Ensemble Clustering for Result Diversification

Dong Nguyen
Human Media Interaction
University of Twente
d.nguyen@utwente.nl

Djoerd Hiemstra
Database Group
University of Twente
d.hiemstra@utwente.nl

ABSTRACT

This paper describes the participation of the University of Twente in the Web track of TREC 2012. Our baseline approach uses the Mirex toolkit, an open source tool that sequantially scans all the documents. For result diversification, we experimented with improving the quality of clusters through ensemble clustering. We combined clusters obtained by different clustering methods (such as LDA and K-means) and clusters obtained by using different types of data (such as document text and anchor text). Our two-layer ensemble run performed better than the LDA based diversification and also better than a non-diversification run.

1. INTRODUCTION

Web queries are often short and ambiguous. Result diversification, which aims to diversify queries to cover the multiple facets or subtopics of a query, can improve the quality of these queries. A common strategy is to estimate the aspects/subtopics of the top ranked documents, and rerank these documents based on the estimated subtopics. Usually the subtopics are discovered by clustering the (top ranked) documents. Two well known methods to rerank results are IA-select [1] and xQuAD [12].

Recently, researchers have explored combining multiple clusterings to improve result diversification. Dou et al. [6] used four methods to obtain subtopics: anchor texts, query logs, search result clusters and hosts. They proposed a reranking framework that incorporated the subtopics from these multiple dimensions. Contrary to our work, they only experimented with clusterings obtained using different data sources, and not with different clustering methods for a particular data source. He et al. [8] proposed a framework to combine clusters of external resources to regularize implicit subtopics based on pLSA using random walks.

In this work, we explore the use of *clustering ensembles* to obtain better clusterings for result diversification. Clustering ensembles can combine arbitrary clusterings, for example based on different data sources (e.g. full document text, anchor text, urls) or by using different clustering methods (such as k-means and LDA [2]). Experiments were done on Category B of ClueWeb09.

We first discuss related work and the track in which we participated. We then describe our experimental setup and discuss the results. We conclude with a summary and suggest future work.

2. WEB TRACK

The Web track of TREC 2012 consists of an adhoc and a diversity track. In this paper we focus on the diversity track. Participants initially only have access to plain queries. However, the evaluation of the runs are evaluated using the full topic descriptions.

Topics are classified either as *ambiguous* or *faceted* [5]. Ambiguous queries have several unrelated interpretations. For example, an ambiguous query in TREC 2010 was *the sun*, which could refer to the newspaper or the star in the solar system. Faceted queries have a primary interpretation. The subtopics then reflect several aspects related to this interpretation. For example, a faceted query was *Neil Young*, with aspects such as Neil Young's albums, biographical information, lyrics and tour dates.

The adhoc task is evaluated using Expected Reciprocal Rank (ERR) [3]. The diversity track is evaluated using an Intent Aware version [1] of Expected Reciprocal Rank (ERR-IA) where the score for the different subtopics are weighted by the probability of that specific subtopic for the given query. In the Web track, these measures are calculated at rank 20. In this paper we also report nDCG@20[10] and α-nDCG@20[4].

3. AD HOC RETRIEVAL

In this section we describe our approach to obtain a baseline ranking. Next, we rerank these results to improve result diversification.

We use Mirex [9],[1] a tool that sequentially scans the documents. Built on Hadoop, sequential scanning becomes a viable approach. In addition, it allows researchers to easily experiment with different retrieval models, because the framework is easy to extend. Documents were scored using a language model with linear interpolation smoothing and a document length prior. We decided to only use anchor text, since previous experiments indicated that this gave high precision and still enough recall for this task.

We use $\lambda = 0.90$ as our baseline for further reranking, after experimenting with different smoothing parameters on data from the Web track of 2009, 2010 and 2011. The baseline run is referred to as utw2012lm09.

[1] http://mirex.sourceforge.net

4. RESULT DIVERSIFICATION

We make the simplifying assumption that a document only belongs to one topic. However, our described methods can easily be extended to support methods where documents belong to multiple topics.

4.1 Clustering

We experiment with several methods to cluster the documents obtained from the baseline ranking.

Methods

I *K-means.* An iterative algorithm where documents are assigned to the cluster with the nearest mean.

II *Ward.* A hierarchical clustering method, where clusters are merged to minimize the total within-cluster variance.

III *LDA* [2]. A generative model that aims to uncover latent topics.

IV *LSA* [11]. A method based on singular value decomposition to uncover latent concepts.

We also vary the data source.

Data

I *Full text.* Cluster documents based on the full text as extracted from the HTML.

II *Anchor.* Cluster documents based on the anchor text.

III *Host.* Documents are assigned to the same cluster when they come from the same host.

In our experiments, we use the same number of clusters for all clustering methods (except for host clustering, for which the number of clusters is dependent on the results). An optimized system that would vary the number of clusters based on the used clustering method or particular query could potentially provide better results.

4.2 Combining Multiple Clusterings

Clustering ensembles combine multiple clusterings into a single clustering. Advantages include more robustness, novelty (a combined solution that may not have been found by the individual clustering algorithms), more stability and confidence, and support of parallelization and scalability [7]. In this paper we cluster the documents using multiple methods and across several dimensions, and combine these into a single, more robust clustering using clustering ensembles.

We apply the most simple method for combining multiple clusterings called the cluster based similarity partitioning algorithm (CSPA) [13]. Two documents have a similarity of 1 if they appear in the same cluster. As a result, for each clustering we are able to create an $n \times n$ binary similarity matrix (with n the number of documents). A similarity matrix for a combined clustering is then just the average of the individual similarity matrices.

We experiment with assigning weights to the specific clusterings. For example, if a certain method has an assigned weight of 0.8, the similarity matrix will have a value of 0.8 if the two documents appear in the same cluster (and zero otherwise). We set the weight such that the total weights for the different clusterings add up to 1. We then apply a clustering method on this induced similarity matrix to make a final clustering. In this paper we use hierarchical clustering using the centroid method, where distances are calculated based on the centroids of the clusters.

The advantage of this approach is that it is independent of the clusterings used. In addition, by combining multiple clusterings into one new clustering, we are also free to choose any reranking algorithm we like to use. And by finding weights for the different clusterings, we obtain insight into what dimension or which clustering methods are effective for result diversification.

I Two-layer Ensemble Clustering

We experiment with an ensemble clustering over ensemble clusterings. The final clustering is an ensemble clustering over three clusterings:

1. *Text clustering.* Ensemble clustering based on clusterings obtained using K-means, Ward, LDA and LSA on the full text.
2. *Anchor clustering.* Ensemble clustering based on clusterings obtained using K-means, Ward, LDA and LSA on the anchor text.
3. *Host clustering.*

II Simple Ensemble Clustering

Preliminary experiments on previous TREC data found LDA to be the most effective of the clustering algorithms. Therefore, in this variant we only use LDA as the clustering method for the text and anchor data:

1. *LDA text clustering.*
2. *LDA anchor clustering.*
3. *Host clustering.*

III One-layer Ensemble Clustering

This ensemble clustering uses the same clusterings as the Two-layer Ensemble Clustering, however the clusters are directly combined into a new clustering, instead of applying two layers. Thus we create an ensemble clustering over the following:

1. *Text - K-means.*

   ...

4. *Text - LSA.*
5. *Anchor - K-means.*

   ...

8. *Anchor - LSA.*
9. *Host clustering.*

Run	nDCG@20	ERR@20	ERR-IA@20	α-nDCG@20
Language modeling baseline (utw2012lm09)	**0.122**	0.218	0.404	0.505
Diversification using LDA (utw2012lda)	0.111	0.215	0.402	0.499
Two-layer ensemble clustering (utw2012c1)	0.120	**0.220**	**0.405**	**0.508**
Simple ensemble clustering (utw2012sc1)	0.107	0.207	0.398	0.498
One-layer ensemble clustering (utw2012fc1)	0.113	0.219	0.400	0.497
Two-layer ensemble clustering (utw2012c2)	0.117	0.219	0.399	0.499

Table 1: Results

4.3 Result Reranking

We use the IA-select algorithm to diversify search results based on clusters [1]. The IA-select algorithm involves computing the conditional probability $P(c|q)$ of a subtopic c given the query q and the quality value of a document d given a query and subtopic, $V(d|q,c)$.

The algorithm then selects documents based on the highest marginal utility:

$$g(d|q,c,S) = \sum_{c \in C(d)} U(c|q,S)V(d|q,c)$$

Where $U(c|q,S)$ is initially set to $P(c|q)$ when no documents are selected yet, and updated for every added document. After preliminary experiments, we decided to calculate $V(d|q,c)$ by the score of the document for query q divided by the total score of all documents in cluster c. $P(c|q)$ is the sum of the quality values of the documents in cluster c divided by the total sum.

4.4 Submitted Runs

For all runs, we rerank the top 1000 documents using clusters with 25 topics. Parameters were selected using a parameter sweep over data from 2009, 2010 and 2011. We submitted the following runs to the adhoc (AH) and diversity (DIV) task:

I [**AH**] **Baseline run (utw2012lm09)** A run using language modeling with $\lambda = 0.9$ and the Mirex toolkit.

II [**DIV**] **LDA (utw2012lda)** A run using LDA clustering based on document text.

III [**DIV**] **Two-layer Ensemble Clustering (utw2012c1)** Clustering based on anchor text (weight 0.8; ensemble cluster of k-means: 0.2, LDA: 0.6, LSA: 0.2) and text (weight: 0.2; ensemble cluster of Ward: 0.2, LDA: 0.8).

IV [**DIV**] **Simple Ensemble Clustering (utw2012sc1)** Clustering based on host (weight 0.8) and LDA based on text (weight 0.2).

V [**AH**] **One-layer Ensemble Clustering (utw2012fc1)** Due to time constraints the weights for this method were not optimized. We used an ensemble clustering over text using LDA (0.4) and anchor text using K-means (0.2) and Ward (0.4).

VI [**AH**] **Two-layer Ensemble Clustering (utw2012c2)** The weights for this run were not optimized. Clustering based on host (weight: 0.4), anchor text (weight 0.2; ensemble cluster of k-means: 0.33, LDA: 0.5, LSA: 0.166) and text (weight: 0.4; ensemble cluster of Ward: 0.2, LDA: 0.8).

5. RESULTS

The results are presented in Table 1. We find that the baseline, with no diversification, performs very well. We suspect that our reranking algorithm is not very effective, since only clustering based on LDA performs worse than the non-diversification run. However, we do find that our two-layer ensemble clustering (utw2012c1) performs better than LDA on all measures, and also better than the non-diversification baseline on all measures except nDCG@20. When comparing based on ERR-IA@20, it performs better than or equal to LDA for 32/50 queries, and for 29/50 queries when comparing with the LM baseline.

The one-layer ensemble clustering (utw2012fc1), performs not as well as the two-layer ensemble clustering, however, since we only did a partial parameter sweep it is hard to draw any conclusions from this. The simple ensemble clustering (utw2012sc1) performs the worst. We would expect this method to perform better than LDA, since LDA is one of the clusterings used. However, we only did a coarse parameter sweep, and perhaps have not find the optimal weights yet. But this also illustrates that the used method is sensitive to the weights that are used. In addition, the performance might be degraded because of the used clustering method to obtain an ensemble clustering based on the similarity matrix.

We will further analyze the performance of our best run (utw2012c1) by comparing with the diversification run using LDA (utw2012lda). The difference in ERR-IA@20 when comparing LDA and the LM baseline (no diversification) can be found in Figure 1. A positive value means that the LDA run performed better. A similar graph comparing the two-layer ensemble model and the LM baseline can be found in Figure 2.

A query that performed well when comparing ERR-IA@20 is query 154 *'figs'* (*Find information on nutritional or health benefits of figs*), with subtopics on nutritional/health benefits, recipes, varieties and growing figs. The LDA run obtained an ERR-IA@20 of 0.384, the LM baseline a score of 0.402 and the utw2012c1 run scored 0.430.

We expect that when using a better reranking algorithm, the results can benefit more from improved clusterings. We also encountered some drawbacks with ensemble clusterings. First, we found it to be sensitive to the weights that were used. In addition, given a similarity matrix, we need to decide on a clustering method. More experiments should be done to assess what clustering method is the most suitable for this task.

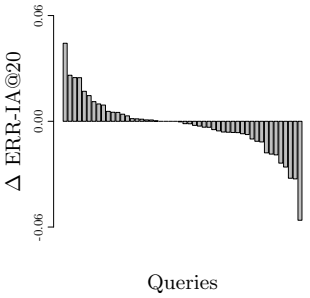

Figure 1: Diff. in performance of LDA (utw2012lda) and the LM baseline.

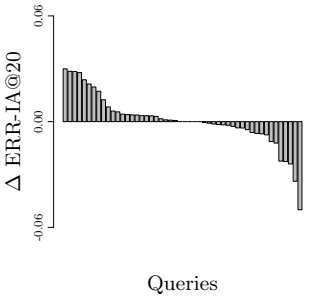

Figure 2: Diff. in performance of two-layer ensemble clustering (utw2012c1) and the LM baseline.

6. CONCLUSION

In this paper we presented the participation of the University of Twente in the Web track of TREC 2012. This year, we focused on the diversity track. We used an ensemble clustering approach aimed to improve the quality of the document clusters. Our ensemble run performed better than the LDA based diversification and also better than a non-diversification run.

The main advantage of this approach is that it is simple, it can be applied on any clustering algorithm, and it is also applicable for any reranking method based on clusters. However, a lot more parameters are introduced, and during development we found the results to be sensitive to the specific parameters used.

Results suggest that the used reranking algorithm might not be effective enough, therefore reducing the possible improvement when better clusters are obtained. For future work other reranking approaches should be explored. In addition, in our experiments we used the same weights across all queries for the different clustering methods and data sources. We expect better results could be obtained by estimating the quality of clusters at query time and adapting the weights per query.

7. ACKNOWLEDGEMENS

This research was supported by the Netherlands Organization for Scientific Research, NWO, grants 640.005.002 and 639.022.809.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[3] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630. ACM, 2009.

[4] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[5] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 Web Track. In *Proceedings of TREC*, 2012.

[6] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 475–484. ACM, 2011.

[7] R. Ghaemi, M. N. Sulaiman, H. Ibrahim, and N. Mustapha. A survey: cluster ensemble techniques. In *Proceedings of World Academy of Science, Engineering andTechnology*, 2009.

[8] J. He, V. Hollink, and A. P. de Vries. Combining implicit and explicit topic representations for result diversification. In *Proceedings of SIGIR*, 2012.

[9] D. Hiemstra and C. Hauff. Mapreduce for information retrieval evaluation: 'let's quickly test this on 12 tb of data'. In *Multilingual and Multimodal Information Access Evaluation. Lecture Notes in Computer Science 6360*, pages 64–69. Springer Verlag, 2010.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, Oct. 2002.

[11] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[12] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890. ACM, 2010.

[13] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, Mar. 2003.