

# Cluster-Based Retrieval Using Language Models

Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2004.

Xiaoyong Liu and W. Bruce Croft.

# A Cluster-Based Resampling Method for Pseudo-Relevance Feedback

Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008.

Kyung Soon Lee, W. Bruce Croft, and James Allan

---

Presented by Matt Chaney

CS 834 - Presentation 4

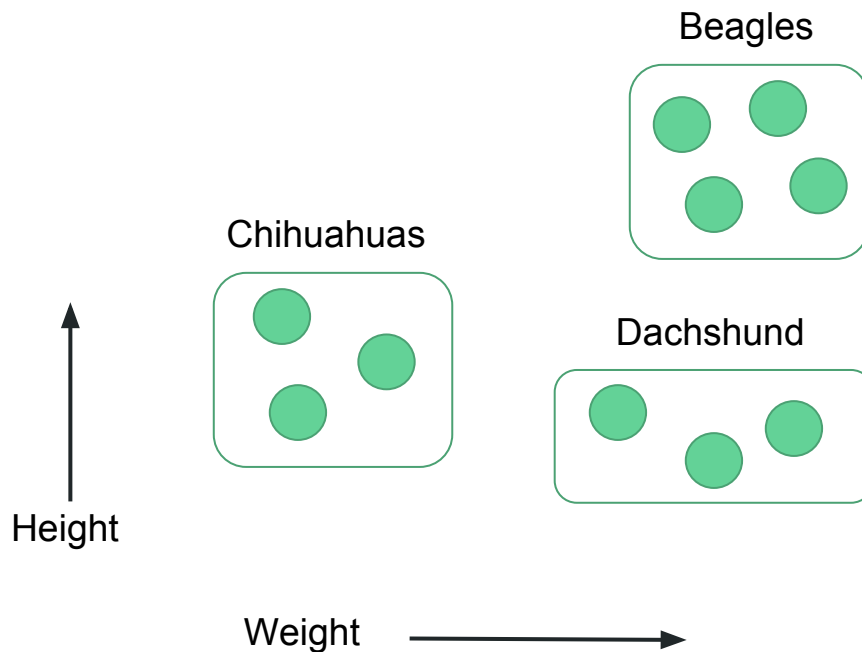
# Motivation

- Similar documents can serve the same information need
- Many studies applying clustering to IR
  - Initially to improve effectiveness/efficiency or categorize documents
- Previous study lacked definition for finding optimal document clusters
  - Automatically
  - Without relevance judgements
  - On collections of realistic size
- Apply new language modeling techniques to cluster-based IR systems
  - Provide principled way to explore document-cluster relationships
  - Language models allow for use of sophisticated smoothing parameters

# Clustering

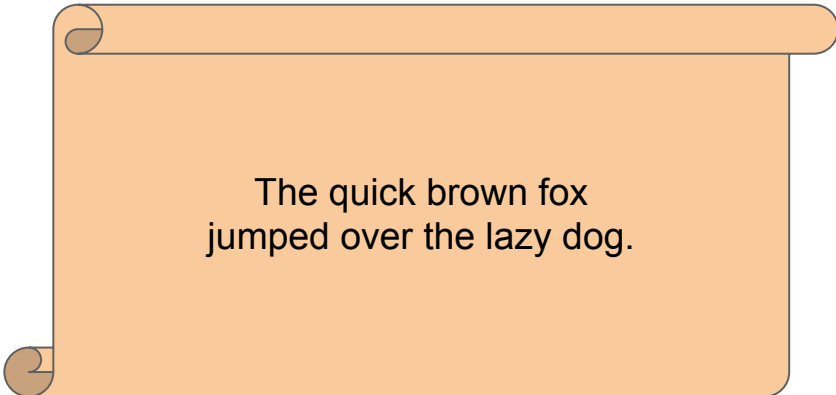
- Grouping things based on similarity of features

- Dogs
  - Height
  - Weight



# Language Models

Probability distribution over all terms in a language vocabulary

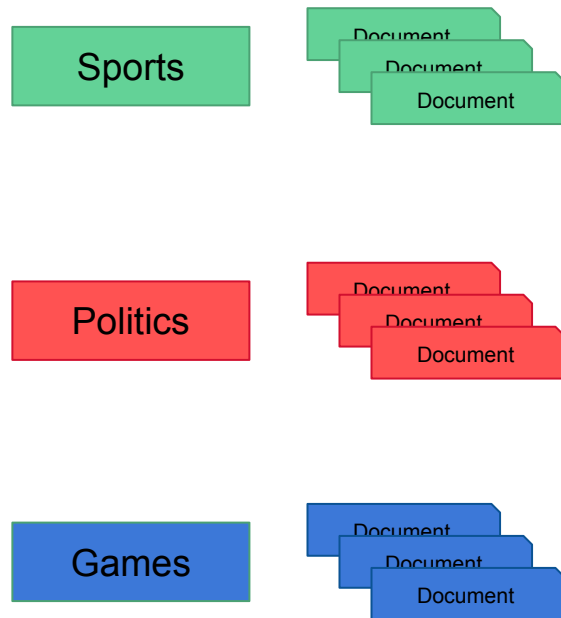


The quick brown fox  
jumped over the lazy dog.

Word	Probability
The	0.222
Quick	0.111
Brown	0.111
Fox	0.111
⋮	⋮

# Cluster-based Language Models

- Organize document collections around topics
- Create language model for topics to use as representation
- Estimate query-likelihood using cluster topics and select collection with best topic



# Cluster-based information retrieval

- Rank clusters in response to query (CQL)
  - By calculating centroid of cluster in comparison to query terms
  - Assumed arbitrary document in higher ranked cluster more relevant
- Use clusters as a form of *document smoothing* (CBDM)
  - Grouping similar documents smooths out differences among individuals

# Traditional Query-Likelihood (QL)

- Standard document language model  $P(Q | D) = \prod_{i=1}^m P(q_i | D)$
- Adding smoothing

$$P(w | D) = \lambda P_{ML}(w | D) + (1 - \lambda) P_{ML}(w | Coll)$$

- Simple Jelinek-Mercer
- Bayesian with Dirichlet prior

$$\lambda = \frac{\sum_{w' \in D} tf(w', D)}{\sum_{w' \in D} tf(w', D) + \mu}$$

# Cluster-Based Query Likelihood (CQL)

- Calculate probability of query given a cluster language model

$$P(Q \mid Cluster) = \prod_{i=1}^m P(q_i \mid Cluster)$$

- Cluster language model

$$\begin{aligned} P(w \mid Cluster) &= \lambda P_{ML}(w \mid Cluster) + (1 - \lambda) P_{ML}(w \mid Coll) \\ &= \lambda \frac{tf(w, Cluster)}{\sum_{w' \in cluster} tf(w', Cluster)} + (1 - \lambda) \frac{tf(w, Coll)}{\sum_{w' \in V} tf(w', Coll)} \end{aligned}$$



# Cluster-Based Document Smoothing (CBDM)

- Smooth document language model based on similar documents

$$\begin{aligned}P(w|D) &= \lambda P_{ML}(w|D) + (1-\lambda)P(w|Cluster) \\ &= \lambda P_{ML}(w|D) + (1-\lambda)[\beta P_{ML}(w|Cluster) + (1-\beta)P_{ML}(w|Coll)]\end{aligned}$$

- Both  $\lambda$  and  $\beta$  are general symbols for smoothing
- Two-stage smoothing
  - Cluster model smoothed with collection model
  - Document model smoothed with smoothed cluster model

# Clustering Approaches

- Distance Measures
  - Dice, Jaccard and overlap coefficients
  - Kullback-Liebler (KL) Divergence
  - **Cosine Measure**
- Partitioning → Static Clustering
  - **Three-pass K-means clustering**
- Hierarchical Agglomerative → Query-specific Clustering
  - Single/Complete Linkage
  - Group Average
  - Centroid
  - Ward's method

# Experimental Data

All queries taken from *title* field of TREC topics

Collection	Contents	# of Docs	Size	Average # of Words/Doc <sup>1</sup>	Queries	# of Queries with Relevant Docs
AP	Associated Press newswire 1988-90	242,918	0.73 Gb	473.6	TREC topics 51-150 (title only)	99
FR	Federal Register 1988-89	45,820	0.47 Gb	873.9	TREC topics 51-100 (title only)	21
WSJ	Wall Street Journal 1987-92	173,252	0.51 Gb	465.8	TREC topics 51-100 & 151-200 (title only)	100
FT	Financial Times 1991-94	210,158	0.56 Gb	412.7	TREC topics 301-400 (title only)	95
SJMN	San Jose Mercury News 1991	90,257	0.29 Gb	453.0	TREC topics 51-150 (title only)	94
LA	LA Times	131,896	0.48 Gb	526.5	TREC topics 301-400 (title only)	98

Table 1. Statistics of data sets.  
Cluster-Based Retrieval Using Language Models. Xiaoyong Liu and W. Bruce Croft. 2004.

# Experimental Design

- Evaluate ranking clusters (CQL method) with the AP and WSJ collections
  - Five clustering algorithms for cluster language model compared to baseline QL+DM
  - Bayesian smoothing w/ Dirichlet prior
  - Jelinek-Mercer smoothing
- Cluster-based retrieval (CBDM)
  - Query-Likelihood (QL+DM)
  - Relevance Model (RM+DM)
- Measured in Average Precision
- Trained parameter values before actual tests
  - Various settings for cluster distance threshold as well as smoothing parameters

# Cluster-based IR by Ranking Clusters (CQL)

1. Document-based, query-likelihood retrieval
2. Cluster top 1,000 results
3. Rank clusters with CQL method
4. Return ordered list of clusters where documents within cluster ranked according to step 1.

Collection	First-stage doc retrieval (QL+DM)	Group-average	Single-linkage	Complete-linkage	Centroid	Ward's
AP (training)	0.2179	0.2161 (t=0.8)	0.2153 (t=0.8)	0.2130 (t=0.8)	0.2164 (t=0.7)	0.2160 (t=0.8)
WSJ	0.2958	0.2902 (t=0.8)	0.2911 (t=0.8)	0.2889 (t=0.8)	0.2936 (t=0.8)	0.2963 (t=0.8)

Table 2. CQL results.

Cluster-Based Retrieval Using Language Models. Xiaoyong Liu and W. Bruce Croft. 2004.

# Cluster-based Document Smoothing (CBDM)

Collection	Simple Okapi	QL+DM	QL+CBDM	%chg	RM+DM	RM+CBDM	%chg
AP (K=2000)	0.2198	0.2179	0.2326 (+)	+6.73*	0.2745	0.2775	+1.08
WSJ (K=2000)	0.2762	0.2958 (+)	0.3006 (+)	+1.62*	0.3422	0.3445	+0.64
FT (K=2000)	0.2556	0.2610	0.2713 (+)	+3.95*	0.2835	0.2845	+0.36
SJMN (K=2000)	0.2098	0.2032	0.2171 (+)	+6.88*	0.2633	0.2673	+1.52*
LA (K=2000)	0.2279	0.2468 (+)	0.2590 (+)	+4.94*	0.2614	0.2621	+0.28
FR (K=1000)	0.2644	0.2875	0.3316	+15.37	0.1486	0.1934	+30.10

Table 5. Evaluation of Cluster-based Retrieval compared with Simple Okapi method<sup>1</sup>  
Cluster-Based Retrieval Using Language Models. Xiaoyong Liu and W. Bruce Croft. 2004.

<sup>1</sup> Sparck et. al. A probabilistic model of information retrieval: development and comparative experiments. 2004.

# Cluster-Based Retrieval Using Language Models

Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2004.

Xiaoyong Liu and W. Bruce Croft.

# A Cluster-Based Resampling Method for Pseudo-Relevance Feedback

Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008.

Kyung Soon Lee, W. Bruce Croft, and James Allan

---

Presented by Matt Chaney

CS 834 - Presentation 4

# Pseudo-Relevance Feedback (PRF)

- PRF assumes top-retrieved documents are relevant
  - Query expansion
  - Re-evaluate initial rankings
- Top retrieved documents can contain “noise” documents
  - Sub-optimal precision allows non-relevant documents to sway relevancy judgment away from initial query
- Two central issues to utilizing PRF to improve query effectiveness
  - How to select relevant documents from initial retrieval set
  - How to select query expansion terms



# Recent Cluster-based IR and PRF Developments

- Cluster theory applied successfully to IR
  - Re-ranking using clusters
  - **Cluster-based retrieval**
  - Score regularization
- Sampling / Resampling - Select more varied and novel top-ranked docs for PRF
  - Random
  - **Selective**
    - Skipping some top-ranked documents
    - Boosting - Focus subsequent training on poor performers
    - Query variant result resampling

# Cluster-Based Selective Resampling

- Resampling method using clusters
  - Document clusters can represent query subtopics
  - *Dominant documents* appear in overlapping clusters
  - Selectively resampling these documents emphasizes core query topics
- Cluster-based resampling can achieve a higher *relevance density*

$$\text{Density} = \frac{\text{the number of relevant feedback documents}}{\text{the number of feedback documents}}$$

# Resampling Process

- Uses language model and relevance model frameworks
- Dominant documents contribute more to expansion terms than other documents

Process:

1. Retrieve initial results using query-likelihood language model

$$P(Q | D) = \prod_{i=1}^m P(q_i | D) \longrightarrow \begin{aligned} P(w | D) &= \frac{|D|}{|D| + \mu} P_{ML}(w | D) + \frac{\mu}{|D| + \mu} P_{ML}(w | Coll) \\ P_{ML}(w | D) &= \frac{freq(w, D)}{|D|}, \quad P_{ML}(w | Coll) = \frac{freq(w, Coll)}{|Coll|} \end{aligned}$$

# Clustering for Dominant Documents

2. Cluster these results using K-nearest neighbors to find dominant documents
  - Assumption: If a document is in several clusters that are highly related to the query it is considered a *dominant document*

To which group do we assign the green point?

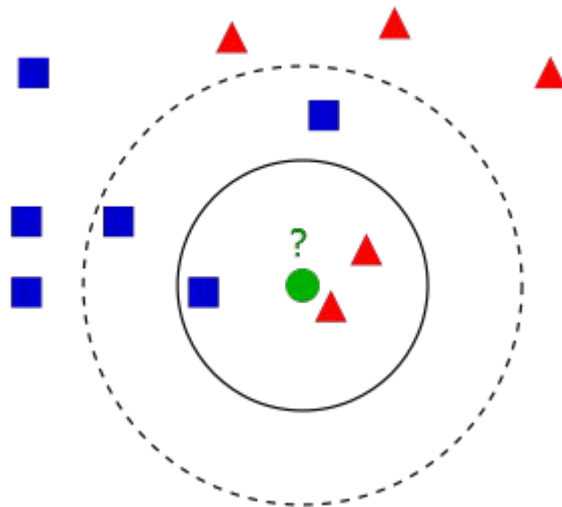


Fig 1. Example of k-NN classification.  
Creative Commons. Public Domain.

# Using cluster-based ranking

## 3. Rank clusters using **cluster-based ranking**

- Treating a document cluster as a large document allows use of the query-likelihood model

$$P(Q | Clu) = \prod_{i=1}^m P(q_i | Clu)$$

$$P(w | Clu) = \frac{|Clu|}{|Clu| + \lambda} P_{ML}(w | Clu) + \frac{\lambda}{|Clu| + \lambda} P_{ML}(w | Coll)$$

$$P_{ML}(w | Clu) = \frac{freq(w, Clu)}{|Clu|}, \quad P_{ML}(w | Coll) = \frac{freq(w, Coll)}{|Coll|}$$

## Query Expansion Term Selection

4. Select query terms from each document in top-ranked clusters using the relevance model

$$\sum_{D \in R} P(D)P(w | D)P(Q | D)$$

# Experimental Setup

- Several Test Collections
- Large heterogeneous web collections
  - GOV2
  - WT10G
- Small homogeneous news-related collections
  - AP
  - WSJ
  - ROBUST
- Topic title field used as query

# Collection Summary

Collection	Description	# of docs	Topics	
			Train	Test
GOV2	2004 crawl of .gov domain	25,205,179	701-750	751-800
WT10g	TREC web collection	1,692,096	451-500	501-550
ROBUST	Robust 2004 collection	528,155	301-450	601-700
AP	Association Press 88-90	242,918	51-150	151-200
WSJ	Wall street Journal 87-92	173,252	51-150	151-200

Table 1. Summary of Test Collections.

*A Cluster-Based Resampling Method for Pseudo-Relevance Feedback.* Lee, Croft, Allan. 2008.



# Training

- Each collection is divided into training and testing topics
- Training used to tune smoothing parameters
  - $\mu$  in initial query-likelihood model  $\mu \in \{ 500, 750, 1000, 1500, 2000, \dots, 5000 \}$
  - Number of feedback documents  $|R| \in \{ 5, 10, 25, 50, 75, 100 \}$
  - Number of expansion terms  $e \in \{ 10, 25, 50, 75, 100 \}$
  - Weight of original query  $\lambda \in \{ 0.1, 0.2, \dots, 0.9 \}$
  - Number of clusters  $|C| \in \{ 1, 2, 5, 10, 15, 20 \}$
- Combined expansion terms with query via *Indri form*  
$$\#weight (\lambda \#combine (q_1 \dots q_m)$$
$$(1 - \lambda) \#weight (p_1 t_1 \dots p_e t_e))$$

# Experimental Comparisons

- Baseline models
  - Language Model (LM)
  - Relevance Model (RM)
- Cluster-based Reranking Method (Rerank)
- **Cluster-based Resampling**
- Upper Bound - True relevance feedback (TrueRF)

# Test Collection Results

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

Table 2. Performance comparisons using MAP for test topics on test collections..

*A Cluster-Based Resampling Method for Pseudo-Relevance Feedback. Lee, Croft, Allan. 2008.*

# Relevance Density

- Explain why this method works
- Compare Cluster-based resampling to PRF without redundant document resampling
- Recall

$$\textit{Density} = \frac{\textit{the number of relevant feedback documents}}{\textit{the number of feedback documents}}$$

# Relevance Density Measure

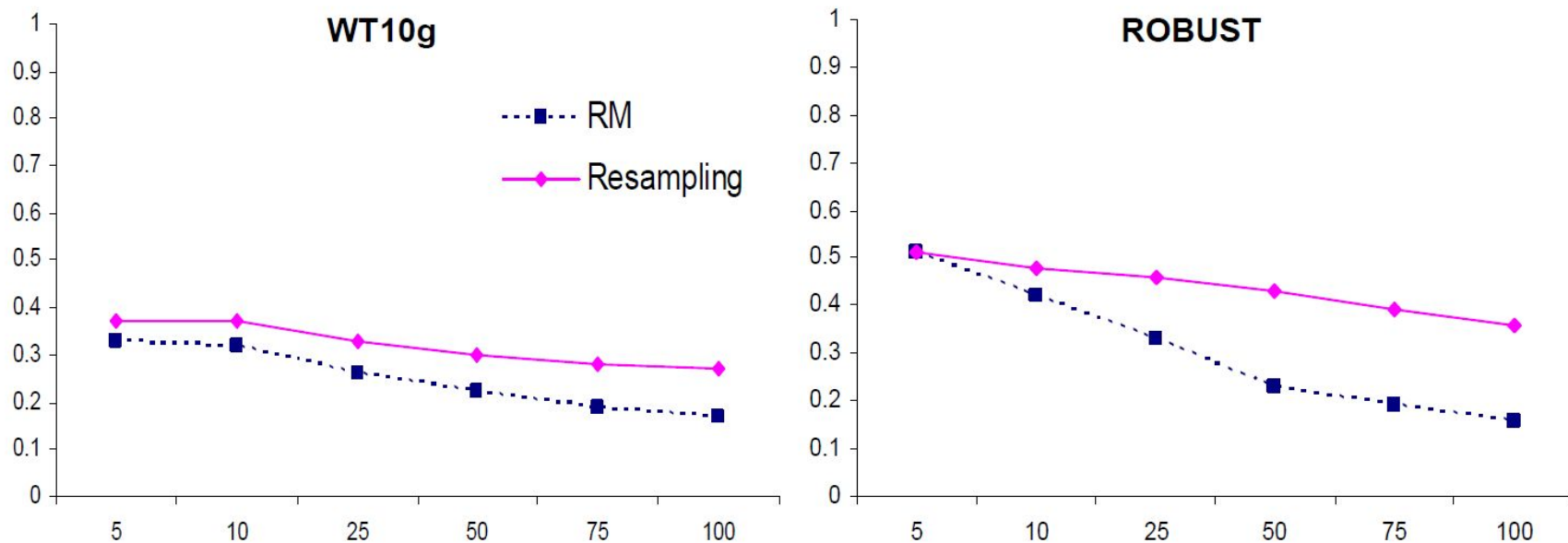


Fig 1. The relevance density for RM and Resampling  
*A Cluster-Based Resampling Method for Pseudo-Relevance Feedback. Lee, Croft, Allan. 2008.*

# Relevance Density Results

	LM	RM	chg%	Resampling	chg%
GOV2	0.3258	0.3519 $^{\alpha}$	8.01	0.3764 $^{\alpha\beta}$	15.53
WT10G	0.1861	0.1886	1.34	0.2072 $^{\alpha}$	11.34
ROBUST	0.2920	0.3262 $^{\alpha}$	11.71	0.3549 $^{\alpha\beta}$	21.54
AP	0.2077	0.2758 $^{\alpha}$	32.79	0.2853 $^{\alpha}$	37.36
WSJ	0.3258	0.3785 $^{\alpha}$	16.18	0.4009 $^{\alpha\beta}$	23.05

Table 3. Performance on fixed feedback (set to 100 documents).  
*A Cluster-Based Resampling Method for Pseudo-Relevance Feedback.* Lee, Croft, Allan. 2008.