Cluster-Based Retrieval from A Language Modeling Perspective

Xiaoyong Liu
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
xliu@cs.umass.edu

ABSTRACT

Cluster-based retrieval (CBR) is based on the hypothesis that similar documents will match the same information needs. The most common approach to CBR is to retrieve one or more clusters in their entirety to a query. The second and less common approach is to smooth documents with information from clusters. Previous research in this area has suggested that "optimal" clusters exist that, if retrieved, would yield very large improvements in effectiveness relative to document-based retrieval. However, it is precisely if and how the good clusters can be automatically identified and used by the retrieval system that has long been an interesting yet challenging problem. Previous research has been inconclusive as to whether a real CBR strategy does bring improved retrieval effectiveness. Recent developments in the language modeling approach to information retrieval have motivated us to re-examine this problem within this new retrieval framework. In the proposed research, I study both approaches to CBR, namely cluster retrieval (which directly ranks clusters) and cluster-based smoothing (which smoothes documents with clusters), and develop a set of techniques that will address several aspects of CBR including systematic modeling of document-cluster relationships, different ways of representing clusters for retrieval, and new retrieval models that are more suitable for CBR. Preliminary results on TREC collections show that, with the proposed techniques, CBR can perform consistently across collections of realistic size, and significant improvements over document-based retrieval can be obtained in a fully automatic manner and without relevance information provided by human.

There are two main contributions of this work. The first contribution is a systematic study of the characteristics of good clusters and ways to identify them. The second contribution is the development of new models for CBR.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models.*

General Terms: Theory, Experimentation

Keywords: Cluster-based Retrieval, Optimal Cluster, Language Model, Representation, Smoothing, Mixture Model.

1. INTRODUCTION

Cluster-based retrieval is based on the hypothesis that similar documents will match the same information needs [16]. In document-based retrieval (DBR), an information retrieval (IR) system matches the query against documents in the collection and returns a ranked list of documents to the user. Cluster-based retrieval (CBR), on the other hand, groups documents into clusters and returns a list of documents based on the clusters that they come from.

The use of document clustering in experimental IR systems dates back to 1960s. It was initially proposed as a means for improving efficiency and also as a way to categorize or classify documents [13, 14]. Later, Jardine and van Rijsbergen [8] suggested that document clustering could be used to improve the effectiveness as well as the efficiency of retrieval. The most common approach to CBR is based on the idea of *cluster-based retrieval* introduced by Jardine and van Rijsbergen [8]. The task for the retrieval system is to retrieve one or more clusters in their entirety to a query, by matching the query against clusters of documents instead of individual documents and ranking clusters based on their similarity to the query. Jardine and van Rijsbergen introduced the notion of an "optimal" cluster. A cluster is considered optimal, if, when retrieved, it would give the maximum possible value for a retrieval criterion out of all clusters. The ideal CBR strategy is that which infallibly finds the optimal clusters. Real CBR strategies are devised with the goal of approximating to this ideal. They and others examined the ideal case where the IR system is able to select the best clusters based on relevance judgments from users or the number of known relevant documents contained in each cluster, and showed that if the good clusters could be retrieved, effectiveness would be far better than a document-based search [8, 7, 15]. The second approach to CBR is to use clusters as a form of document smoothing. Previous studies have suggested that by grouping documents into clusters, differences between representations of individual documents are, in effect, smoothed out. For both approaches, it is whether and how the good clusters can be automatically identified or used by the IR system that has long been an interesting yet challenging problem. A number of cluster retrieval and search methods have been proposed [1, 2, 8, 17, 19], and a variety of clustering algorithms have been investigated [19, 9, 15]. While the experimental results to date have suggested that document clustering may indeed have substantial merits for retrieval purposes, there have been no conclusive findings on whether a real CBR strategy can yield improvements in effectiveness relative to DBR, especially on test collections of realistic size and when no relevance information is available.

Recent developments in statistical language modeling for information retrieval have opened up new ways of thinking about the retrieval process. Research carried out by a number of groups has confirmed that the language modeling approach is a theoretically attractive and potentially very effective probabilistic

Table 1. Statistics of data sets

Collection	Contents	# of Docs	Size	Average # of Words/Doc	Queries (TREC topics, title only)	# of Queries with Relevant Docs
AP	Associated Press newswire 1988-90	242,918	0.73 Gb	473.6	51-150	99
WSJ	Wall Street Journal 1987-92	173,252	0.51 Gb	465.8	51-100 & 151-200	100
TREC12	TREC disks 1 & 2: Wall Street Journal, 1987-89; Associated Press newswire, 1988-89; Computer Selects articles, Ziff-Davis; Federal Register, 1988-89; Abstracts of U.S. DOE publications.	741,856	2.07 Gb	415.7	51-200	150
TREC45	TREC disks 4 & 5: The Financial Times, 1991-94; Federal Register, 1994; Congressional Record, 1993; Foreign Broadcast Information Service (FBIS); The LA Times.	556,077	2.14 Gb	541.9	301-400	100

framework for studying information retrieval problems [3]. This led us to a re-examination of CBR within this new framework.

This thesis intends to address the following research questions: 1). What are the characteristics of good clusters? 2). How to use language models or other statistical methods to capture these characteristics and improve retrieval effectiveness? In order to answer these questions, I study both cluster retrieval (which directly ranks clusters) and cluster-based smoothing (which smoothes documents with clusters), and develop new retrieval techniques that will address several aspects of CBR, including systematic modeling of document-cluster relationships, different ways of representing clusters for retrieval, and retrieval models that are more suitable for CBR. Preliminary results show that CBR can perform consistently across collections of realistic size, and significant improvements over DBR can be obtained on several collections when clusters are used automatically and without relevance information provided by human.

The rest of the paper is organized as follows. I briefly discuss the related work in section 2 and our methodology in section 3. Section 4 and 5 present proposed methods and preliminary results. Conclusions and future work are given in section 6.

2. RELATED WORK

There have been numerous studies on whether or how document clustering can be employed to improve retrieval results. In most early attempts the strategy has been to build a static clustering of the entire collection in advance, independent of the user's query, and clusters are retrieved based on how well their centroids match the query. A hierarchical clustering technique is typically used in these studies as the size of the collection used is small, and different strategies for matching the query against the document hierarchy generated by such clustering algorithm have been proposed, most notably a top-down or a bottom-up search and their variants [8, 17, 2]. More recently, query-specific clustering has been proposed [7, 15, 9] which generates clusters from the set of documents retrieved by an IR system on a query. The task for the IR system is still ranking clusters but with clusters generated in the query-specific manner. While some studies on comparing the effectiveness of CBR with that of the DBR have shown that the former has the potential of outperforming the latter for precision-oriented searches [2, 8], other experimental work [4, 18, 20] has suggested that DBR is generally more effective.

There has been a resurgence of research in CBR in the past few years, including our own work in this area [10] which is also part of the thesis. The main spirit is to use clusters as a form of document smoothing. The language modeling retrieval framework is used. Topic models are constructed from clusters and documents are smoothed with these topic models, to improve document retrieval [10, 23]. A closely related but different approach is reported in [22] which uses query-specific clusters to regularize document scores.

3. METHODOLOGY

3.1 Proposed Work

As in section 1, there are two approaches to CBR – cluster retrieval and cluster-based smoothing. For the first approach, I study the ideal case where the system is provided with relevance information and always finds the best clusters. This is to establish the upperbound in CBR performance with state-of-the-art language modeling retrieval techniques. I then employ a query-likelihood model to retrieve clusters and show how well a real CBR strategy performs as compared to DBR. Through a close examination as to why the CBR strategy fails in ranking the good clusters at the top, I identify several aspects of CBR that can possibly be improved, including modeling of document-cluster relationships, different cluster representations, and ultimately new retrieval models for CBR that will incorporate these qualities. To test the idea that it is useful to model document-cluster relationships, I use a simple heuristic measure called *query-informed within-cluster deviation* that takes into consideration the relative performance of a cluster and its member documents. Based on the findings, I propose two language-model-based methods for representing clusters - one is based on a mixture of term frequencies from member documents and the other is based on a mixture of member document language models. I am in the process of developing a new retrieval model based on these methods. For cluster-based smoothing, I propose a language model of documents that utilizes information from clusters. Sections 4 and 5 discuss the proposed methods for each approach. The proposed techniques are empirically evaluated and I describe the data sets and evaluation measures next.

3.2 Data

The data sets come from the TREC collections. Queries are taken from the "title" field of TREC topics. The data sets used in this paper are: Associated Press newswire (AP) with TREC queries 51-150, Wall Street Journal (WSJ) with queries 51-100 and 151-200, TREC disks 1 & 2 (TREC12) with queries 51-200, and TREC disks 4 & 5 (TREC45) with queries 301-400. TREC12 and TREC45 are large, heterogeneous collections in which both document sizes and topics vary widely. AP and WSJ are examples of homogeneous collections. The statistics of data sets are given in table 1. I also plan to use data from the Terabyte Track (http://trec.nist.gov/tracks.html) for future experiments.

3.3 Evaluation

The experimental results are evaluated using standard IR performance measures: mean average precision (MAP) and precision at 5 documents (PREC-5). For comparing different cluster representations, I employ an additional measure – mean reciprocal rank (MRR). I identify good clusters by whether it will give a 10% or higher precision than that of DBR with the same number of documents (as that in each cluster) taken from the top of the retrieved list. The good clusters form the cluster relevance judgment set. I go through the list of ranked clusters and mark the highest rank at which a good cluster is retrieved. The reciprocal of the rank is computed. The MRR score is the average of reciprocal ranks across all queries. This measure is appropriate because I am interested in whether different cluster representations help improve cluster ranking and this is the direct way to show it.

4. CLUSTER-BASED SMOOTHING

I have proposed and evaluated a new language model for CBR that smoothes representations of individual documents using models of the clusters that they come from. The model and results are discussed in detail in [10]. Due to the space limitations of this paper, I only briefly summarize the approach and findings here and readers can refer to [10] if interested.

The standard document language model for DBR is a mixture of two parts - the document model and the collection model. The collection model plays the role of data smoothing. I propose that, besides the collection model, the document model can be further smoothed by the model of the cluster to which the document belongs. The task for the IR system is to retrieve documents but with help from clusters. Experimental results show that this model of CBR is at least as good as and sometimes significantly more effective than existing models for DBR and CBR. I also attempted at using language models to directly rank clusters but was not able to obtain improved retrieval results. The rest of the thesis, therefore, focuses on investigating how to identify good clusters and how to effectively rank them, which is a harder problem relative to cluster-based smoothing.

5. CLUSTER RETRIEVAL

Using this approach for CBR, the task for the IR system is to retrieve clusters and then form a document list by ordering documents from the best cluster followed by documents from the second best cluster, and so on.

5.1 Performance and Retrieval of Good Clusters

I establish the upperbound on CBR performance, which is the ideal case in which the system is provided with relevance information and produces the ideal ranking of clusters based on the number of relevant documents they contain. I first perform document retrieval using the standard query likelihood (QL) model [11], and then perform query-specific clustering with the K nearest neighbor clustering algorithm [5] on the retrieved documents. K is set to 5. The cosine similarity is used for clustering. The ideal performance of CBR is compared to the performance of DBR in table 2. The average number of good clusters identified across all queries for each data set is also given. We can see that, indeed, if we were able to retrieve good clusters at the top ranks, retrieval performance could be largely improved.

Table 2. Ideal and real performance Cluster-based retrieval with QL. KNN Clustering with K=5.

Collection	Metric	DBR	CBR (ideal)	CBR (QL)
TREC45	PREC-5	0.4140	0.8540	0.3240
	MAP	0.2011	0.4317	0.1580
	Avg. # of opti. clus.	-	47	-
WSJ	PREC-5	0.5060	0.8800	0.4520
	MAP	0.2958	0.5054	0.2262
	Avg. # of opti. clus.	-	53	-

Next I investigate what the actual performance is in retrieval of good clusters using a state-of-the-art retrieval technique. My first model of CBR is based on the QL retrieval model [11, 12]. Clusters are ranked based on the likelihood of generating the query, i.e. $P(Q|Cluster)$. It can be estimated by:

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

where Q is the query, q_i is the ith term in the query, and $P(q_i|Cluster)$ is specified by the cluster language model

$$\begin{aligned} P(w \mid Cluster) &= \lambda P_{ML}(w \mid Cluster) + (1-\lambda) P_{ML}(w \mid Coll) \\ &= \lambda \frac{tf(w, Cluster)}{\sum_{w' \in cluster} tf(w', Cluster)} + (1-\lambda) \frac{tf(w, Coll)}{\sum_{w' \in V} tf(w', Coll)} \end{aligned} \quad (2)$$

where $P_{ML}(w|D)$ is the maximum likelihood estimate of word w in the document, $P_{ML}(w|Coll)$ is the maximum likelihood estimate of word w in the collection, $tf(w, Cluster)$ is the term frequency of w in the cluster, $tf(w, Coll)$ is the term frequency of w occurs in the entire collection, w' is any word, V is the vocabulary, and λ is a general symbol for smoothing which takes different forms when different smoothing methods are used [10]. I use Dirichlet smoothing at 1000 in experiments. The results are shown in the last column of table 2. These results confirm with previous studies [6, 18] that the task of retrieving good clusters is very hard, and despite the fact that there are a decent number of good clusters per query, those clusters are typically not retrieved at the top ranks. The overall performance of retrieving clusters is inferior to that of retrieving documents.

5.2 Analysis

To find out why good clusters are not retrieved at top ranks, I performed the analysis presented in table 3. I show part of the ranked cluster list for query 306 on TREC45 collection. For each cluster on the ranked list, the rank at which the cluster is retrieved, the number of relevant documents in the cluster, the cluster ID, member documents in the cluster, and the respective QL of the cluster and member documents (the log of the QL scores are shown) are given in the first five columns. In the next three columns, the table gives frequencies of each query term in the cluster as a whole and in individual documents in the cluster. The last two columns show the cluster and document length in terms of the number of indexing terms, as well as the number of unique terms contained within the cluster and each member document. The member documents of a cluster that are relevant are marked with an "*" in front the document ID in the 4th column.

Let us take a close look at query 306. The query is "African civilian deaths". After stemming and stopword removal, the query becomes "africa civilian death". Five clusters on the ranked cluster list are shown. The best clusters for the query are C14 and C80. Both clusters have five relevant documents. However, simply based on the term occurrences the system was not able to assign high ranks to them. The top ranked cluster is C636 which has only

Table 3. Analysis of Query 306, "*" means relevant documents.

Cluster Rank	Num. of Rel. Docs	Cluster ID	Member Docs	Cluster/Doc. Log QL	Freq. of Term "death"	Freq. of Term "civilian"	Freq. of Term "africa"	Cluster/Doc. Length in Terms	Num. of Unique Terms
1	1	C636	-	-15.436329	12	12	141	3735	1170
			*CR93H-2896	-16.520723	1	12	107	1905	549
			FR940712-2-00058	-20.346148	3	0	9	315	149
			FT941-12410	-20.981014	4	0	6	563	328
			CR93E-250	-20.996843	2	0	13	627	377
			FR940712-2-00057	-21.119274	2	0	6	325	187
3	3	C2	-	-15.550194	87	4	18	2344	691
			*FBIS3-25118	-16.796564	20	2	4	548	263
			*FBIS3-471	-17.601721	26	1	3	635	313
			FBIS4-24155	-18.422461	11	1	1	207	143
			*FBIS3-602	-19.521227	25	0	6	739	356
			FBIS3-470	-20.388792	5	0	4	215	136
4	0	C208	-	-15.615210	17	2	77	1573	709
			FBIS4-24155	-18.422461	11	1	1	207	143
			FBIS4-48773	-19.985443	0	1	10	128	81
			FBIS4-23488	-20.339703	3	0	5	91	57
			CR93E-2102	-20.836098	1	0	58	1073	524
			FBIS4-1186	-21.120800	2	0	3	74	51
13	5	C14	-	-15.940542	9	15	17	1732	750
			*FT942-7623	-17.831032	1	6	8	531	319
			*FBIS4-28901	-18.292290	5	2	2	329	224
			*FBIS4-23790	-18.964863	2	2	3	409	262
			*FT942-9707	-19.172813	1	2	2	107	91
			*FBIS4-23738	-20.970539	0	3	2	356	233
77	5	C80	-	-16.665789	9	16	13	2258	891
			*FT942-7623	-17.831032	1	6	8	531	319
			*FT942-15976	-19.275204	3	1	2	269	187
			*FBIS4-47810	-19.629265	1	7	1	616	277
			*FT943-15255	-20.335827	3	1	1	511	340
			*FBIS4-912	-20.905777	1	1	1	331	226

Table 4. Query term occurrence in relevant and non-relevant documents. Cell is # of documents.

Query ID	Category	Total # of docs	# in top 1000 ret	# of unique query terms occurred			
				0	1	2	3
306	REL	352	171	5	74	191	82
	NON-rel	-	829	0	17	708	104
301	REL	474	84	69	142	158	105
	Non-rel	-	916	0	47	462	407

one relevant document. I observe that the document that is relevant is very long with lots of occurrences of the query terms. Therefore, even if the other documents are not relevant and with only few occurrences of query terms, the overall cluster QL would still be high. Cluster C2 is ranked the third on the list. The occurrences of the query terms spread more evenly across the member documents than C636. There are 3 documents that are judged relevant but by reading document FBIS4-24155 I found that it is very similar in content to another document that is judged relevant. It may have been misjudged or judged by an assessor with really strict criterion when the relevance judgment set is created. I also observe that the relevant document FBIS3-602 does not have term "civilian". Instead of using "civilian", the article makes use of more descriptive terms such as victims, teachers, women, and children involved in the incidents. The term "casualties" is often used instead of "death". This gives an example of the case when the language used in the document is quite different from the query language (i.e. vocabulary mismatch). The next cluster on the list is C208 which has no relevant documents. I found that even though the individual documents may not have all query terms appearing, the overall cluster QL is high because the cluster model picks up different query terms from different documents. For example, document CR93E-2102 is very long and it contains many occurrences on the query term "africa", thus this document contributes largely to the overall cluster frequencies of that term. The cluster frequencies of term "civilian" come from only two of the five documents. The clusters C14 and C80, while of good quality, have very low frequency counts on the query terms, thus their QL scores are lower than that of the other clusters. Across all clusters for this query, I observe that good clusters tend to not only have a good overall cluster language model but their member documents also have good language models with low variability in performance. It seems that a technique that explicitly considers document-cluster relationship may benefit retrieval. To test this idea, I designed a set of pilot experiments which are described in section 5.3.

From this example, however, one may also get the impression that bad clusters tend to have more documents with zero occurrences of one or more query terms. However, after examining several other queries, I found that this is not typically true. I examined the

occurrences of query terms in relevant and non-relevant documents, and found that relevant documents may have query terms missing whereas non-relevant documents can have all query terms appearing. Query 306 (African civilian deaths) and 301 (international organized crime) are analyzed in table 4 for illustration. "REL" means relevant documents and "NON-rel" stands for non-relevant documents. Both queries have 3 unique query terms. For query 306, there is a total number of 352 relevant documents, of which 171 were retrieved by DBR. There are 829 non-relevant documents also in the top 1000 retrieved by DBR. We observe that there are 5 relevant documents that have zero query terms whereas 104 non-relevant documents have all of the query terms. For query 301, 69 relevant documents don't contain any query terms but 407 non-relevant documents have all three query terms. A seemingly straightforward feature like the number of member documents with missing query terms will not help distinguish good and bad clusters.

5.3 Pilot Experiments

I have observed in the analysis that the member documents in good clusters tend to have similar query likelihood with each other and to that of the cluster (Other matching functions between query and documents/clusters can be used. I use QL here for convenience of discussion). To test whether this observation is valid, I developed a simple heuristic measure that takes into consideration how well a cluster as well as its member documents matches the query. The intuition is that the less the member document QL varies from the cluster QL, the more likely that the documents contribute evenly to the cluster model. Clusters with large variability of member document QL from the cluster QL may mean that only some member documents contribute largely to the cluster model (e.g. cluster C636 in table 3) or the individual documents contribute to different query term occurrences in the cluster model for the cluster QL to be high but the documents tend to have low QL (e.g. cluster C208 in table 3). A popular way to measure variation is variance [21]. It is computed as the average squared deviation of each number in a distribution from its mean. Taking a similar approach, I compute the average squared deviation of the QL of each document in a cluster from the cluster QL. That is,

$$WCD = \frac{\sum_{d \in C} (MS_d - MS_C)^2}{K}$$

where C stands for a cluster, d stands for any document in the cluster, K is the number of documents in the cluster, MS_d is the a measure of closeness of the document to the query, and MS_C is a measure of closeness of the cluster to the query. I call this metric the *query-informed within-cluster deviation (WCD)*. I conjecture that the good clusters would be those with high cluster QL but low WCD. If for a given query, such clusters exist then the system is likely to succeed with CBR. In this case, the system applies CBR and ranks these clusters before others. A document list is created by displaying documents from the first cluster, then those from the second cluster, and so on. Documents from the same cluster are ranked according to their closeness score to the query. If no clusters are considered satisfactory, the system outputs the list of documents produced from DBR.

5.3.1 Results

I compare performance of cluster retrieval using the WCD measure with that of document retrieval. Results are given in table 5. Similar to experiments in section 5.1, I use query-specific clustering with the K Nearest Neighbor method. K is set 5. The cosine measure is used to determine the similarity between documents and top 1000 documents from DBR are clustered.

In order to decide which clusters are potentially good clusters, I need to find a threshold for both the cluster query likelihood and WCD. There are two parameters to determine. A cluster is considered good if its cluster query likelihood falls into the upper x percent of its value range and its WCD falls into the lower y percent of the WCD value range. The value ranges are determined from all clusters for the given query. The WSJ data set is selected as the training collection for determining these parameters. An exhaustive parameter search is applied and the best retrieval performance is obtained at parameters set to 80 (x for cluster query likelihood) and 40 (y for WCD). We then apply these parameters to the TREC45 data set. If the system finds clusters that satisfy the requirement on the cluster query likelihood and WCD, the system performs cluster retrieval. If no such clusters are found, document retrieval is used. From table 5, we can see that, by considering relationship between clusters and their member documents, cluster retrieval can be more effective than document retrieval. There are 18 and 34 queries that used cluster retrieval, on WSJ and TREC45 respectively. Similar results are observed on other data sets, e.g. LA Times.

Table 5. Retrieval performance with proposed selection mechanism compared to using document-based retrieval consistently for all queries. Retrieval is done using QL model. "*" means that there is a significant improvement in performance using Wilcoxon test with 95% confidence. The percentage of improvement in performance is given in parentheses.

Collection	Doc QL		Proposed technique	
	Prec. at 5 docs	MAP	Prec. at 5 docs	MAP
WSJ	0.5060	0.2958	0.5200* (+2.8%)	0.2981* (+0.8%)
TREC45	0.4140	0.2011	0.4500* (+8.7%)	0.2063* (+2.6%)

5.4 Cluster Representations

The pilot experiments show that it would be useful to model the document-cluster relationship and one way to achieve this is through cluster representations. In CBR, clusters are typically represented as simple concatenation of their member documents [10, 2, 15]. Other representations have also been used in the past, e.g. centroid vector [18]. These representations, while being simple and intuitive, may have a number of problems (e.g. the cluster model can be biased by one document). A representation that would allow for a more principled way of taking contributions from member documents is desired. I propose two new ways of representing clusters in this section.

The standard approach to representing clusters is to treat them as if they were big documents formed by concatenating their member documents. Thus, *tf(w, Cluster)* is computed by:

$$tf(w, Cluster) = \sum_{i=1}^{k} tf(w, D_i) \quad (3)$$

where $Cluster = \{D_1, ..., D_k\}$ and k is the number of documents in a cluster. Clusters are ranked by equation (1) with components estimated from equations (2) and (3).

Our first method is to represent clusters by a weighted mixture of term frequencies from member documents, that is,

$$tf(w, Cluster) = \sum_{i=1}^{k} (\alpha_i * tf(w, D_i)) \quad (4)$$

Table 6. Mean average precision comparisons. "+" means significantly better than centroid representation, "*" means significantly better than concatenating docs, and "!" means significantly better than document-based retrieval, with wilcoxon test at 95% confidence. Clustering method is Ward's.

Collection	Document Retrieval	Cluster Retrieval				
		Ideal	Concatenating documents	Centroid (avg. doc)	TF mix	DM mix
AP	0.2179	0.5802	0.2160	0.2039	0.2196 (+)	0.2204 (+*)
WSJ	0.2958	0.7022	0.2924	0.2719	0.2969 (+)	0.2979 (+)
TREC12	0.2300	0.5686	0.2258	0.2049	0.2306 (+*)	0.2309 (+*)
TREC45	0.2011	0.5866	0.1963	0.1776	0.2063 (+*!)	0.2076 (+*!)

Table 7. Comparison of CBR performance using different cluster representations. Evaluation metric is MRR. Clustering method is Ward's.

Cluster Representation	AP	WSJ	TREC12	TREC45
Concatenating documents	0.5089	0.6569	0.5409	0.5457
Centroid	0.5734	0.6708	0.5886	0.5906
TF mixture	0.5680	0.6709	0.5931	0.5907
DM mixture	0.5735	0.6712	0.5959	0.6030

where α is a weighting parameter between 0 and 1, and $\sum_{i=1}^{k} \alpha_i = 1$.

Clusters are ranked by equation (1) with components estimated from equations (2) and (4). This approach is referred to as TF mixture.

Our second way of representing clusters is to build language models for individual member documents and the cluster language model is a weighted mixture of these member document models. Again, λ is a general symbol for smoothing.

$$P(w \mid Cluster) = \sum_{i=1}^{k} \left[\beta_i * (\lambda P_{ML}(w \mid D_i) + (1-\lambda) P_{ML}(w \mid Coll)) \right] \quad (5)$$

where β is a weighting parameter between 0 and 1, and $\sum_{i=1}^{k} \beta_i = 1$.

Clusters are ranked by equation (1) with components estimated from equation (5). We refer to this approach as DM mixture.

5.4.1 Preliminary Results

I first perform document retrieval using the query likelihood (QL) retrieval model with Dirichlet smoothing at 1000. The top 1000 retrieved documents are clustered using Ward's method. The cosine similarity measure is used to determine the similarity between documents. As I discussed in section 2, different ways of estimating the cluster language models are employed when different cluster representations are considered. Clusters are ranked by their query likelihood. Again, Dirichlet smoothing at 1000 is used for TF mixture and concatenating documents, which is the best parameter setting for the latter. We also use this smoothing parameter for setting the λ in DM mixture (equation (5)). Currently, both α and β in equations (4) and (5) are estimated by the first-stage retrieval log QL score of each document divided by the sum of log QL scores of all member documents in a cluster. Note that the log QL scores are negative. Setting α and β this way penalizes clusters with documents that match the query poorly. Results are shown in table 6. AP collection is used for selecting the clustering threshold. The best performance for concatenating documents and the centroid representation is achieved at document similarity threshold set to 0.8 whereas the best performance for TF and DM mixture models is achieved with threshold set to 0.6. The best results are given in table 6. We can see that both TF and DM mixtures consistently outperform document retrieval on all four data sets, and significant improvement is found on TREC45. Both methods are significantly more effective than concatenating documents and the centroid representation. DM mixture is slightly better than TF mixture. Table 7 shows the MRR measure of the rank position of the first good cluster retrieved.

I also experimented with KNN clusters and significant effectiveness of TF and DM mixture models over other representations have been observed. The overall performance of KNN clustering for retrieval is lower than that of Ward's method.

6. CONCLUSIONS AND FUTURE WORK

In this thesis, I study both approaches to CBR, namely cluster-based smoothing and cluster retrieval, in the language modeling framework. For the former approach, I proposed a new model for retrieval and showed that it can improve retrieval performance over DBR. For the latter, I examined cluster retrieval performance on TREC test collections and performed an analysis as to why good clusters are not retrieved at top ranks. The analysis shows that there are several possible reasons. One is that the representation of a cluster created by simply combining term occurrences in member documents may not be best suitable for CBR. Also, the current CBR techniques do not take into consideration of how well member documents match the query. I proposed two new ways of representing clusters for retrieval. Preliminary results show that these techniques are significantly more effective than other representations used in previous studies, and cluster retrieval can consistently outperform document retrieval on standard, large test collections. This is particularly encouraging as results from previous studies on retrieving clusters are, at best, mixed. In addition, the parameters in the proposed models are not tuned and I anticipate further improvement in performance with parameter tuning. I plan to carry out a thorough evaluation of these techniques by experimenting with static clusters, other clustering algorithms, and their possible use in cluster-based smoothing. The proposed techniques addressed some aspects of the identified problems and I plan to investigate other aspects in future work. I will also explore other ways of constructing cluster representations and hope to develop a new retrieval model for cluster retrieval.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #CNS-0454018.

8. REFERENCES

[1] Croft, W. B. (1978). *Organizing and Searching Large Files of Document Descriptions*. Ph.D. dissertation, University of Cambridge.

[2] Croft, W. B. (1980). A model of cluster searching based on classification. *Information Systems*, Vol. 5, pp. 189-195.

[3] Croft W. B., & Lafferty, J (eds.) (2003). *Language Modeling for Information Retrieval*. In Kluwer International Series on Information Retrieval, Volume 13, Kluwer Academic Publishers.

[4] El-Hamdouchi, A. & Willet, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3), pp. 220-227.

[5] Yang, Y, & Liu, X. (1999). A re-examination of text categorization methods. In *SIGIR-99*.

[6] Griffiths, A., Luckhurst, H.C., and Willett, P. (1986). Using interdocument similarity information in document retrieval systems. *JASIS&T*, 37, pp. 3-11.

[7] Hearst, M.A., and Pedersen, J.O. (1996). Re-examining the cluster hypothesis: Scatter/Gather on retrieval results. In *SIGIR 1996*.

[8] Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Info. Stor. and Rer.*, 7:217-240.

[9] Leuski, Anton. (2001). Evaluating Document Clustering for Interactive Information Retrieval. In *CIKM'01*.

[10] Liu, X. and Croft, W. B. (2004). Cluster-based retrieval using language models. In *SIGIR'04*, pp. 186-193.

[11] Miller, D., Leek, T., and Schwartz, R. (1999). A hidden Markov model information retrieval system. In *SIGIR 1999*, pp. 214-221.

[12] Ponte, J., and Croft, W.B. (1998). A language modeling approach to inormation retrieval. In *SIGIR 1998*, pp.275-281.

[13] Rocchio, J. J. (1966). *Document Retrieval Systems – Optimization and Evaluation*. Ph. D. thesis, Harvard University.

[14] Salton, G. (1971). Cluster search strategies and optimaization of retrieval effectiveness. In G. Salton, editor, *The SMART Retrieval System*, pp. 223-242. PrenticeHall, Englewood Cliffs, N. J..

[15] Tombros, A.; Villa, R.; and Van Rijsbergen, C.J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval, *Information Processing and Management*, 38, pp. 559-582.

[16] van Rijsbergen, C.J. (1972). *Automatic Information Structuring and Retrieval*. Ph.D. thesis, University of Cambridge.

[17] van Rijsbergen, C.J. & Croft, W. B. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 collection. *Information Processing & Management*, 11, pp. 171-182.

[18] Voorhees, E.M. (1985). The cluster hypothesis revisited. In *SIGIR 1985*, pp.188-196.

[19] Voorhees, E. M. (1985). *The effectivenss and efficiency of agglomerative hierarchic clustering in document retrieval*. Ph.D. Thesis, Cornell University.

[20] Willet, P. (1985). Query specific automatic document classification. *International Forum on Information and Documentation*, 10(2).

[21] Mukhopadhyay, N. (2000). *Probability and Statistical Inference*, Marcel Dekker Inc.

[22] Diaz, F. (2005). Regularizing ad hoc retrieval scores. In *CIKM 2005*.

[23] Kurland, O. and Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *SIGIR'04*, pp. 194-201.