Low-dimensional Query Projection based on Divergence Minimization Feedback Model for Ad-hoc Retrieval

Javid Dadashkarimi and Masoud Jalili Sabet and Azadeh Shakery and Heshaam Faili
School of ECE, College of Engineering, University of Tehran, Iran
{dadashkarimi,jalili.masoud,shakery,hfaili}@ut.ac.ir

Abstract
Pseudo-relevance feedback (PRF) is known as a technique for updating query language models. In this paper we propose a new technique for PRF, based on an embedded coefficient matrix, whose aim is to improve the vector representation of the query by transforming it to a more reliable space, and then update the query language model. The proposed embedded coefficient divergence minimization model (ECDMM) takes the top-ranked documents retrieved by the query and obtains a couple of positive and negative samples; these samples are used for learning the coefficient matrix which will be used for updating the query language model. Experimental results on several TREC and CLEF data sets demonstrate effectiveness of ECDMM compared to state-of-the-art PRF techniques in language modeling framework.

1 Introduction
Top-ranked documents in response to the query of a user in the Web have long been considered as a useful collection for updating the query, resulting in retrieving more relevant documents (Lv and Zhai, 2014; Lavrenko and Croft, 2001; Zhai and Lafferty, 2001; Rocchio, 1971). Maximum-entropy divergence minimization model (MEDMM) and the Rochio algorithm are among powerful methods for language modeling and vector-space model respectively. The main idea of these methods is to find a model more closer to the top-ranked documents and far away from a noisy collection (Lv and Zhai, 2014). In this paper, we propose a pseudo-relevance feedback technique for language modeling, the state-of-the-art retrieval framework, whose aim is to find a better direction for the vector representation of the query by learning an embedded coefficient matrix on a set of positive and negative samples and then rotate the vector of the query to a more relevant sub-space. Finally, the obtained query vector is used to expand the original query with a number of related words. Experimental results on several non-industrial data sets from TREC and CLEF in English, French, Spanish, German, and Persian demonstrate the effectiveness of the proposed method.

2 Previous Works
2.1 Pseudo-relevance Feedback
Top-ranked documents $F = \{d_1, d_2, ..., d_{|F|}\}$ in response to a query $q$ have long been used as helpful resources for expanding the query (Lavrenko and Croft, 2001; Lv and Zhai, 2014). Lavrenko et al., introduced relevance models for updating the query. The RM1 method models the query as $p(w|\theta_q) \propto \sum_{d \in F} p(w|\theta_d)p(\theta_d)\prod_{i=1}^{q} p(q_i|\theta_d)$ where $\theta_d$ is the language model of document $d \in F$. The RM2 models the query as $p(w|\theta_q) \propto p(w)\prod_{i=1}^{q} \sum_{d \in F} p(q_i|\theta_d)\frac{p(w|\theta_d)p(\theta_d)}{p(w)}$. The obtained relevance models can be interpolated with the original query as follows:

$$p(w|\theta'_q) = (1-\alpha)p(w|\theta_q) + \alpha p_{ml}(w|q)$$ (1)

where $p_{ml}(w|q)$ is the maximum likelihood estimation of the original query. The interpolated model for RM1 and RM2 are known as RM3 and RM4 respectively (Jaleel et al., 2004). Zhai and Lafferty introduced the mixture model (MIXTURE) based on an expectation maximization algorithm in which topicality of a word estimated in E-step.
(i.e., \( f^{(n)}(w) = \frac{(1-\lambda)p^{(n)}_\lambda(w|\theta_F)}{(1-\lambda)p^{(n)}_\lambda(w|\theta_F) + \lambda p(u|c)} \)) is used for estimating the feedback model in M-step (i.e., \( p^{(n+1)}_\lambda(w|\theta_F) = \frac{\sum_{i=1}^n c(w,d_i)f^{(n)}(w)}{\sum_i \sum_j c(w,d_i)p^{(n)}_\lambda(w_j)} \)) (Zhai and Lafferty, 2001). MEDMM, introduced by Lv and Zhai, is another powerful feedback model that aims at minimizing a linear combination of a number of entropy components as follows:

\[
\arg \min_{\theta} \sum_{d \in F} \alpha_d H(\theta_F, \theta_d) - \lambda H(\theta_F, \theta_C) - \beta H(\theta_F) \quad (2)
\]

where \( H(\theta_1, \theta_2) \) is the cross-entropy between \( \theta_1 \) and \( \theta_2 \) and \( H(\theta) \) is the entropy of \( \theta \).

### 2.2 Low-dimensional Word Vectors

Low-dimensional representations of words are tailored in a variety of tasks in natural language processing (Collobert et al., 2011). To learn these vectors a common approach is to predict the context of each word and then aim at minimizing a loss function. The method is known as skip-gram negative sampling and can be interpreted as a binary regression task with \( \mathcal{L}(\theta) = \sum_{(w,c) \in \mathcal{D} \cup \mathcal{D}'} \log \left( \frac{1}{1 + \exp(v_c^T w)} \right) \) where \( v_w \in \mathbb{R}^{n \times 1} \) and \( v_c \in \mathbb{R}^{n \times 1} \) are the vectors of a word and its context respectively. \( z \) indicates if this sample \((w, c)\) is a valid sample \((z = 1)\) or not \((z = 0)\) (Goldberg and Levy, 2014).

### 3 Embedded Coefficients for Query Projection

In this section we propose embedded coefficient divergence minimization model (EC-DMM) for updating the query language model. To this end, we aim at finding a coefficient matrix \( W \in \mathbb{R}^{n \times n} \) for projecting the query model \( v_q \) to a more relevant space by minimizing \( f \) as follows:

\[
f(W) = \sum_{w_n \in F} \frac{\alpha}{2} \|W^T v_q - v_{w_n}\|^2 - \frac{\lambda}{2} \|W^T v_q - v_{w_n}\|^2 - \frac{\beta}{2} \|W^T W\|
\]

where \( v_q \in \mathbb{R}^{n \times 1} \) is the query vector, \( v_{w_n} \in \mathbb{R}^{n \times 1} \) is the vector of a relevant sample from \( F \), and \( v_{w_n} \in \mathbb{R}^{n \times 1} \) is the vector of a non-relevant sample from \( F \). The query vector is built from averaging the vectors of query words (i.e., \( [v_q]_j = \frac{1}{m} \sum_{1 \leq i \leq m = |q|} [v_{q_i}]_j \)).

Equation 3 has similar components to MEDMM (Lv and Zhai, 2014). The first part captures the same essence as the cross-entropy between the feedback model \( \theta_F \) and the positive sample model \( \theta_w \) in MEDMM. It tries to minimize the distance between the query model and the positive sample model. The second part also captures the effect of the negative samples on \( \theta_F \). It tries to maximize the distance between the query model and the negative sample model. And, \( W^T W \) acts as a regularization term for \( W \) in the model.

### 3.1 Building Embedded Query Model

After learning \( W \) we project the query vector \( v_q \) to the new space (i.e., \( W^T v_q \)). The obtained vector represents a better direction of the query semantically and therefore, we aim at defining a feedback model based on similarity of the projected query with the words from the feedback documents:

\[
p(w_n | \hat{\theta}_F) \propto \sum_{j=1}^n c(w_n, d_j) e^{\frac{W^T v_q v_{w_n}}{\|v_q\| \|v_{w_n}\|}}
\]

### 4 Experiments

#### 4.1 Experimental Setup

Overview of the used collections are provided in Table 1. In all experiments, we use the language modeling framework with the KL-divergence retrieval model and Dirichlet smoothing with \( \mu = 1000 \). All European documents and queries are stemmed by the Porter stemmer; the Persian collection are remained intact (Hashemi and Shakery, 2014; Dadashkarimi et al., 2014; Rahimi et al., 2016). Stopwords are removed in all the experiments. The Lemur toolkit is employed as the re-
Table 1: Collections Characteristics

| ID  | Lang. | Collection                          | Queries (title only)            | #docs  | #qrels |
|-----|-------|------------------------------------|---------------------------------|--------|--------|
| AP  | English | Associated Press 88-89            | TREC 1-3 Ad-Hoc Track, Q:51-200 | 164,597| 15,838  |
| ROB356 | English | Los Angeles Times 1994, plus Glasgow Herald 1995 | CLEF 2003-2004-2006, Q:141-350 | 170,153| 4,327  |
| SP  | Spanish | EFE 1994                          | CLEF 2002, Q:91-140             | 215,738| 2,854  |
| DE  | German | Frankfurter Rundschau 94, SDA 94, Der Spiegel 94-95 | CLEF 2002-03, Q:91-140           | 225,371| 1,938  |
| FR  | French | Le Monde 94, SDA French 94-95     | CLEF 2002-03, Q:251-350         | 129,806| 3,524  |
| FA  | Persian | Hamshahri 1996-2002               | CLEF 2008-09, Q:551-650         | 166,774| 9,625  |

Table 2: Comparison of different feedback methods. Superscripts 1/2/3/4/5 indicate that the MAP improvements are statistically significant (2-tail t-test, $p \leq 0.05$). * indicates $0.05 \leq p \leq 0.1$.

| ID | MAP   | P@5 | P@10 | MAP   | P@5 | P@10 | MAP   | P@5 | P@10 | MAP   | P@5 | P@10 |
|----|-------|-----|------|-------|-----|------|-------|-----|------|-------|-----|------|
| AP |        |     |      |        |     |      |        |     |      |        |     |      |
| MLE | 0.2643 | 0.451 | 0.4262 | 0.3721 | 0.4366 | 0.3719 | 0.348 | 0.532 | 0.458 |
| MIXTURE | 0.2962 | 0.4725 | 0.4362 | 0.3781 | 0.4366 | 0.3758 | 0.3933 | 0.532 | 0.48  |
| RM3 | 0.2751 | 0.4577 | 0.4302 | 0.3783 | 0.4366 | 0.3739 | 0.3697 | 0.548 | 0.464 |
| RM4 | 0.2681 | 0.451  | 0.4282 | 0.3847 | 0.4379 | 0.3758 | 0.3185 | 0.496 | 0.43  |
| MEDMM | 0.2702 | 0.4617 | 0.4262 | 0.375* | 0.434 | 0.3719 | 0.3551 | 0.54  | 0.466 |
| ECDMM | 0.3222 | 0.4765 | 0.4544 | 0.3863 | 0.4405 | 0.383 | 0.4321 | 0.532 | 0.506 |

| ROB356 |        |     |      |        |     |      |        |     |      |        |     |      |
| MLE | 0.3554 | 0.584 | 0.561 | 0.3936 | 0.5212 | 0.4556 | 0.488 | 0.664 | 0.578 |
| MIXTURE | 0.3738 | 0.596 | 0.575 | 0.4172 | 0.5293 | 0.4717 | 0.5197 | 0.676 | 0.586 |
| RM3 | 0.3646 | 0.59 | 0.565 | 0.4051 | 0.5253 | 0.4636 | 0.5027 | 0.68  | 0.596 |
| RM4 | 0.3721 | 0.596 | 0.586 | 0.3662 | 0.4848 | 0.4374 | 0.4759 | 0.652 | 0.58  |
| MEDMM | 0.3585 | 0.59 | 0.56 | 0.3982* | 0.5172 | 0.4566 | 0.4963 | 0.664 | 0.584 |
| ECDMM | 0.3911 | 0.608 | 0.576 | 0.4217 | 0.5152 | 0.4717 | 0.5384 | 0.652 | 0.602 |

| DE |        |     |      |        |     |      |        |     |      |
| MLE | 0.2643 | 0.451 | 0.4262 | 0.3721 | 0.4366 | 0.3719 | 0.348 | 0.532 | 0.458 |
| MIXTURE | 0.2962 | 0.4725 | 0.4362 | 0.3781 | 0.4366 | 0.3758 | 0.3933 | 0.532 | 0.48  |
| RM3 | 0.2751 | 0.4577 | 0.4302 | 0.3783 | 0.4366 | 0.3739 | 0.3697 | 0.548 | 0.464 |
| RM4 | 0.2681 | 0.451  | 0.4282 | 0.3847 | 0.4379 | 0.3758 | 0.3185 | 0.496 | 0.43  |
| MEDMM | 0.2702 | 0.4617 | 0.4262 | 0.375* | 0.434 | 0.3719 | 0.3551 | 0.54  | 0.466 |
| ECDMM | 0.3222 | 0.4765 | 0.4544 | 0.3863 | 0.4405 | 0.383 | 0.4321 | 0.532 | 0.506 |

trieval engine in our experiments. ECDMM is compared with the following methods: (1) maximum likelihood estimation of query (MLE): $p(w|\theta_q) = \frac{c(w,q)}{|q|}$ where $c(w,q)$ is the count of term $w$ in the query; (2) RM3, (3) RM4, and (4) the MEDMM model explained in Section 2.1.

$\alpha$ in Equation [2] is set to the default value 0.5 and number of blind relevant documents is assumed $|F'|=15$. All free parameters $\alpha$, $\lambda$, $\beta$, $n^+$, and $n^-$ are fixed for all experiments after learning on a small sub-set of topics from the AP collection. Empirically we fixed the parameters to $\alpha = 0.8$, $\lambda = 0.05$, $\beta = 0.01$, $n^+ = 40$, and $n^- = 100$ in all the experiments. $W$ in Equation [4] is initialized with random values in $[-1, 1]$; $\eta$ is set to a small value which also decreases after each iteration. The iterations terminate when the changes are very small or the number of iterations meets 1000. $v_{wn}$ and $\bar{v}_{wn}$ are computed with negative sampling skip-gram introduced in (Mikolov et al., 2013); size of the window, number of negative samples, and size of the vectors are set to typical values of 10, 45, and 100 respectively.

4.2 Performance Comparison and Discussion

All the experimental results are provided in Table 2. As shown in the table, the proposed ECDMM outperforms all the baselines in terms of MAP, P@5, P@10 in quite all the collections. The results show 21.9%, 3.8%, 25.3%, 10.0%, 2.8%, 9.3% improvements in AP, ROB356, DE, FA, FR, and SP respectively in terms of MAP compared to MLE. As dis-
Table 3: Query language models for "airbus subsidy" by different feedback models. Terms are stemmed by the Porter stemmer.

| MIXTURE  | RM3    | RM4    | MEDMM   | ECDMM   |
|----------|--------|--------|---------|---------|
| airbu    | 0.0382 | said   | 0.0101  | airbu   | 0.0882  |
| subsidi  | 0.0232 | s      | 0.0074  | subsidi | 0.0552  |
| us       | 0.0185 | airbu  | 0.0068  | us      | 0.0322  |
| govern   | 0.0182 | us     | 0.0055  | subsidi | 0.0120  |
| west     | 0.0164 | subsidi | 0.0046 | us      | 0.0116  |
| trade    | 0.0143 | govern | 0.0044 | govern  | 0.0080  |
| industri | 0.0130 | west   | 0.0035 | trade   | 0.0078  |
| daimler  | 0.0114 | trade  | 0.0033 | state   | 0.0071  |
| yeutter  | 0.0113 | will   | 0.0033 | will    | 0.0059  |
| econom   | 0.0111 | state  | 0.0030 | aircraft | 0.0052 |

ECDMM takes advantage of the first step of MIXTURE for positive sampling and idea of MEDMM for divergence minimization. But, the experimental results reveal that ECDMM is more effective and captures both topicality and entropy.

Table 3 shows top-10 stemmed expansion terms and the weights corresponding to obtained query model by the methods. It is more clear that MIXTURE and ECDMM purify the feedback model from common words more successfully. On the other hand, MEDMM and ECDMM captures the original query more than other methods; this is the reason that both the models do not strongly dependant on the interpolation with original query (see (Lv and Zhai, 2014) and Figure 1 where $0.7 \leq \alpha \leq 0.8$ works well in all the collections). Generally speaking, the proposed ECDMM weights semantically related words more than others.

4.3 Parameter Sensitivity

In this section we investigate the sensitivity of the proposed method to the number of positive and negative samples. To this aim, one parameter is fixed to its optimum value and then try to get optimum value of the other one. As shown in Figure 1 both parameters $n^+$ and $n^-$ work stable in all the collections.

5 Conclusion and Future Works

In this paper, we presented a PRF model using low-dimensional query projection. We used a set of positive and negative samples from the top-ranked documents, retrieved by the query, to learn an embedded coefficient matrix. The query vector, which got transformed by the coefficient matrix, is then used to expand the original query. We showed that the proposed model is robust regarding the parameters. Our model, in terms of MAP, has significant improvements up to 3.8% compared to the other state-of-the-art models.

For our future work, we want to study the usage of low-dimensional representations in recommendation systems by applying a similar method on the users’ profiles. We will also test our models with industrial data sets.
References

[Collobert et al.2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493–2537.

[Dashkarimi et al.2014] Javid Dadashkarimi, Azadeh Shakery, and Hesham Faili. 2014. A Probabilistic Translation Method for Dictionary-based Cross-lingual Information Retrieval in Agglutinative Languages. In Conference of Computational Linguistic, Tehran, Iran.

[Goldberg and Levy2014] Yoav Goldberg and Omer Levy. 2014. word2vec explained: Deriving mikolov et al.’s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

[Hashemi and Shakery2014] Homa B. Hashemi and Azadeh Shakery. 2014. Mining a Persian-English Comparable Corpus for Cross-language Information Retrieval. Inf. Proc. Man., 50(2):384–398.

[Jaleel et al.2004] Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. In TREC.

[Lavrenko and Croft2001] Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In SIGIR ’01, pages 120–127.

[Lv and Zhai2014] Yuanhua Lv and ChengXiang Zhai. 2014. Revisiting the divergence minimization feedback model. In CIKM ’14, pages 1863–1866.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Adv. in Neur. Inf. Proc. Sys., pages 3111–3119.

[Rahimi et al.2016] Negin Rahimi, Azadeh Shakery, Javid Dadashkarimi, Mozdeh Aryannejad, Mostafa Dehghani, and Hossein Nasr Esfahani. 2016. Building a multi-domain comparable corpus using a learning to rank method. Natural Language Engineering.

[Rocchio1971] Joseph John Rocchio. 1971. Relevance feedback in information retrieval.

[Zhai and Lafferty2001] ChengXiang Zhai and John Lafferty. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In Proceedings of the 10th International Conference on Information and Knowledge Management, pages 403–410, Atlanta, Georgia, USA.