

User Models, Metrics and Measures of Search: A Tutorial on the CWL Evaluation Framework

ACM CHIIR UMMMS 2021

by

Leif Azzopardi, Alistair Moffat, Paul Thomas and Guido Zuccon



Who are we?



Leif Azzopardi, @leifos

Associate Professor, University of Strathclyde

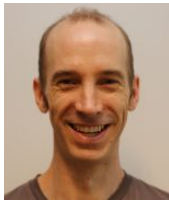
Studying how information systems shape and influence people and society with models of user behavior, interaction and performance.



Alistair Moffat

Professor, University of Melbourne

Searching for better information retrieval metrics, text and index compression methods, and information retrieval heuristics.



Paul Thomas, @pt_ir

Senior Applied Scientist, Microsoft Bing

Studying how people use search systems, and using that to evaluate current systems and build new ones.



Guido Zuccon, @guidozuc

Associate Professor, University of Queensland

Researching and developing formal models of search, ranking, and research diversification, especially in the domain of e-health.

But, what about you?

Evaluation

- What is evaluation?
 - Measure the **effectiveness**, **efficiency** and **cost** of a system
- Search **Effectiveness**: how **good** a system is in retrieving relevant documents
 - This is the focus of this tutorial
- Search **Efficiency**: how **fast** a system is in retrieving documents
- Often there is **trade-off** between effectiveness and efficiency
- **Cost**: how much does it cost to run the system (\$\$, Kw/h, etc)
- Usually cost is determined by the desired level of effectiveness and efficiency

Why do we want to Evaluate?

- Say whether the system is any **good**
- **Compare** two systems, so as to choose the “best” (or best fit)
- Understand where the system **succeeds** and where it **fails** (diagnostic)
- (What about evaluating users?)

Test your evaluation intuition

Which SERP is better?

SERP one

Good

Good

Bad

Bad

Bad

Bad

SERP two

Bad

Bad

Bad

Good

Bad

Bad

Test your evaluation intuition

Which SERP is better?

SERP one

Good

Good

Bad

Bad

Bad

Bad

SERP two

Good

Good

Bad

Good

Bad

Bad

Test your evaluation intuition

Which SERP is better?

SERP one

- Bad
- Bad
- Bad
- Good
- Good
- Good

SERP two

- Good
- Bad
- Bad
- Bad
- Bad
- Bad

Metric Choices

There are lots!

- **P@k**, precision at depth k: fraction of the top k which are relevant
- **RBP**, rank-biased precision: geometrically-decreasing importance
- **RR**, reciprocal rank: $1 / \text{rank of the first relevant result}$
- **NDCG**, normalized discounted cumulative gain: logarithmically-decreasing importance, and scaled by available relevance
- **AP**: average of precision values, at positions where there's relevance

And don't forget: TBG, ERR, U-measure, Bejewelled, BPref, INSQ, INST...

Metric Choices

- Lots of metrics:
 - which metric is best?
 - which metric should I use?
 - anything in common, any coherent way to talk about these?
- Any way to discuss, trade off, choose?
- Yes: examine the model underlying each
 - In this tutorial, we introduce you to the C/W/L framework that allows you to understand, analyse and compare metrics

Tutorial's Goals

Give you the knowledge and skills to:

- **Explain** the **C/W/L framework** and the different measurements it incorporates;
- **Explain** the **User Browsing Models** (continuation functions)
- **Analyse existing metrics** in light of C/W/L
- **Design a metric** given the **C/W/L framework**
- Learn to **use** the “**cwl_eval**” toolkit

Schedule

- 2 hours of **live presentation** (this)
 - Part 1: Welcome + Introduction to Evaluation
 - Part 2: The C/W/L Framework
 - Part 3: Open problems/research directions
 - Part 4: C/W/L in practice
- Followed by “***office hour***” session (1 hour)
 - starts 1 hour after the end of live session
- All is **repeated** again 12 hours later

Course Resources

- Website: <http://ielab.io/tutorials/ummmms2021>
- Online videos:
<https://www.youtube.com/playlist?list=PLgrOo-AsKcmV3pzpvFbd5SUpUMWUv2fQr>

Introduction to Evaluation

Preliminaries

Tasks and User Models

- An evaluation **metric** is typically **grounded on a task and a user model**
 - Robertson, SIGIR 2008: “If we can interpret a measure (. . .) in terms of an explicit user model (. . .), this can only improve our understanding of what exactly the measure is measuring”
- **Task:** what the objective of the user is
- **User Model:** how the user behaves
- Example: precision
 - Task: find relevant documents, without finding non-relevant ones
 - User Model: examine all documents retrieved by the search engine, without order

What is a model?

Online Video

- “a deliberate **simplification** of something complicated with the objective of making it more tractable”

Frigg, Roman and Hartmann, Stephan, 2018. “Models in Science”, in The Stanford Encyclopedia of Philosophy.

- “a physical, conceptual, or mathematical representation of a real phenomenon that is **difficult to observe** directly”

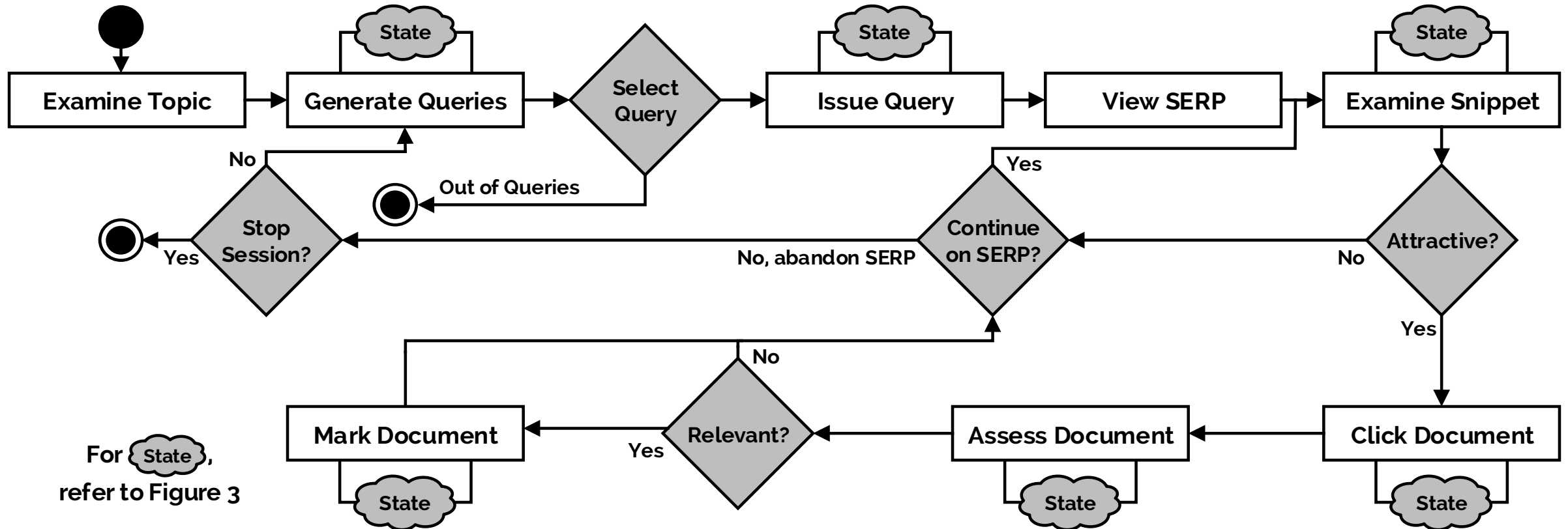
Rogers, 2012. “Scientific modeling”, in Encyclopædia Britannica.

- “A simplified description, especially a mathematical one, of a system or process, to assist calculations and **predictions**”

OED, 2019.

Complex User Models

Online Video



Maxwell, Azzopardi. "Agents, simulated users and humans: An analysis of performance and behaviour." CIKM, 2016.

*All models are wrong
but some are useful*



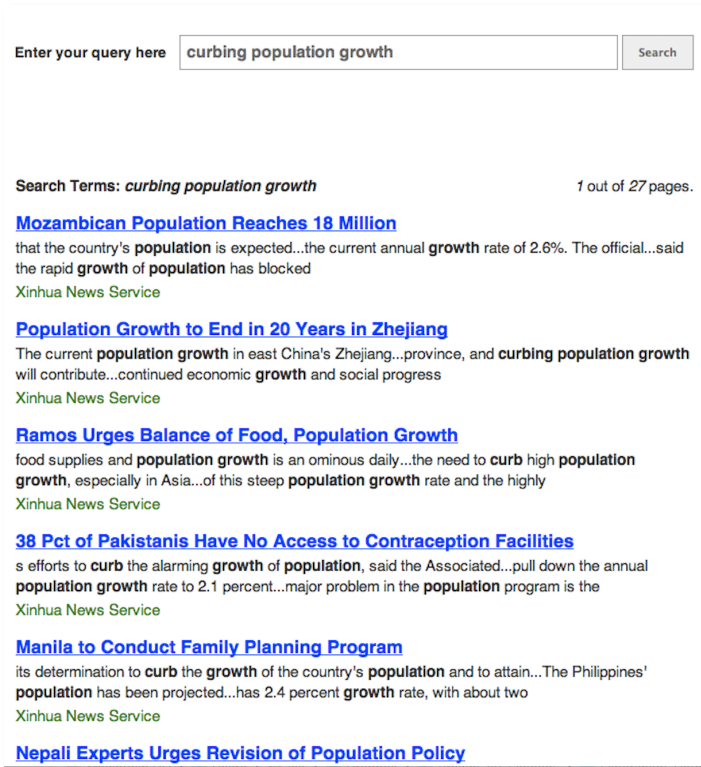
George E.P. Box

Offline, Online Evaluation; User Studies

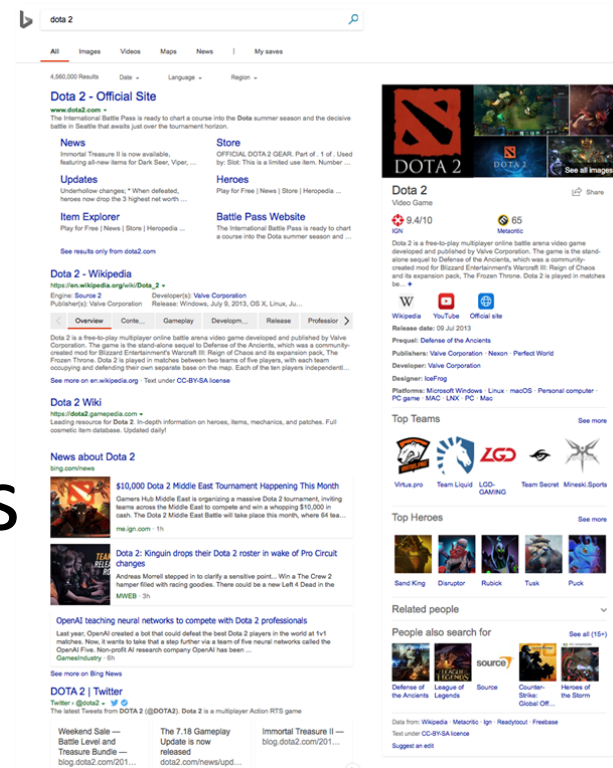
- Effectiveness evaluation categorized in three broad classes: **Offline, Online, User Studies**
- The evaluation methods we see here are **offline** methods: the **system is not live**, experiments are done through **simulations**.
 - They do not involve actual users
 - The topic of this tutorial
- What if we want to test a **system in production**, live, as it gets used? **Online evaluation!**
- Online evaluation: Test (or even train) using live traffic on a search engine
- Offline evaluation with users: **User Studies** — non-production systems, **careful control** on user, task, interactions, feedback

Measuring a SERP: Offline/test collection metrics

Our approximation of a SERP



sort of
approximates

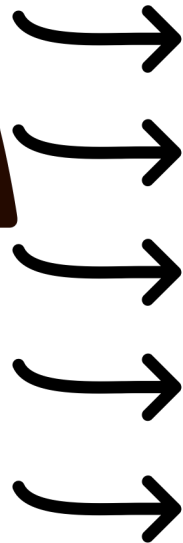


Ranked list, top down

Homogenous elements

Web SERP

How we model assessing a ranking



Enter your query here

curbing population growth

Search

Search Terms: *curbing population growth*

1 out of 27 pages.

Bad

[African Population Reaches 18 Million](#)

country's **population** is expected...the current annual **growth** rate of 2.6%. The official...said **growth** of **population** has blocked

Xinhua News Service

Bad

[Population Growth to End in 20 Years in Zhejiang](#)

current **population growth** in east China's Zhejiang...province, and **curbing population growth** contribute...continued economic **growth** and social progress

Xinhua News Service

Good

[Urges Balance of Food, Population Growth](#)

polies and **population growth** is an ominous daily...the need to **curb** high **population** especially in Asia...of this steep **population growth** rate and the highly

Xinhua News Service

Fair

[Most of Pakistanis Have No Access to Contraception Facilities](#)

s to **curb** the alarming **growth** of **population**, said the Associated...pull down the annual **population growth** rate to 2.1 percent...major problem in the **population** program is the

Xinhua News Service

Good

[Must to Conduct Family Planning Program](#)

mination to **curb** the **growth** of the country's **population** and to attain...The Philippines' **on** has been projected...has 2.4 percent **growth** rate, with about two

Xinhua News Service

[Nepali Experts Urges Revision of Population Policy](#)

Weight

×

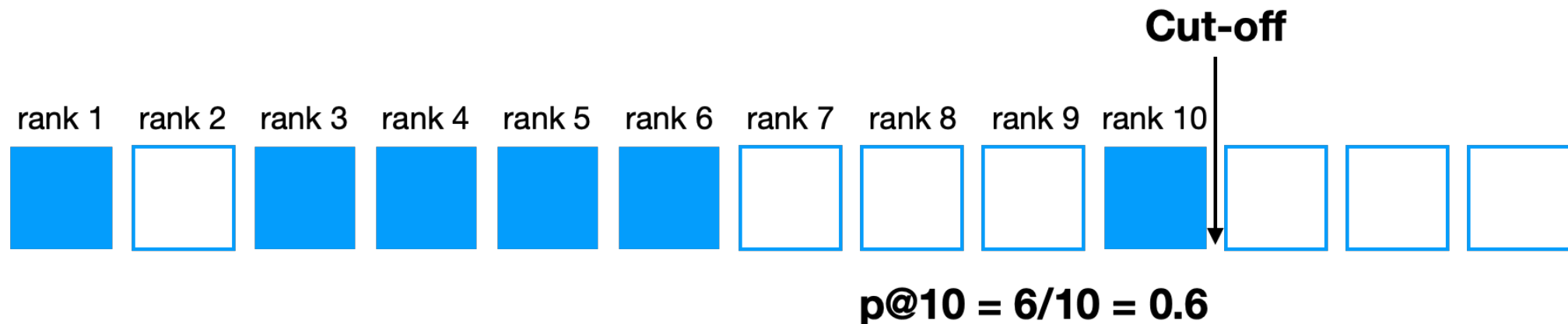
Gain

Basic Evaluation Metrics in IR

Precision at cutoff

Online Video

- Define the set of retrieved documents in function of ranking, i.e. fix a rank cutoff
- then, compute precision up to that cutoff
- e.g. **p@10**: precision up to rank 10 = (relevant docs retrieved up to rank 10) / 10



The user model of p@k

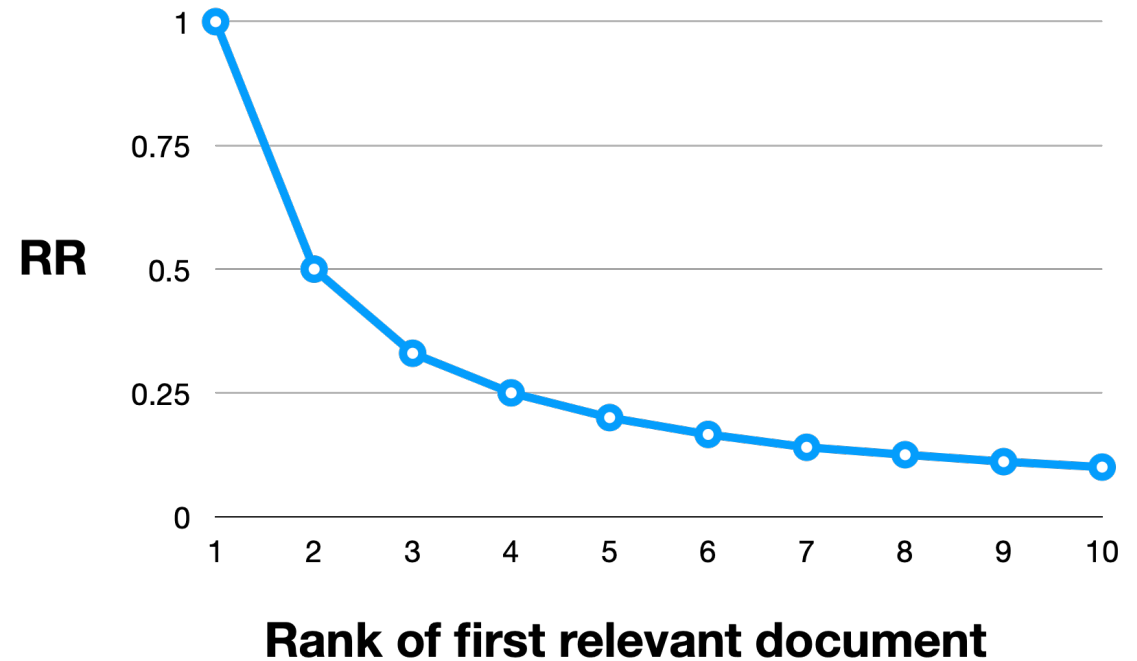
Online Video

- The user
 - examines all documents from position 1 to cut-off k
 - puts equal importance to any of the first k docs
 - wants as many relevant documents as possible
- Thus:
 - The goal of the system is to find the highest number of relevant documents among the first k retrieved.
 - No distinction between differences in the rankings at positions 1 to k

Reciprocal Rank (RR)

Online Video

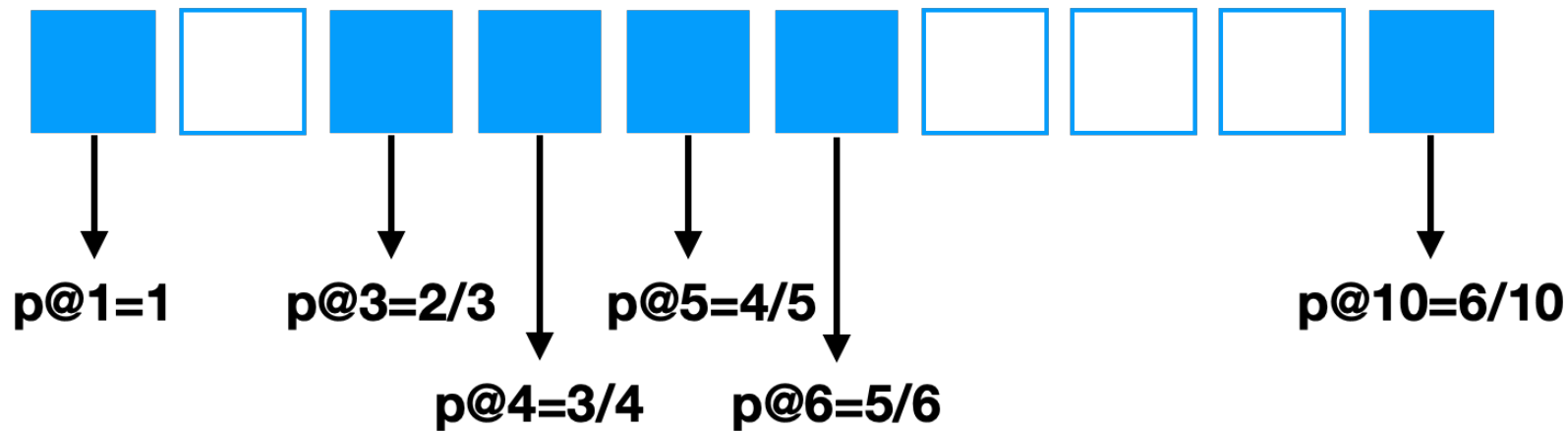
- $RR = 1 / \text{rank_first_relevant}$
 - Increasingly lower RR value obtained by larger ranks
- Task: find...
 - just one relevant document, OR
 - the only relevant document



- User model: the user examines documents in order, and stops when finds the first relevant document

Average Precision (AP)

- AP has 2 key steps
 - **Sum** the **precision** from the rank at which a relevant document is retrieved (each relevant doc produces an increase in recall)
 - and **normalise** by the number of known relevant documents



$$AP = (p@1 + p@3 + p@4 + p@5 + p@6 + p@10) / \text{num_rel} = 0.78$$

- AP value depends heavily on the highly ranked relevant documents: top-ranked documents are the most important
 - that is, AP is a **top-heavy measure**
- Unless otherwise specified, it is usually computed over the entire ranking (i.e. commonly AP@1000)

AP: relevant documents not retrieved

Online Video

- What if a relevant document in the collection is never retrieved by a system?
- Contribution of the relevant, non-retrieved document to the sums of precisions is 0
- But we still need to account for the relevant document when normalizing
- *e.g. assume 6 relevant document for query:*



$$AP = (0.5 + 0.4 + 0.5 + 0.57 + 0.5 + 0.0) / 6$$

User models of AP

Online Video

Moffat&Zobel, Rank-biased precision for measurement of retrieval effectiveness, TOIS 2008

- **Every** time a **relevant document** is encountered, the user asks “Over the documents I have seen so far, on average **how satisfied am I**”
- User writes a number on a piece of paper.
- **User continues to examine every document** in the collection (only way to ensure all relevant docs have been seen)
- At the end, user **computes the average** of the values they have written.

User models of (Probabilistic) AP

Online Video

Dupret&Piwowarski, A User Behavior Model for Average Precision and its Generalization to Graded Judgments, SIGIR'10

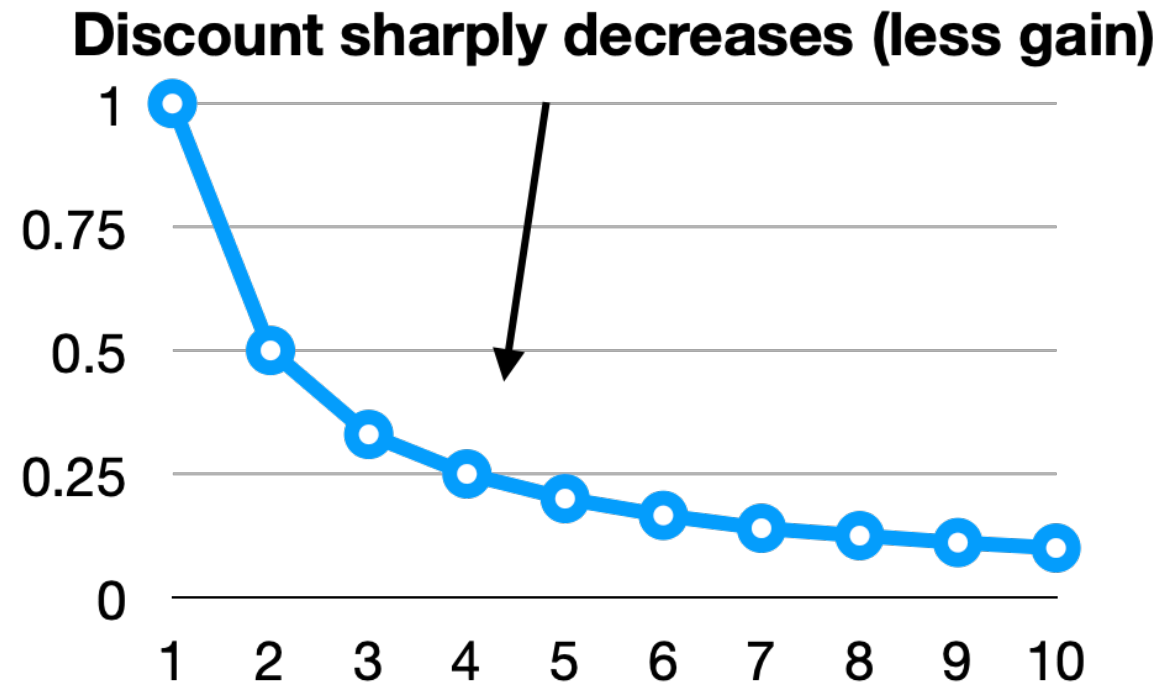
- The user decides the number n of relevant documents she needs to meet her information need.
- She browses the result list sequentially.
- She clicks on a document she examines with a probability that depends on the relevance of the document
 - 0/1 in case of binary relevant
- She ends her search as soon as she clicked on n relevant documents.

Gain&Discount metrics

Online Video

Let's revisit RR

- The first relevant document can be thought as contributing a gain of 1
- Every other relevant document retrieved contribute no gain (gain = 0)
- Each rank is associated to a discount ($d = 1/\text{rank}$)
- The user experience the gain of finding the relevant document, but decreased by the discount



Gain&Discount Framework

Online Video

$$M = \frac{1}{N} \sum_{i=1}^k gain(rel_i) \cdot discount(i)$$

- A metric may be defined in this framework
- The metric is expressed as
 - the **sum** of the **gain** generated by relevant documents
 - **weighted** by the **discount** of the **rank** at which each relevant document is retrieved
 - the sum is up to rank k
 - may be **normalised** ($1/N$)

Carterette. "System effectiveness, user models, and user utility: a conceptual framework for investigation." SIGIR, 2011.

normalised Discounted Cumulative Gain (nDCG)

Online Video

$$M = \frac{1}{\mathcal{N}} \sum_{i=1}^k \textit{gain}(\textit{rel}_i) \cdot \textit{discount}(i)$$

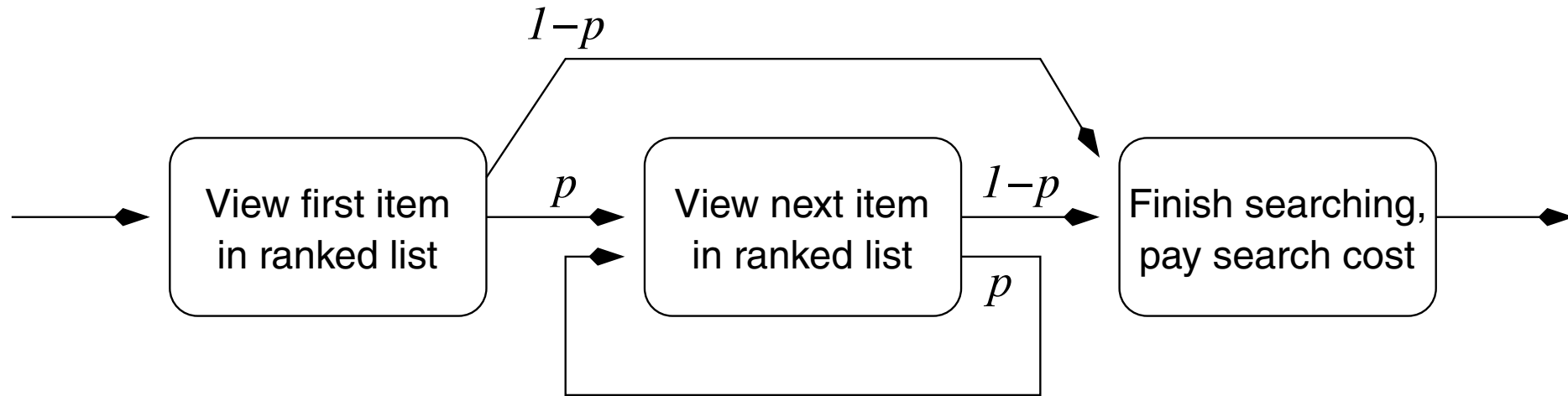
$2^{\textit{rel_k}} - 1$

$1 / \log_a (1 + k)$

- Normalisation
 - obtained by dividing DCG by ideal gain (perfect ranking for the query)
 - is useful when averaging across queries

Rank Biased Precision (RBP)

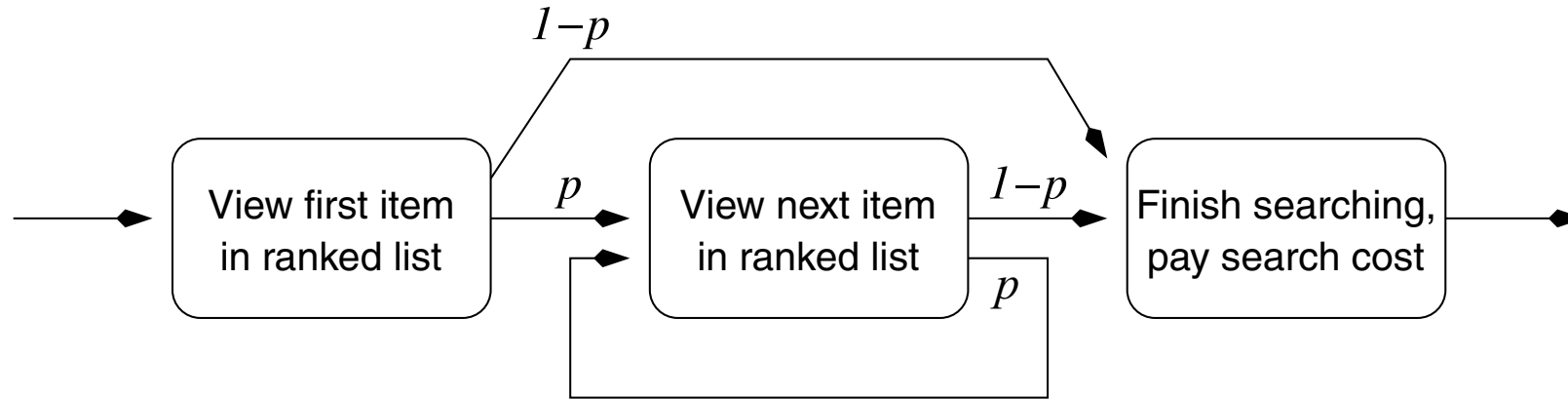
Online Video



- User Model of RBP:
 - a user always examines the first document in the list
 - then
 - examines the next with probability p
 - or stops the search with probability $1-p$

Rank Biased Precision (RBP)

Online Video



- The user model can be used to define a **discount**: function of the probability of examining a document at rank i :

$$d = p_{i-1}$$

- User receives a gain when examining a relevant document
- **Gain function**: $gain=1$ when doc relevant; $gain=0$ when non-relevant

$$g = r_i$$

Rank Biased Precision (RBP)

Online Video

normalisation

$$RBP = (1 - p) \sum_{i=1}^k r_i \cdot p^{i-1}$$

gain for doc at rank i

discount for doc at rank i

The diagram shows the RBP formula with three annotations. An arrow points from the word 'normalisation' to the term (1 - p). Another arrow points from the text 'gain for doc at rank i' to the term r_i. A third arrow points from the text 'discount for doc at rank i' to the term p^{i-1}.

- Parameter $0 \leq p \leq 1$ encodes user **persistence** or patience: the probability of continuing to the next rank
- High p : persistent user; Low p : impatient user

The Cranfield Paradigm of Test Collections

Framework for evaluation: Cranfield/TREC

- In practice, how do we go about using these measures?

The Cranfield/TREC experiments:

- Formalises a **way to experimentally evaluate IR systems**
- Predicates the development of **test collections** to measure IR effectiveness
 - A set of **queries**: sufficiently large & representative
 - A set of **documents**: large & representative
 - A set of **relevance assessments** for query-doc pairs
 - Need for completeness/exhaustivity?

TREC (and its sisters)

Online Video

- TREC (TExt Retrieval Conference - <http://trec.nist.gov/>) is an initiative from NIST (US gov agency) for the evaluation of IR systems
- Other initiatives exist:
 - CLEF: based in Europe, initial focus on cross-lingual IR
 - NTCIR: based in Japan, focus on Asian languages
 - FIRE: based in India, focus on Indian languages
- TREC is probably the most thorough and reliable: sizeable budgets for assessments; attract many participants; diversity in participants submissions and efforts

Example TREC Collections

[Online Video](#)

Collection	Tasks	# documents	# queries
Clueweb09	TREC Web Search	~1B	TREC Web09-12: 200
Clueweb12	TREC Web Search, CLEF eHealth 2016	~870M	TREC Web13-14: 100, CLEF2016: 300
New York Times Annotated Corpus	TREC Common Core'17,	~1.8M	Common Core'17: 50
Washington Post Corpus (WAPO)	TREC Common Core'18, TREC News	671,947	Common Core'18: 50 News: 150
MS MARCO Passage Ranking	MS MARCO, TREC Deep Learning, TREC CAST	~8.8M	MS MARCO: ~59K/6.9K dev in qrels, ~6.8K test in qrrels DL: 97 CAST: 50+25 topics (multiple sequential queries)

TREC Topic Example

Online Video

<top>

<num> Number: 794

<title> pet therapy

<desc> Description:

How are pets or animals used in therapy for humans and what are the benefits?

<narr> Narrative:

Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

</top>

Relevance Assessments

Online Video

- Obtaining relevance assessments is an expensive, time-consuming process
 - who does it?
 - what are the instructions?
 - what is the level of agreement?
- TREC judgments
 - depend on task being evaluated (e.g., NIST assessors, medical experts, crowd)
 - Early collections had binary assessments; recent ones are graded
 - agreement good because of “narrative”

A qrel file

Online Video

```
101 0 AP880212-0047 1
101 0 AP880219-0139 0
101 0 AP880219-0166 0
101 0 AP880222-0172 0
101 0 AP880223-0104 0
101 0 AP880229-0146 0
101 0 AP880314-0113 0
101 0 AP880314-0121 0
101 0 AP880314-0145 0
101 0 AP880320-0041 0
101 0 AP880321-0117 0
101 0 AP880323-0210 0
101 0 AP880323-0211 0
101 0 AP880324-0256 0
101 0 AP880326-0149 0
101 0 AP880329-0195 0
101 0 AP880329-0201 0
101 0 AP880330-0014 1
101 0 AP880330-0182 0
101 0 AP880404-0207 0
101 0 AP880414-0171 0
```

.....

An example assessment exercise

Online Video


Relevation! for CLEF 2017

eHealth: <http://clef2017relevation.ielab.webfactional.com>

108 - Information should discuss the different treatment modalities, including risks and benefits, for hypothyroidism.

[Back to Query](#)

[Prev Document](#) **Document "clueweb12-0008wb-15-08709" (12 / 500)** [Next Document](#)



SEARCH

HomeHealth TopicsArticles

Mental Health
By eMedTV

Health Topics

Medications

Advertisement

View All

Related eMedTV
Health Channels

Advertisement

Tyrosine Drug Interactions

If certain medications are taken with tyrosine, drug interactions can possibly affect how the medications are absorbed into your bloodstream or compound the effects. These and other interactions include those containing levodopa and thyroid medications. If you are taking these and other interactions, tell your healthcare provider about all drugs, vitamins, and supplements you are taking prior to taking tyrosine.

Judgement

☒ Unjudged
☐ Highly relevant
☐ Somewhat relevant
☐ Not relevant

Understandability

Easy Neutral Hard

Trustworthiness

Low Medium High

A TREC Result File

[Online Video](#)

101	Q0	WSJ870226-0091	1	0.7194	Brkly3
101	Q0	WSJ861216-0134	2	0.7078	Brkly3
101	Q0	AP890130-0077	3	0.7005	Brkly3
101	Q0	WSJ880523-0063	4	0.6999	Brkly3
101	Q0	WSJ881007-0136	5	0.6932	Brkly3
101	Q0	AP881030-0049	6	0.6912	Brkly3
101	Q0	AP880714-0012	7	0.6844	Brkly3
101	Q0	AP890426-0036	8	0.6844	Brkly3
101	Q0	AP881024-0011	9	0.6800	Brkly3
101	Q0	AP880608-0123	10	0.6766	Brkly3
101	Q0	WSJ870408-0045	11	0.6745	Brkly3
101	Q0	AP880314-0145	12	0.6743	Brkly3
101	Q0	AP890717-0130	13	0.6683	Brkly3
101	Q0	WSJ870715-0122	14	0.6663	Brkly3
101	Q0	AP891215-0115	15	0.6651	Brkly3
101	Q0	WSJ880712-0128	16	0.6614	Brkly3
101	Q0	AP890718-0020	17	0.6609	Brkly3
101	Q0	AP880611-0055	18	0.6601	Brkly3
101	Q0	DOE1-76-0712	19	0.6598	Brkly3
101	Q0	AP880610-0262	20	0.6585	Brkly3

.....

The trec_eval Tool

Online Video

- IR has a large number of evaluation measures: different measures for different domains, tasks, user models
- There are standard/reference implementations
- **trec_eval** is one such implementation of a number of IR measures:
http://trec.nist.gov/trec_eval/
- Usage:

`trec_eval qrels run`

Where the relevance assessments are
(in format: <qid, docid, rel>)

Where the document rankings are
(in TREC format)

A more complex
usage:

`trec_eval -q -c -M1000 qrels run`

trec_eval Tips & Tricks

Online Video

- **-q**: give evaluation for each query/topic
- **-J**: Calculate all values only over the judged (either relevant or nonrelevant) documents
- **-l** (labels): minimum relevance judgement value needed for a document to be relevant. Default is 1; larger values would make the measure more restrictive (e.g. 3 for only highly relevant)
- **-m**: allows to select a measure, or a subset of measures
- **-m relstring**: relevance values for first N (default 10, otherwise relstring.k) retrieved docs printed as string, e.g. 01010-11-0

open-source Python library for TREC-like campaigns; implements common activities:

- **Querying IR Systems:** Benchmark runs from Indri, Terrier, PISA
- **Pooling Techniques:** create pools using Depth@K, Comb[Min/Max/Med/Sum/ANZ/MNZ], Take@N, RRFTake@N, RBPTake@N
- **Evaluation Measures:** P@k, R@k, AP, nDCG, Bpref, uBpref, RBP, uRBP. Break ties options: doc score, doc ranking. Allows computation of residuals & unassessed documents, and standard evaluation plots for analysis
- **Correlation and Agreement Analysis:** Pearson, Spearman, Kendall and τ -ap correlation between system rankings; Agreement between relevance assessment sets: Kappa or Jaccard
- **Fusion Techniques.** For run fusion: Comb[Max/Min/Sum/Mnz/Anz/Med], RBPFFusion, RRFFusion, BordaCountFusion.

<https://github.com/joaopalotti/trectools>

Pooling

- **Exhaustive assessments** for all documents in a collection is **not practical**
- A simple top-k pooling
 - **top k results** (for past TRECs, k varied between 50 and 200) from the rankings obtained by different search engines (or retrieval algorithms) are **merged into a pool**
 - duplicates are removed
 - documents are presented in some random order to the relevance judges
- Produces a **large number of relevance judgments for each query**, although still **incomplete**

Incomplete Relevance Assessments

- Modern test collections are formed by pooling a (hopefully) large & diverse set of runs from different systems, and assessing the relevance of these documents
- Relevance assessment are incomplete: not all documents in the collection are assessed for each and every query
- Relevant documents may exist that none of the systems that participated in the pool managed to retrieve
- Questions to consider:
 - Is systems comparison reliable? if test collection is less incomplete, would the comparison b/w two systems be the same?
 - Is it reliable to compare a system that has been pooled and not-pooled system?
 - Is an incomplete test collection reusable?

Alistair will talk more about this

Why offline evaluation



Cheap



Fast



Repeatable



Tells us about the real experience



**Meta-Evaluation: is our
evaluation good?**

Is our evaluation good?

- How do we know if our measure is good?
- How do we know if our collection is good?
- Does our evaluation setup predict user behavior / user satisfaction?

Paul will discuss these, and other issues
Leif will show how to try this out in practice

Metric Choices

- Lots of metrics:
 - which metric is best?
 - which metric should I use?
 - anything in common, any coherent way to talk about these?
- Any way to discuss, trade off, choose?
- Yes: examine the model underlying each
 - C/W/L framework!

Alistair will next provide a framework to explore these questions