Web Track

Jaime Arguello
INLS 890-190: Experimental IR
jarguell@email.unc.edu

February 3, 2014

Goal

- Retrieval from large web collections
  - Lots of documents (~1B)
  - High-precision important
  - Hyperlinked documents
  - Lots of spam/junk/non-English documents
  - Domain general
  - Low ambiguity of user intent (tail queries)
  - High ambiguity of user intent (head queries)

Two Tasks

- Ad-hoc retrieval task (same task, different measures)
  - single-facet topics: queries associated with a single sub-intent. tail-like queries.
  - faceted topics: queries associated with multiple sub-intents. head-like queries.
- Risk-sensitive retrieval task
  - maximize average performance, minimize probability of failure compared to a baseline

Evaluation Methodology

ClueWeb12 Collection

- Category A (full dataset)
  - 733,019,372 webpages
  - crawled from 2/10/2012-5/10/2012
  - 5TB compressed, 27TB uncompressed
- Category B (smaller option)
  - Experiments with 1B documents requires a cluster
  - 52,343,021 webpages
  - 7% of Category A

Single-Facet Topic

```
<topic number="227" type="single">
<query>i will survive lyrics</query>
<description>Find the lyrics to the song "I Will Survive".</description>
</topic>
```

- Query: input to the system
- Description: hypothetical information need used by assessors to judge relevance

Multiple-Facet Topic

```
<topic number="235" type="faceted">
<query>ham radio</query>
<description>How do you get a ham radio license?</description>
<subtopic number="1" type="inf">How do you get a ham radio license?</subtopic>
<subtopic number="2" type="nav">What are the ham radio license classes?</subtopic>
<subtopic number="3" type="inf">How do you build a ham radio station?</subtopic>
<subtopic number="4" type="inf">Find information on ham radio antennas.</subtopic>
<subtopic number="5" type="nav">What are the ham radio call signs?</subtopic>
<subtopic number="6" type="nav">Find the web site of Ham Radio Outlet.</subtopic>
</topic>
```

- Informational sub-intent (INF): multiple relevant documents
- Navigational sub-intent (NAV): a single or a few relevant documents

Faceted Topic Construction

- Created by NIST assessors by considering at the output of Bing's query suggestions and auto-complete suggestions in response to the query
- Based on query-term co-occurrence information in the Bing query-log
  - "ham radio" --> "ham radio license"
  - "ham radio" --> "building a ham radio"

Ad-hoc Evaluation

single-facet measures

- NDCG@K

$$NDCG@k = \frac{DCG@k}{iDCG@k}$$

$$DCG@k = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2 (1 + i)}$$

Ad-hoc Evaluation
gain definitions

- NAV(4): represents a homepage of the entity named in the query
- Key(3): dedicated primarily to the topic
- HRel(2): provides substantial information about the topic
- Rel(1): provides some information about the topic
- Non (0): provides no information about the topic
- Junk (0): is not useful for any purpose

Ad-hoc Evaluation

single-facet measures

- Expected Reciprocal Rank (ERR)

$$ERR := \sum_{r=1}^{n} \frac{1}{r} P(\text{user stops at position } r)$$

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

$$\mathcal{R}(g) := \frac{2^g - 1}{2^{g_{\max}}}$$

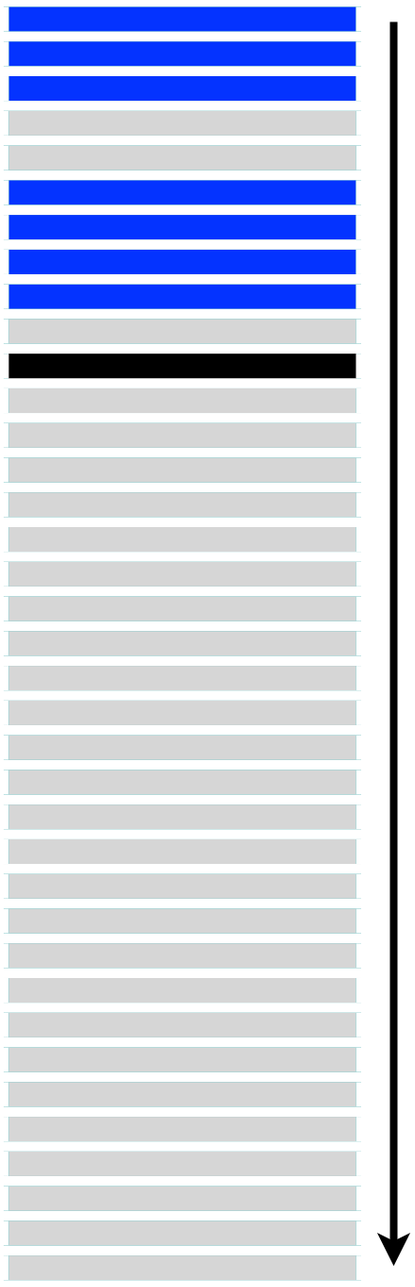
Ad-hoc Evaluation
multiple-facet measures

- IR metrics can be easily extended to evaluate diversity
  - judge documents with respect to <u>each</u> facet independently
  - compute the metric value for each facet
  - take weighted average

$$\sum_t P(t|Q) \times \text{metric}(\mathcal{R}_Q)$$

Ad-hoc Evaluation

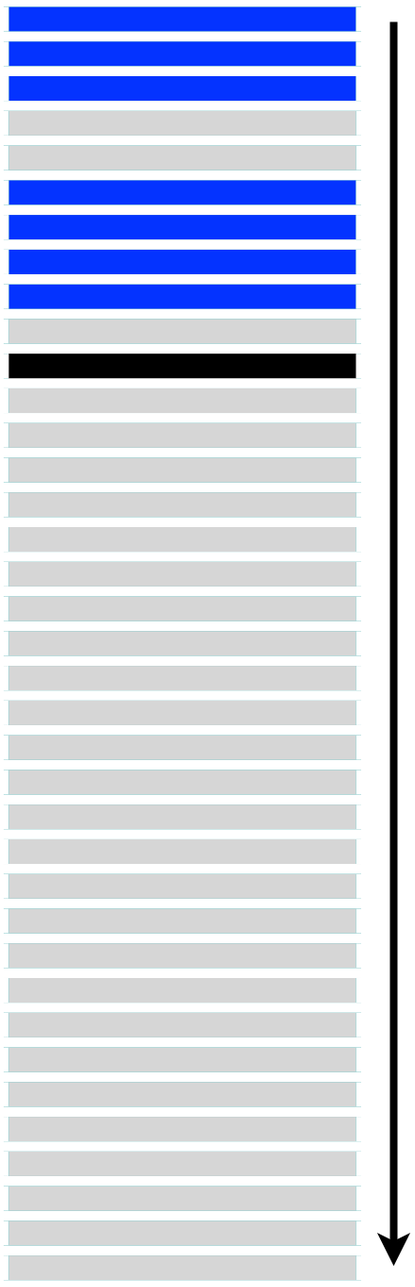
multiple-facet measures



P@10 = ?

Ad-hoc Evaluation

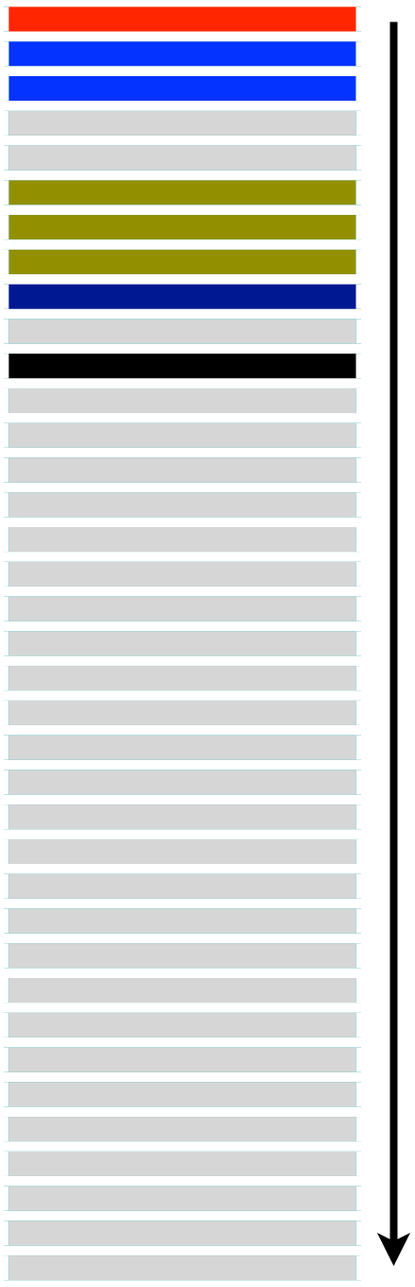
multiple-facet measures



P@10 = 0.70

Ad-hoc Evaluation

multiple-facet measures



P@10 = ?

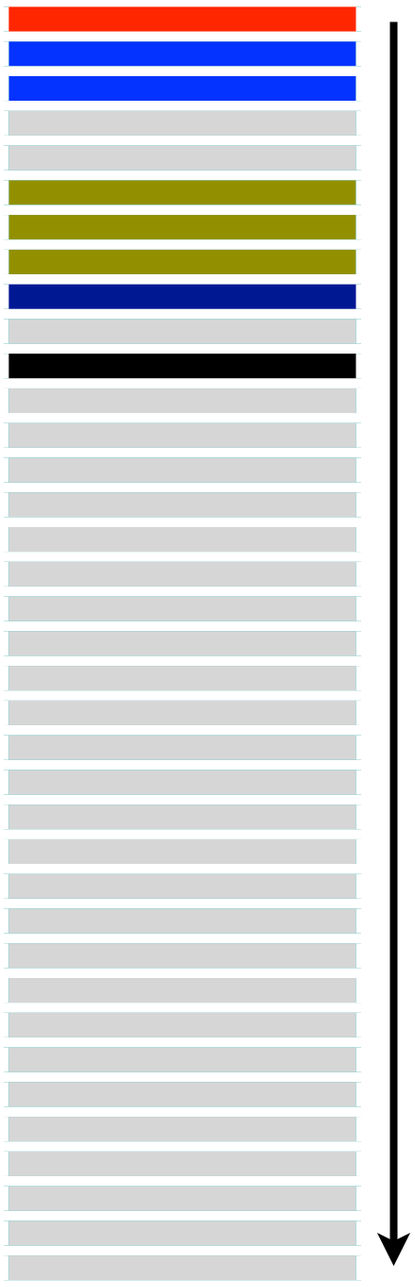
P@10 = ?

P@10 = ?

P@10 = ?

Ad-hoc Evaluation

multiple-facet measures



P@10 = 0.10

P@10 = 0.30

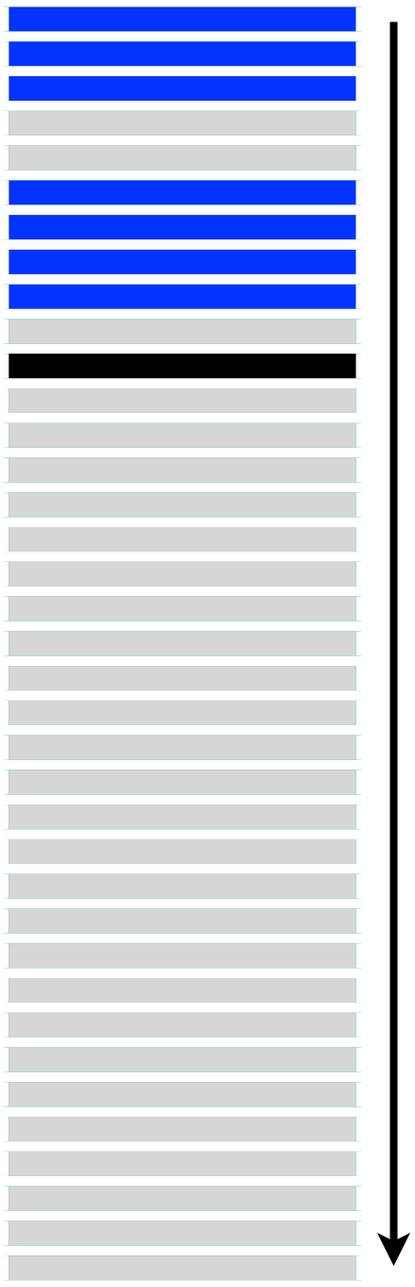
P@10 = 0.10

P@10 = 0.20

Intent P@10 =
(0.10 + 0.30 + 0.10 + 0.20) / 4 = 0.175

Ad-hoc Evaluation

multiple-facet measures



P@10 = 0.00

P@10 = 0.00

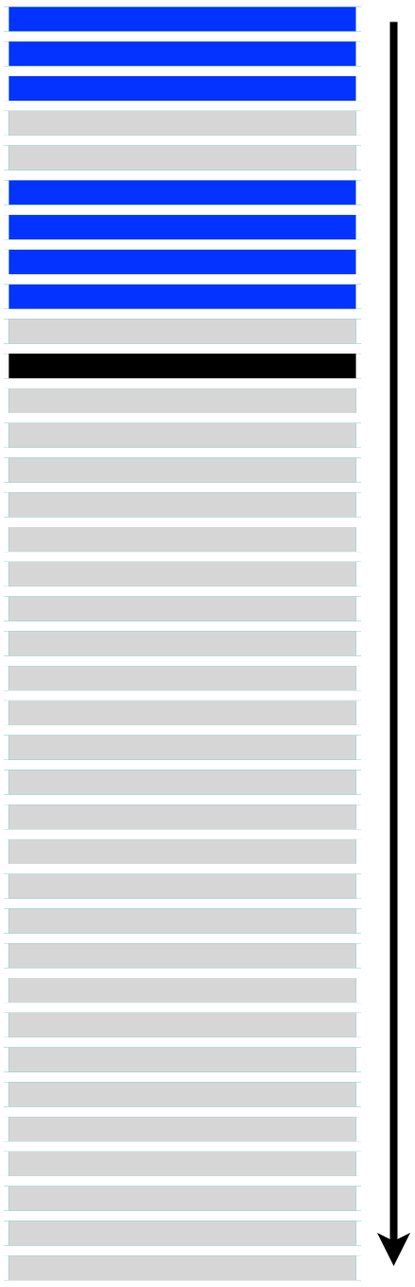
P@10 = 0.00

P@10 = 0.70

Intent P@10 =
(0.00 + 0.00 + 0.00 + 0.70) / 4 = 0.175

Ad-hoc Evaluation

multiple-facet measures



P@10 = 0.00 P(t|Q) = 0.10

P@10 = 0.00 P(t|Q) = 0.50

P@10 = 0.00 P(t|Q) = 0.10

P@10 = 0.70 P(t|Q) = 0.30

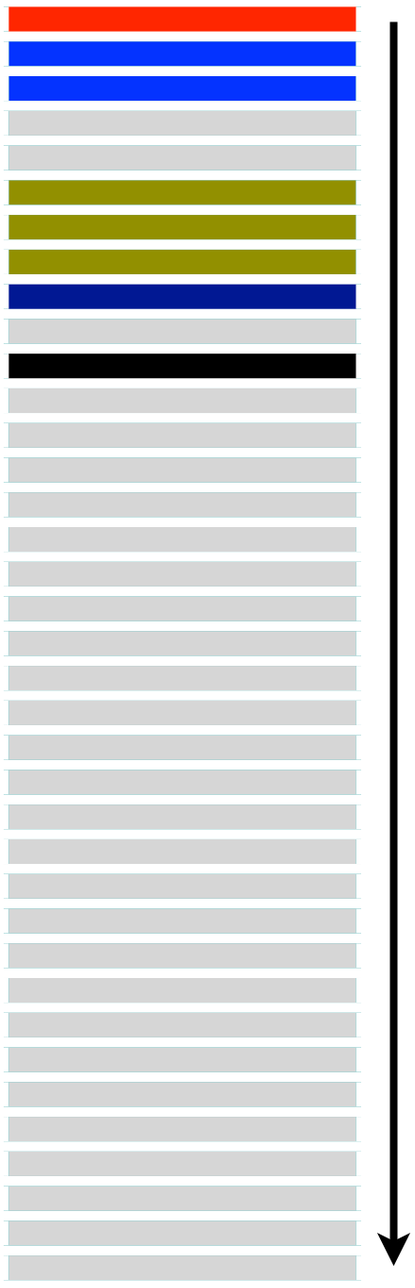
Intent P@10 =

(0.10 x 0.00) + (0.50 x 0.00) +

(0.10 x 0.00) + (0.30 x 0.70) = 0.21

Ad-hoc Evaluation

multiple-facet measures



P@10 = 0.10 P(t|Q) = 0.10

P@10 = 0.30 P(t|Q) = 0.50

P@10 = 0.10 P(t|Q) = 0.10

P@10 = 0.20 P(t|Q) = 0.30

Intent P@10 =

(0.10 x 0.10) + (0.50 x 0.30) +

(0.10 x 0.10) + (0.30 x 0.20) = 0.23

Better performance on the more probably facet!

Ad-hoc Evaluation

multiple-facet measures

- Expected Reciprocal Rank:

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

- Intent-Aware Expected Reciprocal Rank:

$$\sum_{r=1}^{n} \frac{1}{r} \sum_{t} P(t|q) \prod_{i=1}^{r-1} (1 - R_i^t) R_r^t .$$

Risk-Sensitive Evaluation

- Parameter controls the influence from poor performing queries relative to the baseline performance
- Between -1 and +1

$$U_{risk}(Q) = \frac{1}{N}\left(\sum_{q \in Q_+} \Delta(q) - (\alpha - 1) \sum_{q \in Q_-} \Delta(q)\right)$$

Algorithmic Solutions

Results

ad-hoc with single-facet

Group	Run	Cat	Type	ERR@20	nDCG@20
Technion	clustmrfaf	A	auto	0.184	0.310
udel_fang	UDInfolabWEB2	A	auto	0.176	0.282
uogTr	uogTrAIwLmb	A	auto	0.160	0.259
ICTNET	ICTNET13RSR2	A	auto	0.158	0.236
udel	udelManExp	A	manual	0.157	0.246
ut	ut22xact	A	auto	0.152	0.228
diro_web_13	udemQlm1lFbWiki	A	auto	0.152	0.254
wistud	wistud.runD	A	manual	0.134	0.225
CWI	cwiwt13cps	A	auto	0.128	0.218
UJS	UJS13LCRAd2	B	auto	0.107	0.148
RMIT	RMITSCTh	A	auto	0.102	0.179
webis	webisrandom	A	auto	0.101	0.181
MSR_Redmond	msr_alpha0_95_4	A	manual	0.097	0.175
Organizers	baseline	A	auto	0.096	0.168
UWaterlooCLAC	UWCWEB13RISK02	A	auto	0.085	0.132
DLDE	dlde	B	manual	0.008	0.007

Ad-hoc Retrieval
technion

- Retrieval top-10K docs using baseline (MRF) score
- Re-rank the top-1K using “learning to rank”
- Re-rank the top-50 using cluster-based retrieval
  - assumption: similar documents are relevant to same information needs (or, similar documents should have similar scores)
  - cluster top-50 documents
  - rank clusters based on its (average) document scores
  - re-rank documents based on cluster rank

Ad-hoc Retrieval

university of delaware

$$S(Q,D) = \sum_{t \in Q \bigcap D} C(t,Q) \times \frac{C(t,Q)}{C(t,Q) + s + s \cdot \frac{|D|}{avdl}} \times ln \frac{N+1}{df(t)}$$

- Sum of TF.IDF values from query-terms
- Expand query using terms with a high average mutual information with the query-terms

Mutual Information

$$MI(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right)$$

- $P(w_1, w_2)$: probability that words w_1 and w_2 both appear in a text
- $P(w_1)$: probability that word w_1 appears in a text, with or without w_2
- $P(w_2)$: probability that word w_2 appears in a text, with or without w_1
- The definition of "a text" is up to you (e.g., a sentence, a paragraph, a document)

Mutual Information

$$MI(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right)$$

- If $P(w_1, w_2) = P(w_1)\,P(w_2)$, it means that the words are independent: knowing that one appears conveys no information that the other one appears

- If $P(w_1, w_2) > P(w_1)\,P(w_2)$, it means that the words are not independent: knowing that one appears conveys some information that the other one appears

Mutual Information

estimation (using documents as units of analysis)

	word w_1 appears	word w_1 does not appear
word w_2 appears	a	b
word w_2 does not appear	c	d

every document falls under one of these quadrants

total # of documents
N = a + b + c +d

$P(w_1, w_2) = ?$

$P(w_1) = ?$

$P(w_2) = ?$

Mutual Information

estimation (using documents as units of analysis)

	word w_1 appears	word w_1 does not appear
word w_2 appears	a	b
word w_2 does not appear	c	d

every document falls under one of these quadrants

total # of documents
N = a + b + c +d

$P(w_1, w_2) = a / N$

$P(w_1) = (a + c) / N$

$P(w_2) = (a + b) / N$

Ad-hoc Retrieval
university of glasgow

- Learning to rank approach (LambdaMart)
- Learns to predict a documents rank as a function of a set of features
- Query-document features + document features

Features	Total
Sample: DPH, DFIC or BM25	1
Weighting models on the whole document [11] (DFRee, DPH [1], PL2 [1], BM25, Dirichlet LM, MQT [10], LGD, DFIC [6], DFIZ [6])	8
Weighting models as above on each field, namely: title, URL, body and anchor text; + PL2F	37
Term-dependence proximity models (MRF [15], pBiL [17])	2
URL (e.g. length) link (e.g. inlink counts) & content quality (e.g., fraction of stopwords, table text [2], spam classification [5]) features	15
TOTAL	63

Results

ad-hoc with multiple facets

Group	Run	Cat	Type	ERR-IA@20	α-nDCG@20	NRBP
udel_fang	UDInfolabWEB2	A	auto	0.582	0.654	0.547
Technion	clustmrfaf	A	auto	0.567	0.668	0.521
ICTNET	ICTNET13RSR3	A	auto	0.551	0.627	0.512
uogTr	uogTrAIwLmb	A	auto	0.548	0.637	0.498
udel	udelPseudo2	A	auto	0.539	0.637	0.486
ut	ut22base	A	auto	0.513	0.596	0.470
wistud	wistud.runD	A	manual	0.512	0.589	0.466
CWI	cwiwt13cps	A	auto	0.480	0.557	0.439
diro_web_13	udemQlm1lFbWiki	A	auto	0.480	0.576	0.433
UJS	UJS13Risk2	B	auto	0.468	0.539	0.434
webis	webismixed	A	auto	0.423	0.516	0.374
RMIT	RMITSCTh	A	auto	0.388	0.489	0.330
MSR_Redmond	msr_alpha1	A	manual	0.368	0.476	0.308
Organizers	baseline	A	auto	0.352	0.451	0.294
UWaterlooCLAC	UWCWEB13RISK02	A	auto	0.323	0.399	0.283
DLDE	dlde	B	manual	0.045	0.058	0.038

Intent-Aware Retrieval
baseline solution

- Maximal Marginal Relevance (MMR)
- Assumption: diversification = redundancy reduction

$$MMR = \arg \max_{D_i \in R/S} \left(\lambda \text{SIM}(Q, D_i) - (1 - \lambda) \max_{D_j \in S} \text{SIM}(D_i, D_j) \right)$$

- Simple
- Intuitive
- Doesn't explicitly model the different possible facets or senses associated with the query

Intent-Aware Retrieval

Explicit Query-Aspect Diversification (xQuAD)

$$(1-\lambda)P(D|Q) + \lambda \left(\sum_{Q_i \in Q} \left(P(Q_i|Q)P(D|Q_i) \prod_{D_j \in \mathcal{S}} (1 - P(D_j|Q_i)) \right) \right)$$

Intent-Aware Retrieval

Explicit Query-Aspect Diversification (xQuAD)

$$(1-\lambda)P(D|Q) + \lambda \left(\sum_{Q_i \in Q} \left(P(Q_i|Q) P(D|Q_i) \prod_{D_j \in \mathcal{S}} (1 - P(D_j|Q_i)) \right) \right)$$

- Simple (even if looks intimidating)
- Most components can be estimated using language modeling
- Explicitly models the different possible facets or senses associated with the query
- Requires producing the different "subqueries" for Q

Intent-Aware Retrieval

Explicit Query-Aspect Diversification (xQuAD)

- Subqueries
  - query expansion
  - document clustering (from the retrieved set)
  - query recommendations from a commercial service (e.g., Google, Bing, Yahoo)
  - any other ideas?

Intent-Aware Retrieval

technion

$$MMR = \arg \max_{D_i \in R/S} \left(\lambda \mathrm{SIM}(Q, D_i) - (1 - \lambda) \max_{D_j \in S} \mathrm{SIM}(D_i, D_j) \right)$$

- For SIM(Q,D_i) use 1/rank(D_i) as produced by cluster-based retrieval.

Results

risk-sensitive retrieval

Group	ERR@10	$\Delta, \alpha = 0$	$\Delta, \alpha = 1$	$\Delta, \alpha = 5$	$\Delta, \alpha = 10$
Technion	0.175	0.087	0.076	0.033	-0.020
udel_fang	0.167	0.078	0.059	-0.018	-0.114
udel	0.150	0.061	0.047	-0.011	-0.084
diro_web_13	0.143	0.055	0.034	-0.051	-0.158
uogTr	0.151	0.062	0.030	-0.101	-0.265
ICTNET	0.149	0.060	0.028	-0.079	-0.209
ut	0.144	0.056	0.025	-0.098	-0.248
wistud	0.125	0.037	0.005	-0.063	-0.143
CWI	0.121	0.033	0.003	-0.115	-0.263
Organizers	0.088	0.000	0.000	0.000	0.000
MSR_Redmond	0.087	-0.001	-0.009	-0.042	-0.084
RMIT	0.093	0.005	-0.027	-0.156	-0.317
UJS	0.100	0.012	-0.027	-0.184	-0.379
webis	0.093	0.005	-0.029	-0.163	-0.332
UWaterlooCLAC	0.080	-0.009	-0.040	-0.164	-0.319
DLDE	0.008	-0.081	-0.162	-0.486	-0.891

Risk-Aware Retrieval
related work

- Many "special sauces" improve average performance
- Improve a few queries by a lot; hurt many queries by a little
- Recent interest in risk-analysis for IR
- Query-expansion relies on an effective baseline retrieval
- Risk = instability
- How can we estimate retrieval effectiveness?
- Perturbation approaches: make small modifications to different components of the search process and measure the difference in the output ranking

Risk-Aware Retrieval
related work

- Perturbation approaches: make small modifications to different components of the search process and measure the difference in the output ranking
  - the query
  - documents
  - model

Risk-Aware Retrieval

university of glasgow

- Use two retrieval models/index-configurations
- Generate query-difficulty features from each retrieval
- Given a set of training queries (with relevance judgements), learn to select the model with the best retrieval

Brain-Storming Topics

- What are the possible intents for the query?
- What is the probability distribution across intents?
- Risk-management requires predicting retrieval effectiveness (aka query difficulty). What are sources of evidence for predicting retrieval effectiveness?
- How important is robustness from the user's perspective?