Improving Result Diversity using Probabilistic Latent Semantic Analysis

Peter Lubell-Doughtie
University of Amsterdam
Amsterdam, Netherlands
lubell@science.uva.nl

Katja Hofmann
University of Amsterdam
Amsterdam, Netherlands
khofmann@uva.nl

ABSTRACT

IA-SELECT is a recently developed algorithm for increasing the diversity of a search result set by reordering an original document list based on manually generated clusters. In this paper we extend this approach to create a diversification framework in which arbitrary clustering methods can be used, and where the influence of clusters can be balanced against the original rank of documents. We study whether clusters that are automatically generated using probabilistic latent semantic analysis (PLSA) can compete with manually created clusters, and investigate how balancing the influence of clusters and original document rank affects diversity scores. As there are currently few datasets for evaluating diversity, we develop a new dataset, which is released with this paper. Our results show that diversification using PLSA can improve diversity, but that there is a large gap in performance between automatically and manually created clusters.

Keywords

result diversification, latent semantic indexing, clustering

1. INTRODUCTION

As search engines struggle to return well-ranked and relevant information for ambiguous queries from large and growing sets of documents, interest in improving accuracy through alternative methods has increased [7]. One approach is to ensure that documents representing multiple topics, or aspects of a query are highly ranked, by reducing redundancy within the same topics. This can be measured by the *diversity* of a set of search results, which reflects that set's coverage of multiple interpretations of a query.

We present a result diversification system that extends the recently developed IA-SELECT method [1]. IA-SELECT reorders documents based on manually created clusters reflecting different interpretations of a query. We create a diversification framework which can use arbitrary clustering methods, and where the influence of clusters is balanced against the original rank of documents.

We assume that the interpretations of a query contain documents that are conditionally independent given this interpretation, and can therefore be represented by a mixture of conditionally independent clusters. Additionally, in order to cope with ambiguity in query term meaning, we desire a model that can represent polysemy. These conditions justify our use of a conditionally independent latent class model, such as PLSA.

Because there are currently few datasets for evaluating diversification approaches, we contribute a new small scale dataset to complement the TREC ClueWeb09 dataset[1]. It is created from a question answering corpus in which ideal clusters are given by human judges. We are releasing this dataset with our paper.[2]

Our results show that while diversification using PLSA can improve diversity, there is a significant gap between the performance of automatically and manually created clusters. The best diversity scores were achieved with a non-linear function that weights a document's original rank higher for highly ranked documents and places more importance on cluster structure at lower ranks.

2. RELATED WORK

Our result diversification system is a continuation of related research focusing on measuring the diversity of a list of search results and designing algorithms that optimize result order to increase diversity. The earliest diversity metric and algorithm formally explored is maximum marginal relevance (MMR), which maximizes a linear interpolation of the similarity between each document and the query, minus the similarity between that document and previously returned documents [2]. In [8] the authors apply MMR to subtopic retrieval and find that gains obtained by increasing the rank of novel documents are offset by the cost of increasing the rank of non-relevant documents, as is confirmed in our experiments. Both the original [2] and modified [8] MMR do not directly measure subtopic coverage and assume document novelty is independent from document relevance.

Clarke et al. [4] address the shortcomings of MMR by explicitly measuring subtopic retrieval . The summarization and question answering community defines *information nuggets* (or nuggets) as representations of facts, topicality, or any binary property of a document or information need. Clarke et al. assign nuggets to the query and its returned documents and define the probability that a document is relevant based on the intersection of its nuggets and the query's nuggets. Based on nDCG, the authors define α-nDCG, which rewards novelty through a gain vector accounting for the relevant nuggets within a document.

DIR'2011, February 4, 2011, Amsterdam.

[1] `http://boston.lti.cs.cmu.edu/Data/clueweb09/`
[2] `http://code.helioid.com/diversity/` and `http://ilps.science.uva.nl/webclef_diversity`

α-nDCG unrealistically assumes all nuggets are equally relevant. Agrawal et al. [1] address this by defining *intent aware* (IA) metrics, which sum evaluation scores over categories, weighted by the probability of a category given a query. Categories are defined as locations in a taxonomy of information and user intents, for our purposes they can be seen as equivalent to nuggets. Agrawal et al. also present the IA-SELECT algorithm, which reorders results to maximize the likelihood that the top k results will covers all the query's categories relative to their likelihood.

Dou et al. [5] present a more general algorithm which combines various subtopic indicators and further improves diversity scores. It is possible that greedy strategies exclude a relevant but rare nugget which co-occurs only in documents containing other nuggets already returned. To address this [3] assumes one "correct" interpretation of a query and returns documents covering all its *facets* (which are defined similarly to nuggets or categories).

3. APPROACH

Our diversification system performs three steps: (i) retrieval, (ii) clustering, and (iii) reordering. We use clusters generated from an initial ranked document list to ensure documents from different clusters are highly ranked and a single cluster is not overrepresented. A function of rank and cluster membership likelihood balances the importance of rank and cluster diversity. Clusters represent query interpretations and diversifying over clusters will diversify over interpretations.

We use a modified version of IA-SELECT to reorder documents [1]. Given the query q, a category (nugget) c, and a document d, IA-SELECT builds a set of documents, labeled S, which maximizes utility. Using a measure of document quality, $V(d|q,c)$, and the conditional distribution over categories for the query and current S, $U(c|q,S)$, the algorithm calculates utility as:

$$g(d|q,c,S) = \sum_{c \in C(d)} U(c|q,S)V(d|q,c) \quad (1)$$

where $C(d)$ is the set of categories for document d. $U(c|q,S)$ is initialized as $P(c|q)$, defined below, and then at each iteration the algorithm adds the d with greatest utility and updates $U(c|q,S)$. This allows us to approximate the S which maximizes utility.

Given a query, in step (i) our implementation uses the successful BM25 formula to create an ordered set of documents. To model each cluster as a possible interpretation of the query, we assume clusters are independent and that documents can belong to an arbitrary number of clusters. This motivates using PLSA to assign cluster membership probabilities to the top k documents in step (ii).

Using the cluster probabilities for the top k documents we calculate our initial conditional category distribution as:

$$P(c|q) = \sum_{d \in D} p(c|d)^{\phi_p(\text{rank}(q,d))}, \quad (2)$$

where $p(c|d)$ is the probability that document d is a member of cluster c, $\text{rank}(q,d)$ returns the rank of d for q, and ϕ_p determines the importance rank plays in calculating cluster to query relevance. Document quality is similarly calculated as:

$$V(d|q,c) = p(c|d)^{\phi_v(\text{rank}(q,d))}, \quad (3)$$

where ϕ_v is the importance of rank in calculating relevance. When ϕ_p and ϕ_v are constant rank is irrelevant, otherwise the greater their convexity the greater the influence of rank.

Using these definitions of document quality and conditional category distribution we perform step (iii) and reorder the top k documents. As opposed to IA-SELECT, we explicitly define the value and conditional category distributions in terms of rank and cluster membership probability. This allows us to adjust their influence to benefit the system's goals, in our case, increased diversity scores.

4. EXPERIMENTS AND RESULTS

There are few standard information retrieval evaluation sets that can be used to evaluate diversification because most do not define ground truth categories for documents. In addition, many evaluation sets used in diversity research are either proprietary and unreleased, or are incompletely evaluated versions of question answering corpuses, and can be used only after preprocessing and result extrapolation. The recent ClueWeb09 dataset provides a large diversification task. To complement this, we develop a smaller scale dataset based on the WebCLEF 2007 question answering corpus [6].

In this corpus, information nuggets are assigned to each document and defined by a set of passages taken directly from the document text. To convert this dataset into a retrieval task with subtopics we parse the assessments file, letting the topic of each question form the query and the answer nuggets form the query's subtopics. We then search for the nuggets in the corpus' documents to generate a subtopic document list.

We run experiments with the following settings. The baseline is generated by retrieving the top 200 documents using BM25 with $k_1 = 1$ and $b = 0.3$.

To test the influence of induced clusters, we apply PLSA to the top 20 and 200 returned results to create 20 subtopics, and then reorder with $\phi_p(x) = 1 + \log(x)$ and $\phi_v(x) = x^2$. After experimenting with different functions we found that these produce the best results by appropriately weighting the influence of rank and cluster membership. We evaluate our system using α-nDCG and P-IA, following [1, 4].

The results of our experiments are shown in Table 1. We see that reordering based on 20 documents has the best performance for α-nDCG@{5,10,20} with scores of 0.151, 0.157, and 0.180 respectively. It is unable to beat the baseline for P-IA@10 but does so for P-IA@{5,20} with scores of 0.055 and 0.049 respectively, where the P-IA@20 score is significant at the 0.001 level using a paired student's t-test. Reordering based on 200 documents has the worst performance on all metrics.

Increasing the influence of rank by increasing the convexity of ϕ_v increases diversity scores up to a point. In addition to the results displayed in Table 1, we tested $\phi_v(x) = \{1, 1 + \log(x), x, x^2, x^3\}$. Excluding P-IA@10, $\phi_v(x) = x^2$ produces the best scores. Ignoring rank, with a constant $\phi_v(x) = 1$, produces the lowest scores in all runs except P-IA@20. Up to and including $\phi_v(x) = x^2$, α-nDCG scores increase as function convexity increases, but further increasing convexity decreases scores.

To determine how reordering with 20 results is able to improve on the baseline scores, we plot the α-nDCG@5 scores per query in Fig. 1. Considering individual queries, the reordered list produces better results by matching or im-

Table 1: Diversity scores for all diversification systems. Significant differences from the baseline are marked with $^{\triangledown}$(decrease, $p = 0.01$) and $^{\triangle}$(improvement).

Experiment	α-nDCG@5	α-nDCG@10	α-nDCG@20	P-IA@5	P-IA@10	P-IA@20
Baseline	0.145	0.155	0.175	0.051	**0.046**	0.031
PLSA 20	**0.151**	**0.157**	**0.180**	**0.055**	0.044	**0.049**$^{\triangle}$
PLSA 200	0.136	0.152$^{\triangledown}$	0.173	0.049	0.040	0.032
QRELS 20	0.324	0.305	0.287	0.080	0.050	0.033
QRELS 200	0.611	0.632	0.621	0.134	0.102	0.079

proving over the baseline on most queries and substantially beating the baseline on a few queries (2 and 11). The algorithm takes a conservative approach and maximizes utility by reordering in cases where both document rank and subtopic relevance are high.

In order to measure the effect of the unequal distribution of subtopics per query, Fig. 1 also plots the number of subtopics in each query. The correlation between reordering performance and the number of subtopics is low, with a Pearson correlation coefficient between the number of subtopics and the baseline, rank2, and constant runs of 0.08, 0.03, and -0.17 respectively. This indicates that performance is not directly related to the number of subtopics in a query.

The IA measures have an undefined upper bound which is less than 1 unless there is a single perfect ordering for all subtopics [1]. In addition, if the best ordering relies on a document outside the top k reordered documents it will be impossible to achieve the maximum score. To estimate this upper bound we reorder based on the ground truth subtopics and assignments in the dataset. The results are labeled as QRELS and shown in the lower part of Table 1. Except for P-IA@20 these scores are substantially higher than either the baseline scores or those achieved when reordering using PLSA clusters. In this case, ignoring rank with constant $\phi_v(x) = 1$ gives the best scores and increasing the influence of rank decreases scores, the opposite of what occurs when using induced subtopics. This is expected if subtopics are more relevant to improving diversity than rank, and provides anecdotal evidence that these estimates may form a reasonable upper bound.

5. DISCUSSION

In our diversification system, changes in the importance of diversity are expressed by changing the influence of rank through the ϕ_v function. We would expect the influence of document rank to be inversely correlated with the diversity of reordered results. However, this is not strictly the case as we achieve maximum diversity scores by balancing the influence of rank and relevance.

In an approach based on reordering results according to subtopics, the effectiveness of the system depends on generating subtopics aligned with those used by the scoring function. Putting significant emphasis on a document's rank appears to be successful primarily due to the poor quality of induced clusters. Experiments generating an upper bound demonstrate that increasing cluster quality and decreasing rank's influence correlate with higher diversity scores.

This is exemplified by results for the query "plastic tableware and the environment" (topic 26), in which the PLSA and QRELS orderings disagree for the second document returned. QRELS returns a document about restrictions on plastic products, fitting the nugget "restricted use of disposable plastic," while PLSA returns a document listing various plastic products for sale, including biodegradable products. Although the document returned by PLSA does not fit any given subtopics it could arguably fit an appropriate subtopic, such as "environmentally friendly plastic products." Here the diversity score is decreased by an understanding of the query's subtopics that is discordant with the subtopics used in evaluation, although not necessarily incorrect.

Concerning the diversification algorithm, other variations in the influence of rank, i.e. $\phi_v(x) = x^n$ for $1 < n < 3$, may further improve scores. That increasing n — the influence of rank — eventually leads to decreasing scores shows that clusters provide valuable information about how to best reorder documents. A more effective strategy would bias towards the original ranking when doing so benefits diversity scores and away when it does not. Figure 1 presents the α-nDCG@5 score per query using a method that heavily weights rank, *rank2* with $\phi_v(x) = x^2$, and a method that ignores rank, *constant* with $\phi_v(x) = 1$. Although the overall score of constant is much lower, on certain queries (2 and 16) it significantly outperforms the baseline and rank2. We suspect that constant performs well on these queries because the clusters generated for them closely match the known clusters.

In our PLSA 20 experiments, which reorder 20 documents using IA-SELECT with 20 PLSA topics, our implementation may reduce to a maximum likelihood estimator by assigning one document to each class, leading to equivalent class and document language models. Work remains to be done in testing that one document is indeed assigned to each class. In this case our implementation would be very similar to the original MMR algorithm and future work could investigate this connection and its implications for the usefulness of PLSA in search result diversification.

6. CONCLUSIONS AND FUTURE WORK

The expansion of online documents and users has led to increases in the number of documents a query is applicable to and in the number of users using the same or similar queries to express different information needs. This, in turn, has led to an increase the number of valid yet differing ways in which we can interpret queries. A complementary challenge arises when different queries express similar information needs. This has also been exacerbated by increasing numbers of documents and users. Search result diversification methods address these challenges by satisfying users' multifarious needs. Diversity research has moved beyond independent analysis of document novelty and relevance (as in MMR) to measuring a document's contribution in relation to the additional information it provides.

In this paper we have shown that using PLSA to create

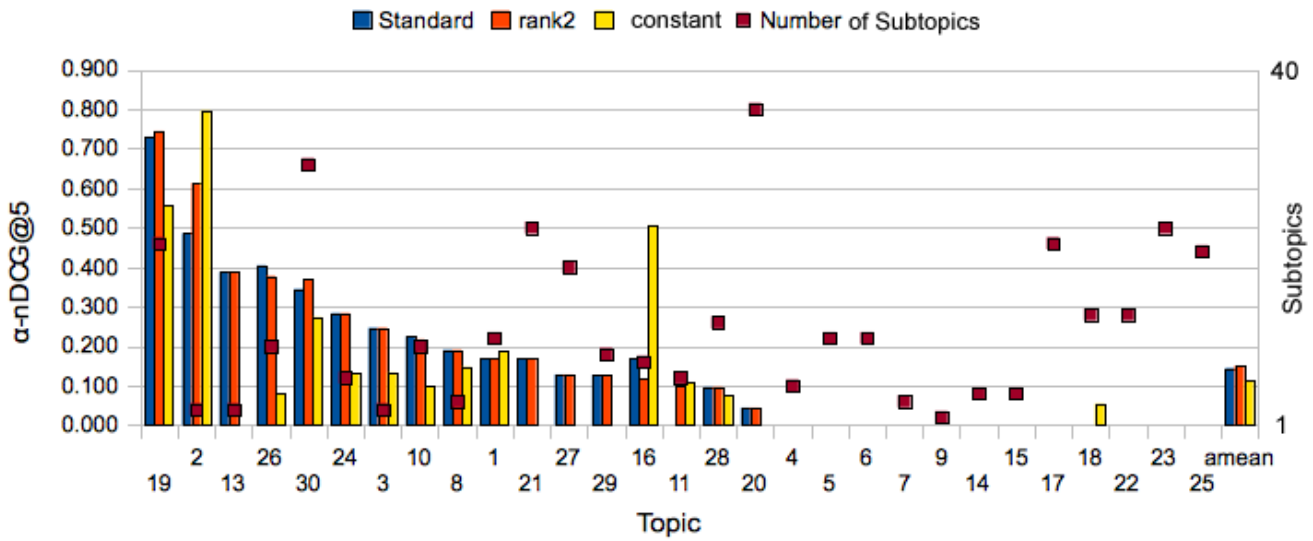

Figure 1: α-nDCG@5 per query for 20 documents, the final column is the arithmetic mean. Topics are ordered by decreasing score for rank2 ($\phi_v(x) = x^2$).

an external partition for reordering search results can improve diversity. The functioning of the reordering algorithm is sensitive to, and can be tuned through, changes in the influence of a document's original rank. Decreasing the influence of rank puts more trust in the accuracy of induced clusters and vice versa.

Future work includes inducing clusters with alternative algorithms and adapting to specific queries. In our experiments we use a fixed number of clusters. This could be improved by changing the number of clusters relative to vocabulary cardinality or other heuristics. In addition, our results show that some queries greatly benefit from diversification while for others the original ranking performs better. Diversification could be applied selectively, for example based on measures developed for query performance prediction or topic models.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*, pages 5–14, 2009.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, 1998.

[3] B. Carterette and P. Chandar. Probabilistic Models of Novel Document Rankings for Faceted Topic Retrieval. *CIKM '09*, pages 1287–1296, 2009.

[4] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666, 2008.

[5] Z. Dou, K. Chen, R. Song, Y. Ma, S. Shi, and J. Wen. Microsoft Research Asia at the Web Track of TREC 2009. In *TREC '09*. NIST, 2009.

[6] V. Jijkoun and M. de Rijke. Overview of webclef 2007. *Advances in Multilingual and Multimodal Information Retrieval*, pages 725–731, 2008.

[7] M. Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR '08*, pages 499–506, 2008.

[8] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03*, pages 10–17, 2003.