9. Evaluation

Outline

9.1. Cranfield Paradigm & TREC

9.2. Non-Traditional Measures

9.3. Incomplete Judgments

9.4. Low-Cost Evaluation

9.5. Crowdsourcing

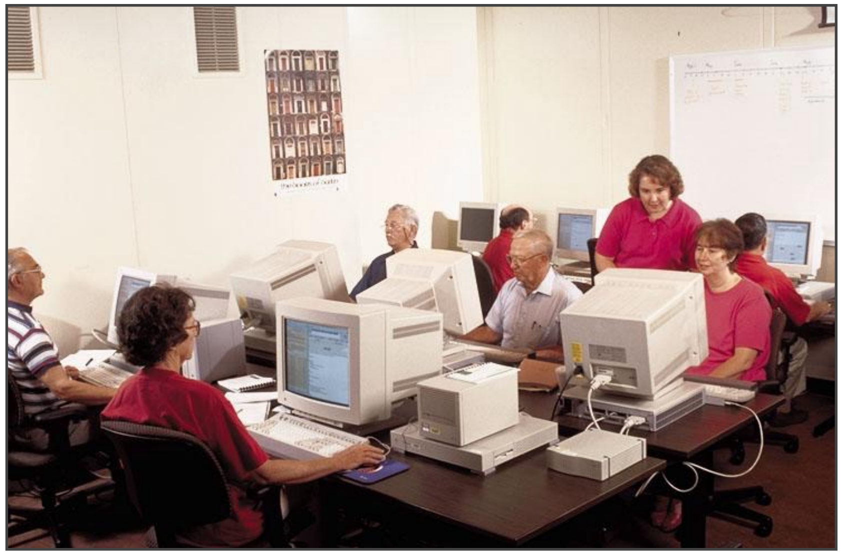
9.6. Online Evaluation

9.1. Cranfield Paradigm & TREC

- IR evaluation typically follows **Cranfield paradigm**
  - named after two studies conducted by **Cyril Cleverdon** in the 1960s who was a librarian at the College of Aeronautics, Cranfield, England
  - Key Ideas:
    - provide a **document collection**
    - define a **set of topics** (queries) upfront
    - obtain **results** for topics from different participating systems (runs)
    - collect **relevance assessments** for topic-result pairs
    - **measure** system effectiveness (e.g., using MAP)

TREC

- **Text Retrieval Evaluation Conference** (TREC) organized by the National Institute of Standards and Technology (NIST) since 1992
  - from 1992–1999 **focus on ad-hoc information retrieval** (TREC 1–8) and document collections mostly consisting of news articles (Disks 1–5)
  - **topic development** and **relevance assessment** conducted by retired information analysts from the National Security Agency (NSA)
  - nowadays **much broader scope** including tracks on web retrieval, question answering, blogs, temporal summarization



Evaluation Process

- **TREC process** to evaluate participating systems
  - (1) Release of **document collection** and **topics**
  - (2) Participants submit **runs**, i.e., results obtained for the topics using a specific system configuration
  - (3) Runs are **pooled** an a per-topic basis, i.e., merge documents returned (within top-*k*) by any run
  - (4) **Relevance assessments** are conducted; each (topic, document) pair judged by one assessor
  - (5) **Runs ranked** according to their overall performance across all topics using an agreed-upon effectiveness measure

Document Collection

Topics

Pooling

Relevance Assessments

Run Ranking

9.2. Non-Traditional Measures

- Traditional effectiveness measures (e.g., Precision, Recall, MAP) assume **binary relevance assessments** (relevant/irrelevant)

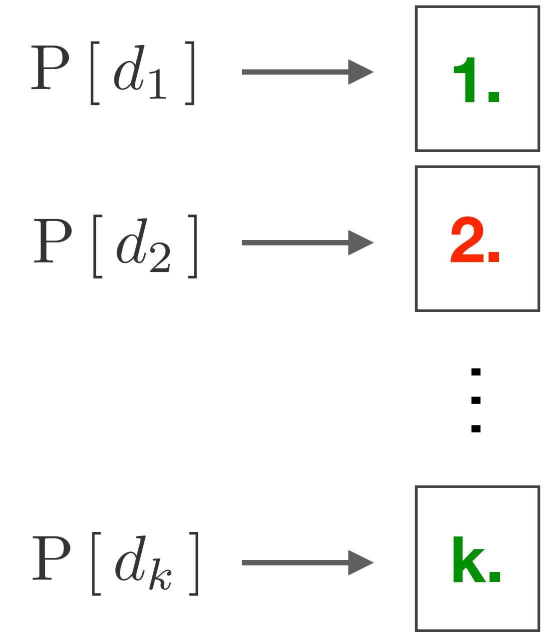
- Heterogeneous document collections like the Web and complex information needs demand **graded relevance assessments**

- User behavior exhibits **strong click bias** in favor of top-ranked results and tendency not to go beyond first few relevant results

- **Non-traditional effectiveness measures** (e.g., RBP, nDCG, ERR) consider graded relevance assessments and/or are based on more complex models of user behavior

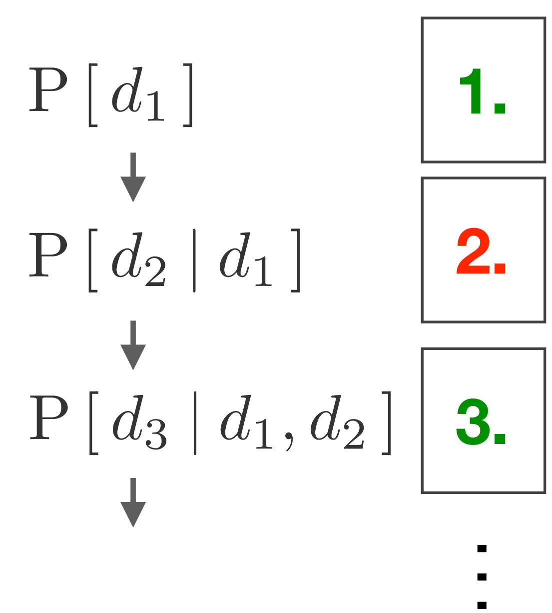
Position Models vs. Cascade Models

- **Position models** assume that user inspects each rank with **fixed probability** that is **independent** of other ranks



- Example: Precision@k corresponds to user inspecting each rank 1…k with uniform probability 1/k

- **Cascade models** assume that user inspects each rank with probability that **depends on** relevance of **documents at higher ranks**



- Example: α-nDCG assumes that user inspects rank k with probability P[n ∉ d1] x … x P[n ∉ d_{k-1}]

Rank-Biased Precision

- Moffat and Zobel [9] propose **rank-biased precision** (RBP) as an effectiveness measure based on a more realistic user model

- **Persistence parameter** p: User moves on to inspect next result with probability p and stops with probability (1-p)

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

with $r_i \in \{0,1\}$ indicating relevance of result at rank i

Normalized Discounted Cumulative Gain

- **Discounted Cumulative Gain** (DCG) considers
  - graded relevance judgments (e.g., 2: relevant, 1: marginal, 0: irrelevant)
  - position bias (i.e., results close to the top are preferred)
- Considering top-**k** result with **R(q,m)** as grade of **m**-th result

$$DCG(q,k) = \sum_{m=1}^{k} \frac{2^{R(q,m)} - 1}{\log(1+m)}$$

- **Normalized DCG** (nDCG) obtained through normalization with **idealized DCG** (iDCG) of fictitious optimal top-**k** result

$$nDCG(q,k) = \frac{DCG(q,k)}{iDCG(q,k)}$$

Expected Reciprocal Rank

- Chapelle et al. [6] propose **expected reciprocal rank** (ERR) as the expected reciprocal time to find a relevant result

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \left(\prod_{i=1}^{r-1} (1 - R_i) \right) R_r$$

with R_i as probability that user sees a relevant result at rank **i** and decides to stop inspecting result

- R_i can be estimated from **graded relevance assessments** as

$$R_i = \frac{2^{g(i)} - 1}{2^{g_{max}}}$$

- ERR equivalent to RR for binary estimates of R_i

9.3. Incomplete Judgments

- **TREC** and other initiatives typically **make** their document collections, topics, and relevance assessments **available** to foster **further research**

- Problem: When evaluating a **new system** which did not contribute to the pool of assessed results, one typically also retrieves **results** which have **not been judged**

- Naïve Solution: Results without assessment **assumed irrelevant**
  - corresponds to applying a **majority classifier** (most irrelevant)
  - induces a **bias against new systems**

Bpref

- Bpref assumes **binary relevance assessments** and evaluates a system **only based on judged results**

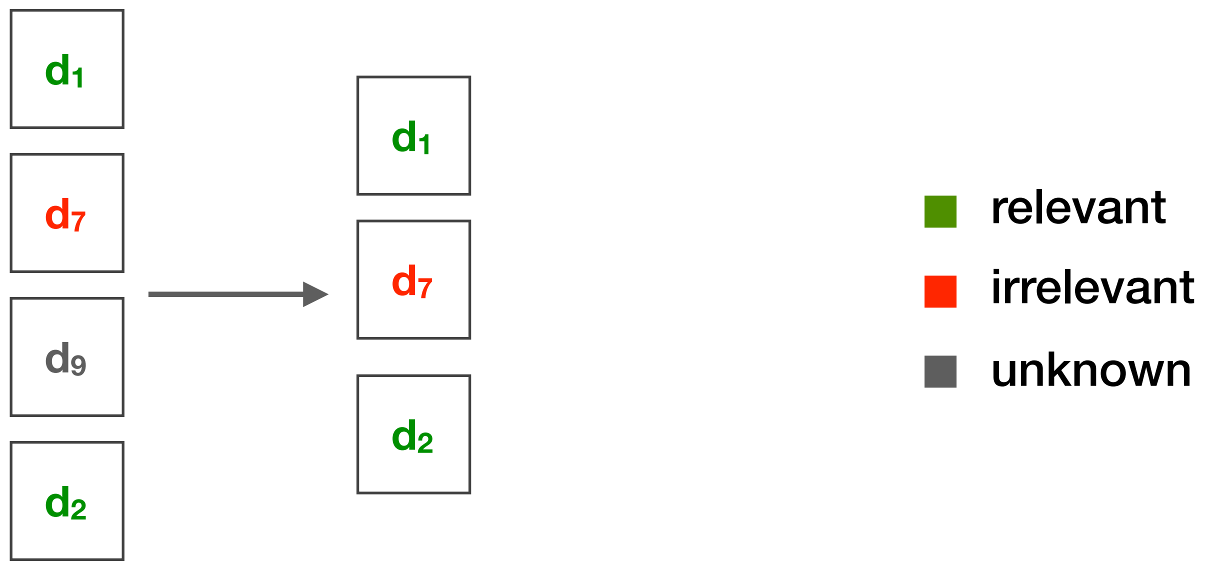
$$bpref = \frac{1}{|R|} \sum_{d \in R} \left(1 - \frac{min(|d' \in N \text{ ranked higher than } d|, |R|)}{min(|R|, |N|)} \right)$$

with R and N as sets of **relevant** and **irrelevant** results

- <u>Intuition</u>: For every retrieved relevant result compute a **penalty** reflecting how many irrelevant results were ranked higher

Condensed Lists

- Sakai [10] proposes a **more general approach** to the problem of incomplete judgments, namely to **condense result lists** by **removing all unjudged results**
    - can be used with any effectiveness measure (e.g., MAP, nDCG)



- Experiments on runs submitted to the Cross-Lingual Information Retrieval tracks of NTCIR 3&5 suggest that the **condensed list** approach is **at least as robust as bpref** and its variants

Kendall's τ

- Kendall's τ coefficient measures the **rank correlation** between **two permutations** π_i and π_j of the same set of elements

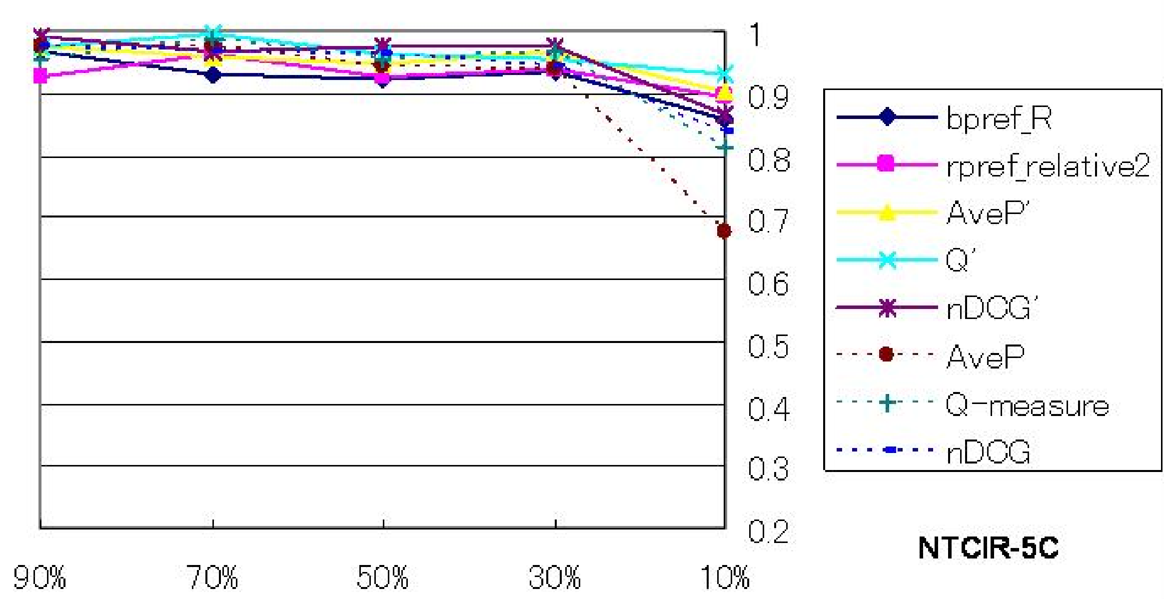
$$\tau = \frac{(\#\text{ concordant pairs}) - (\#\text{ discordant pairs})}{\frac{1}{2} \cdot n \cdot (n-1)}$$

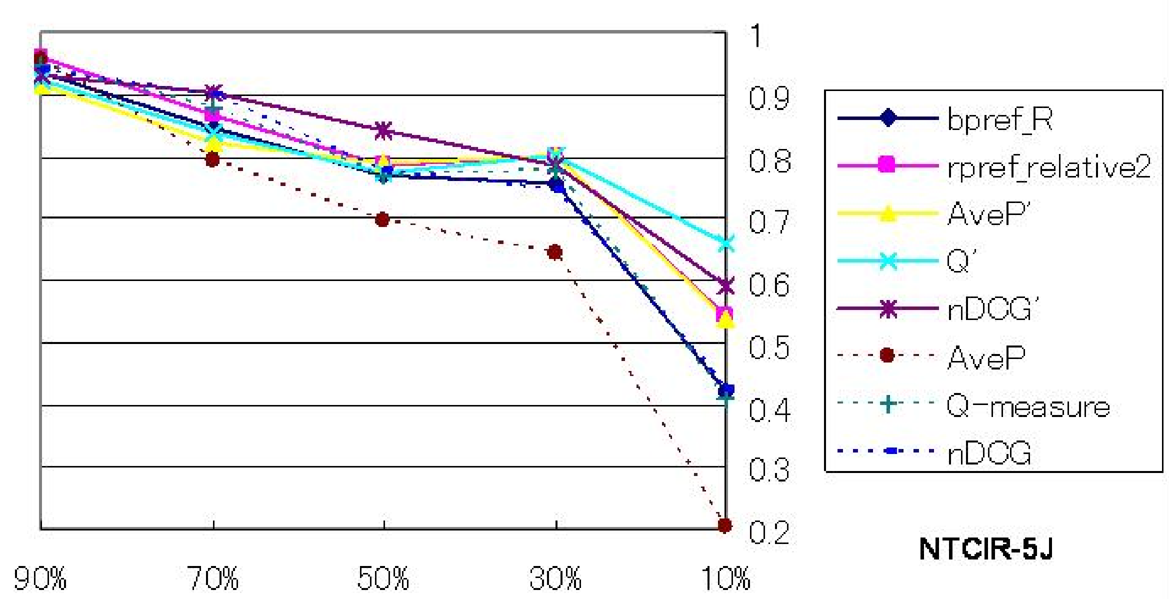
  with **n** as the number of elements

- Example: $\pi_1 = \langle$**a b c d**$\rangle$ and $\pi_2 = \langle$**d b a c**$\rangle$
  - concordant pairs: (**a**,**c**) (**b**,**c**)
  - discordant pairs: (**a**,**b**) (**a**,**d**) (**b**,**d**) (**c**,**d**)
  - Kendall's τ: -2/6

Experiments

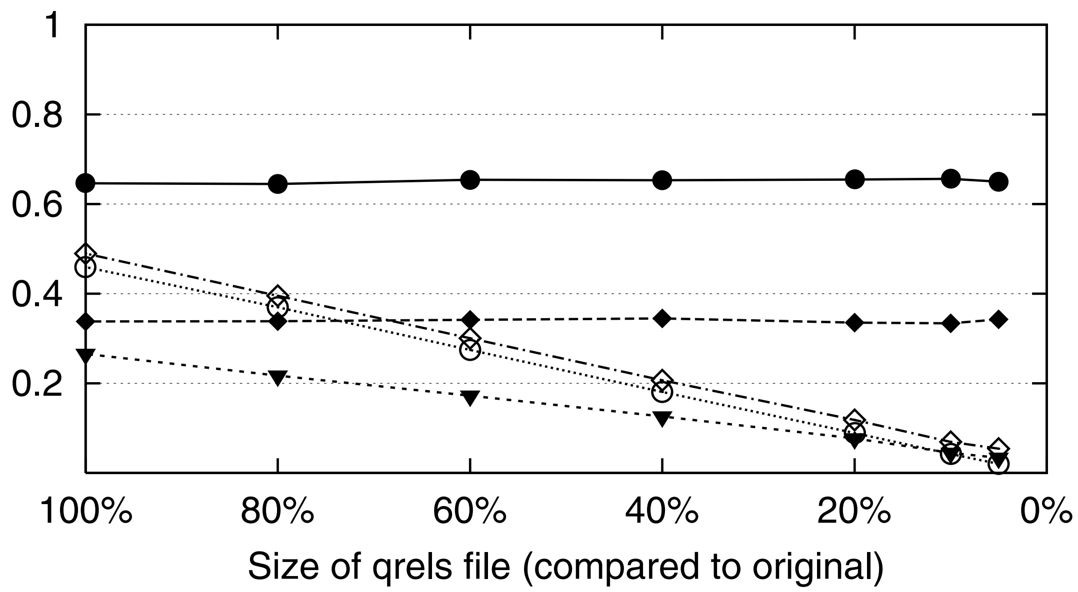
- Sakai [10] compares the **condensed list** approach on several effectiveness measures against bpref in terms of **robustness**
- Setup: Remove a **random fraction of relevance assessments** and compare the resulting system ranking in terms of Kendall's τ against the original system ranking with all relevance assessments

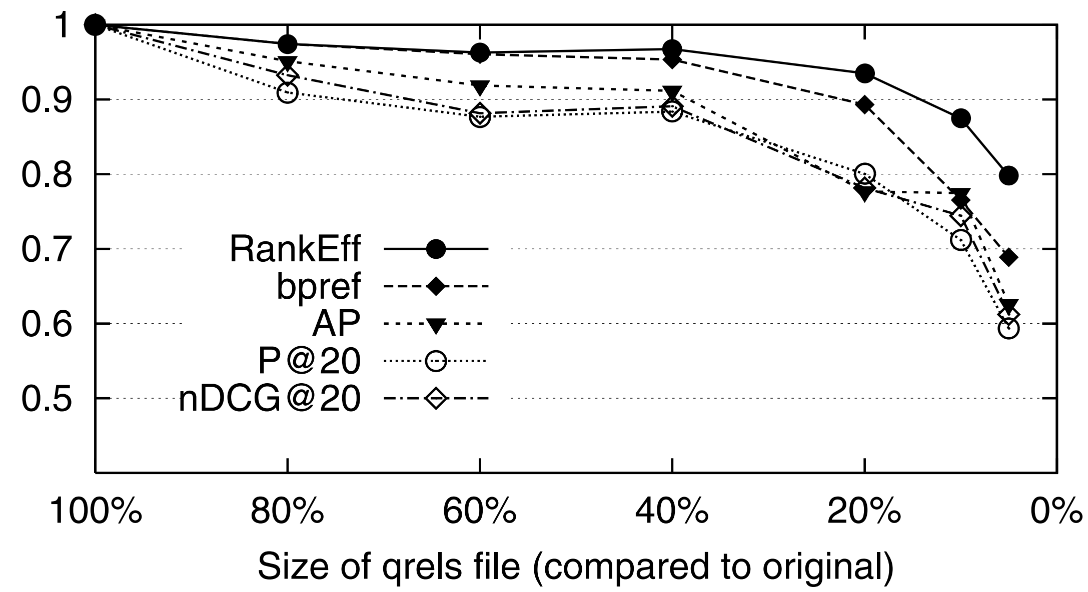




Label Prediction

- Büttcher et al. [3] examine the **effect of incomplete judgments** based on runs submitted to the TREC 2006 Terabyte track





- They also examine the amount of **bias against new systems** by removing judged results solely contributed by one system

	MRR	P@10	P@20	nDCG@20	Avg. Prec.	bpref	P@20(j)	RankEff
Avg. absolute rank difference	0.905	1.738	2.095	2.143	1.524	2.000	2.452	0.857
Max. rank difference	0↑/15↓	1↑/16↓	0↑/12↓	0↑/14↓	0↑/10↓	14↑/1↓	22↑/1↓	4↑/3↓
RMS Error	0.0130	0.0207	0.0243	0.0223	0.0105	0.0346	0.0258	0.0143
Runs with significant diff. ($p < 0.05$)	4.8%	38.1%	50.0%	54.8%	95.2%	90.5%	61.9%	81.0%

Label Prediction

- Idea: **Predict missing labels** using classification methods

- Classifier based on **Kullback-Leibler divergence**
  - estimate unigram language model θ_R from relevant documents
  - document **d** with language model θ_d is considered **relevant** if

$$KL(\theta_d \| \theta_R) < \psi$$

  with **threshold** ψ estimated such that **exactly** |R| documents in the training data exceed it and are thus considered relevant

Label Prediction

- Classifier based on **Support Vector Machine** (SVM)

$$\mathrm{sign}(\mathbf{w}^T \cdot \mathbf{x} + b)$$

  with **w** $\in R^n$ and **b** $\in R$ as parameters and **x** as document vector

  - consider the 10^6 **globally most frequent terms** as features
  - features determined using **tf.idf weighting**

Label Prediction

- **Prediction performance** for varying amounts of training data

Training data	Test data	KLD classifier			SVM classifier		
		Precision	Recall	F_1 measure	Precision	Recall	F_1 measure
5%	95%	0.718	0.195	0.238	0.777	0.162	0.174
10%	90%	0.549	0.252	0.293	0.760	0.212	0.243
20%	80%	0.455	0.291	0.327	0.742	0.246	0.307
40%	60%	0.403	0.329	0.356	0.754	0.354	0.420
60%	40%	0.403	0.353	0.370	0.792	0.386	0.455
80%	20%	0.413	0.338	0.355	0.812	0.413	0.474
Automatic-only	Rest	0.331	0.318	0.262	0.613	0.339	0.355
Manual-only	Rest	0.233	0.400	0.231	0.503	0.419	0.364

- **Bias against new systems** when predicting relevance of results solely contributed by one system

		MRR	P@10	P@20	nDCG@20	Avg. Prec.	bpref	P@20(j)	RankEff
KLD	Avg. absolute rank diff.	0.976	0.929	1.000	1.214	0.667	1.119	1.000	1.071
KLD	Max. rank difference	9↑/8↓	2↑/11↓	7↑/7↓	7↑/8↓	3↑/8↓	5↑/9↓	7↑/7↓	5↑/5↓
KLD	RMS Error	0.0499	0.0245	0.0238	0.0442	0.0067	0.0179	0.0238	0.0103
KLD	% significant ($p < 0.05$)	14.3%	19.1%	28.6%	40.5%	54.8%	64.3%	28.6%	52.4%
SVM	Avg. absolute rank diff.	0.595	0.500	0.619	0.691	0.691	0.667	0.619	0.643
SVM	Max. rank difference	1↑/7↓	0↑/4↓	1↑/6↓	4↑/5↓	3↑/7↓	2↑/5↓	1↑/6↓	1↑/4↓
SVM	RMS Error	0.0071	0.0086	0.0088	0.0078	0.0046	0.0068	0.0088	0.0028
SVM	% significant ($p < 0.05$)	2.4%	7.1%	16.7%	33.3%	35.7%	16.7%	16.7%	26.2%

9.4. Low-Cost Evaluation

- Collecting relevance assessments is **laborious and expensive**

- Assuming that we **know returned results**, have decided on an **effectiveness measure** (e.g., P@k), and are **only interested in the relative order** of (two) systems: Can we pick a minimal-size set of results to judge?

- Can we **avoid collecting relevance assessments** altogether?

Minimal Test Collections

- Carterette et al. [4] show how a minimal set of results to judge can be selected so as to determine the **relative order** of two systems

- <u>Example</u>: System 1 and System 2 compared under P@3

  - determine **sign of** ΔP@3(S_1, S_2)

$$\Delta P@k = \frac{1}{k}\sum_{i=1}^{n} x_i \cdot \mathbb{1}(rank_1(i) \leq k) - \frac{1}{k}\sum_{i=1}^{n} x_i \cdot \mathbb{1}(rank_2(i) \leq k)$$

$$= \frac{1}{k}\sum_{i=1}^{n} x_i \cdot [\mathbb{1}(rank_1(i) \leq k) - \mathbb{1}(rank_2(i) \leq k)]$$

S_1	S_2
A	C
B	B
E	D
D	A
C	E

  - judging a document only **provides additional information** if it is within the top-k of **exactly one** of the two systems

Minimal Test Collections

- **iteratively judge** documents with

$$\mathbb{1}(rank_1(i) \leq k) - \mathbb{1}(rank_2(i) \leq k) \neq 0$$

- determine **upper and lower bound** of ΔP@k(S_1, S_2) after every judgment

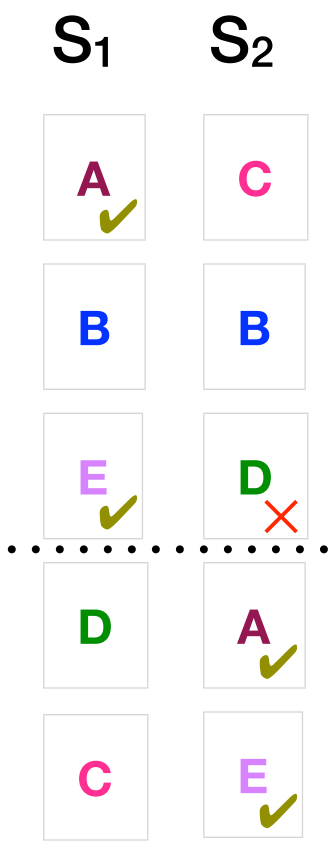
$$\Delta P@3(S_1, S_2) = 2/3 - 0/3$$

upper bound (if C is irrelevant)

$$\Delta P@3(S_1, S_2) \leq 2/3 - 0/3$$

lower bound (if C is relevant)

$$2/3 - 1/3 \leq \Delta P@3(S_1, S_2)$$



- **terminate collecting relevance** assessments as soon as upper bound smaller than -1 or lower bound larger than +1

Automatic Assessments

- Efron [8] proposes to assess relevance of results **automatically**

- Key Idea: **Same information need** can be expressed by **many query articulations** (aspects)

- Approach:
  - Determine for each topic t a set of **aspects** $a_1 \ldots a_m$
  - Retrieve **top-k results** $R_k(a_i)$ with **baseline system** for each a_i
  - Consider **all results in union** of $R_k(a_i)$ **relevant**

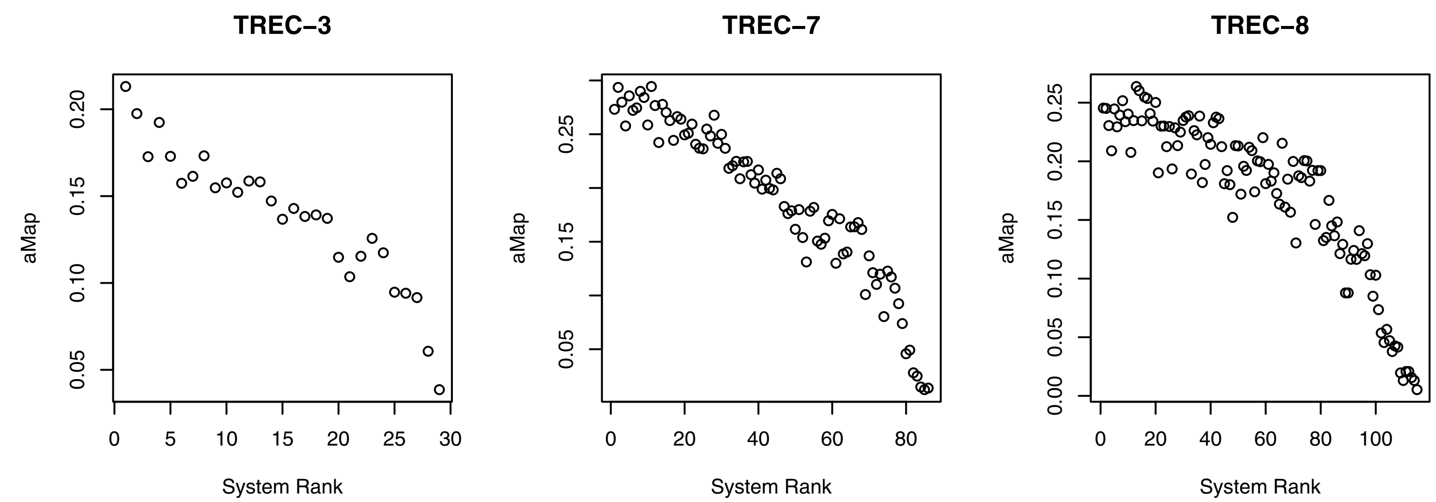
Automatic Assessments

- How to determine **query articulations** (aspects)?
  - **manually** by giving users the topic description, letting them search on Google, Yahoo, and Wikipedia, and **recording their query terms**
  - **automatically** by using **automatic query expansion methods** based on pseudo-relevance feedback
- Experiments on TREC-3, TREC-7, TREC-8 with
  - **two manual aspects** (A_1, A_2) per topic (by author and assistant)
  - **two automatic aspects** (A_3, A_4) derived from A_1 and A_2
  - **Okapi BM25** as baseline retrieval model

Automatic Assessments

- **Kendall's τ** between original system ranking under **MAP** and system ranking determined with automatic assessments

Data	tau
TREC-3	0.852
TREC-7	0.867
TREC-8	0.77



- Performance of query aspects $A_1 \ldots A_4$ when used in **isolation**

Data	A_1	A_2	A_3	A_4	$Union$
TREC-3	0.773	**0.857**	0.778	0.827	0.852
TREC-7	0.78	0.796	0.772	0.801	**0.867**
TREC-8	0.747	**0.77**	0.72	0.709	**0.77**

9.5. Crowdsourcing

- Crowdsourcing platforms provide a **cheap and readily available alternative** to hiring skilled workers for relevance assessments
  - Amazon Mechanical Turk (AMT) (mturk.com)
  - CrowdFlower (crowdflower.com)
  - oDesk (odesk.com)
- **Human Intelligence Tasks** (HITs) are small tasks that are easy for humans but difficult for machines (e.g., labeling an image)
  - workers are paid a **small amount** (often \$0.01–\$0.05) per HIT
  - workers from **all-over-the-globe** with **different demographics**

Example HIT

Judge the Relevance of a Document to a Query

We are interested in cases where **temporal information** is important to satisfy an information need. By temporal information we mean any time reference (e.g., “August 1999”, “last week”, “20th century”, or “January 1 2002”) contained in documents.

Instructions

- **Read** the document (do not just look at the title)
- **Judge** whether the document is relevant or not relevant to the query
- **Explain** your judgment in your own words (i.e., briefly tell us why you think the document is relevant or not relevant)

Tips

- Each document should be judged **on its own merits**, i.e., a document is still relevant even if you've seen other documents containing the same information
- A document is considered relevant if it contains **both textual and temporal information matching the query**
- Only work with **meaningful explanations** will be accepted (i.e., do not just write "relevant" or "not relevant")

Task

Please judge the relevance of the following document to the query **musket 16th century**. Remember, a document is considered relevant if it contains **both textual and temporal information** matching the query.

Your continued donations keep Wikipedia running! Try Beta Log in / create account

article | discussion | edit this page | history

Pike and shot

From Wikipedia, the free encyclopedia

Example HIT

Task

Please judge the relevance of the following document to the query **musket 16th century**. Remember, a document is considered relevant if it contains **both textual and temporal information** matching the query.

Please judge the relevance of the above document to the query **musket 16th century** as follows.

- ◯ **Relevant**. A relevant document containing both textual and temporal information relevant to the query.
- ◯ **Not relevant**. The document is not good because it doesn't contain any relevant information.
- ◯ **I don't know**. I don't have enough information to evaluate this document.

Please explain why you think the document is relevant or not relevant!

Submit

Crowdsourcing Best Practices

- Alonso [1] describes **best practices** for crowdsourcing
  - **clear instructions** and description of task in **simple language**
  - use **highlighting** (bold, italics) and show **examples**
  - ask for **justification of input** (e.g., why do you think it is relevant?)
  - provide **“I don’t know”** option

Crowdsourcing Best Practices

- assign **same task to multiple workers** use **majority voting**
- continuous **quality monitoring** and **control of workforce**
  - before launch: use **qualification test** or **approval rate threshold**
  - during execution: use **honey pots** (tasks with known answer), **ban workers** who provide unsatisfactory input
  - after execution: check **assessor agreement** (if applicable), **filter out** input that was provided **too quickly**

Cohen's Kappa

- Cohen's kappa measures **agreement** between **two assessors**
- Intuition: How much does the **actual agreement** P[A] deviate from **expected agreement** P[E]

$$\kappa = \frac{\mathrm{P}\,[\,A\,] - \mathrm{P}\,[\,E\,]}{1 - \mathrm{P}\,[\,E\,]}$$

- Example: Assessors A_i, Categories C_j
  - actual agreement:
    20 / 35
  - expected agreement:
    10 / 35*8 / 35 + 10/35*11/35 + 15/35*16/35
  - Cohen's kappa: ~ 0.34

		A_2		
		C_1	C_2	C_3
	C_1	**5**	**2**	**3**
A_1	C_2	**2**	**5**	**3**
	C_3	**1**	**4**	**10**

Fleiss' Kappa

- Fleiss' kappa measures **agreement** between a **fixed number of assessors**
- Intuition: How much does the **actual agreement** P[A] deviate from **expected agreement** P[E]

$$\kappa = \frac{\mathrm{P}\left[A\right] - \mathrm{P}\left[E\right]}{1 - \mathrm{P}\left[E\right]}$$

- Definition: Assessors A_i, Subjects S_j, Categories C_k and n_{jk} as the **number of assessors** who assigned S_j to C_k
- Probability p_k that category C_k is assigned

$$p_k = \frac{1}{|S||A|} \sum_{j=1}^{|S|} n_{jk}$$

Fleiss' Kappa

- Probability **P_j** that two assessors agree on category for subject **S_j**

$$P_j = \frac{1}{|A|(|A|-1)} \sum_{k=1}^{|C|} n_{jk}(n_{jk}-1)$$

- **Actual agreement** as average agreement over all subjects

$$P[A] = \frac{1}{|S|} \sum_{j=1}^{|S|} P_j$$

- **Expected agreement** between two assessors

$$P[E] = \sum_{k=1}^{|C|} p_k^2$$

Crowdsourcing vs. TREC

- Alonso and Mizzaro [2] investigate whether crowdsourced relevance assessments can **replace TREC assessors**
  - **10 topics** from TREC-7 and TREC-8, **22 documents** per topic
  - **5 binary assessments** per (topic,document) pair from AMT
  - Fleiss' kappa **among AMT workers**: 0.195 (slight)
  - Fleiss' kappa **among AMT workers and TREC assessor**: 0.229 (fair)
  - Cohen's kappa between **majority vote among AMT workers** and **TREC assessor**: 0.478 (moderate)

9.6. Online Evaluation

- Cranfield paradigm **not suitable** when evaluating online systems
  - need for **rapid testing** of small innovations
  - some innovations (e.g., result layout) **do not affect ranking**
  - some innovations (e.g., personalization) **hard to assess** by others
  - hard to **represent user population** in 50, 100, 500 queries

A/B Testing

- **A/B testing** exposes two large-enough user populations to products **A** and **B** and measures differences in behavior
  - has its **roots in marketing** (e.g., pick best box for cereals)
  - deploy innovation on **small fraction of users** (e.g., 1%)
  - define **performance indicator** (e.g., click-through on first result)
  - **compare performance** against rest of users (the other 99%) and test for **statistical significance**

Interleaving

- Idea: Given result rankings $A = (a_1 \ldots a_k)$ and $B = (b_1 \ldots b_k)$
  - construct an **interleaved ranking** I which mixes A and B
  - show I to users and **record number of clicks** on individual results
  - click on result **scores A, B, or both a point**
  - derive **users' preference** for A or B based on total number of clicks
- Team-Draft Interleaving Algorithm:
  - flip coin whether A or B starts selecting results (players)
  - A and B take turns and select yet-unselected results
  - interleaved result I based on order in which results are picked

Summary

- **Cranfield paradigm** for IR evaluation (provide documents, topics, and relevance assessments) goes back to 1960s
- **Non-traditional effectiveness** measures handle graded relevance assessments and implement more realistic user models
- **Incomplete judgments** can be dealt with by using (modified) effectiveness measures or by predicting assessments
- **Low-cost evaluation** seeks to reduce the amount of relevance assessments that is required to determine system ranking
- **Crowdsourcing** as a possible alternative to skilled assessors which requires redundancy and careful test design
- **A/B testing** and **interleaving** as forms of online evaluation

References

[1] **O. Alonso:** *Implementing crowdsourcing-based relevance experimentation: an industrial perspective*, Information Retrieval 16:101–120, 2013

[2] **O. Alonso and S. Mizzaro:** *Using crowdsourcing for TREC relevance assessment*, Information Processing & Management 48:1053–1066, 2012

[3] **S. Büttcher, C. L. A. Clarke, P. C. K. Yeung:** *Reliable Information Retrieval Evaluation with Incomplete and Biased Judgments*, SIGIR 2007

[4] **B. Carterette, J. Allan, R. Sitaraman:** *Minimal Test Collections for Information Retrieval,* SIGIR 2006

[5] **B. Carterette and J. Allan:** *Semiautomatic Evaluation of Retrieval Systems Using Document Similarities, CIKM 2007*

References

[6] **O. Chapelle, D. Metzler, Y. Zhang, P. Grinspan:** *Expected Reciprocal Rank for Graded Relevance*, CIKM 2009

[7] **O. Chapelle, T. Joachims, F. Radlinski, Y. Yue:** *Large-Scale Validation and Analysis of Interleaved Search Evaluation*, ACM TOIS 30(1), 2012

[8] **M. Efron:** *Using Multiple Query Aspects to Build Test Collections without Human Relevance Judgments,* ECIR 2009

[9] **A. Moffat and J. Zobel:** *Rank-Biased Precision for Measurement of Retrieval Effectiveness,* ACM TOIS 27(1), 2008

[10] **T. Sakai:** *Alternatives to Bpref, SIGIR 2007*