Discussion of papers

- Implicit relevance feedback:

  R.W. White, J.M. Jose and I. Ruthven. 2006.

  An implicit feedback approach for interactive information retrieval.

  Information Processing and Management, 42(1):166-190.

- Pseudo-relevance feedback:

  K.S. Lee, W.B. Croft and J. Allen. 2008.

  A cluster-based resampling method for pseudo-relevance feedback.

  In *Proceedings of SIGIR*, pp. 235-242.

White et al. (2006)

- Unobtrusive monitoring to derive relevance feedback
  - Explicit judgments are expensive: time consuming and difficult
    - Searchers often don’t quite know what they’re looking for
  - Can we just monitor user behavior to harvest *implicit* info?
    - What kind of behavior indicates relevance?
- Information needs are dynamic
  - Even users may not be clear initially, change over time
  - How can one detect change and adapt search?
- This paper explores an implicit, dynamic RF model

Motivation

- Many possible surrogate (implicit) measures of relevance
  - Dwell time, click-through, scrolling
  - Often defined with respect to whole document
- Relevance is taken to be a static assessment
- White et al. (2006) approach:
  - Segment information based on utility for RF
  - Update interface based on RF-derived changes
  - Continue to monitor for updated needs

Experimental interface

- Created an experimental interface to explore user behavior
- Multiple regions on the screen, linked to documents

  1. Top-ranking *sentences* from the collection
  2. Document titles
  3. Query focused summary
  4. Sentences in summary
  5. Sentences in context in the document
  6. Full document

- Termed the *relevance path*; importance weighted accordingly

Experimental interface

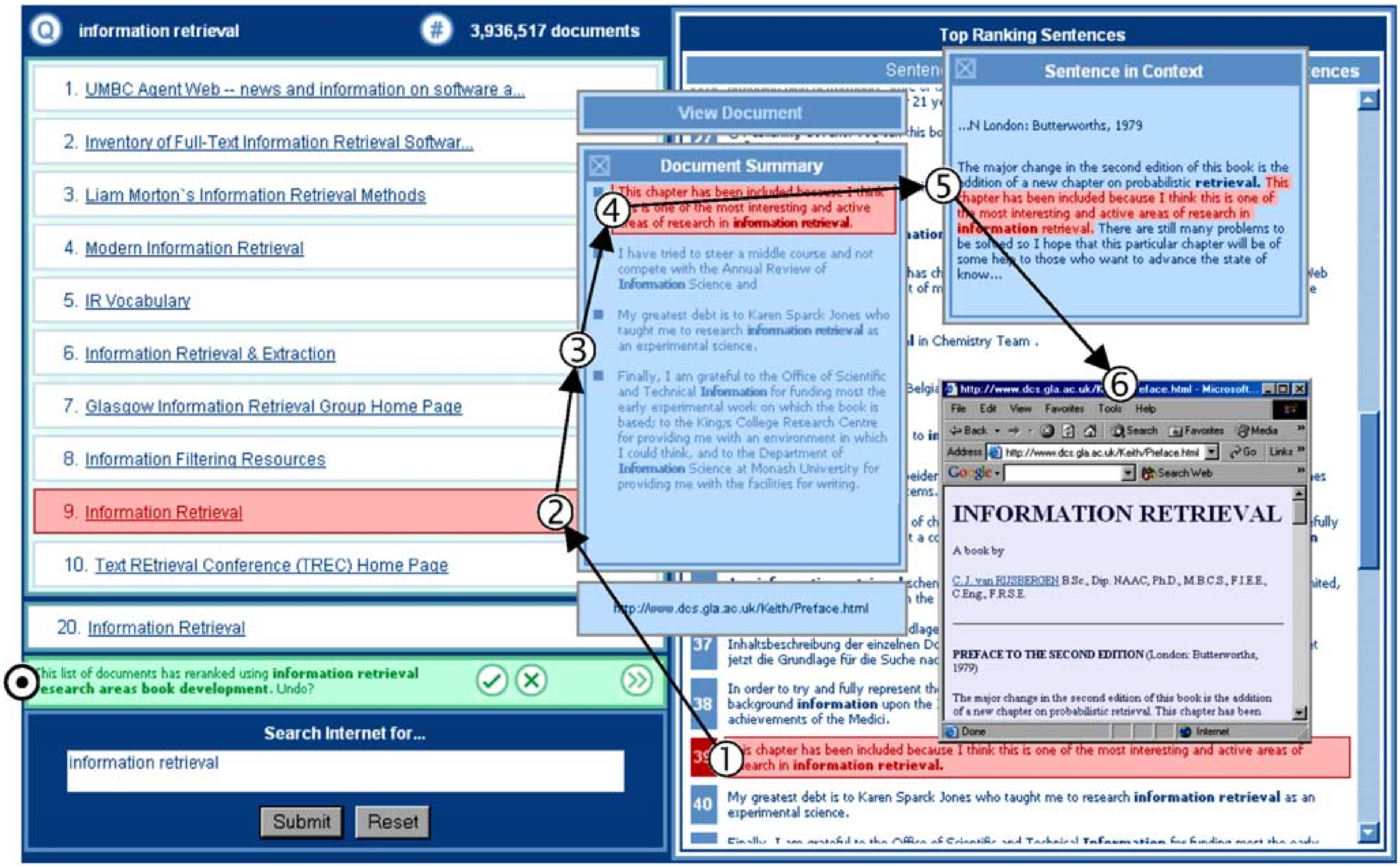
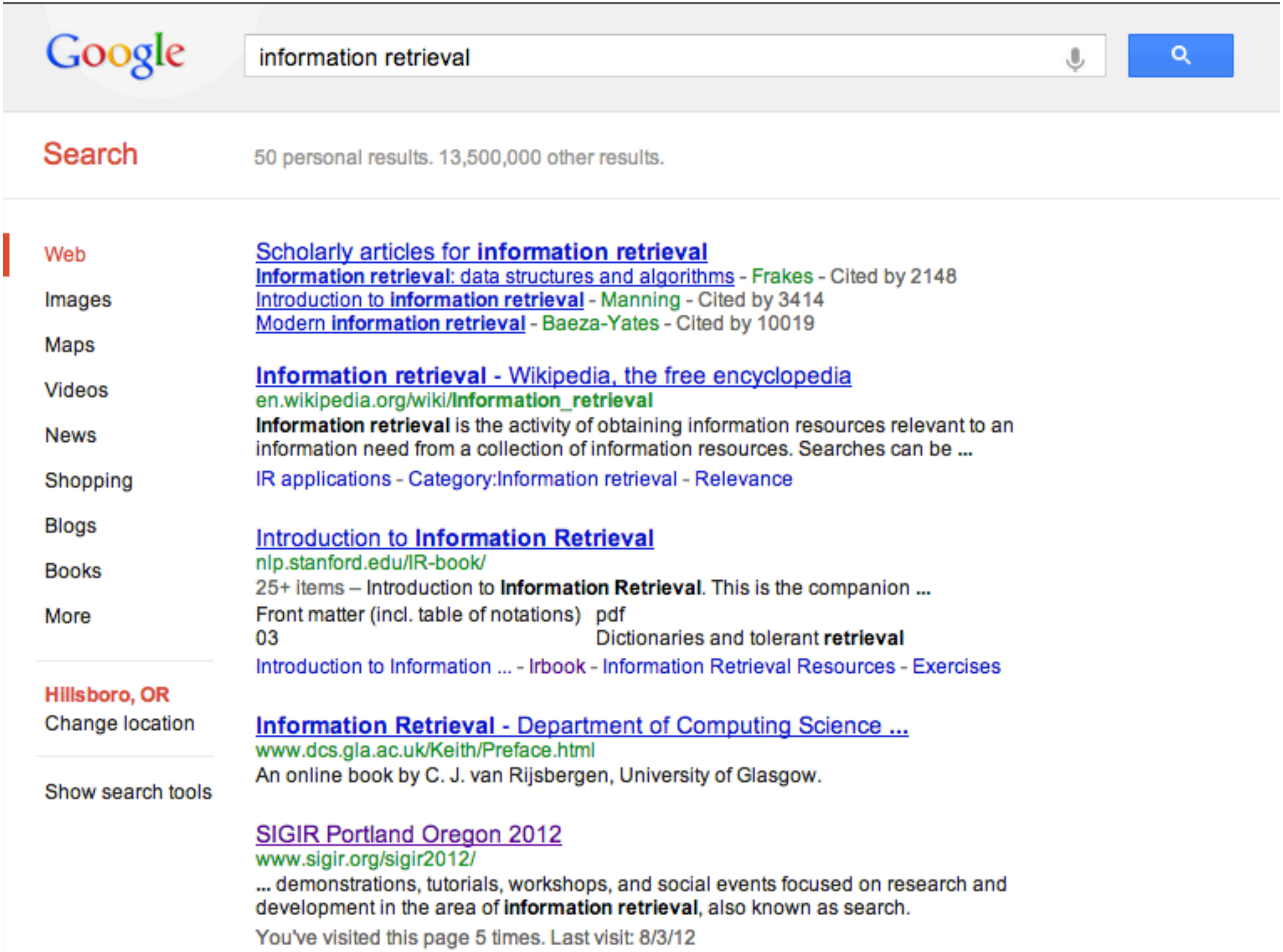


Fig. 1. Search interface with relevance path marked.



Google

information retrieval

Search

50 personal results. 13,500,000 other results.

Web
Images
Maps
Videos
News
Shopping
Blogs
Books
More

Hillsboro, OR
Change location

Show search tools

Scholarly articles for **information retrieval**
Information retrieval: data structures and algorithms - Frakes - Cited by 2148
Introduction to **information retrieval** - Manning - Cited by 3414
Modern **information retrieval** - Baeza-Yates - Cited by 10019

Information retrieval - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Information_retrieval**
Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be ...
IR applications - Category:Information retrieval - Relevance

Introduction to **Information Retrieval**
nlp.stanford.edu/IR-book/
25+ items – Introduction to **Information Retrieval**. This is the companion ...
Front matter (incl. table of notations) pdf
03 Dictionaries and tolerant **retrieval**
Introduction to Information ... - Irbook - Information Retrieval Resources - Exercises

Information Retrieval - Department of Computing Science ...
www.dcs.gla.ac.uk/Keith/Preface.html
An online book by C. J. van Rijsbergen, University of Glasgow.

SIGIR Portland Oregon 2012
www.sigir.org/sigir2012/
... demonstrations, tutorials, workshops, and social events focused on research and development in the area of **information retrieval**, also known as search.
You've visited this page 5 times. Last visit: 8/3/12

Search system

- System accesses existing search engine
- Downloads and summarizes top 30 ranked documents
  - Query-focused extractive summarizer on set of documents
  - (What is this?)
- Uses a novel query expansion method using implicit RF
  - Query expansion: add new terms to original query
- If updated model changes sufficiently, re-search new query
  - Otherwise, simply re-rank

Binary voting model

- Disclaimer: I don't 100% get how this is working
- Version of a term-document matrix, with query as a document
  - Initialized with query word cells in query row set to 1
  - Terms normalized so that query row sums to 1
  - When a document is visited, row is created, initialized to 0
  - For all (non-stop) terms, weight added based on relevance path

region:	TRS	Title	Sum	SumSent	Context	Document
weight:	0.2	0.1	0.3	0.2	0.2	0

  - Revise query every 5 "paths" through the interface
    * Sum score for each term and choose top 6 for query

What to do with new query

- Several possible actions with new query word set
  - Re-order the top-ranked sentences from documents
  - Re-order the documents
  - Re-search and derive a new 30 document set
- Calculate Spearman rank-order correlation coefficient
  - How different are the rankings?
- Pursue most disruptive/expensive actions only if list substantially reordered

Comments so far

- Interface strikes me as very busy
  - Extensive user evaluations, not much HCI
- Updating method (binary voting model) is overly heuristic
  - Heuristic weighting of sections of the relevance path
  - Weird query row normalization; then no normalization
  - Cumulative weighting may make it hard to influence later
  - Picking top 6, why?
- Wonder if there's not a probabilistic version
- Lots of empirical results for too little punchline, IMO

Meta-comments

- My presentation differs from the authors' presentation
  - I am synthesizing, not just presenting everything from the paper
  - What's important?
- I am omitting details that I don't think are central or interesting
  - Others will likely have other opinions about what to talk about
  - They should not be inhibited from bringing up the point just because I omitted it
- I didn't glean much from the experimental results
  - Rather than exhaustive slog through results, I'll present some stuff and invite comments

Experiments

- 24 subjects performing search tasks, e.g., investment strategies
  - Both experienced and inexperienced searchers
- Baseline system allowed subjects to expand initial query manually
  - Their system expanded queries automatically
- Auto query expansion substantially overlaps with manual
- Some difference in the number of 'iterations' for fact tasks
  - Not sure how to interpret this
- Overall, subject perceptions were on the positive end of the scale
  - Experienced users liked it more

Lee et al. (2008): Pseudo-relevance feedback

- Implicit RF, as with previous paper, tries to capture relevance from searcher behavior
- Pseudo-RF derives info from the document set rather than user
  - Select query expansion terms from the retrieved documents
- How to choose extra terms from the documents?
  - One baseline approach: choose from highest ranked documents
  - Narrow down selected text from highest ranked documents
  - e.g., choose from key passages or sentences
- This paper looks at new methods for choosing best documents

Related work

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes a cluster-based resampling framework. Section 4 shows experimental results on TREC test collections, results analyses and justification of the results. We will conclude in Section 5.

2. RELATED WORK

Our approach is related to previous work on pseudo-relevance feedback, resampling approaches, and the cluster hypothesis in information retrieval.

Relevance feedback (RF) and pseudo-relevance feedback (PRF) have been shown to be effective ways of improving retrieval accuracy by reformulating an original query using relevant or pseudo-relevance documents from the initial retrieval result. New interest in relevance feedback has resulted in the establishment of a relevance feedback track at TREC 2008 [30]. This track will provide a framework for exploring the effects of different factors on relevance feedback, such as initial retrieval, judgment procedure, core reformulation algorithm, and multiple iterations on large scale collection. The motivation of the track shows the current state of research: that relevance feedback is one of the successes of information retrieval over the past 30 years, in that it is applied in a wide variety of settings as both explicit and implicit feedback; however there is surprisingly little new basic research [4]. At the RIA workshop [2], there were comparative experiments on the effects of several factors for pseudo-relevance feedback. The report provides the effects of the number of documents, the number and source of terms used, the initial set of documents, and the effects of swapping documents or terms across systems. In some aspects it is not easy to see real effects, since some factors are mixed up with other effects.

Traditional pseudo-relevance feedback algorithms such as Okapi BM25 [22] and Lavrenko and Croft's relevance model [15] are based on the assumption of relevancy for the top-retrieved documents. Research has been conducted to improve traditional PRF by using passages [33] instead of documents, by using a

in terms of mean average precision (MAP) is lower than the baseline relevance model on TREC collections. Our approach primarily focuses on the effects of resampling the top-retrieved documents.

On the other hand, many information retrieval techniques have adopted the cluster hypothesis to improve effectiveness. The cluster hypothesis states that *closely related documents tend to be relevant to the same request* [11]. Re-ranking using clusters [16, 17] based on the vector space model has shown successful results. A cluster-based retrieval model [18] based on language modeling ranks clusters by the likelihood of generating the query. The results show improvements over the query-likelihood retrieval model on TREC collections. A local score regularization method [6] uses a document affinity matrix to adjust initial retrieval scores so that topically related documents receive similar scores. The results on TREC collections show that regularized scores are significantly better than the initial scores. Our work is closely related to document re-ranking using cluster validation and label propagation [32], document-based language models by the incorporation of cluster information [12], re-ranking method using cluster-based language models within a graph-based framework [14], re-ranking using affinity graph [34], and iterative pseudo-query processing using cluster-based language models [13].

There has also been work on term expansion using clustering in the vector space model [3, 2, 33, 19]. At TREC 6, Buckley *et al* [3] used document clustering on SMART though the results of using clusters did not show improvements over the baseline feedback method. At the RIA workshop to investigate the effects criteria for pseudo-relevance feedback [2], there are comparisons to investigate the effects of swapping documents and clusters by document clustering and passage-level clustering. The experimental setup is too complex to see the individual effects of clusters, since an outside source factor is mixed up with the clustering factor [33]: using outside sources for feedback itself affects the performance. Thus the analysis for the comparative experiments is inconclusive.

Clustering idea

- Estimate soft clusters over the set of retrieved documents
- Use these clusters to find *dominant documents* for query
  - Those that occur in multiple highly-ranked clusters
- Clustering works as follows
  - Documents retrieved using smoothed query-likelihood LM
  - k-NN clustering is used for top-100 documents
    - tf-idf weighting, cosine normalization and similarity
  - Then use the *relevance model* to determine query terms

$$\mathrm{R}(w) = \sum_{D \in R} P(D) P(w \mid D) P(Q \mid D)$$

    where $P(D)$ is uniform over the set
- Pick the top scoring terms in the set, add them to query

Experiments

- Five different test domains: webcrawl of .gov, TREC web, and three newswire domains
- Evaluating with mean average precision
- Two baselines
  - Baseline language model based retrieval system (LM)
  - Baseline relevance model (RM)
- Also comparing with two other systems
  - True relevance feedback (a sort of upper bound)
  - Just re-ranking, no re-search, using cluster model

Results

Table 2. Performance comparisons using mean average precision for the test topics on test collections. The superscripts α, β, γ and δ indicate statistically significant improvements over LM, Rerank, RM and Resampling, respectively. We use the paired *t*-test with significance at $p <$ 0.05.

	LM	Rerank	RM	Resampling	TrueRF
GOV2	0.3258	0.3406^{α}	$0.3581^{\alpha\beta}$	$0.3806^{\alpha\beta\gamma}$	$0.4315^{\alpha\beta\gamma\delta}$
WT10g	0.1861	0.2044^{α}	0.1966	$0.2352^{\alpha\beta\gamma}$	$0.4030^{\alpha\beta\gamma\delta}$
ROBUST	0.2920	0.3206^{α}	$0.3591^{\alpha\beta}$	$0.3515^{\alpha\beta}$	$0.5351^{\alpha\beta\gamma\delta}$
AP	0.2077	0.2361^{α}	$0.2803^{\alpha\beta}$	$0.2906^{\alpha\beta}$	$0.4253^{\alpha\beta\gamma\delta}$
WSJ	0.3258	0.3611^{α}	$0.3967^{\alpha\beta}$	$0.4033^{\alpha\beta}$	$0.5306^{\alpha\beta\gamma\delta}$

Further analysis

- Also analyzed a few other characteristics of the systems
- Current ‘resampling’ method achieves higher ‘relevance density’
  - Many graphs showing this, not clear in paper what ‘density’ is
- Also showed what happens if sampling is without redundancy
  - Not as good
  - Important to sample documents that are in multiple clusters
- Also show that the approach is robust across domains

Comments

- Strong intuition that method of picking documents for pseudo-RF should rely on centroids of clusters
  - Baseline RM just picks top-ranked candidates
- This sort of "consensus" ranking is not uncommon in other areas
  - Minimizing the empirical risk
  - Exploiting diversity in output to yield more robust solutions
- Should be a more principled way to get same benefit
  - Relevance model equation a bit unclear

Final meta-comments

- Please synthesize the papers in your presentations
  - What are the main points?
  - Situate within other topics that have been presented in class
- Cover the key important points, not necessarily every point made
  - An overly detailed march through a paper is not helpful
  - Others will speak up when something not adequately covered
- Even if you are not the lead, be prepared to speak up!
- If there are details you are not clear on, say so
- Details can be worked out on the whiteboard, too