Cluster-Based Retrieval Using Language Models

Xiaoyong Liu and W. Bruce Croft

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

A Cluster-Based Resampling Method for Pseudo-Relevance Feedback

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

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

Presented by Polykarpos Thomadakis
CS 834 - Introduction to Information Retrieval
Fall 2017
Old Dominion University

Cluster Based Retrieval

- Based on the hypothesis that similar documents will match the same information needs
- Groups documents into clusters and returns a list of documents based on the clusters they come from
- Two approaches:
  - Clustering the entire collection and then perform document based retrieval (Static clustering)
  - Query-specific clustering. Cluster only the documents retrieved from document based retrieval

Clustering Methods

- Static Clustering
- Two ways to do use it
  - Use clusters as a way to identify the subset of documents that are more likely to be relevant, so that only those documents will be matched to the query
  - Match the query to all clusters in the collection and rank clusters based on their similarity to the query. Documents in clusters with higher rank are considered more relevant
- Query-specific Clustering
  - Perform the clustering to the set of documents returned from document based retrieval. This will smooth out the differences between representations of individual documents
- If an optimal cluster exists and IR retrieves it then it will always perform better than a document-based retrieval
- Clusters put in use “manually” based on already known relevance judgements

Cluster-based Language Models

- Language Model (LM): Probability distribution over all terms in a language vocabulary
- Organize documents around topics
- Each cluster represents a topic containing only relevant documents
- Language models are estimated per cluster
- Language models are used as a representation of each topic and to select the right topics for a given story

Cluster-based Retrieval using LM (Standard model)

- The basic approach for using language models for IR is to model the query generation process
- Rank the documents according to how likely the query Q could have been generated from the document models (referred as query-likelihood)

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

Where q_i is the i_{th} term in the query and $P(q_i|D)$ is specified by

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

Cluster-based Retrieval using LM (CQL model)

- Rank the clusters (instead of individual documents) according to how likely the query Q could have been generated from their models
- Automates relevance extraction

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

Where $P(q_i|Cluster)$ is specified by the cluster language model CQL

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

Cluster-based Retrieval using LM (CBDM model)

- Smooth documents language models using models of the clusters that they come from

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

Where P(W|Cluster) is given from the CQL model

- The cluster model is first smoothed with the collection model, and the document model is then smoothed using the smoothed cluster model

Clustering Algorithms

- First define similarity measures:
  - Cosine measure
  - Dice coefficient
  - Jaccard coefficient
  - Overlap coefficient
  - Kullback-Liebler divergence
- For static clustering:
  - 3-pass K-means algorithm
- For query-specific clustering:
  - Single linkage
  - Complete linkage
  - Group average
  - Centroid
  - Ward's method

Experiment Setup

Parameter tuning was executed on AP before the actual tests took place.
All sets but FR are homogenous, for this reason FR is the most difficult case and extra tuning was performed specifically for it

Collection	Contents	# of Docs	Size	Average # of Words/Doc[1]	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

Experimental Design (CQL)

- Investigate whether CQL can perform good quality cluster ranking
  - Five different clustering algorithms
  - Various similarity thresholds
  - Two smoothing techniques
  - Results compared to the standard document based query-likelihood
- First perform document based retrieval using standard model, then use CQL to cluster the top 1000 documents retrieved

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)

Experimental Design (CBDM)

- Examine the effectiveness of cluster-based retrieval CBDM
  - Query likelihood (QL) retrieval context
  - Relevance model (RM) context
  - For static clustering and query-specific clustering
  - Compared with that used in the original document model

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

Cluster-Based Retrieval Using Language Models

Xiaoyong Liu and W. Bruce Croft

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

A Cluster-Based Resampling Method for Pseudo-Relevance Feedback

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

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

Relevance Feedback

- Take the results that are initially returned from a given query
- Gather user feedback
- Use the feedback to find out whether those results are relevant
- Perform a new query based on this information to have better results

Pseudo-Relevance Feedback

- Same as before but..
- Use top-retrieved documents as feedback instead of user's response
  - Assuming that top-retrieved documents are relevant
- However, top retrieved documents may contain non-relevant documents ("noise")
  - Resulting in drifting the query representation away from the original query
- Selecting the appropriate documents is crucial for effective pseudo-relevance feedback

Cluster-based IR and Pseudo-Relevance Feedback

- Selective Resampling Approach
  - Select randomly from the original sample
  - Selective sampling
    - Boosting: Adaptively change the distribution of training examples focusing on weak learners
    - Skipping some top retrieved documents
    - Using a query-regularized estimation method
    - Leaving a single term of the query out as a noisy term
- Cluster-based Approaches
  - Re-ranking using clusters
  - Cluster-based resampling

Cluster-Based Selective Resampling

- Based on the language and relevance model frameworks
- A dominant document for a query is one with good representation of the topics of a query
  - For example, one with several nearest neighbors with high similarity
  - In overlapped clusters, will appear in multiple highly-ranked clusters
- From such a dominant document, expansion terms that retrieve related documents can be selected

Resampling Process

1. Documents are retrieved for a given query by the query likelihood language model

$$P(w \mid D) = \frac{|D|}{|D| + \mu} P_{ML}(w \mid D) + \frac{\mu}{|D| + \mu} P_{ML}(w \mid Coll)$$

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

2. Clusters are generated by *k*-nearest neighbors clustering method for the top-retrieved *N* documents

3. The clusters are ranked by a cluster-based query-likelihood language model (CQL) and the dominant documents are retrieved
   - The dominant documents are repeatedly being fed for the resampling process

Resampling Process (cont.)

4. Expansion terms are selected using the relevance model for each dominant document in the top-ranked clusters.

5. The probability of a word in the distribution is estimated by

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

R: set of pseudo-relevant documents

6. Finally, the most likely terms are chosen and combined with the original query

Experimental Setup

- Each test collection, topics divided into training and test topics
- Training topics are used for parameter estimation and the test topics are used for evaluation

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
ROUBST	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

Test Collection Results

LM: Language Model
Rerank: Reranking using clusters
RM: Relevance Model
TrueRF: True Relevance Feedback

	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 $^{\alpha\beta\gamma\delta}$

Justification by Relevance Density

- Relevance density is measured to justify the assumption that dominant documents are relevant to the query

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