unified model for metasearch, pooling and system evaluation

Javed Aslam
Virgil Pavlu
Robert Savell
motivation
motivation

avg precision

best system

metasearch
better relevant rate

Hedge

RELEVANT FINDING

LOSS FUNCTION

DOCUMENT SELECTOR

JUDGE

FEEDBACK MODIFY WEIGHTS

SYSTEM WEIGHTS

% relevant found

# docs judged

Depth-n

Hedge
motivation

RELEVANT FINDINGS

better relevant rate

% relevant found

Hedge

LOSS FUNCTION
DOCUMENT SELECTOR
JUDGE
SYSTEM WEIGHTS
FEEDBACK MODIFY WEIGHTS

USER

# docs judged

Depth-n
Hedge
problem setup

- Set of underlying systems
  - On the same query
- User feedback
- Goal
  - Find relevant documents
  - Produce metasearch lists
  - Do partial system evaluation (distinction)
- We are looking for an adaptive approach
this talk

- Hedge algorithm
- The new model
- Loss function
- Pooling
- System evaluation
- Metasearch
- Experiments
online allocation - hedge algorithm
online allocation - hedge algorithm

Hedge

\[ \beta \in [0,1]^N \text{ strategies (systems)} \]

initial weights \( w^1 \in [0,1]^N; \sum_{i=1}^{N} w_i^1 = 1 \)
online allocation - hedge algorithm

\[ \beta \in [0,1]^N \text{ strategies (systems)} \]
\[ \text{initial weights } \ w^1 \in [0,1]^N; \sum_{i=1}^{N} w_i^1 = 1 \]

LOOP for episode \( t=1,2,\ldots,T \)
online allocation - hedge algorithm

\[ \beta \in [0,1] \] \( N \) strategies (systems)

\[ w^1 \in [0,1]^N; \sum_{i=1}^N w_i^1 = 1 \]

LOOP for episode \( t=1,2,\ldots,T \)

- Choose allocation

\[ p_i^t = \frac{w_i^t}{\sum_{i=1}^N w_i^t} \]
online allocation - hedge algorithm

\[ \beta \in [0,1] \text{ } N \text{ strategies (systems)} \]

initial weights \( w^1 \in [0,1]^N \); \( \sum_{i=1}^{N} w_i^1 = 1 \)

- Choose allocation
- Receive loss vector

LOOP for episode \( t = 1, 2, \ldots, T \)

\[ p_i^t = \frac{w_i^t}{\sum_{i=1}^{N} w_i^t} \]

\( l^t \in [0,1]^N \)
online allocation - hedge algorithm

LOOP for episode $t=1,2,\ldots,T$

- Choose allocation
- Receive loss vector
- Suffer loss
online allocation - hedge algorithm

LOOP for episode \( t=1,2,\ldots,T \)
- Choose allocation
- Receive loss vector
- Suffer loss
- Update weights

\[ \beta \in [0,1] \]
\( N \) strategies (systems)
initial weights \( w^1 \in [0,1]^N \); \( \sum_{i=1}^N w_i^1 = 1 \)

\[ p_i^t = \frac{w_i^t}{\sum_{j=1}^N w_j^t} \]

\[ l^t \in [0,1]^N \]

\[ p^t \propto l^t \]

\[ w_{i+1}^t = w_i^t \times \beta l_i^t \]
online allocation - hedge algorithm

LOOP for episode $t=1,2,\ldots,T$
- Choose allocation
- Receive loss vector
- Suffer loss
- Update weights

\[ L_{HEDGE} = \sum_{t=1}^{T} p_t^t \mathcal{A}^t \]

- Hedge loss

$N$ strategies (systems)
- \[ \beta \in [0,1] \]
- \[ \sum_{i=1}^{N} w_i^1 = 1 \]

\[ p_i^t = \frac{w_i^t}{\sum_{i=1}^{N} w_i^t} \]

\[ l^t \in [0,1]^N \]

\[ p^t \propto l^t \]

\[ w_i^{t+1} = w_i^t \times \beta l_i^t \]
Why hedge [schapire, freund ’96]

\[
\sum_{i=1}^{N} w_{i}^{T+1} \leq \left( \sum_{i=1}^{N} w_{i}^{T} \right) \left( 1 - (1 - \beta) p^{T} \mathcal{X}^{T} \right) \leq \ldots
\]
Why hedge [schapire, freund ’96]

\[
\sum_{i=1}^{N} w_i^{T+1} \leq \left( \sum_{i=1}^{N} w_i^{T} \right) \cdot \left( 1 - (1 - \beta) p^T \cdot \mathcal{A}^T \right) \leq \ldots
\]

*hedge loss at episode T = p^T \cdot \mathcal{A}^T*
Why hedge [schapire, freund ’96]

\[
\sum_{i=1}^{N} w_{i}^{T+1} \leq \left( \sum_{i=1}^{N} w_{i}^{T} \right) \left( 1 - (1 - \beta) p^{T} \cdot \mathcal{A}^{T} \right) \leq \ldots
\]

\[
\ldots \leq \exp\left( -(1 - \beta) \sum_{t=1}^{T} p^{t} \cdot \mathcal{A}^{t} \right)
\]

*hedge* loss at episode \( T = p^{T} \cdot \mathcal{A}^{T} \)
Why hedge [schapire, freund ’96]

\[ \sum_{i=1}^{N} w_{i}^{T+1} \leq \left( \sum_{i=1}^{N} w_{i}^{T} \right) \frac{1}{1 - (1 - \beta) p^{T} \times \mathcal{A}^{T}} \leq \ldots \]

\[ \leq \exp\left( -(1 - \beta) \sum_{t=1}^{T} p^{t} \times \mathcal{A}^{t} \right) \]

\[ \text{hedge loss at episode } T = p^{T} \times \mathcal{A}^{T} \]

\[ \text{cumulative loss } L_{HEDGE} \]
Why hedge [schapire, freund ’96]

\[
\sum_{i=1}^{N} w_i^{T+1} \leq \left( \sum_{i=1}^{N} w_i^T \right) \left( 1 - (1 - \beta) p^T \mathcal{A}^T \right) \leq \ldots
\]

\[
\ldots \leq \exp(-(1 - \beta) \sum_{t=1}^{T} p^t \mathcal{A}^t)
\]

hedge loss at episode \( T = p^T \mathcal{A}^T \)

cumulative loss \( L_{HEDGE} \)

\[
L_{HEDGE} \leq \frac{\ln\left( \frac{1}{\beta} \right) L_{SYSTEM} + \ln N}{1 - \beta}
\]
Our model

\[
\text{param } \beta \in [0,1]; \text{ } N \text{ systems; } T \text{ trials}
\]

\[
\text{init } w_s^0 = \frac{1}{N} \quad \forall s \in \{1,2,\ldots,N\};
\]

FOR \( t = 1,2,\ldots,T \):

- select \( d_t = \arg \max_{d \text{ not labeled}} \left( \sum_{s=1}^{N} w_{s}^{t-1} \ast \text{LOSS}(d,s \mid d = \text{NR}) \right) \)

- judge \( d_t \): find out \( \text{label}(d_t) \)

- apply feedback \( w_s^t = w_{s}^{t-1} \ast \beta \ast \text{LOSS}(d,s) \)

FACT: if we label all docs: \( w_s^{\text{final}} = \beta \ast C \ast (Z - 2 \ast \text{TP}(s)) \)
Our model

\[ \text{param } \beta \in [0, 1]; \text{ } N \text{ systems; } T \text{ trials} \]

\[ \text{init } w_s^0 = \frac{1}{N} \quad \forall s \in \{1, 2, \ldots, N\}; \]

\[ \text{FOR } t = 1, 2, \ldots, T : \]

- select \( d_t = \arg \max_{d \text{ not labeled}} \left[ \sum_{s=1}^{N} w_s^{t-1} \times \text{LOSS}(d, s \mid d = NR) \right] \)

- judge \( d_t \): find out label(\( d_t \))

- apply feedback \( w_s^t = w_s^{t-1} \times \beta \text{LOSS}(d,s) \)

FACT: if we label all docs: \( w_s^{\text{final}} = \beta^{C*(Z-2*TP(s))} \)
Our model

\[
\text{param } \beta \in [0,1]; \ N \text{ systems}; \ T \text{ trials}
\]

\[
\text{init } w^0_s = \frac{1}{N} \quad \forall s \in \{1,2,\ldots,N\};
\]

\[
\text{FOR } t = 1,2,\ldots,T :
\]

- select \( d_t = \arg \max_{d \text{ not labeled}} \sum_{s=1}^{N} w^{t-1}_s \times \text{LOSS}(d, s | d = \text{NR}) \)

- judge \( d_t : \text{find out } \text{label}(d_t) \)

- apply feedback \( w^t_s = w^{t-1}_s \times \beta^{\text{LOSS}(d,s)} \)

\[
\text{FACT : if we label all docs : } w^{\text{final}}_s = \beta^{C*(Z-2*TP(s))}
\]
Our model

\text{param } \beta \in [0,1]; N \text{ systems}; T \text{ trials}

\text{init } w_s^0 = \frac{1}{N} \quad \forall s \in \{1,2,...,N\};

\text{FOR } t = 1,2,...,T:\n
\begin{itemize}
  \item select \(d_t = \arg \max_{d \text{ not labeled}} \sum_{s=1}^{N} w_s^{t-1} \cdot \text{LOSS}(d, s | d = \text{NR})\)
  \item judge \(d_t\): find out \(\text{label}(d_t)\)
  \item apply feedback \(w_s^t = w_s^{t-1} \cdot \beta^{\text{LOSS}(d,s)}\)
\end{itemize}

\text{FACT: if we label all docs: } w_s^{\text{final}} = \beta^{C \cdot (Z - 2 \cdot TP(s))}
pooling - howto

- select $d_t = \arg\max_{d \text{ not labeled}} \left[ \sum_{s=1}^{N} w_{s}^{t-1} \ast \text{LOSS}(d, s | d = NR) \right]$
pooling - howto

- select \( d_t = \arg \max_{d \text{ not labeled}} \left[ \sum_{s=1}^{N} w_s^{t-1} \times \text{LOSS}(d, s \mid d = NR) \right] \)

pooling value(d)
- select $d_t = \arg \max_{\text{d not labeled}} \left[ \sum_{s=1}^{N} w_{s}^{t-1} \cdot LOSS(d, s \mid d = NR) \right]$

$LOSS(d, s \mid d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}$
- select $d_t = \arg \max_{d \text{ not labeled}} \left[ \sum_{s=1}^{N} w_s^{t-1} * LOSS(d, s \mid d = NR) \right]$

$LOSS(d, s \mid d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}$

- Naturally “want” top ranks
- If NON RELEVANT, then a NR in top ranks of the system lists
- If RELEVANT, bingo.
Loss function

\[ \text{LOSS}(d, s) = \text{label}(d) \times \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \times \ln \frac{Z}{r} \]
\[ \text{LOSS}(d,s) = \text{label}(d) \times \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \times \ln \frac{Z}{r} \]
**Loss function**

\[
\text{LOSS}(d, s) = \text{label}(d) \ast \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \ast \ln \frac{Z}{r}
\]

**if** \(d\) **is not judged**, **assuming** \(d\) **NONRELEVANT** **we write**:

\[
\text{LOSS}(d, s \mid d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}
\]
Loss function

\[
\text{LOSS}(d, s) = \text{label}(d) \times \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \times \ln \frac{Z}{r}
\]

If \( d \) is not judged, assuming \( d \) NONRELEVANT we write:

\[
\text{LOSS}(d, s \mid d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}
\]
Loss function

\[
\text{LOSS}(d, s) = \text{label}(d) \times \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \times \ln \frac{Z}{r}
\]

if \( d \) is not judged, assuming \( d \) NONRELEVANT we write:

\[
\text{LOSS}(d, s \mid d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}
\]

\( r = \) rank of doc \( d \) in system \( s \)

\( Z = \# \) of documents returned by system \( s \)

\( \text{label}(d) = -1 \) if \( d \) is RELEVANT ; \( +1 \) if \( d \) is NONRELEVANT
\[ \text{LOSS}(d, s) = \text{label}(d) \times \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \times \ln \frac{Z}{r} \]

If \( d \) is not judged, assuming \( d \) NONRELEVANT, we write:

\[ \text{LOSS}(d, s \mid d = \text{NR}) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r} \]

\( r = \) rank of doc \( d \) in system \( s \)

\( Z = \# \) of documents returned by system \( s \)

\( \text{label}(d) = -1 \) if \( d \) is RELEVANT; \( +1 \) if \( d \) is NONRELEVANT

\[ TP(s) = \text{total precision of } s = \text{average precision at all ranks} \]

\[ \text{FACT : } \sum_{\text{all docs}} \text{LOSS}(d, s) = C \times (Z - 2 \times TP(s)) \]
Loss function

\[
\text{LOSS}(d, s) = \text{label}(d) \times \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \text{label}(d) \times \ln \frac{Z}{r}
\]

if \( d \) is not judged, assuming \( d \) NONRELEVANT we write:

\[
\text{LOSS}(d, s | d = NR) = \left( \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \right) \approx \ln \frac{Z}{r}
\]

\( r = \text{rank of doc } d \text{ in system } s \)

\( Z = \# \text{ of documents returned by system } s \)

\( \text{label}(d) = -1 \text{ if } d \text{ is RELEVANT} \quad +1 \text{ if } d \text{ is NONRELEVANT} \)

\( TP(s) = \text{total precision of } s = \text{average precision at all ranks} \)

\[
\text{FACT: } \sum_{\text{all docs}} \text{LOSS}(d, s) = C \times (Z - 2 \times TP(s))
\]
“total” precision

- Average the precision at **ALL** ranks
  - Normalize so ideal system gets TP=1
- Math is more simple
  - We still work on it though
- Bad with “long tails”
pooling - comparison with Cormack
[partial] system evaluation – howto
Before the next episode
  - Assume all docs not judged (so many ?) to be NR
  - Compute AvegPrecision for every system

For comparison with depth-pooling we use average number of pools (over queries)

Two situations
  - One (or few) very good systems – use small $\beta$
  - No singles
metasearch – howto

• Before the next episode
  • Compute “pooling value” for each doc
  $$\sum_{s=1}^{N} w_s^{t-1} \ast LOSS(d, s \mid d = NR)$$
  • Instead of “select the top doc for pooling” do “select the top 1000 doc” for metasearch

• In fact almost 1000 – docs already pooled are automatically in top of metasearch list
  • Fair both ways
experiments

- TREC
  - ~100 systems
  - 50 queries each competition
- Use TREC qrels as user feedback
  - incomplete feedback
experiments

- **TREC**
  - ~100 systems
  - 50 queries each competition
- **Use TREC qrels as user feedback**
  - incomplete feedback

- **Goal**
  - Find relevant documents
  - Produce metasearch lists
  - Do partial system evaluation (distinction)
experiments - system evaluation
experiments - system evaluation
experiments - system evaluation
experiments - system evaluation
system evaluation – kendall’s tau

![Graph showing system ordering, Hedge vs Depth-n, and TREC8 with Kendall's tau values.]

Kendall's tau vs # docs judged

- Depth-n
- Hedge

Graph showing the performance comparison of different system ordering techniques against TREC8.
system evaluation – kendall’s tau
system evaluation – kendall’s tau

![Graph showing system evaluation with Kendall's tau.](image-url)
system evaluation – kendall’s tau

![Graph showing system evaluation with Kendall's tau](image)
experiments - relevant docs found

![Graph showing recall percentage vs number of documents judged for 'Depth-n' and 'Hedge' methods in TREC8 experiments.](image)
### metasearch - no feedback (yet)

<table>
<thead>
<tr>
<th>TREC</th>
<th>MNZ</th>
<th>COND</th>
<th>Hedge-0</th>
<th>%MNZ</th>
<th>%COND</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.423</td>
<td>0.403</td>
<td>0.418</td>
<td>-0.012</td>
<td>+0.037</td>
</tr>
<tr>
<td>5</td>
<td>0.294</td>
<td>0.307</td>
<td>0.309</td>
<td>+0.051</td>
<td>+0.006</td>
</tr>
<tr>
<td>6</td>
<td>0.341</td>
<td>0.315</td>
<td>0.345</td>
<td>+0.012</td>
<td>+0.095</td>
</tr>
<tr>
<td>7</td>
<td>0.320</td>
<td>0.308</td>
<td>0.323</td>
<td>+0.009</td>
<td>+0.049</td>
</tr>
<tr>
<td>8</td>
<td>0.350</td>
<td>0.343</td>
<td>0.352</td>
<td>+0.014</td>
<td>+0.026</td>
</tr>
</tbody>
</table>
metasearch - no feedback (yet)

<table>
<thead>
<tr>
<th>TREC</th>
<th>MNZ</th>
<th>COND</th>
<th>Hedge-0</th>
<th>%MNZ</th>
<th>%COND</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.423</td>
<td>0.403</td>
<td>0.418</td>
<td>-0.012</td>
<td>+0.037</td>
</tr>
<tr>
<td>5</td>
<td>0.294</td>
<td>0.307</td>
<td>0.309</td>
<td>+0.051</td>
<td>+0.006</td>
</tr>
<tr>
<td>6</td>
<td>0.341</td>
<td>0.315</td>
<td>0.345</td>
<td>+0.012</td>
<td>+0.095</td>
</tr>
<tr>
<td>7</td>
<td>0.320</td>
<td>0.308</td>
<td>0.323</td>
<td>+0.009</td>
<td>+0.049</td>
</tr>
<tr>
<td>8</td>
<td>0.350</td>
<td>0.343</td>
<td>0.352</td>
<td>+0.014</td>
<td>+0.026</td>
</tr>
</tbody>
</table>
### metasearch - no feedback (yet)

<table>
<thead>
<tr>
<th>TREC</th>
<th>MNZ</th>
<th>COND</th>
<th>Hedge-0</th>
<th>%MNZ</th>
<th>%COND</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.423</td>
<td>0.403</td>
<td>0.418</td>
<td>-0.012</td>
<td>+0.037</td>
</tr>
<tr>
<td>5</td>
<td>0.294</td>
<td>0.307</td>
<td>0.309</td>
<td>+0.051</td>
<td>+0.006</td>
</tr>
<tr>
<td>6</td>
<td>0.341</td>
<td>0.315</td>
<td>0.345</td>
<td>+0.012</td>
<td>+0.095</td>
</tr>
<tr>
<td>7</td>
<td>0.320</td>
<td>0.308</td>
<td>0.323</td>
<td>+0.009</td>
<td>+0.049</td>
</tr>
<tr>
<td>8</td>
<td>0.350</td>
<td>0.343</td>
<td>0.352</td>
<td>+0.014</td>
<td>+0.026</td>
</tr>
</tbody>
</table>

**no relevant judgements**
experiments - metasearch

![Graph showing metasearch performance for Hedge and TREC8 with MAP on the y-axis and # docs judged on the x-axis. The graph compares Hedge (solid blue line), CombMNZ (dashed pink line), and the best system (green dotted line).]
experiments - metasearch

Metasearch performance  Hedge  TREC8

MAP

# docs judged
conclusion

- A powerful machine learning approach
  - Hedge = AdaBoost core
- Works [usually] better than anything else we’ve seen
- True, it uses feedback
  - But without feedback there are provable limitations
- It is missing a rigorous analysis
  - We are not very far away with that
  - Need a model assumption
differentiate (classify) the search engines

<table>
<thead>
<tr>
<th>RANK</th>
<th>SYS1</th>
<th>SYS2</th>
<th>SYS3</th>
<th>SYS129</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>111</td>
<td>2</td>
<td>1013</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>821</td>
<td>466</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>134</td>
<td>84</td>
</tr>
<tr>
<td>4</td>
<td>134</td>
<td>7284</td>
<td>9725</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>14215</td>
<td>769</td>
<td>106</td>
<td>2038</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>84</td>
<td>439</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>84=RELEVANT</td>
<td>821=NON RELEVANT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TREC8 competition
129 search engines (systems)
~1000 docs returned /query
TREC Evaluation DEPTH 100
-judge top 100 docs (each sys); use MAP to rank systems

the big picture

fuse the lists (metasearch)

toronto sigir

LOSS FUNCTION

DOCUMENT SELECTOR

SYSTEM WEIGHTS

JUDGE

FEEDBACK MODIFY WEIGHTS

USER

Hedge

KEY: FIND RELEVANT DOCS

1. SIGIR 2003
2. Information Retrieval Conferences
3. SIGIR Information Server
4. SIGIR Upcoming events
5. PROPOSAL FOR A SIGIR 2003 WORKSHOP
6. Andreas S. WEIGEND, Ph.D
7. SIGIR '03 SIGIR Conference -
8. CF: Multimedia Information Retrieval
9. SIGIR '03 ACM/IEEE 2003 Conference
10. Mark Grimson's home page -
11. Conferences on Information Retrieval
12. OntoWeb - SIGIR 2003 - 2003-01-
13. Yahoo Groups : web Messages
14. DBWorld Message /2002
15. Publications and Presentations of Thomas
16. SIGIR 2003 Workshop on the Evaluation
17. David Cow: Upcoming Conferences
18. Upcoming conferences for the W4ME Lab
19. SIGIR 99 Preliminary Program
20. SIGIR 99 CF/PC
22. ACM SIGIR 2002 Tutorial Bibliography
23. SIGIR <http://www.informatik.uni-
RELEVANT FOUND

better relevant rate

% relevant found

# docs judged

Deep-n
Hedge

1 2 3 4 5 6 7 8 9 10 15 20