Representing Clusters For Retrieval

Xiaoyong Liu and W. Bruce Croft
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
xliu@cs.umass.edu

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, Cluster Representation

1. INTRODUCTION

In cluster-based retrieval (CBR), documents are grouped into clusters which are then retrieved directly in their entirety to the query or used to smooth document language models [1]. The clusters are typically represented as simple concatenation of their member documents [1, 2, 3]. This representation, while being simple and intuitive, may have a number of problems. For example, if one of the member documents is very long and has many occurrences of the query terms while other member documents are short with only few query terms appearing, then simply concatenating these documents would result in a representation that is largely biased by one particular document. This is what we would want to avoid because the quality of clusters is usually judged by the total number of relevant documents they contain rather than how good one of the documents is [4]. Clusters with more relevant documents are considered better. Based on this, a representation that would allow for a more principled way of taking contributions from member documents is desired. Other representations have also been used in the past, e.g., centroid vector [5], but again they do not explicitly consider individual documents. This work describes an ongoing effort toward developing new cluster representations that would yield improved effectiveness in CBR. We investigate two methods – 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. These representations are compared with the standard approach of concatenating documents in the context of CBR using query-specific clustering. Early results show that these methods are promising.

2. CLUSTER REPRESENTATIONS

The language modeling (LM) framework has been shown to be theoretically attractive and very effective for studying information retrieval problems, including CBR [6, 1]. In this work, we focus on the traditional approach to CBR that is to retrieve clusters in their entirety to a query. To use LM approach for retrieving clusters, we first need to derive language models from cluster representations and then apply retrieval models. Let's take the query likelihood retrieval model for example. 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

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

where $P_{ML}(w|Cluster)$ 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 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 [1]. Commonly used smoothing methods include Dirichlet and Jelinek-Mercer smoothing, among others.

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,

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

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.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

3. DATA

Three data sets are used in the experiments. They are: TREC topics 51-150 (title only) with the whole disks of TREC disk 1 and 2 (TREC12), TREC topics 301-400 (title only) with the whole disks of TREC disk 4 and 5 (TREC45), and TREC topics 51-150 (title only) with the Associated Press newswire (AP) collection. Both the queries and collections have been stemmed and stopwords have been removed using the standard INQUERY list of 418 words. The data sets are summarized in table 1. TREC12 and TREC45 are large, heterogeneous collections in which both document sizes and topics vary widely. AP is an example of homogeneous collections.

Table 1. Summary of data sets.

Collection	Contents	Size	Queries
TREC12	TREC disks 1 & 2: WSJ, 1987-89; AP, 1988-89; Computer Selects articles, Ziff-Davis; FR, 1988-89; DOE abstracts	2.07 Gb	TREC topics 51-200
TREC45	TREC disks 4 & 5 : FT, 1991-94; FR, 1994; CR, 1993; FBIS; LA.	2.14 Gb	TREC topics 301-400
AP	AP 1988-90	0.73 Gb	TREC topics 51-150

4. EXPERIMENTS AND RESULTS

We first perform document-based retrieval using the query likelihood (QL) retrieval model [6] with Dirichlet smoothing at 1000. Next, we take the top 1000 retrieved documents and cluster them using the K Nearest Neighbor clustering method. K is set to 5. The cosine similarity measure is used to determine the similarity between documents. As we 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 the standard approach of 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.

Retrieving clusters that have most relevant documents in the top ranks is the goal of any CBR system. We are most interested in studying whether the proposed cluster representations can help improve the ranking of relevant clusters. Here, we take relevant clusters to be those that give a precision that is better than document-based retrieval with the same number of documents (as that in each cluster) taken from the top of the retrieved list. If the top K documents from document-based retrieval are all relevant, then we consider any cluster with all relevant documents to be relevant. Relevant clusters are identified based on the relevance judgments of documents provided by NIST (http://trec.nist.gov/). Evaluation is done using the Mean Reciprocal Rank (MRR). We go through the list of ranked clusters and mark the highest rank at which a relevant cluster is retrieved. The reciprocal of the rank is computed. The MRR score is the average of reciprocal ranks across all queries on a data set.

Table 2. Comparison of cluster-based retrieval performance using different cluster representations. Evaluation metric is MRR. Percentage improvement over "concatenating documents" is given in parentheses. "*" means that a significant improvement is achieved over "concatenating documents" with a 2-tail t-test at 95% confidence.

Cluster Representation	TREC12	TREC45	AP
Concatenating documents	0.4657	0.4161	0.4801
TF mixture	0.4980 (+6.9%)*	0.4207 (+1.1%)	0.4846 (+0.9%)
DM mixture	0.5247 (+11.3%)*	0.4323 (+3.9%)	0.4875 (+1.5%)

From table 2, we observe that, in general, both TF mixture and DM mixture perform better than the standard approach. DM mixture consistently gives the best performance on all three data sets. TM mixture and DM mixture are more effective for large heterogeneous collections such as TREC12 and TREC45, than for a homogeneous collection like AP. Significant improvements are obtained on the TREC12 collection. A further examination on AP reveals that the member documents in a cluster are similar in length and tend to contribute evenly to the query term distribution in the cluster model, so the proposed methods do not have much advantage over the standard approach in this case.

5. CONCLUSIONS AND FUTURE WORK

We developed and evaluated novel cluster representations for cluster-based retrieval in this work. Early results show that the proposed methods generally perform better than the standard approach. The DM mixture method performs best for all three data sets. These methods seem to be more effective for heterogeneous collections. For future work, we plan to continue exploring different ways of representing clusters and carry out more thorough evaluations on these methods. We will also look into features that can be used to improve cluster representations for homogeneous collections.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

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

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

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

[4] 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.

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

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