Profile-based Translation in Multilingual Expertise Retrieval

Hossein Nasr Esfahani, Javid Dadashkarimi, and Azadeh Shakery
{h_nasr,dadashkarimi,shakery}@ut.ac.ir

School of ECE, College of Engineering, University of Tehran, Iran

Abstract. In the current multilingual environment of the web, authors contribute through a variety of languages. Therefor retrieving a number of specialists, who have publications in different languages, in response to a user-specified query is a challenging task. In this paper we try to answer the following questions: (1) How does eliminating the documents of the authors written in languages other than the query language affect the performance of a multilingual expertise retrieval (MLER) system? (2) Are the profiles of the multilingual experts helpful to improve the quality of the document translation task? (3) What constitutes a good profile and how should it be used to improve the quality of translation? In this paper we show that authors' documents are usually related topically in different languages. Interestingly, it has been shown that such multilingual contributions can help us to construct profile-based translation models in order to improve the quality of document translation. We further provide an effective profile-based translation model based on topicality of translations in other publications of the authors. Experimental results on a MLER collection reveal that the proposed method provides significant improvements compared to the baselines.

Keywords: Expert retrieval, multilingual information retrieval, profiles.

1 Introduction

Expert retrieval has achieved growing attention during the past decade. Users in the web aim at retrieving a number of specialists in specific areas [3]. A couple of methods have been introduced for this purpose; retrieving the experts based on their profiles (the candidate-based model), and retrieving the experts based on their published contributions (the document-based model) [3]. The latter approach is usually opted in the literature due to its better performance and its robustness to free parameters [1].

Since there exist a lot of authors who contribute through a variety of languages, using documents written in other languages than the query should intuitively be able to improve the performance of the expertise retrieval system. However scoring documents in such a multilingual environment is challenging. Multilingual information retrieval (MLIR) is a well-known research problem and has been extensively studied in the literature [10]. There are two options for scoring documents written in languages other than the language of the query; translating the query into all the languages of the documents, or representing all the documents in the language of the query. In MLIR it has been shown that the second approach outperforms the first one in the language modeling framework [9]. In the current paper we are going to cast such an approach to multilingual expert retrieval (MLER). Indeed, our new problem is to retrieve experts who are contributing in multiple languages.

In this research we choose the document translation approach for our problem. It is noteworthy that no translated document in the traditional sense is produced, but rather a multilingual representation of the underlying original document that is suitable for retrieval, but not for consumption by a reader, is constructed.

Furthermore, proper weighting of translations has always had a major effect on MLIR performance. Therefore improving the translation model based on user profile can supposedly lead to better MLER performance.

We are trying to answer the following research questions in this paper:

1. How does eliminating the documents of the authors written in languages other than the query language affect the performance of an MLER system?
2. Are the profiles of the multilingual experts helpful to improve the quality of the document translation task?
3. What constitutes a good profile and how should it be used to improve the quality of translation?

Our findings in this paper reveal that multilingual profiles of the experts are useful resources for extraction of expert-centric translation models. To this aim we propose two profile-based translation models using (1) maximum likelihood estimation (PBML), and (2) topicality of the terms (PBT). Indeed translations are chosen based on their contributions in the target language documents of an expert. Our experimental results on a multilingual collection of researchers, specialists, and employees at Tuilberg University [5] reveal that the proposed method achieves better performance on a variety of query topics, particularly in ambiguous ones.

In Section 2 we provide brief history of studies in the literature of MLER and MLIR. In Section 3 the proposed profile-based document translation method is introduced. In Section 4 we provide experimental results of the proposed method and several baselines and then we conclude the paper in Section 5.

2 Previous Work

There have been multiple attempts in the expert finding literature. Most of the research studies aim at retrieving a number of experts in response to a query [4]. Usually a couple of models are employed in an expert retrieval system; candidate-based model and document-based model. Although the former model takes advantage of lower costs in terms of space by providing brief representations for the experts, the latter one achieves better results in some collections [1]. A number of frameworks have been proposed for this aim; model-based frameworks based on statistical language modeling, and frameworks based on topic modeling [2,8]. Balog et al. proposed a language modeling framework in which they first retrieve a number of documents in response to a query and then rank the documents based on their likelihood to the user-specified query. After employing an aggregation module, experts are ranked based on their contributions in the retrieved documents. Theoretically in such a module, there are two factors affecting the retrieval performance; the query likelihood of the documents of the experts, and the prior knowledge about the documents. In the lack of prior knowledge about documents, the documents of an expert are assumed to have uniform distribution. Deng et al. introduced a citation-based model to improve the accuracy of the knowledge about the documents [8]. Nevertheless, the former approach due to its simplicity and its promising results is a popular one in the literature.

In the current multilingual environment of the web, experts are contributing in a variety of languages. In such an environment, a reliable strategy should be employed to bridge the gap between the languages [10,14,7]. A couple of methods for acheiving this goal are proposed; posing either multiple translated queries to the system or retrieving multiple translated documents in response to a query [10]. Although the former method demands an effective rank-aggregation strategy [12], the latter one achieves promising performance in the language modeling framework [10]. These approaches in MLIR can also be adapted to MLER.

3 Profile-based Document Translation

In this section we introduce the proposed expert finding system. The system is going to be used in a multilingual environment to retrieve a number of experts in response to a user-specified query. In this environment the documents of the experts are not necessarily represented in the language of the query.

In MLIR two major approaches are used to overcome this issue. the first approach translates the query into all the languages of the documents and then executes multiple retrieval processes and finally aggregates the results; the second approach represents the documents in all the languages that the query can be posed in and then executes a single retrieval process. Since superiority of the latter approach compared to the former one has been shown in the literature [10], the strategy of the proposed framework lies also on the same road.

To this aim, we use the documents in the profile of an expert to disambiguate translations of terms in the document. Our assumption is that an expert usually publishes articles in one area. So we expect to be able to estimate a robust translation model using the documents of an expert from other languages. In Section 3.1 we delve into the problem by introducing a novel method to build a profile for each expert to improve the translation disambiguation quality, in Section 3.2 we use the proposed profiles to disambiguate translations, and in Section 3.3 we explain the whole expertise retrieval process.

3.1 Building Profiles for Translation Disambiguation

The main goal of the proposed PDT framework is to use local information of the experts' documents to improve the quality of translations. In order to intuitively explain the key idea, consider the following example: suppose an expert has 2 document sets D_1 and D_2 in languages l_1 and l_2 respectively and we want to translate term w_s from one of the documents of D_1 to language l_2. If w_s has two translations w_{t_1} and w_{t_2}, we investigate how these translations are contributing in D_2 documents. The higher the contribution of a translation in D_2, the more likely it is to be the correct translation of w_s. To this end we first construct multiple term distributions in different languages for each expert. We explore two methods to compute the contribution of each term: maximum likelihood and topicality.

Maximum Likelihood Estimation of Contribution of Each Term: In this method we assume that the terms that are more frequent in each expert's documents are more contributing to the whole profile, so we estimate the contribution of each term in a set of documents D as follows:

$$C(w|D) = \frac{\sum_{d \in D} c(w;d)}{\sum_{d \in D} |d|} \tag{1}$$

In Equation 1, $C(w|D)$ indicates the contribution of term w to document set D, $c(w;d)$ indicates the number of occurrences of term w in document d and $N(d)$ is the number of terms in document d.

Topicality Estimation of the Contribution of Each Term: We can use topicality of each term as the measure of contribution of that term to a document set. Zhai Lafferty in [15] proposed an EM based method to compute topicality of terms for pseudo-relevance feedback. We use a similar method: let $\theta_{e_i}^{l_k}$ be the estimated profile model of expert e_i in language l_k based on the relevant document set $D_{e_i}^{l_k} = \{d_1, d_2, .., d_n\}$. According to Zhai & Lafferty we also set λ to some constant to estimate $\theta_{e_i}^{l_k}$. Similar to the model-based PRF we estimate the model with an expectation maximization (EM) method:

$$t^{(n)}(w;l_k) = \frac{(1-\lambda)p_\lambda^{(n)}(w|\theta_{e_i}^{l_k})}{(1-\lambda)p_\lambda^{(n)}(w|\theta_{e_i}^{l_k}) + \lambda p(w|\mathcal{C}^{l_k})} \tag{2}$$

$$p_\lambda^{(n+1)}(w|\theta_{e_i}^{l_k}) = \frac{\sum_{j=1}^{n} c(w;d_j) t^{(n)}(w;l_k)}{\sum_{w'} \sum_{j=1}^{n} c(w';d_j) t^{(n)}(w';l_k)} \tag{3}$$

in which l_k is the k-th language of the expert e_i. λ indicating amount of background noise when generating documents d_j. The obtained language model for expert e_i in language l_k is based on topicality of the words. If a word frequently occurrs in the publications of the expert and also if it is a non-common term through the collection $\mathcal{C}^{l_k}$, it will get a high weight in the profile $\theta_{e_i}^{l_k}$. Our main contribution is to use the language models of the experts in different languages to construct a robust translation model for document translation. Therefore contribution of each term in document set D_{l_k} would be:

$$C(w|D_{e_i}^{l_k}) = p_\lambda(w|\theta_{e_i}^{l_k}) \tag{4}$$

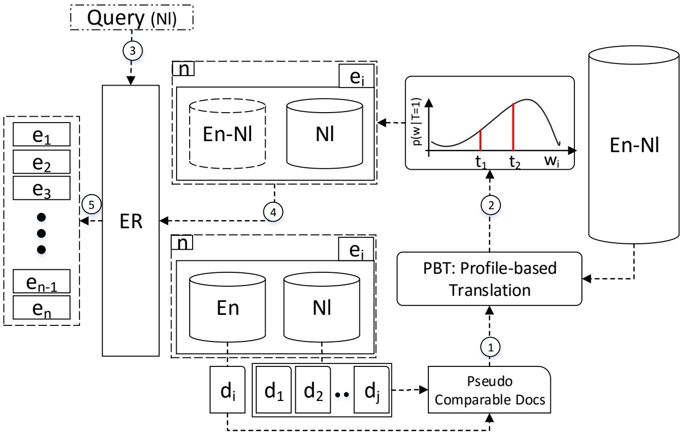

Fig. 1: The proposed expert retrieval framework in multilingual environments.

3.2 Document Translation Based on Cross-lingual Profiles

In this section we introduce the proposed document translation method based on the constructed profiles for each expert. Our goal is to construct translation models for the experts and then to build multilingual documents for them. The translation model for expert e_i is computed as follows:

$$p(w_{t_j}|w_s;e_i) \approx \frac{C(w_{t_j}|D_{e_i}^{l_t})}{\sum_{j'} C(w_{t_{j'}}|D_{e_i}^{l_t})} \tag{5}$$

in which $w_T = \{w_{t_1}, w_{t_2}, .., w_{t_m}\}$ is the set of translation candidates for term w_s from the dictionary. Translations are in language l_t and since we have document translation, w_t is in the source language l_s.

Combining with Other Translation Models: As shown in the cross-lingual information retrieval (CLIR) literature, combining different translation techniques can be useful to obtain a robust translation model [13]. In the proposed framework we also use a general probabilistic dictionary and aim at adapting it to the domain of each expert. We exploit a simple linear interpolation technique:

$$p_\alpha(w_{t_j}|w_s;e_i) = \alpha p(w_{t_j}|w_s;\theta_{par}) + (1-\alpha)p(w_{t_j}|w_s;e_i) \tag{6}$$

where $p(w_{t_j}|w_s;\theta_{par})$ is the translation probability of w_s to w_{t_j} regarding the model obtained from a probabilistic dictionary, and α is a controlling constant.

3.3 The Proposed Expert Retrieval Process

Figure 1 shows the whole process of the proposed expert retrieval system. As shown in the figure, in the first step documents whose languages are different from the query are translated using the PDT framework. This translation technique is based on Rahimi et al. [10] in which all the translations are considered in the retrieval process. Indeed documents are scored based on their relevance to the query. The relevance is computed based on $p_\alpha(w_{t_j}|w_s;e_i)$ obtained in Equation 6. Finally experts are scored based on a document-based model:

$$p(q|e_i) = \sum_d p(q|d)p(d|e_i) \tag{7}$$

For simplicity we estimate $p(d|e_i)$ with a uniform distribution over all the publications of e_i. Moreover we estimate $p(q|d)$ as follow:

$$p(q|d) = \prod_{w \in q} p(w|\theta_d) \tag{8}$$

Similar to [10] we compute $p(w|\theta_d)$ in a multilingual environment as follows:

$$p(w|\theta_d;e_i) = \lambda p_{ml}(w|\theta_d;e_i) + (1-\lambda)p'(w|\mathcal{C}) \tag{9}$$

in which:

$$p'(w|\mathcal{C}) = \frac{\sum_{d \in \mathcal{C}} c_p(w,d)}{N \sum_{d \in \mathcal{C}} |d|}, \quad p_{ml}(w|\theta_d;e_i) = \frac{c_p(w,d)}{N|d|},$$
$$c_p(w,d) = \sum_{u \in d} p(w|u;\theta_{e_i}^{l_k})c(u,d). \tag{10}$$

and N is the number of languages in the collection.

Time Complexity: Although document translation could be time consuming, and profiled based translation exacerbates the problem, but it is worth mentioning that we only translate the terms which are likely to be translated to a query term. Furthermore the EM process is to be computed once per expert and could be done offline, hence this process is totally practical. Nevertheless, the translation model for each expert must be updated when a new document is inserted.

4 Experiments

In this section we provide experimental results of the proposed PDT framework and a number of baselines on a multilingual expert retrieval collection.

4.1 Experimental Setups

We used the bilingual TU expert collection [5] in our experiments. This collection contains a number of documents written by scientists, researchers, and support staff from Tilburg University The collection is provided in an English-Dutch environment. Table 1 shows some statistics one the dataset and Figure 2 shows the contribution of each expert on the set. As shown in Figure 2, experts have enough documents in both languages which makes the dataset suitable for our tests.

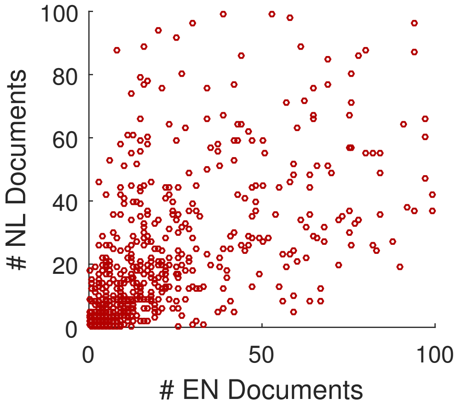

Fig. 2: Distribution of number of documents for each expert.

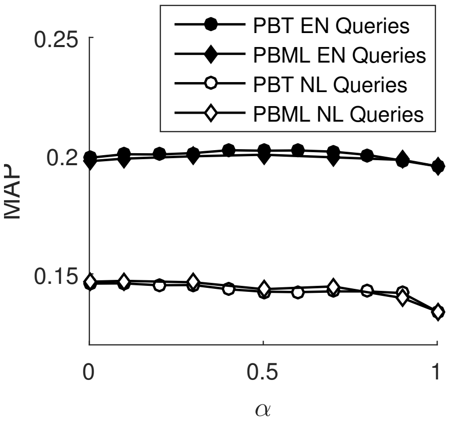

Fig. 3: Sensitivity of the interpolation framework to α

ID	collection	Queries	#queries	#experts	#docs	μ_d	#qrels
TU	Researchers, Scientists, and Employees at Tuilberg University	EN	1,673	893	16,237	1,336	3,936
		NL	2,470	881	20,356	1,204	4,868

Table 1: Collection Statistics. μ_d is the average document length.

Parameter Settings: In all experiments, the Jelinek-Mercer smoothing parameter λ is set to the typical value of 0.9. All free parameters, particularly the constant controlling values of the linear interpolations, are set using 2-fold cross validation over the collection. The noise constant in the EM algorithm is set to 0.7 according to [15].

Evaluation Metrics: We evaluate all the methods based on Mean Average Precision (MAP) of all the retrieved experts as the main evaluation metric. We also report the precision of the top 5 (P@5) and top 10 (P@10) retrieved documents. Statistical differences between the performance of the proposed PDT method and all the baselines are also computed based on two-tailed paired t-test with 95% confidence level on the main evaluation metric [11]. We also provide robustness index (RI) [6] for the last set of our experiments for all the competitive baselines computed as $\frac{N_+ - N_-}{|Q|}$ where $|Q|$ is the number of queries in the collection. N_+ shows the number of queries we have improvements by the proposed method and N_- shows the number of queries in which we have performed worse. Indeed, RI represents the robustness of the method among the query topics.

4.2 Results and Discussions

In this section we report the experimental results of the proposed method and some MLER and CLER baselines. The baselines include MLER based on document translation using (1) top-ranked translations in a probabilistic dictionary

English (EN)						Dutch (NL)					
	TOP-1	MT	PAR	PBT	EN-EN		TOP-1	MT	PAR	PBT	NL-NL
MAP	0.2898	0.2740	0.2898	**0.2911**^{1}	0.2633	MAP	0.2656	0.2458	0.2668	**0.2674**$^{012^*}$	0.2504
P@5	0.1782	0.1637	0.1782	**0.1787**	0.1723	P@5	0.1559	0.1392	0.1568	**0.1571**	0.1474
P@10	0.1208	**0.1244**	0.1208	0.1212	0.1164	P@10	0.1007	0.0981	0.1016	**0.1016**	0.0942

Table 2: Using different translation methods for multilingual expert retrieval. Indicators 0/1/2 denote statistical differences between TOP-1/MT/PAR with confidence of 95%. $*$ shows the confidence is above 90%.

(TOP-1)[1], (2) document translation based on machine translation (MT), (3) weighted translation provided by a probabilistic dictionary (PAR), (4) monolingual retrieval by eliminating documents in out-of-the-context languages (the EN-EN run or the NL-NL one), (5-6) profile-based document translation where profiles are computed w.r.t maximum likelihood (PBML) and topicality (PBT).

Table 2 shows all the results. As shown in the table, all the MLER baselines outperform the simple mono-lingual one. This demonstrates that all the publications of an author, either those in the language of the query or those in other languages, are helpful in our retrieval performance. Although the proposed PBT method outperforms all the baselines in terms of MAP, P@5, and P@10, the improvements in English queries are marginal. The reason for marginal improvements in this dataset goes back to the high performance of the monolingual results. As shown in the table the results of the mono-lingual runs are competitive to the MLER ones (90.45% and 93.64% of PBT in EN-EN and NL-NL runs respectively).

We did further experiments to directly study the effect of the proposed profile-based document translation method. We opted CLER instead of MLER for this purpose. In Table 3 experimental results of a number of CLER runs are provided. These experiments are done only on the documents which are in out-of-the-context languages. To shed light on the effectiveness of the profile-based translation model, we experiment on a subset of the queries which are ambiguous. A query is considered to be ambiguous if at least one of its terms is ambiguous. A term w_t is ambiguous if there exists a term w_s such that $p(w_t|w_s) > 0$ and there exist at least 2 term $w_{t'}$ which $p(w_{t'}|w_s) > \delta$, where δ is a constant value (empirically we set $\delta = 0.2$). As shown in the table, PBT outperforms all the TOP-1, PAR, and PBML baselines in all the evaluation metrics. In the Dutch queries improvements in terms of MAP are also robust (0.2215 out of $[-1, 1]$).

Figure 3 shows the sensitivity of the interpolation framework to α (see Equation 6). As shown in the figure, although the proposed PBT takes advantage of the interpolation approach in both English and Dutch queries, the overall changes are very robust to the parameter. Nevertheless, the results of the PAR baseline without any interpolation with the profile-based translation model drop considerably in Dutch.

[1] We have used a probabilistic dictionary provided by the Google machine translator.

English (EN)					Dutch (NL)				
	TOP-1	PAR	PBML	PBT		TOP-1	PAR	PBML	PBT
MAP	0.1945	0.1955	0.1949	**0.2026**^{012}	MAP	0.1221	0.1341	0.1455	**0.1458**^{01}
P@5	0.1195	0.1208	0.1229	**0.1275**	P@5	0.0712	0.0829	0.0883	**0.093325**
P@10	0.087	0.0867	0.0885	**0.09015**	P@10	0.0585	0.0647	0.0669	**0.0691**
RI	-	-0.1566	-0.0482	**0.0172**	RI	-	0.0196	0.1538	**0.2215**

Table 3: Experimental results for different translation methods for cross-lingual expert retrieval over ambiguous queries.

To sum up our findings we answer the following research questions:

1. How does eliminating the documents of the authors written in languages other than the query language affect the performance of an MLER system? Regarding the competitive mono-lingual results in Table 2 in the TU dataset we can claim that the authors repeat majority of their contributions through languages and so their publications in only one language are almost good but not complete indicators of their expertise. However this kind of conclusion is not valid in real-world data and sometimes authors contribute mainly in a language other than the language of the query.
2. Are the profiles of the multilingual experts helpful to improve the quality of the document translation task? When we want to translate a document of an expert, documents of the expert written in the target language help us to find topical terms. Since correct translations are more likely to be the topical ones we expect to reach a better translation (see Figure 3).
3. What constitutes a good profile and how should it be used to improve the quality of translation? According to Table 3 the proposed PBT method outperforms PBML. This shows that topicality of translations instead of their simple maximum likelihood probabilities are helpful for the document translation task. Further results reveal that interpolating the topical probabilities with values from parallel dictionaries are also useful.

5 Conclusion and Future Work

In this paper we elaborate on the subject of MLER by introducing a novel profile-based document translation method. We have set a number of research questions to this aim and our findings supported the following views: (1) According to our observations, although authors contribute almost similarly in multiple languages, considering all the contributions in different languages can be helpful for expertise retrieval system. Since authors usually repeat their contributions through languages, eliminating documents in out-of-the-context languages does not harm the retrieval performance considerably. (2) Document translation in MLER takes advantage of profile-based translation models. The profile of each expert helps us to opt for topical translations which usually contributes to correct translations. Experimental results on the TU dataset, demonstrate that the proposed profile-based translation approach outperforms a variety of baselines.

An interesting future work of this paper is dynamically learning the interpolation weight between topical probabilities and values from dictionaries based on generality of words. Constructing profiles for a number of expert clusters and employing them in the document translation process will be another future work for this paper.

References

1. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Inf. Ret. pp. 43–50. ACM (2006)
2. Balog, K., Azzopardi, L., de Rijke, M.: A language modeling framework for expert finding. Inf. Proc. & Man. 45(1), 1–19 (2009)
3. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L.: Expertise retrieval. Found. Trends Inf. Retr. 6, 127–256 (Feb 2012)
4. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L.: Expertise retrieval. Foundations and Trends in If. Ret. 6(2–3), 127–256 (2012)
5. Berendsen, R., Rijke, M., Balog, K., Bogers, T., Bosch, A.: On the assessment of expertise profiles. Journal of the American Society for Inf. Sci. and Tec. 64(10), 2024–2044 (2013)
6. Collins-Thompson, K.: Reducing the risk of query expansion via robust constrained optimization. In: Proceedings of the 18th ACM Conference on Inf. and Know. Manag. pp. 837–846. ACM (2009)
7. Dadashkarimi, J., Shakery, A., Faili, H.: A Probabilistic Translation Method for Dictionary-based Cross-lingual Information Retrieval in Agglutinative Languages. In: Conference of Computational Linguistic (2014)
8. Deng, H., King, I., Lyu, M.R.: Formal models for expert finding on dblp bibliography data. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. pp. 163–172. IEEE (2008)
9. Nie, J.Y.: Cross-language information retrieval. Synthesis Lectures on Human Language Technologies 3(1), 1–125 (2010)
10. Rahimi, R., Shakery, A., King, I.: Multilingual information retrieval in the language modeling framework. Inf. Ret. Journal 18(3), 246–281 (2015)
11. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Res. and Dev. in Inf. Ret. pp. 162–169. ACM (2005)
12. Tabrizi, S.A., Dadashkarimi, J., Dehghani, M., Esfahani, H.N., Shakery, A.: Revisiting optimal rank aggregation: A dynamic programming approach. In: International Conference on the Theo. of Inf. Ret., ICTIR, September (2015)
13. Türe, F., Lin, J.J., Oard, D.W.: Combining statistical translation techniques for cross-language information retrieval. In: COLING. pp. 2685–2702 (2012)
14. Vulic, I., Smet, W.D., Tang, J., Moens, M.: Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. Inf. Process. Manage. 51(1), 111–147 (2015)
15. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the Tenth International Conference on Inf. and Kno. Man. pp. 403–410. ACM (2001)