Modelling A User Population for Designing Information Retrieval Metrics

Tetsuya Sakai
(NewsWatch, Inc.)
tetsuyasakai@acm.org
Stephen Robertson
(Microsoft Research Cambridge)
ser@microsoft.com

EVIA 2008, December 16, 2008@NII, Tokyo
TALK OUTLINE

1. Objectives
2. Normalised Cumulative Precision (NCP)
3. Normalised Cumulative Utility (NCU)
4. Evaluating Evaluation Metrics: Resemblance to Average Precision
5. Evaluating Evaluation Metrics: Discriminative Power
6. Conclusions
Average Precision (AP)

\[
AP = \frac{1}{R} \sum_{r} I(r) \frac{C(r)}{r}
\]

- Used widely since the advent of TREC
- Mean over topics is referred to as “MAP”
- Cannot handle graded relevance (but many IR researchers just love it)
Criticisms of (Mean) Average Precision ((M)AP)

• AP may be a poor measure of user performance/satisfaction
  [Turpin/Scholer SIGIR 06 etc.]
• AP lacks a user model
  “there is no single user application that directly motivates MAP” [Buckley/Voorhees TREC book]
  “there is no plausible search model that corresponds to MAP, because no user knows in advance the number of relevant answers present in the collection…” [Moffat/Webber/Zobel SIGIR 07]
“AP lacks a user model?”

Rubbish!
[Robertson SIGIR 08]
Objectives

Robertson showed that AP is a special case of Normalised Cumulative Precision (NCP) which models a population of users.

- We generalise NCP and introduce Normalised Cumulative Utility (NCU), and show that
  - AP and Q-measure are in fact reasonable metrics!
  - A version of NCU, which utilises graded relevance in a novel way, has high discriminative power!
I need the latest information on EVIA!
**L1 (partially relevant)**

**EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)**
provisional schedule of the 6th NTCIR Workshop, evaluation of IR, QA, Opinion, Trend and Cross-lingual information access
research.nii.ac.jp/ntcir/ntcir-ws6/pmw-en.html • キャッシュページ

**L3 (highly relevant)**

**NTCIR-7 Information Retrieval Evaluation Site - EVIA**
NTCIR-7 Information Retrieval Evaluation Workshop in Tokyo, Japan ... Workshop Date: December 16, 2008 (Day 1 of the NTCIR-7 Meeting) Workshop Venue: 
ntcir.nii.ac.jp/index.php/ja/EVIA-2008 • キャッシュページ

**L0 (not relevant)**

**Eviaのホテル**
Evia ホテル予約 - 直前格安価格スペシャル ... ホテル名称 ロケーション ホテルランク 価格(〜から)
www.ratestogo.com/City/JP/Evia • キャッシュページ

**L2 (relevant)**

[ntcir:240] **EVIA 2007: Call for Participation**
To: ntcir@xxxxxxxxx Subject: [ntcir:240] EVIA 2007: Call for Participation From: Atsushi Fujii <fujii@xxxxxxxxxxxxxxxxx> Date: Wed, 02 May 2007 15:46:00 +0900 (JST)
research.nii.ac.jp/ntcir/archive/ntcir/msg00199.html • キャッシュページ

**L0 (not relevant)**

**アンフィニ・コメルス株式会社 高級輸入家具リース ...**
Evia
www.infini-commerce.co.jp/collection/modern/catteran/evia.ht... • キャッシュページ

**L0 (not relevant)**

**Υπό κατασκευή - Under construction**
Υπό κατασκευή - Under construction
www.eviaweb.gr • キャッシュページ

**L3 (highly relevant)**

**検索メニアックス! (Information Retrieval Maniac!): EVIA 2008 ...**
フレッシュアイトップ 情報検索の研究者・酒井哲也が奏でる、研究とビジネスと私生活のハーモニー。不協和音？
voice.fresheye.com/sakai/2003/08/evia_2008.html • キャッシュページ
Where do users stop scanning the list?

I stop at rank 1

I stop at rank 2

I stop at rank 4

I stop at rank 7

EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)
provisional schedule of the 6th NTCIR Workshop, evaluation of IR, QA, Opinion, Trend and Cross-lingual information access
research.nii.ac.jp/ntcir/ntcir-ws6/pmww-en.html

NTCIR-7 Information Retrieval Evaluation Site - EVIA
NTCIR-7 Information Retrieval Evaluation Workshop in Tokyo, Japan ... Workshop Date: December 16, 2008 (Day 1 of the NTCIR-7 Meeting) Workshop Venue:
ntcir.nii.ac.jp/index.php/ja/evia-2008

Eviaのホテル
Evia ホテル予約 - 直前格安価格スペシャル ... ホテル名称 ロケーション ホテルランク 価格（〜から）
www.ratestogo.com/City/JP/Evia

[ntcir:240] EVIA 2007: Call for Participation
To: ntcir@xxxxxxxxxx Subject: [ntcir:240] EVIA 2007: Call for Participation From: Atsushi Fujii
Date: Wed, 02 May 2007 15:46:00 +0900 (JST)
research.nii.ac.jp/ntcir/archive/ntcir/msg00199.html

アンフィニ・コメルス株式会社 高級輸入家具リース...
www.infini-commerce.co.jp/collection/modern/catteran/evia.htm...

Υπό κατασκευή - Under construction
Υπό κατασκευή - Under construction
www.eviaweb.gr

検索メイニアックル (Information Retrieval Maniac!): EVIA 2008...
フレッシュアイトップ 情報検索の研究者・酒井哲也が奉える、研究とビジネスと私生活のハーモニー。不協和音？
**pu**: Uniform Distribution over Relevant Documents

**ASSUMPTIONS:**
- Users stop at a relevant doc;
- The stopping probability is uniform across all relevant docs.
**prb**: Rank-Biased Distribution over Relevant Docs

**ASSUMPTIONS:**
- Users stop at a relevant doc;
- Users tend to stop near the top than near the bottom

**EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)**
provisional schedule of the 6th NTCIR Workshop, evaluation of IR, QA, Opinion, Trend and Cross-lingual information access
research.nii.ac.jp/ntcir/ntcir-ws6/pmww-en.html ・ キャッシュページ

**NTCIR-7 Information Retrieval Evaluation Site - EVIA**
NTCIR-7 Information Retrieval Evaluation Workshop in Tokyo, Japan ... Workshop Date: December 16, 2007
ntcir.nii.ac.jp/ntcir/ntcir-ws7/evia-en.html ・ キャッシュページ

**Eviaの感想**
Eviaの感想のページを訪問
www.rat.hiroshin.net/ntcir2/ntcir/ntcir-ws7/evia-en.html ・ キャッシュページ

**アンフィニ・コメルス株式会社 高級輸入家具リース ...**
Eviaの感想のページを訪問
www.infini-commerce.co.jp/collection/modern/catteran/evia.html ・ キャッシュページ

**ュパカサクイ・ - Under construction**
ュパカサクイ・のページを訪問
www.eviaweb.gr ・ キャッシュページ

**検索メニューアック! (Information Retrieval Maniac!): EVIA 2008 ...**
フレッシュアイトップ 情報検索の研究者・酒井哲也さんが考で、研究とビジネスと私生活のハーモニー（不協和音？）
voice.fresheye.com/sakai/2003/08/evia_2008.html ・ キャッシュページ
**pgu: Graded-Uniform Distribution over Relevant Docs**

**ASSUMPTIONS:**
- Users stop at a relevant doc;
- Users tend to stop at a highly relevant doc than at a partially relevant doc.
TALK OUTLINE

1. Objectives
2. Normalised Cumulative Precision (NCP)
3. Normalised Cumulative Utility (NCU)
4. Evaluating Evaluation Metrics: Resemblance to Average Precision
5. Evaluating Evaluation Metrics: Discriminative Power
6. Conclusions
Robertson’s Normalised Cumulative Precision (NCP)

Expectation over a user population

Utility/Cost given the stopping point (precision at $n$)

Probability that the user stops at the (relevant) document at rank $n$

\[ NCP = \sum_{n=1}^{\infty} p_s(n) P(n) \]

\[ P(n) = \frac{C(n)}{n} \]

Cost: #docs seen so far

Utility: #relevant seen so far
**ASSUMPTIONS:**
- Users stop at a relevant doc; 
- The stopping probability is uniform across all relevant docs

Let $p_s(n) = p_u(n) = \frac{1}{R}$ for every rank $n$ that has a relevant doc.

Then NCP reduces to AP ($\approx \text{NCPU}$)!
That is,

- AP is a special case of NCP.
- It is an expectation of utility/cost over a user population whose stopping probability is uniform across all relevant documents.
- Hence, AP is in fact a reasonable metric!
1. Objectives
2. Normalised Cumulative Precision (NCP)
3. Normalised Cumulative Utility (NCU)
4. Evaluating Evaluation Metrics: Resemblance to Average Precision
5. Evaluating Evaluation Metrics: Discriminative Power
6. Conclusions
We generalise NCP in two ways

\[ NCP = \sum_{n=1}^{\infty} p_s(n)P(n) \]

\[ NCU = \sum_{n=1}^{\infty} p_s(n)NU(n) \]

Stopping probability:
- \( pu \) (uniform)
- \( prb \) (rank-biased)
- \( pgu \) (graded-uniform)

Normalised Utility:
- \( BR(n) \) (blended ratio) which generalises \( P(n) \)
**pu: Uniform Distribution over Relevant Documents**

ASSUMPTIONS:
- Users stop at a relevant doc;
- The stopping probability is uniform across all relevant docs.
**prb: Rank-Biased Distribution over Relevant Docs**

**ASSUMPTIONS:**
- Users stop at a relevant doc;
- Users tend to stop near the top than near the bottom

**EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)**  
provisional schedule of the 6th NTCIR Workshop, evaluation of IR, QA, Opinion, Trend and Cross-lingual information access  
research.nii.ac.jp/ntcir/ntcir WS6/pmww-en.html  

**NTCIR-7 Information Retrieval Evaluation Site - EVIA**  
NTCIR-7 Information Retrieval Evaluation Workshop in Tokyo, Japan ... Workshop Date: December 16, 2007  
ntcir.nii.ac.jp/evia/  

---

**Evia**  
www.rattler.com/evia  

---

**Evia**  
www.infini-commerce.co.jp/collection/modern/catteran/evia.html  

---

**Evia**  
www.eviaweb.gr  

---

**Evia**  

\( \gamma \): top-heaviness parameter for \( prb \)

\( \gamma = 1 \) reduces \( prb \) to \( pu \)
pgu: Graded-Uniform Distribution over Relevant Docs

ASSUMPTIONS:
- Users stop at a relevant doc;
- Users tend to stop at a highly relevant doc than at a partially relevant doc

Stopping weights
stop(L3):stop(L2):stop(L1) = 3:2:1
(stop(L3):stop(L2):stop(L1) = 10:5:1
reduces pgu to pu)
Blended Ratio (BR)

\[
BR(n) = \frac{C(n)}{n + \frac{\beta \sum_{i=1}^{n} g(i)}{\beta \sum_{i=1}^{n} g^*(i)}}
\]

A large \( \beta \) represents a very persistent user; \( \beta = 0 \) reduces BR to P.

BR is suitable as a utility/cost function because, given the stopping point \( n \), it does NOT matter where the relevant documents are within top \( n \).
NCU family

\[ NCU = \sum_{n=1}^{\infty} p_s(n) NU(n) \]

Stopping probability:
- \( prb \) (rank-biased) with top-heaviness parameter \( \gamma \) (\( \gamma = 1 \) reduces \( prb \) to \( pu \))
- \( pgu \) (graded-uniform) with stopping weights (flat weights reduces \( pgu \) to \( pu \))

Normalised Utility given the stopping point: \( BR(n) \) (blended ratio) with persistence parameter \( \beta \) (\( \beta = 0 \) reduces \( BR(n) \) to \( P(n) \))
TALK OUTLINE

1. Objectives
2. Normalised Cumulative Precision (NCP)
3. Normalised Cumulative Utility (NCU)
4. Evaluating Evaluation Metrics: Resemblance to Average Precision
5. Evaluating Evaluation Metrics: Discriminative Power
6. Conclusions
Comparing a system ranking by Metric $M$ to that by AP

- Kendall’s rank correlation
  Monotonic function of the probability that a randomly chosen system pair is ordered identically in two rankings
- Yilmaz/Aslam/Robertson (YAR) rank correlation
  [SIGIR 08]
  Monotonic function of the probability that a randomly chosen system and one ranked above it are ordered identically in two rankings
  Assumes that the top ranks are the most important
  Not symmetrical, but is almost symmetrical in practice
YAR rank correlation with AP (NCU \( u, \beta =0 \): NTCIR-6J

Q, NCU \( gu, \beta =0 \) and NCU \( gu, \beta =1 \) produce rankings that are very similar to that by AP.

Heavy rank bias produces very unconventional system rankings.
YAR rank correlation with AP (NCU \( u, \beta = 0 \): TREC03

Q, NCU \( gu, \beta = 0 \) and NCU \( gu, \beta = 1 \) produce rankings that are very similar to that by AP

Stop weights=3:2:1

Heavy rank bias produces very unconventional system rankings
1. Objectives
2. Normalised Cumulative Precision (NCP)
3. Normalised Cumulative Utility (NCU)
4. Evaluating Evaluation Metrics:
   Resemblance to Average Precision
5. Evaluating Evaluation Metrics:
   Discriminative Power
6. Conclusions
Measuring discriminative power of metrics [Sakai SIGIR06]

• Given a set of systems and a significance level $\alpha$, for how many system pairs can a metric detect statistical significance?

Probability of Type I error $\alpha = 0.05 \Leftrightarrow 95\%$ confidence

• Sakai’s method uses the bootstrap test, and can also estimate the absolute performance difference required to achieve statistical significance (e.g. “a difference of 0.20 is usually statistically significant”)

• Sakai’s method and the Voorhees/Buckley swap method [SIGIR 02] give similar results in practice
Discriminative power at $\alpha = 0.05$: NTCIR-6J

Heavy rank bias hurts discriminative power (by ignoring low-ranked docs)

AP, Q, NCU $gu, \beta = 0$ and NCU $gu, \beta = 1$ have high discriminative power

$\gamma = 1$, $\gamma = 0.9$, $\gamma = 0.7$, $\gamma = 0.5$
Discriminative power at $\alpha = 0.05$: TREC03

AP, Q, NCU $gu, \beta = 0$ and NCU $gu, \beta = 1$ have high discriminative power.

Heavy rank bias hurts discriminative power (by ignoring low-ranked docs).
Effect of $\gamma$ on discriminative power: TREC03

Heavy rank bias hurts discriminative power (by ignoring low-ranked docs)
TALK OUTLINE

1. Objectives
2. Normalised Cumulative Precision (NCP)
3. Normalised Cumulative Utility (NCU)
4. Evaluating Evaluation Metrics: Resemblance to Average Precision
5. Evaluating Evaluation Metrics: Discriminative Power
6. Conclusions
Conclusions

We defined NCU, whose components are:
- Probability distribution of the user’s stopping behaviour $(pu, prb, pgu)$
- Blended Ratio (BR) as the utility/cost function given the stopping point

and showed that:
- Heavy rank-bias (small $\gamma$) is not desirable
- AP and Q, which rely on $pu$, are reasonable metrics – they emphasize long-tail users who tend to dig deep into the ranked list and achieve high discriminative power
- NCU $gu, \beta = 1$ has high consistency with AP and has the highest discriminative power (utilises graded relevance for both probability distribution $pgu$ and utility/cost BR)
Uniform distribution of (AP and Q) can be interpreted as…

Long-tail user
ir4qa evaluation scripts

• Simple scripts for computing AP, Q, nDCG, RBP, NCU etc. are available at:
  http://research.nii.ac.jp/ntcir/tools/ir4qa_eval-en

ir4qa_eval

These information retrieval (IR) evaluation scripts were developed for the NTCIR-7 AQLIA 3IR QA subtask. They can be used for other IR tasks at NTCIR, TREC, etc.

The scripts can compute average precision, Q-measure, nDCG and some other evaluation metrics.

For more details, please read the README file included in the tar file.

Download
http://research.nii.ac.jp/ntcir/tools/ir4qa_eval.tar.gz
Thank you!