Incorporating Statistical Topic Information in Relevance Feedback

Karla Caballero*
UC, Santa Cruz
Santa Cruz CA, USA
karla@soe.ucsc.edu

Ram Akella
UC, Santa Cruz
Santa Cruz CA, USA
akella@soe.ucsc.edu

ABSTRACT

Most of the relevance feedback algorithms only use document terms as feedback (local features) in order to update the query and re-rank the documents to show to the user. This approach is limited by the terms of those documents without any global context. We propose to use statistical topic modeling techniques in relevance feedback to incorporate a better estimate of context by including global information about the document. This is particularly helpful for difficult queries where learning the context from the interactions with the user is crucial. We propose to use the topic mixture information obtained to characterize the documents and learn their topics. Then, we rank documents incorporating positive and negative feedback by fitting a latent distribution for each class of documents online and combining all the features using Bayesian Logistic Regression. We show results using the OHSUMED dataset for 3 different variants and obtain higher performance, up to 12.5% in Mean Average Precision (MAP).

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance Feedback, Document Filtering*

General Terms

Algorithms, Experimentation

Keywords

Relevance Feedback,Topic Models, Language Models

1. INTRODUCTION

Relevance feedback has been studied extensively in Information Retrieval as a form of incorporating feedback from the user to refine the results retrieved. The authors in [5] concluded that negative feedback is also valuable to improve the ranking. However, the need to capture broader context in difficult queries is still a challenge. The authors in [2] have showed that including global features and using clusters can improve the retrieval performance significantly. Thus, statistical topic modeling provides a robust and automatic method to incorporate context to the user feedback.

Previous approaches have used statistical topic models to represent documents according to their latent topic content and use this representation in information retrieval [1, 4]. Authors in [4], use topics as a form of smoothing the Language Model used in retrieval. However, this approach does not address the incorporation of relevance feedback. Recent work from [1] explores the use of topics as a form to perform the query expansion for relevance feedback. However, this action might make the query noisier because the top topic terms might not contribute to a better discrimination of the relevant documents. In addition, these terms might not be distinctive across different topics.

We propose to include the topic information as feedback using the document topic mixture instead of the document word mixture. We first estimate the topic mixture for each document in the corpus using LDA and save it as meta data. Given an initial query we use a standard retrieval engine, Language Models for this case, to show the first set of documents to the user and obtain relevance judgments. Then, we assume that topic mixtures for feedback documents are observed. We then define two latent Dirichlet distributions: one for relevant documents and another for non-relevant documents. We fit these distributions iteratively, by finding a sufficient statistic and maximizing the likelihood of observing this statistic. To score the documents, we use Bayesian Logistic Regression. This function results in a very efficient scoring function, and incorporates the benefits of active learning. Under this model, we incorporate positive and negative feedback, and context based on topics in the interaction without changing the query. We also provide efficient updates of the latent distributions based on topics.

2. METHODOLOGY

In this section we describe how we incorporate the topic mixture of feedback documents as a global measure in contrast to query expansion. To achieve this, we estimate the topic mixtures of the documents, θ_i, for K topics in the corpus using LDA off-line. Given a initial ranking, the user provides relevance feedback which is used to fit the latent relevant/non-relevant distributions of topic mixtures. Thus, we assume two Dirichlet distributions: one for the relevant set of documents α_R, and one for the non-relevant documents $\alpha_{\bar{R}}$. Therefore, we have:

$$P(\theta_i|\alpha_R) \sim Dirichlet(\alpha_R) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_{k,R})}{\prod_{k=1}^{K} \Gamma(\alpha_{k,R})} \prod_{k=1}^{K} \theta_{i,k}^{\alpha_{k,R}-1}$$

for α_R and $\alpha_{\bar{R}}$. Then, we calculate the log-probability of the document being generated by those distributions:

*Main contact.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

$$\text{Log}P(\theta_i|\alpha) = \log\Gamma(\sum_{k=1}^{K}\alpha_k) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\log(\theta_{i,k})$$

We denote these scores as PR_i and $P\bar{R}_i$ respectively. To update the latent distributions of the relevant α_R and non-relevant topics $\alpha_{\bar{R}}$, we use the topic content from the documents labeled by the user as relevant, θ_R, and non-relevant, $\theta_{\bar{R}}$, after each interaction. The Dirichlet distribution guarantees a unique maximum when the Maximum Likelihood (ML) is estimated for α. Moreover, a sufficient statistics, SS, can be estimated to update this distribution as more observations are available. We can update SS efficiently without keeping previous document feedback. The initial value of the sufficient statistic $SS^{(0)}_{k,R}$ for the relevant topic k and its update from the interaction j is described by:

$$SS^{(0)}_{k,R} = \frac{1}{N^{(0)}_R}\sum_{i\in R_0}\log\theta_{i,k}$$

$$SS^{(j)}_{k,R} = \frac{N^{(j-1)}_R}{N^{(j)}_R}SS^{(j-1)}_{k,R} + \frac{1}{N^{(j)}_R}\sum_{i\in R_j}\log\theta_{i,k}$$

where $N^{(j)}_R = N^{(j-1)}_R + |R_j|$, and $|R_j|$ is the total number of relevant documents at the j-th interaction. Given $SS^{(j)}_R$ and $SS^{(j)}_{\bar{R}}$, we use the method proposed in [3] to calculate the ML estimator for $\alpha^{(j)}_R$ and $\alpha^{(j)}_{\bar{R}}$. In addition to these distributions, we use the topic-based Language Model P_{TW} for document i as follows:

$$P_{TW,i}(w|\theta_i,\hat{\phi}) = \sum_{k=1}^{K}P(w|z=k,\hat{\phi}_k)P(z=k|\theta_i)$$

where $\hat{\phi}$ are the word mixture for the topics obtained from LDA. Thus, the score $S_{TW,i}$ for query Q with terms q is defined as:

$$S_{TW,i} = \prod_{q\in Q}P_{TW,i}(w=q|\theta_i,\hat{\phi})$$

To combine the scores from the latent relevant/non-relevant topic mixtures and the topic-based Language Model, we use the Bayesian Logistic Regression approach [5]. Let $y_i = \{+1,-1\}$ be the relevant/non-relevant label for document, we have the score function:

$$P(y_i|\beta,\mathbf{d_i}) = \frac{1}{1+\exp(-\beta^T\mathbf{d_i}y_i)}$$

where $\mathbf{d_i}$ is the feature vector scores: PR_i, $P\bar{R}_i$, $S_{TW,i}$. β is a parameter vector assumed to be normally distributed and updated in a Bayesian form. Here, the distribution of β from the j-th iteration is taken as prior distribution for the next iteration. To approximate the posterior distribution we use the Laplace approximation as discussed in [5].

3. RESULTS

We test our method using the OHSUMED dataset which consists of 196,000 medical abstracts and 3,506 relevance labels for 63 queries from the Document Filtering Track from TREC 4. As suggested in the track, we assume unobserved labels as non-relevant. We fit the LDA model using $K = 50$ topics, which is the number of topics with highest performance based on Empirical Likelihood. To test the impact of topic information, we use standard Language Model (LM) with Dirichlet smoothing described in [4] as baseline. This score is used with Bayesian Logistic Regression. We test 3 variants of the model and the baseline: LM as baseline; LM+S_{TW}; LM+S_{TW}+PR; LM+S_{TW}+PR+$P\bar{R}$. We calculate the initial ranking using LM and asked for feedback until we have at least one relevant and one non-relevant documents. We use 10 feedback documents and estimate precision at 10 (P@10), Mean Average Precision (MAP), and Discounted Gain (DiscGain). There are two relevance level labels available in the dataset, $\{1,2\}$, that are assumed equally, $\{+1\}$ for P@10 and MAP. However for DiscGain, we use both labels in the evaluation.

Table 1: Results of Topic feedback using 50 topics in the OHSUMED dataset

Method	P@10	MAP	DiscGain
LM	0.3968	0.4286	0.5660
LM+S_{TW}	0.4206	0.4557	0.6315
LM+S_{TW}+PR	0.2968	0.3307	0.5590
LM+S_{TW}+PR+$P\bar{R}$	**0.4698**	**0.5141**	**0.6580**

Table 1 shows the results for the variants tested. We observe that the LM+S_{TW} performs better than the baseline. This score is similar to the LDA-based retrieval proposed in [4] but the value of the linear combination parameters β is fitted based on the feedback as opposed to a corpus-wide parameter. When we incorporate only the score from the relevant distribution of topics PR, the performance decreases. However, when the score for the non-relevant distribution $P\bar{R}$ is incorporated, the performance is the highest. This shows the value of negative feedback reported previously in [5]. We notice that, the combination of PR and $P\bar{R}$ is equivalent to the log of the likelihood ratio test (probabilistic ranking principle) weighted by β. This also explains why both scores should be included in the model.

We observe that the combination of the four scores improves the general performance by: **11.6**% P@10, **12.8**% MAP, and **4.6**% DiscGain respect topic-based language model (LM+S_{TW}). This demonstrates, the power of statistical topic modeling in relevance feedback.

4. CONCLUSION AND FUTURE WORK

We have presented a method to incorporate statistical topic information in relevance feedback without changing the query. Results show that including the mixture of topics in relevance feedback improves the performance by pruning the search space, and adding context to the query. As future work, we plan to incorporate a policy to decide when to update the parameters of the relevant and non-relevant topic distribution optimally.

5. ACKNOWLEDGMENTS

This work is partially funded by CONACYT grant 207751 and SAP Gift Support

6. REFERENCES

[1] D. Andrzejewski and D. Buttler. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD conference*, KDD '11, pages 600–608, 2011.

[2] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the SIGIR conference*, pages 235–242, 2008.

[3] T. Minka. Estimating a dirichlet distribution. Technical report, 2003.

[4] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th ACM SIGIR conference*, SIGIR '06, pages 178–185, 2006.

[5] Z. Xu and R. Akella. A bayesian logistic regression model for active relevance feedback. In *Proceedings of the 31st ACM SIGIR conference*, SIGIR '08, pages 227–234, 2008.