CS 646 - IR

P3

Patrick Verga

November 30, 2014

1 Background

My P3 project was based on a paper that presented a clustering based approach to the problem of selecting good documents for pseudo-relevant feedback and query expansion [**?**]. The typical pseudo-relevance feedback model selects the top few documents of an initial retrieval, and then chooses expansion terms from this document set. However, this requires assuming that the top ranked documents are relevant and can often lead to selecting bad documents and bad expansion terms. The clustering method attempts to address this problem by selecting resampling documents.

2 Methods

Experiments were run using the Galago search engine.

2.1 Data

- Robust-Community : 250 queries from TREC on the robust dataset.
- Robust-Class : Queries generated by the class for the robust dataset
- Books : Queries generated by the class for the book dataset

2.2 Clustering Method

The Clustering method was proposed in [**?**]. In this method, the top 100 documents are initially retrieved. Knn (k=5) clustering with cosine similarity is then performed on the tfidf

representations of the documents. Documents are then ranked using a cluster based query liklihood where the documents in the cluster are treated as a single document. Documents from the top ranked clusters are then used to select query expansion terms. In these experiments, we select the top 5 clusters and choose 50 terms.

2.3 BASELINES

- Query Liklihood : Default Galago query liklihood using SDM.
- Relevance Model : Standard relevance model using top 10 documents and selecting 50 expansion terms.
- LDA clustering : Same as the above clustering method but using the LDA topic distribution vectors instead of tfidf.
- Query-LDA : Choosing the documents with the most similar LDA topic distribution vectors to the query.

3 RESULTS

For each of our three data sets we performed three evaluations : mean average precision, normalized discounted cumulative gain, and precision at 10.

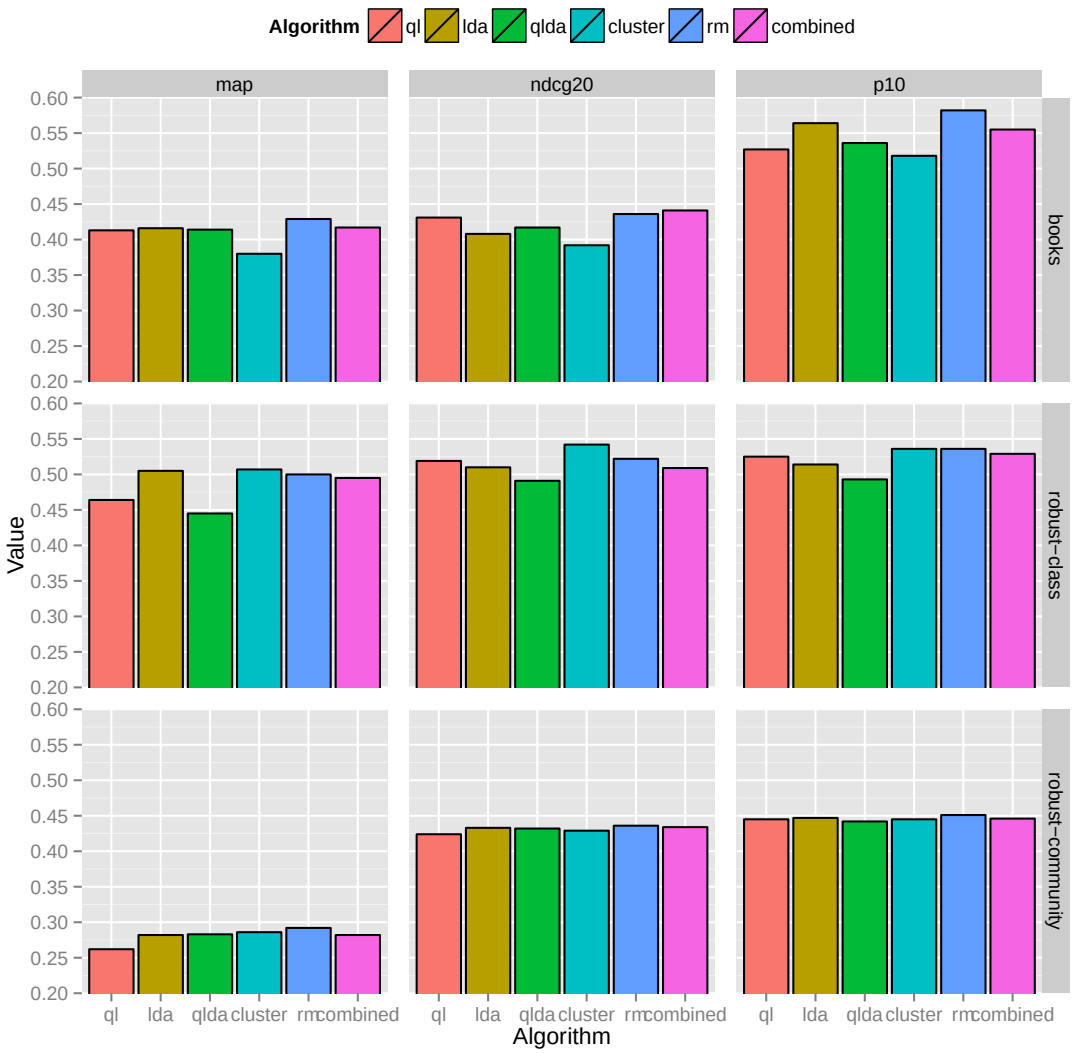

Figure 3.1: ql = Query Liklihood, lda = clustering using lda vectors, qlda = choose documents with most similar lda vectors to query, cluster is the clustering method from [**?**], rm = relevance model, and combined is clustering combining the tfidf vectors and the lda vectors.

		map	ndcg20	p10
Books	Cluster	0.3800	0.3920	0.5180
	RM	0.4290	0.4360	0.5820
	QL	0.4130	0.4310	0.5270
	LDA	0.4160	0.4080	0.5640
	QLDA	0.4140	0.4170	0.5360
	Combined	0.4170	0.4410	0.5550
Robust Class	Cluster	0.5070	0.5420	0.5360
	RM	0.5000	0.5220	0.5360
	QL	0.4640	0.5190	0.5250
	LDA	0.5050	0.5100	0.5140
	QLDA	0.4450	0.4910	0.4930
	Combined	0.4950	0.5090	0.5290
Robust Community	Cluster	0.2860	0.4290	0.4450
	RM	0.2920	0.4360	0.4510
	QL	0.2620	0.4240	0.4450
	LDA	0.2820	0.4330	0.4470
	QLDA	0.2830	0.4320	0.4420
	Combined	0.2820	0.4340	0.4460

4 Discussion

References

[1] Kyung Soon Lee, W. Bruce Croft, and James Allan. 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 - SIGIR '08*, page 235, 2008.

References