorcet-fuse: Sorting

- Insertion-sort suggested by proof.
- Quicksort too; $O(n \log n)$ comparisons.
  - $n$ documents.
- Each comparison: $O(m)$.
  - $m$ input systems.
- Total: $O(m \, n \log n)$.
- Need not compute entire graph.
Condorcet-fuse on TREC3

TREC 3: avg precision over 200 random sets of systems.

- CombMNZ
- CombMNZ (relevance scores simulated with ranks, unret: 0)
- Quicksort Condorcet

Avg precision vs. Number of randomly chosen input systems
Condorcet-fuse on TREC5

TREC 5: avg precision over 200 random sets of systems.
Condorcet-fuse on Vogt
Condorcet-fuse on TREC9

TREC 9: avg precision over 200 random sets of systems.

- CombMNZ
- CombMNZ (relevance scores simulated with ranks, unret: 0)
- Quicksort Condorcet

Number of randomly chosen input systems vs. Avg precision
Breaking Cycles

SCCs are properly ordered.

How are ties within an SCC broken? (Quicksort)
Outline

- Introduce problem
- Characterize problem
- Survey current techniques
- Describe new approaches
  - decision theory, social choice theory
  - experiments with TREC data
- Upper bounds for metasearch
- Future work
Upper Bounds on Metasearch

- How good can metasearch be?
- Are there fundamental limits that methods are approaching?
- Need an analog to running time lower bounds...
Upper Bounds on Metasearch

- Constrained oracle model:
  - omniscient metasearch oracle,
  - constraints placed on oracle that any reasonable metasearch technique must obey.

- What are “reasonable” constraints?
Naïve Constraint

- **Naïve constraint:**
  - Oracle may only return docs from underlying lists.
  - Oracle may return these docs in any order.
  - Omniscient oracle will return relevant docs above irrelevant docs.
TREC5: Naïve Bound

The graph shows the average precision over 200 random sets of systems. The x-axis represents the number of randomly chosen input systems, ranging from 2 to 12. The y-axis represents the average precision, ranging from 0.3 to 1.1. The graph includes three lines:

- **Naïve Bound**: Dashed line with black dots.
- **Condorcet-fuse**: Dotted line with gray dots.
- **The best input system**: Solid line with black squares.

The Naïve Bound line starts at 0.3 and increases to approximately 0.8 as the number of input systems increases. Condorcet-fuse starts at a lower average precision and increases more gradually compared to the Naïve Bound. The best input system line shows a consistent increase, reaching close to 1.1 at the highest number of input systems.
Pareto Constraint

- Pareto constraint:
  - Oracle may only return docs from underlying lists.
  - Oracle must respect *unanimous* will of underlying systems.
  - Omniscient oracle will return relevant docs above irrelevant docs, subject to the above constraint.
TREC5: Pareto Bound
Majoritarian Constraint

- **Majoritarian constraint:**
  - Oracle may only return docs from underlying lists.
  - Oracle must respect *majority* will of underlying systems.
  - Omniscient oracle will return relevant docs above irrelevant docs and break cycles optimally, subject to the above constraint.
TREC5: Majoritarian Bound
Upper Bounds: TREC3
Upper Bounds: Vogt
Upper Bounds: TREC9
TREC8: Avg Prec vs Feedback
TREC8: System Assessments vs TREC
Metasearch Engines

- Query multiple search engines.
- May or may not combine results.
Metasearch: Dogpile

Search engine: Looksmart found 117 results.
The query string sent was "chili -peppers"

1. The Red Hot Chili Peppers
Find photos, lyrics, updates, tour info, and news on alternative-funk-rock band the Red Hot Chili Peppers.
Looksmart category – Red Hot Chili Pepper

2. Red Hot Chili Peppers Audio and Video
Watch videos and listen to music by this rock/funk band.
Looksmart category – Red Hot Chili Peppers

3. Chili and Hot Sauces
Shop for mouth-burning chili sauces, Tabasco, hot salas and other pepper-inspired sauces.
Looksmart category – Chili & Hot Sauces

4. Chili and Hot Sauces
Find chili and other hot sauce recipes, including salsas, dips, spices, and rubs, and visit the Pepper Fool.
Looksmart category – Chili & Hot Sauces

5. Red Hot Chili Peppers – Screens and Themes
Promotional screensaver for the funk-rock band features falling chili peppers.
LookSmart category – Red Hot Chili Peppers Multimedia

Search engine: GoTo.com found 10 or more results.
Metasearch: Metacrawler
Metasearch: Profusion
Characterizing Metasearch

- Three axes:
  - common vs. disjoint database,
  - relevance scores vs. ranks,
  - training data vs. no training data.
Axis 1: DB Overlap

- High overlap
  - data fusion.

- Low overlap
  - collection fusion (distributed retrieval).

- Very different techniques for each...

- This work: data fusion.
CombMNZ on TREC3

TREC 3: avg precision over 200 random sets of systems.
CombMNZ on Vogt

TREC 5 subset: avg precision over between 1 and 200 random sets of systems.
CombMNZ on TREC9

TREC 9: avg precision over 200 random sets of systems.

- CombSUM
- CombMNZ
- The best input system

Number of randomly chosen input systems
Borda Count

- Consider an \( n \) candidate election.
- For each ballot:
  - assign \( n \) points to top candidate,
  - assign \( n-1 \) points to next candidate,
  - ...
- Rank candidates by point sum.
Borda Count: Election 2000

- Ideological order: Nader, Gore, Bush.
- Ideological voting:
  - Nader voter: Nader, Gore, Bush.
  - Gore voter:
    - Gore, Bush, Nader.
    - Gore, Nader, Bush.

\[
\begin{aligned}
\{ & 50/50, 100/0 \\
\end{aligned}
\]
Election 2000: Ideological Florida Voting

<table>
<thead>
<tr>
<th></th>
<th>Gore</th>
<th>Bush</th>
<th>Nader</th>
</tr>
</thead>
<tbody>
<tr>
<td>50/50</td>
<td>14,734,379</td>
<td>13,185,542</td>
<td>7,560,864</td>
</tr>
<tr>
<td>100/0</td>
<td>14,734,379</td>
<td>14,639,267</td>
<td>6,107,138</td>
</tr>
</tbody>
</table>

Gore Wins
Borda Count: Election 2000

- Ideological order: Nader, Gore, Bush.
- Manipulative voting:
  - Gore voter: Gore, Nader, Bush.
  - Nader voter: Nader, Gore, Bush.
Election 2000: Manipulative Florida Voting

<table>
<thead>
<tr>
<th></th>
<th>Gore</th>
<th>Bush</th>
<th>Nader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Votes</td>
<td>11,825,203</td>
<td>11,731,816</td>
<td>11,923,765</td>
</tr>
</tbody>
</table>

Nader Wins
Future Work

- Bayes
  - approximate dependence.
- Condorcet
  - weighting, dependence.
- Upper bounds
  - other constraints.
- Meta-retrieval
  - Metasearch is approaching fundamental limits.
  - Need to incorporate user feedback: learning...