Machine Learning and Information Retrieval

The Combination of Expert Advice Problem

Jay Aslam

College of Computer and Information Science

Northeastern University

Combination of Expert Advice

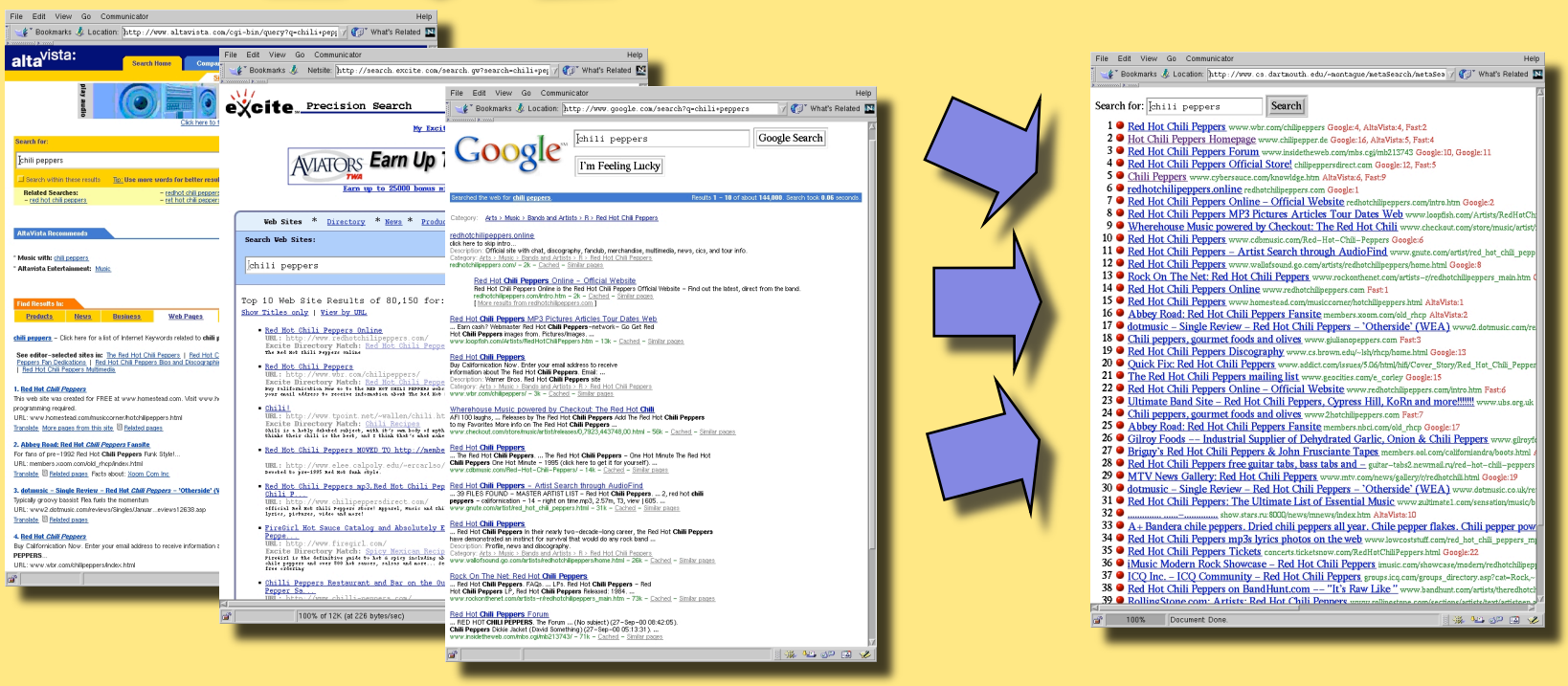
- Machine Learning
  - *Problems:*
    - medical diagnostics, stock market prediction, etc.
  - *Solutions:*
    - boosting, bagging, weighted majority, etc.

Combination of Expert Advice

- Information Retrieval
  - *Problems:*
    - filtering (news/spam), recommendation systems, *metasearch*, etc.
  - *Solutions:*
    - Linear classifiers, collaborative filtering, *on-line learning*, etc.

The Metasearch Problem

Search for: chili peppers

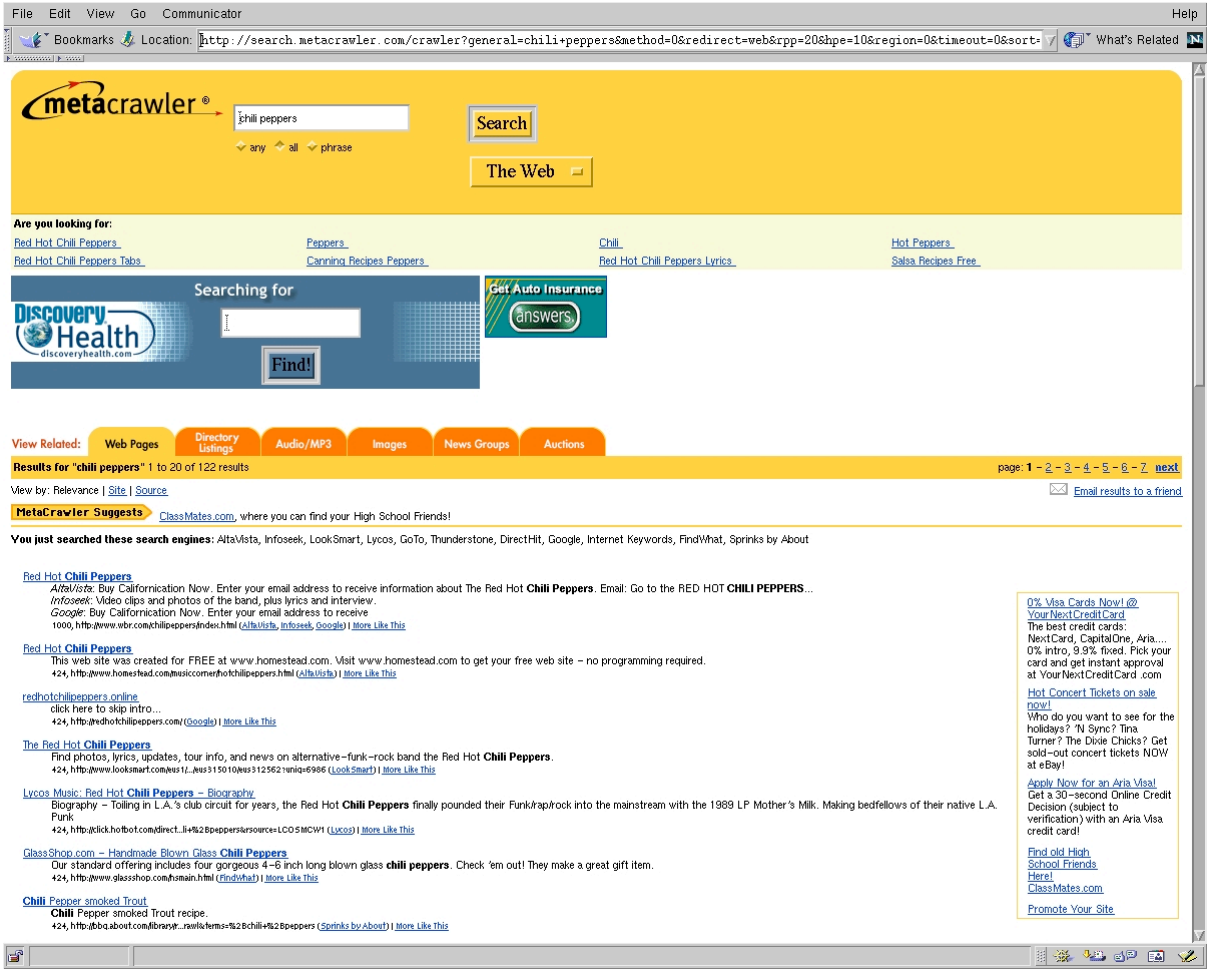


Search and Metasearch

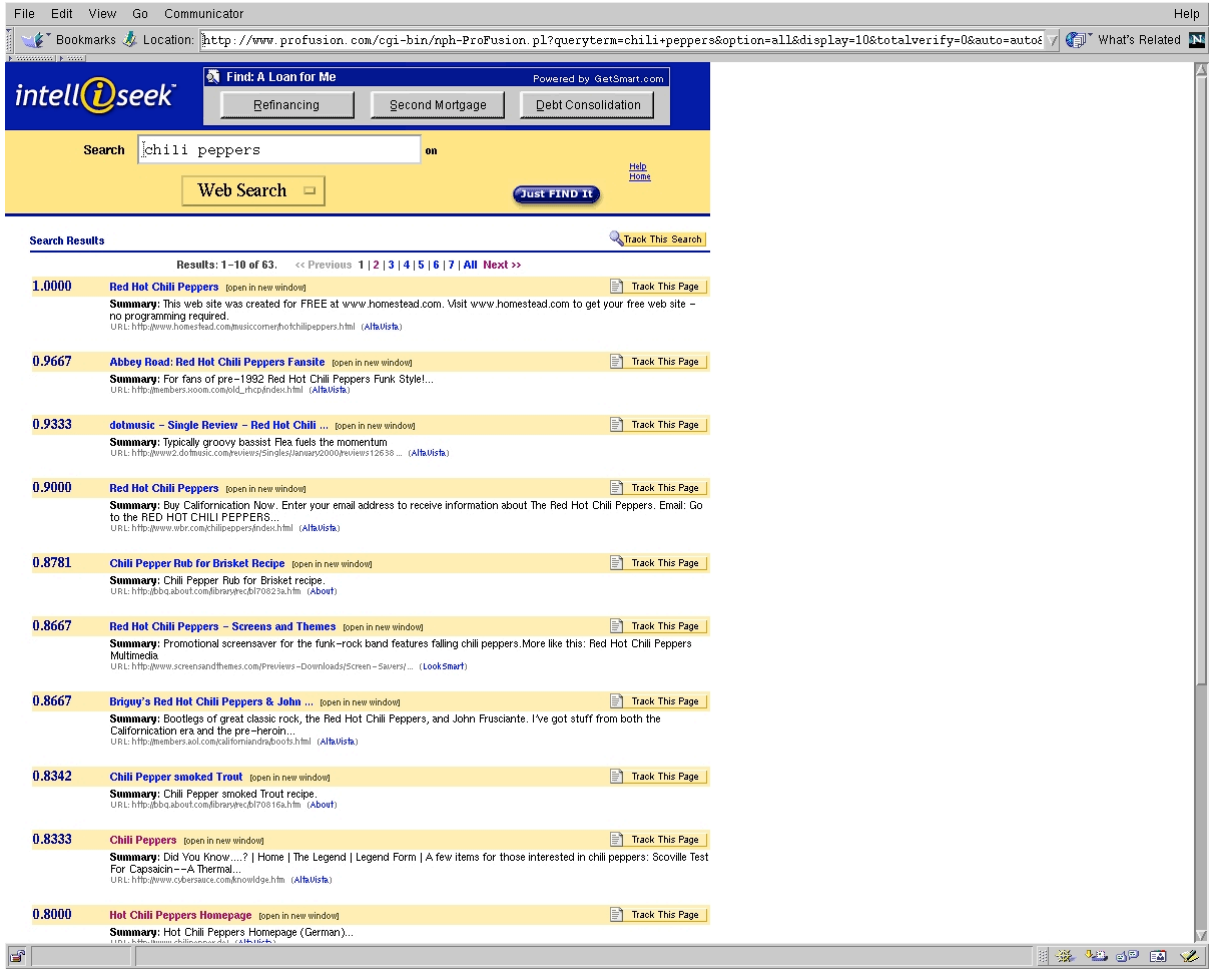
- Search engines:
  - Provide a ranked list of documents.
  - May provide relevance scores.
- Metasearch engines:
  - Query multiple search engines.
  - May or may not combine results.

Metasearch: Dogpile

Metasearch: Metacrawler



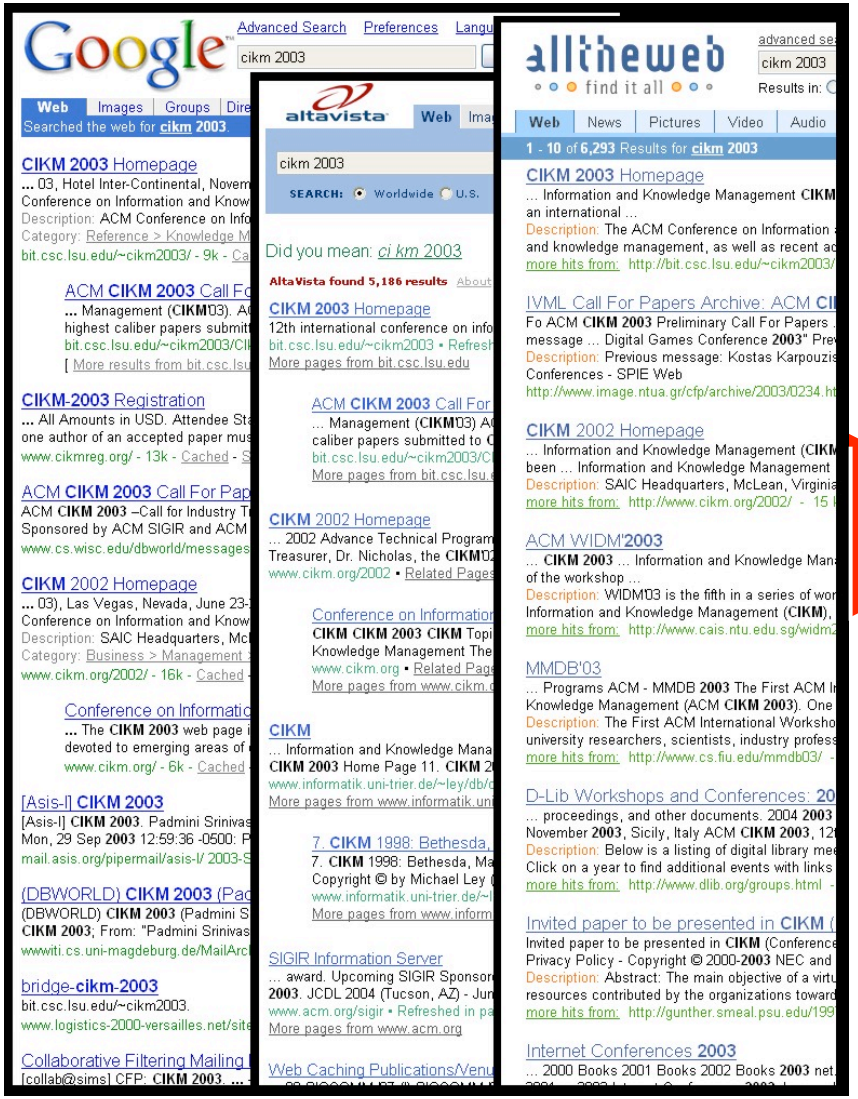
Metasearch: Profusion



Metasearch Algorithms

- Heuristics and hacks:
  - Interleave, average rank, sum scores, etc.
- Principled models:
  - Bayesian inference, election theory, etc.
  - *On-line combination of expert advice.*

On-line Metasearch [Aslam, Pavlu, Savell]



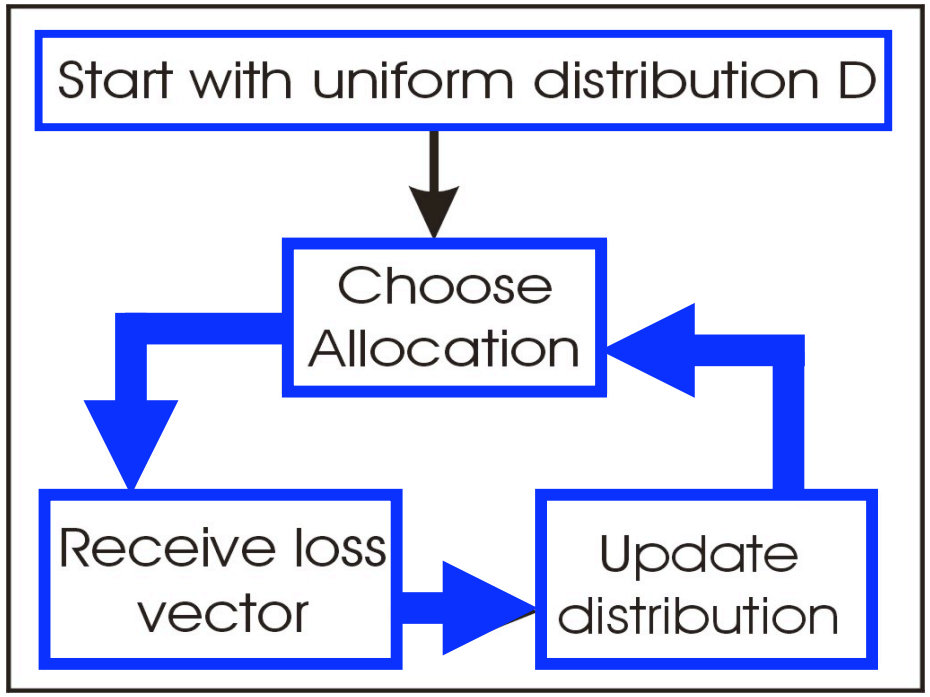
metasearch
CIKM 2003 Homepage
ACM CIKM 2003 Call For Pape
CIKM-2003 Registration
ACM CIKM 2003 Call For Pape
CIKM 2002 Homepage
Conference on Information an
[Asis-l] CIKM 2003
(DBWORLD) CIKM 2003 (Padm
bridge-cikm-2003
Collaborative Filtering Mailing
CIKM Home Page ACM DL: CIK
Yahoo! Groups : webir Messag
dBforums - Cfp: Cikm 03
Mailing List ARL-ERESERVE@a
Received: from cni.org by b.cr
ACM WIDM 2003
ACM - MMDB 2003
ACM CIKM 2003 PRELIMINARY
Selected Publications
Web Caching Publications/Ver
cikm '03
Call for papers
Mario A. Nascimento - Person
Iceved
Conferences and Journals on T
Conferences On Information V
Calendrier des manifestations
(DBWORLD) Final Call for ACM
Collaborative Filtering Mailing

metasearch
CIKM 2003 Homepage
CIKM-2003 Registration
CIKM 2002 Homepage
ACM CIKM 2003 Call For Pape
ACM CIKM 2003 Call For Pape
CIKM Home Page ACM DL: CIK
Conference on Information an
ACM CIKM 2003 PRELIMINARY
CIKM Home Page ACM DL: CIK
ACM CIKM 2003 PRELIMINARY
[Asis-l] CIKM 2003
(DBWORLD) CIKM 2003 (Padm
bridge-cikm-2003
Collaborative Filtering Mailing
Yahoo! Groups : webir Messag
dBforums - Cfp: Cikm 03
Mailing List ARL-ERESERVE@a
Received: from cni.org by b.cr
ACM WIDM 2003
ACM - MMDB 2003
Selected Publications
Web Caching Publications/Ver
cikm '03
Call for papers
Mario A. Nascimento - Person
Iceved
Conferences and Journals on T
Conferences On Information V
Calendrier des manifestations

On-line Allocation Problem

- Investment problem:
  - Monthly paycheck…
  - Many stock brokers & mutual funds…
  - *How to invest?*
  - *How to incorporate feedback?*

On-line Allocation: The Hedge Algorithm



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

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

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

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

$L_H^t = p^t \cdot l^t$

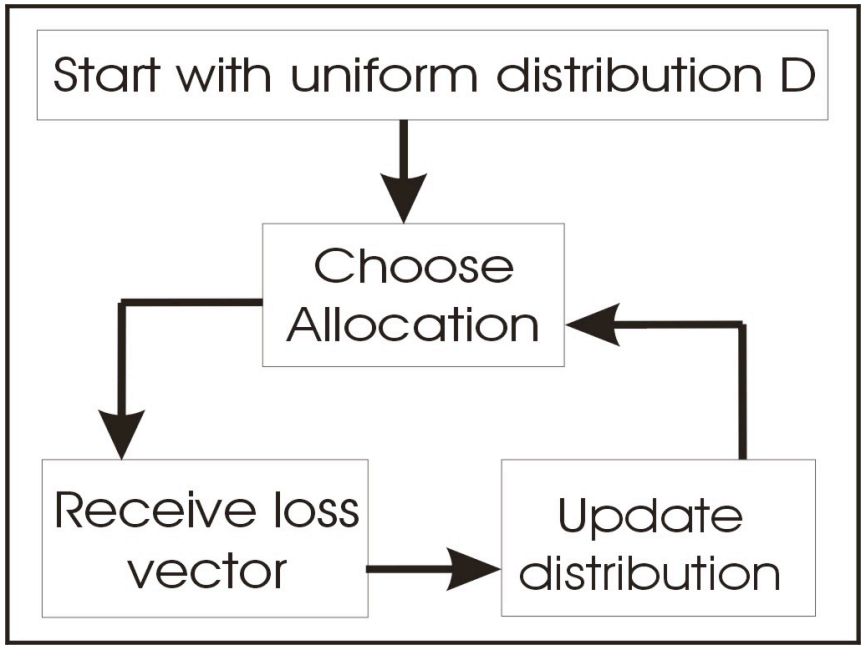
$w_i^{t+1} = w_i^t \cdot \beta^{l_i^t}$

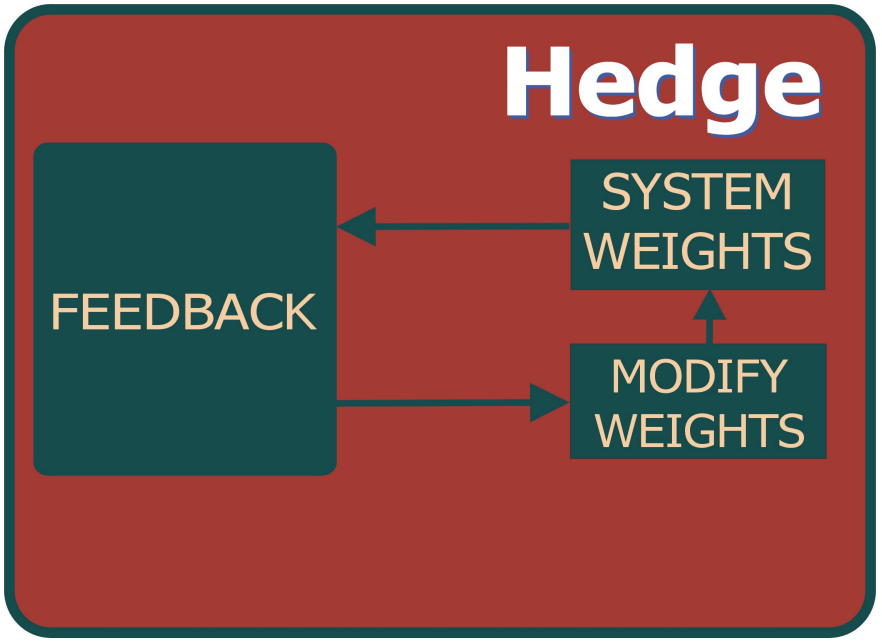
$$L_{HEDGE} = \sum_{t=1}^{T} p^t \cdot l^t$$

Why Hedge? [Schapire, Freund]

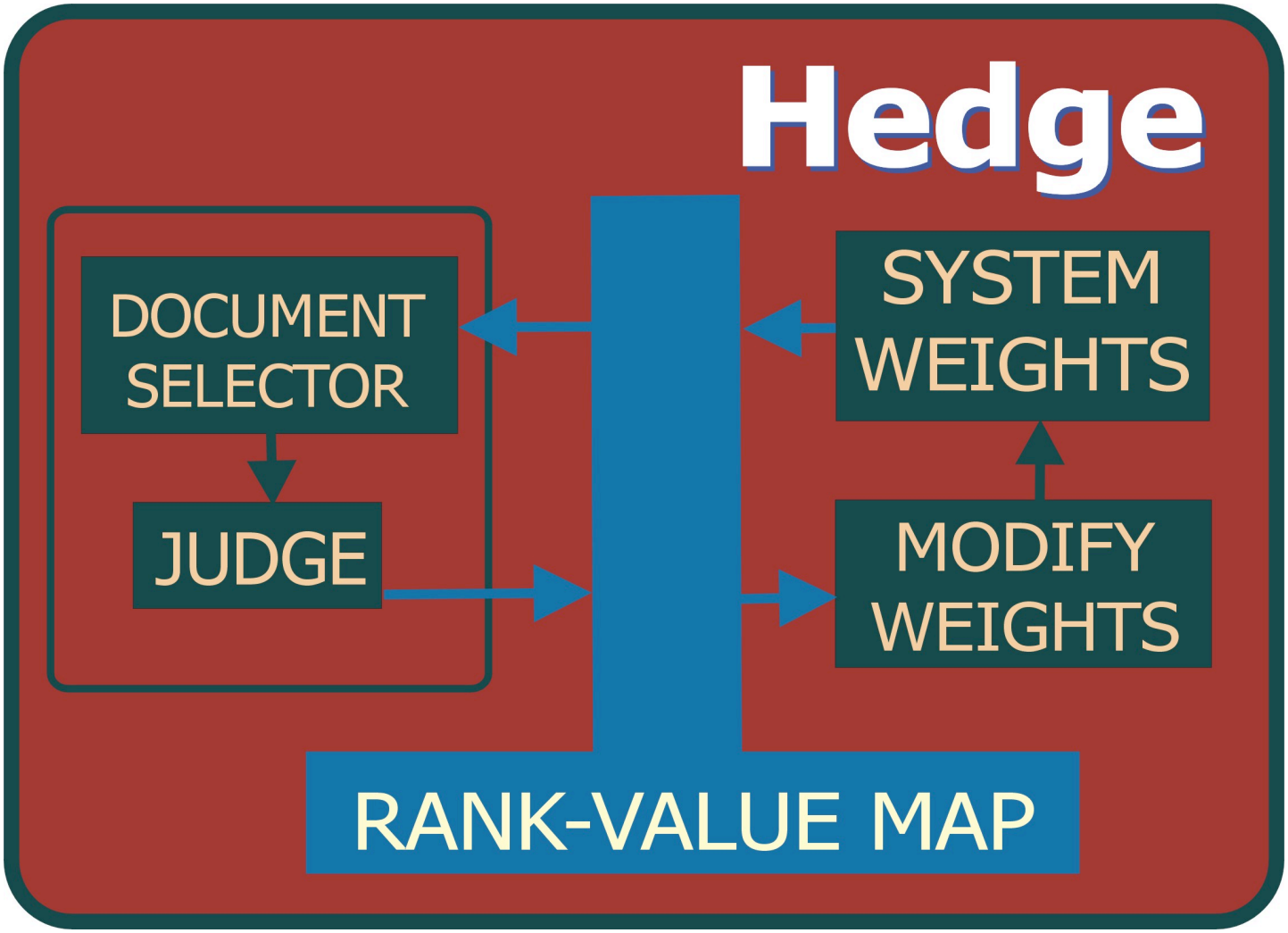
$$L_{HEDGE} \leq \frac{\ln\left(\frac{1}{\beta}\right) L_{BEST\ SYSTEM} + \ln N}{1-\beta}$$

On-line Setup

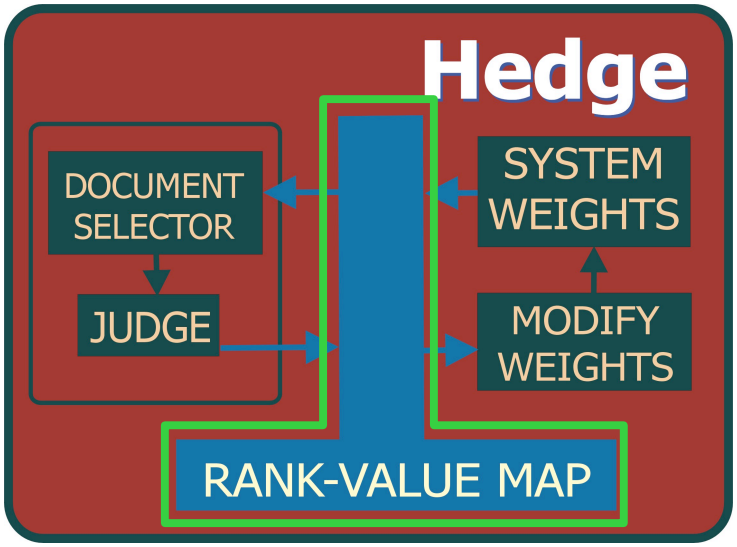




A Unified Model



Implementation



- map ranks to values

$$value(r) = \frac{1}{r} + \frac{1}{r+1} + \ldots + \frac{1}{Z} \approx \ln \frac{Z}{r}$$

$$RELEVANT = -1$$
$$NONRELEVANT = +1$$

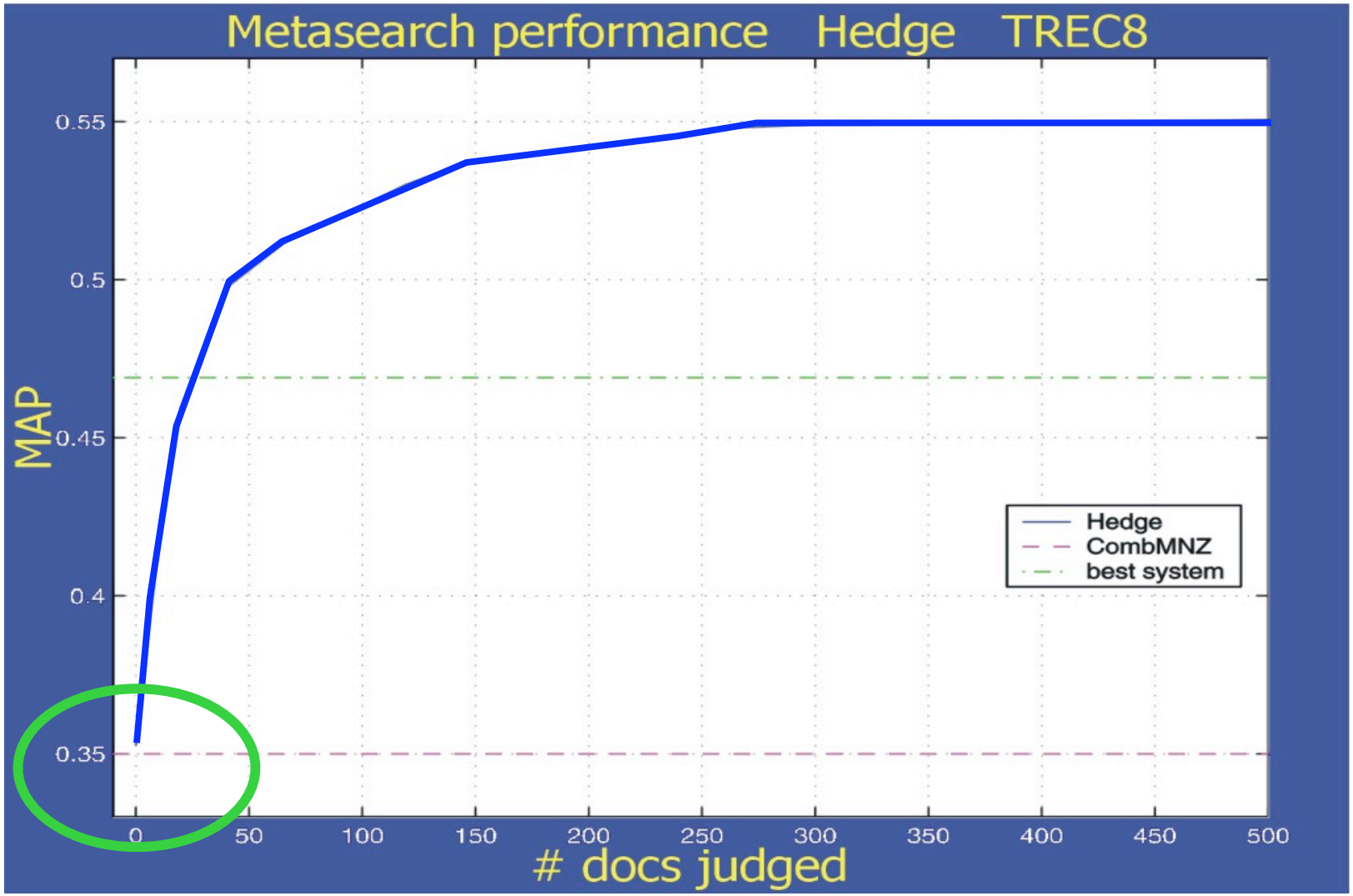
$$LOSS(d,s) = label(d) \cdot value(rank_{d,s}) \approx label(d) \cdot \ln \frac{Z}{r}$$

total loss vs. total precision vs. average precision

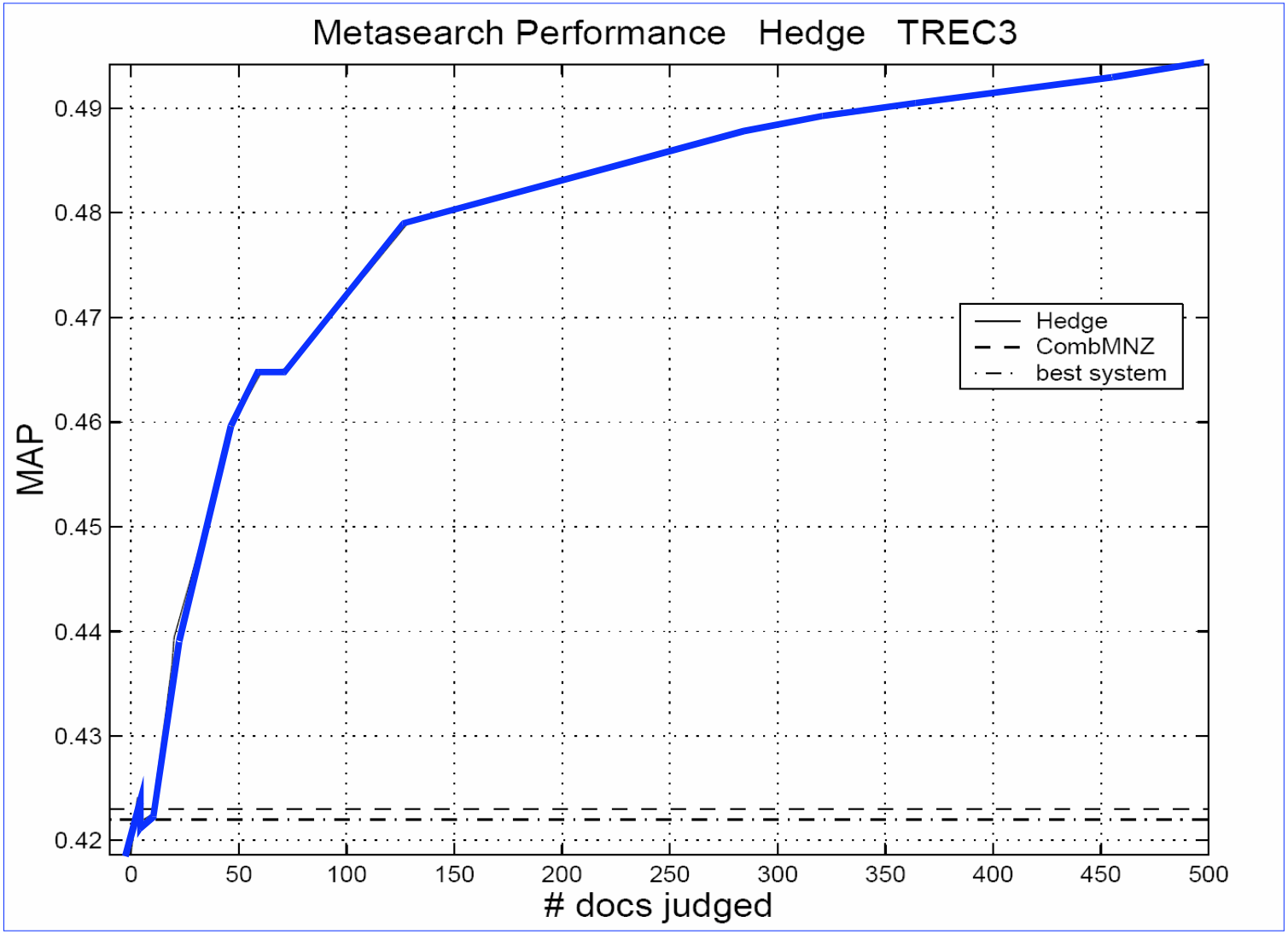
Experiments

- TREC 3,5,6,7,8
  - 41-129 systems
  - 50 queries per TREC
  - Metasearch combines *all* systems
- Use TREC judgments as user feedback

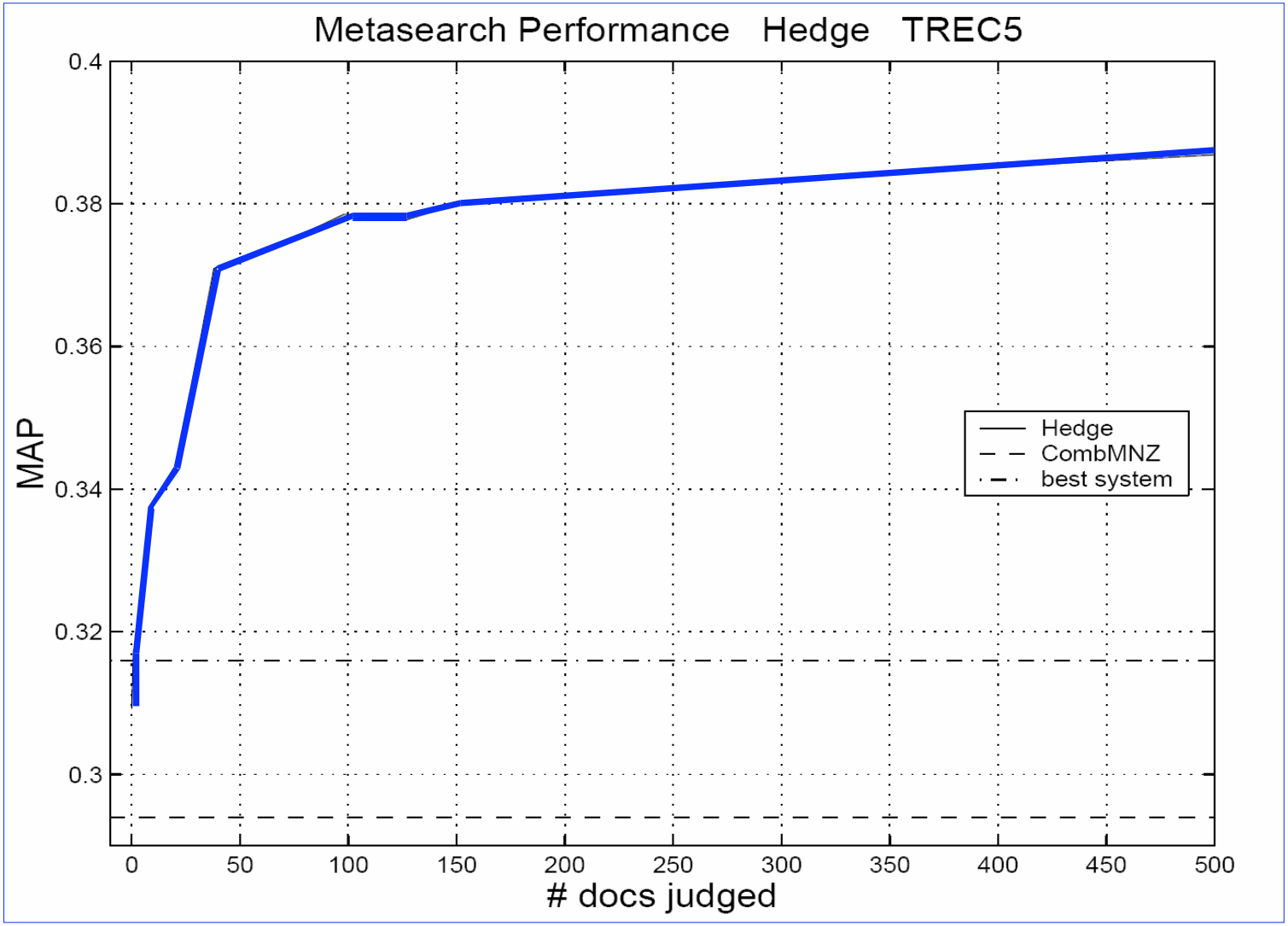
Metasearch – TREC8



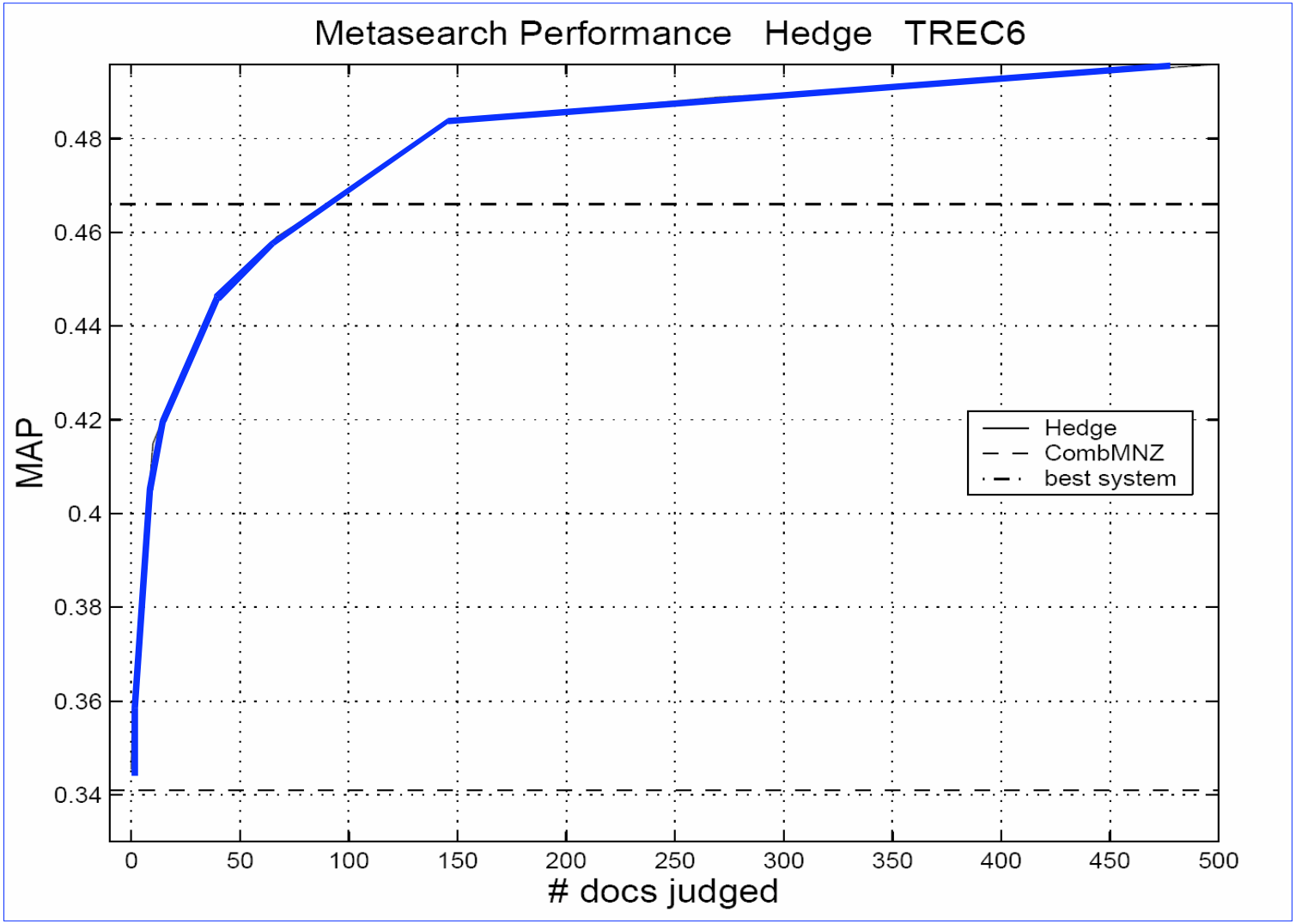
Metasearch – TREC 3



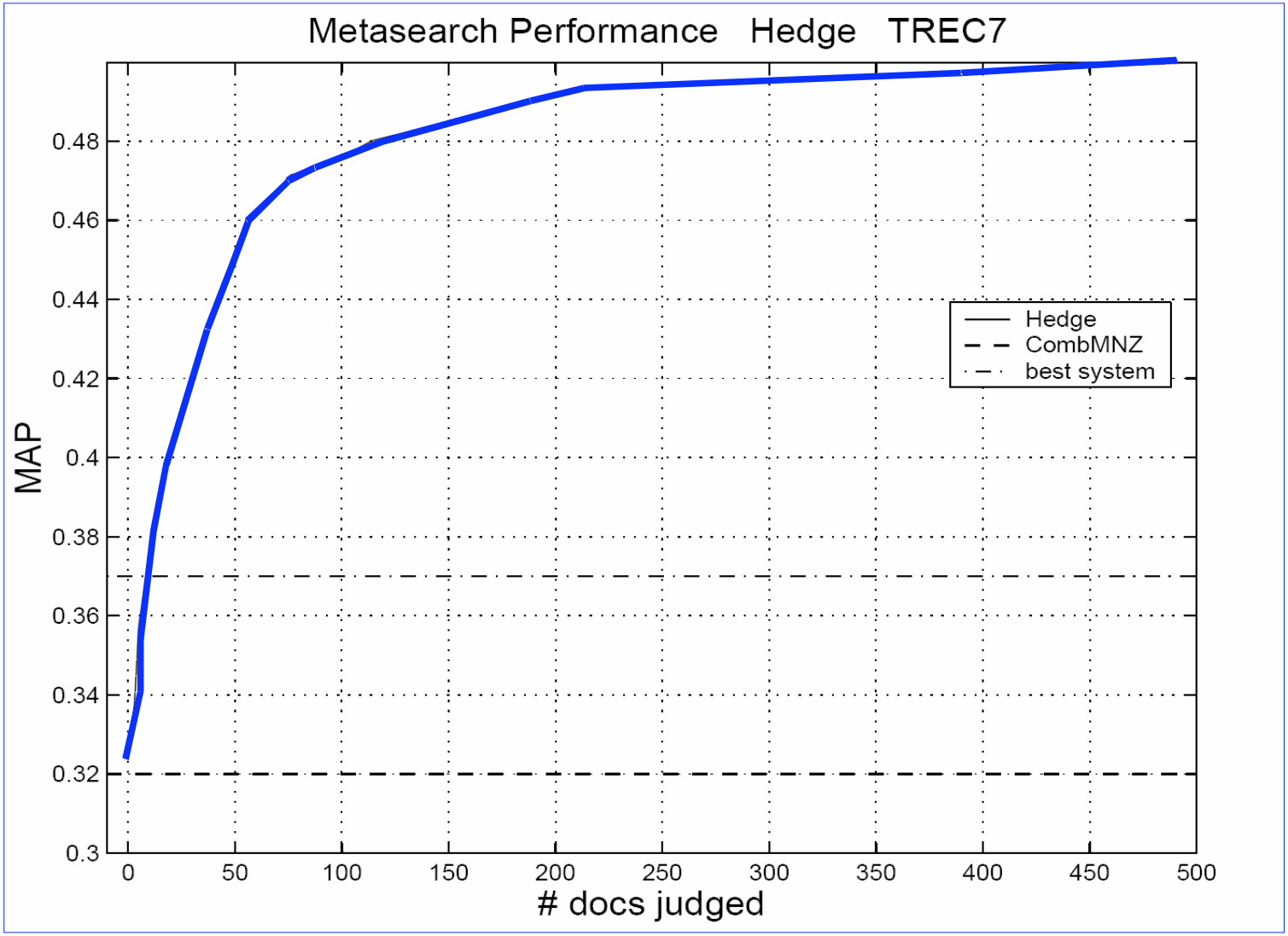
Metasearch – TREC 5



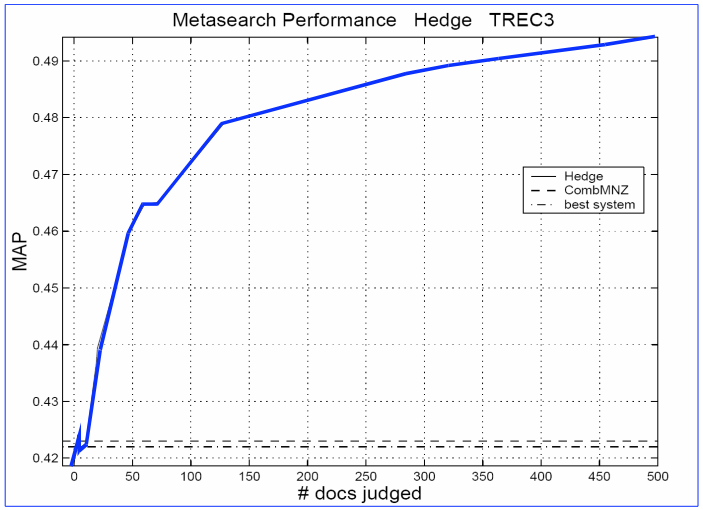
Metasearch – TREC 6

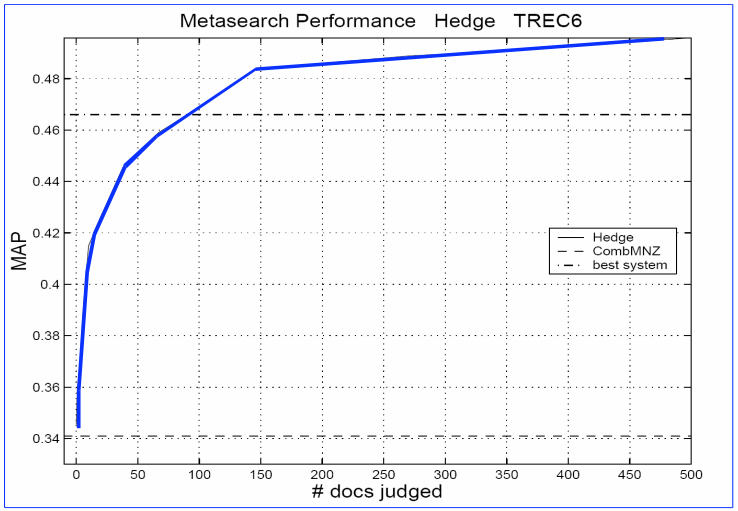


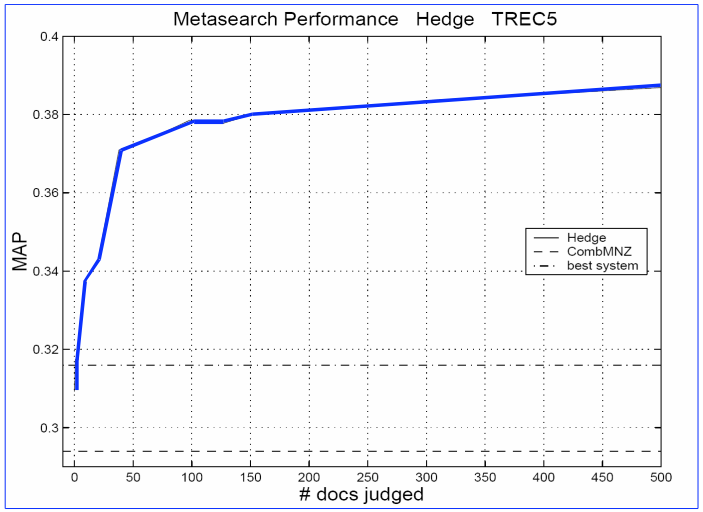
Metasearch – TREC 7

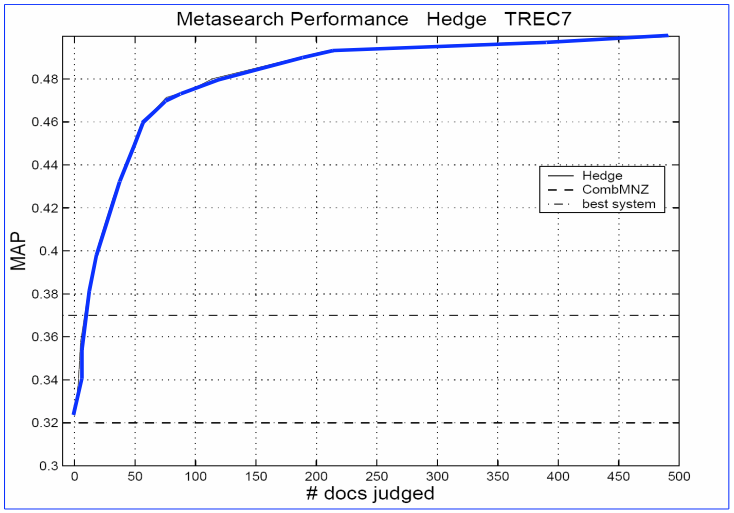


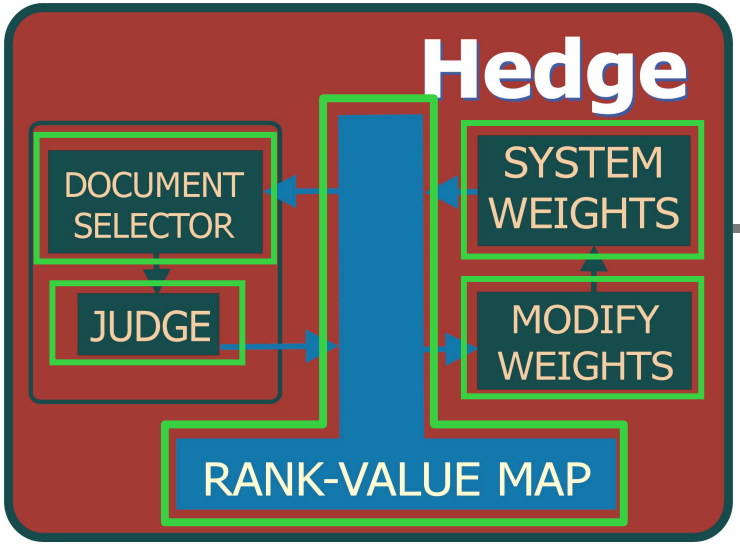
Metasearch – TREC 3,5,6,7





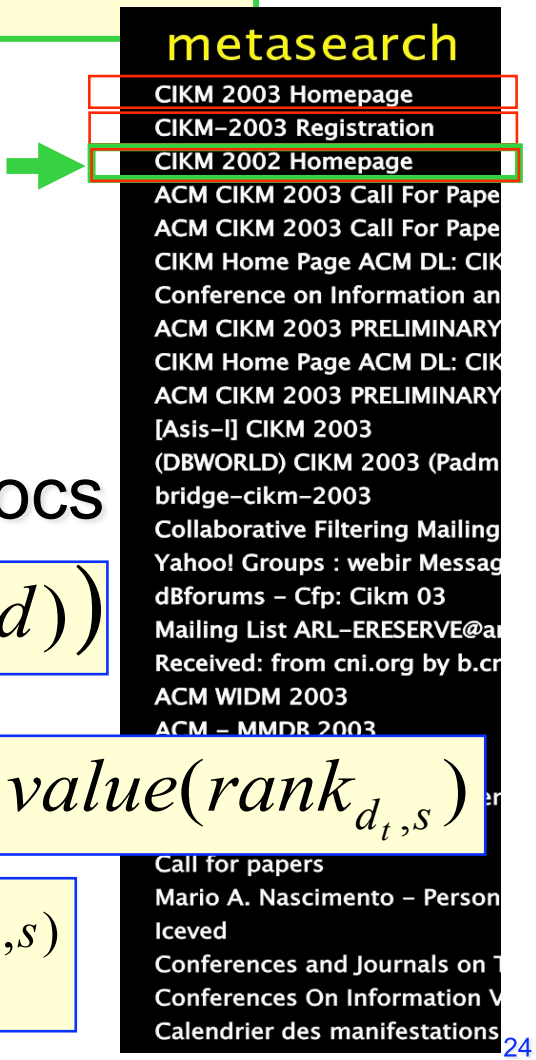






$$average\_value_t(d) = \sum_{s=1}^{N} w_s^{t-1} \cdot value(rank_{d,s})$$

Loop



- average_value at episode t
  - "trust" in systems change with t
- metasearch : include already judged docs
- pooling $d_t = \arg\max_d \left(average\_value(d)\right)$
  - system evaluation
- get feedback $LOSS\ (d_t, s) = label(d_t) \cdot value(rank_{d_t,s})$
- modify weights $w_s^{t+1} = w_s^t \cdot \beta^{LOSS_t\ (d_t,s)}$