ISSN 2278-3091
Volume 8, No.1.4, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse0381.42019.pdf
https://doi.org/10.30534/ijatcse/2019/0381.42019

Information Retrieval Framework for Digital Resource Objects

Wafa' Za'al Alma'aitah [1], Abdullah Zawawi Talib [2], Mohd Azam Osman [3]

[1] School of Computer Sciences, Universiti Sains Malaysia, Malaysia, wzma15_com083@student.usm.my
[1] Department of Basic Sciences, Hashemite University, Jordan, wafaa_maitah@hu.edu.jo
[2] School of Computer Sciences, Universiti Sains Malaysia, Malaysia, azht@usm.my
[3] School of Computer Sciences, Universiti Sains Malaysia, Malaysia, azam@usm.my

ABSTRACT

Basically, digital resource objects (DRO) suffer from two fundamental issues, namely lack of quality of metadata content and difficulty in accessing metadata content. These lead to decrease in the performance of the DRO retrieval. With a view to increase the performance of the DRO retrieval, many components of information retrieval have been enhanced such as document expansion (DE), retrieval model such as Dirichlet smoothing (DS) model, and query expansion (QE). Most of these studies have shown that employing IR components (DE, QE or DS) independently to enhance the DROs retrieval has helped to increase the performance of the retrieval. It is assumed that IR components can enhance the performance of the DRO retrieval. Based on this assumption, an information retrieval framework (IRF) for DROs is presented in this paper. The proposed IRF is to address the retrieval problems in DROs and provide an environment for retrieving information from DROs with the highest possible performance. The principle task of IRF is to make all components of IR (DE, DS, and QE) work together to achieve the greatest benefit in improving the retrieval performance. Several experiments were conducted on CHiC2013 which is a collection on cultural heritage. The results show a considerable enhancement over other IR approaches that use the DE method, DS model and QE method independently.

Key words: Digital resource objects, Dirichlet smoothing model, Expansion methods.

1. INTRODUCTION

Digital resource objects **(**DRO) refers to information that are structured which elaborates, describes and eases the retrieval, usage and management of information resources [1]. Apart from the content storage, DROs offer platforms to seek, retrieve and organise contents from databases. Recently, the need for enriching and accessing DROs has been addressed differently by the information retrieval (IR) research communities. IR is composed of two expansion methods and many retrieval models. The two expansion methods refer to the document expansion (DE) method [2] and query expansion (QE) method [3] while the Dirichlet smoothing (DS) model is for the retrieval model. The DE method for IR incorporates modification of documents found in the collection by embedding extra terms into the documents. The additional terms enhance the description of document contents, thus easing retrieval of noisy and short documents. A document that is short may not be adequate compared to a long document. Therefore, external resources are needed to provide vocabulary terms information so that additional high quality data can be produced for enlargement of document sample [4].

QE refers to the process that updates the original query with additional terms [5], [6], phrases [7] and sentences [8]. It is a promising way to enhance retrieval efficiency in IR [9]-[12]. Basically, in QE, a query is expanded by including associated terms based on the query originally submitted by the user [13].

In the IR literature, there are different retrieval models based on varying notions of the relevancy of a document to a query. Historically, some models are important. However, nowadays, ranked retrieval is the most common form of IR. A query is regarded as a set of keywords which is unordered ("bag of words") [14]. Similarity measure between the query and each document is calculated by the IR system using the statistics of the distribution of terms in the documents and across the entire collection. Next, documents are returned in decreasing order of similarity score [15]. Similarity calculation in various models is calculated in many different ways. The three highly popular and successful families of models [16] are vector space, probabilistic and language models. This paper emphasises on the language model (LM) as some successful studies [17]-[20], have proven that the LM approaches are very effective probabilistic framework for IR. Bennett, Scholer and Uitdenbogerd [17] reported that LM outperformed other IR models.

Various studies that focused on the DRO retrieval performance by enhancing the DRO's contents utilize the DE method while some other studies are concerned with the retrieval model especially LM. Many studies tried to improve the DRO retrieval performance by enhancing the user's query. Most previous studies have shown that employing the IR components (DE, DS or QE) independently has helped to increase the retrieval performance of DROs. Based on this fact, the IR framework (IRF) for DROs is proposed in this

paper. The idea of IRF is to build a complete IRF that serves the DRO retrieval. The proposed IRF consists of three main stages: document expansion (DE), Dirichlet smoothing (DS) model and query expansion (QE). The aim of the proposed IRF is to increase the performance of the DRO retrieval by improving the quality of the content of the retrieved documents, improving the retrieval model and improving the query. Based on the literature, there is no existing integrated IRF for DRO retrieval. Therefore, the proposed IRF aims to improve the retrieval performance of DROs, and it comprises all IR components (DE, DS and QE) rather than the improvement of each individual IR component.

The rest of the paper is consists of Section 2 in which the related work is discussed, Section 3 which presents the IRF proposed in this paper, and Section 4 which presents the results of the experiments and discussions on the results. Finally, the conclusion of the work is given in Section 5.

2. RELATED WORK

In overcoming short document problem, the DE method is found to be highly effective in giving satisfying or convincing response during retrieval as proven by many researchers. Some studies have solved the DRO retrieval issues by applying the DE method. Kando and Adachi [18] recommended that in expanding the metadata content, topic is used instead of the title of the document. Min et al. [19] proposed a DE method which regards text-based image as a short document with a metadata unit that describes its content to increase the retrieval performance. Their proposed method employs DBpedia dataset to expand the metadata content. In their method, a query is formed by selecting a few informative terms from the document, and to obtain appropriate terms, it is then sent to DBpedia dataset. Liang, Ren and De Rijke [20] proposed a semantic DE-based fusion method to seek micro-blog posts to enrich the contents. The proposed method makes use of the outcome pool contributed by micro-blog search algorithms to determine the semantic elements for every post in the list to be fused. It also incorporates central sentences from articles found in Wikipedia that are linked with tweet, and lastly, to improve the performance of the fusion, it uses the resultant clusters retrieved from the expanded micro-blog based on the cluster. In Mizzaro et al. [21] a method that uses information extracted from the web derived from the same temporal context was proposed. In this method, Wikipedia which acts as an external resource is used to query the words.

Many DS models have been reported in the literature for IR. Kurland and Krikon [22] proposed a novel LM approach which ranks query-specific clusters via presumed percentage of relevant documents in the clusters. Two types of information are integrated in the proposed model. They are produced based on the clusters and their associated documents. The estimated similarity to the query is the first type of information while the centrality of a document or cluster is the second. The experiments involved three text datasets: AP, TREC8 and WT10G. The experimental outcomes show that the proposed model is substantially more effective in determining clusters with a high percentage of relevant documents compared to other existing methods. Zou et al.[23] extended LMs for IR by adding concepts induced from the query as well as terms from the document. In their model, query terms are provided by combining the document probability to produce concept incorporated from the query with conventional LM probability. Concepts were expressed by word-embedding space. Weighted cosine distance is used to predict similarity of two vectors in the space that enhances the discrimination between vectors. The experiments assessed the TREC collection, and from the result, their model performs better than other existing models.

The problem of accessing DRO has been addressed by many researchers by using and developing the traditional QE method to improve the retrieval of the content of this collection. ALMasri et al. [24] solved the problem of term mismatch by proposing a semantic model which exploits semantic correlations between indexing terms. The model modifies documents based on the query and semantic term relations. The document is extended based on query terms that are absent from the document but semantically related with one term in the document at the very least. Next, two smoothing methods i.e. Dirichlet and Jelinek-Mercer (JM) from LMs, are integrated with the modified document. The experiment was carried out on varied CLEF corpora which are from the medical domain. The experimental results exhibit substantial improvement over traditional LMs and seem to be better than the translation models. Akasereh [25] described various retrieval approaches and looked at the relative merits of techniques for QE and semantic enrichment. A range of strategies was applied based on blind-QE using Wikipedia as the external resource. Different strategies were employed to provide the best concepts for incorporation into the user's query. A CH content test collection was used in the experiment. The experimental outcomes show that the retrieval performance is increased by expanding both pseudo-relevant documents and external resources, and external resources can be used for a more significant query expansion. In AlMasri et al. [26], a semantic enrichment of a query that incorporates term links into the language model using Wikipedia as an external resource is proposed. It deals with short queries which could not provide specific information need. The method aims to find the best terms given a query to enrich the topic semantically and predict the information need or the user's intent based on the original query. A CH content test collection was used in the experiment on this method. The experimental results show that the results of applying Porter stemming method on the term links are not much different since the difference between the two results is very small. In the semantic enrichment method, the results show that links out is better than mixing links in and out.

3. PROPOSED IRF FOR DRO

The proposed IRF involves three stages namely: DE, DS and QE. The principle task of IRF is to make all components of IR (DE, DS, and QE) work together to achieve the greatest benefit in improving the retrieval performance. This IRF

addresses the DRO retrieval issues and provides an environment for retrieving information with the highest possible performance. Each IRF stage will be discussed in detail in the next subsection. Figure 1 shows the stages of the proposed IRF.

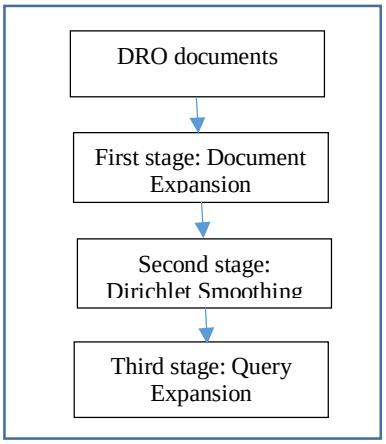

Figure 1: Proposed IR framework for DRO documents

3.1 First stage

This stage is concerned with addressing the shortage of content problem in DROs. The DE method adopted in the proposed IRF is the procedure by Akasereh [25] which introduced the title query to enrich the metadata content. Wikipedia is used as an external resource to provide extra information which is the title of articles retrieved from Wikipedia which are among the top ten. The titles are then put as a term under the tag <WikiTitle>. The pseudocode of the DE method employed in the proposed IRF is shown in Algorithm 1.

Algorithm 1: Document Expansion (DE) method [25]

Input: Document D, Wikipedia
For all D do {
For all metadata unit MD do {
Formulate query Q by taking the title terms T_t :

$$Q = T_t$$

Send Q to Wikipedia
Return articles
}
From top 10 articles do {
Select the title of articles t

$$T = \{t_1, t_2, t_3, \ldots t_n\}$$

Return T
}
Add T to original MD content under new tag <WikiTitle>
Output: expanded MD

3.2 Second stage

The aim of this stage is to apply DS model as a retrieval model to improve the matching between the queries with DRO documents. DS model is a model that is used to re-calculate the zero probability for the unseen terms by giving them small values derived from the probabilistic values of the seen terms. LM model is made up of two parts namely Query Likelihood Estimation Model (QLEM) and Dirichlet Prior (DP). QLEM is the basic approach for estimating documents using probabilistic language where q_i is a query term in query Q, and D is the document. Query terms are independent and unigram language model is employed. So, the probability of Q given D $p(Q|D)$ is calculated by applying the Bayesian theorem that is the product of the probabilities of q_i in D. The probability estimates is written as

$$p(Q|D) = \prod_{i=1}^{n} p(q_i|D) \quad (1)$$

$$p(q_i|D) = \frac{f_{qi}, D}{|D|} \quad (2)$$

where

f_{qi}, D: Number of occurrences of query term q_i in document D

$|D|$: Length of the document

$p(q_i|D)$ will be zero if q_i is missing in D. Missing 1 out of 5 query terms is the same as missing 3 out of 5. Zero probability on single query term leads to a zero probability on the all queries and causes the reduction of retrieved documents belonging to the query. Since the individual document sets lead to more unseen terms in the held-out data, the search process is moved to the whole collection to find the probability of the unseen terms that are not found in each of the documents. To estimate how much probability is needed to shift for the unseen terms, the total frequency of terms that occur only once is used. We can write the equation for the probability estimation as

$$p(q_i|D) = \frac{f_{qi}, D + p(qi|c)}{|D|} \quad (3)$$

where

$p(qi|c)$: Probability of query qi given collection c

$$p(qi|c) = \frac{f_{q_i}, c}{|c|} \quad (4)$$

f_{qi}, c: Number of occurrences of term q_i in collection c

$|c|$: Length of the collection

The contribution of the probability that comes from the

collection leads to sharp changes in probabilities and the probability mass moving too much from the seen terms to the unseen terms. So, the probability needs to be redistributed. Multinomial distribution can be used to summarize collection-wide information. DP is applied between the document and collection to minimize the probability of the unseen terms by utilizing it to estimate $p(Q|D)$. It uses α parameter that depends on the length of the document. α can be written as

$$\alpha = \frac{\mu}{|D| + \mu} \qquad (5)$$

where

μ: Smoothing prior value

$|D|$: Length of the document

Combining QLEM and DP forms the DS model. The model makes smoothing depend on the size of the document because longer documents allow estimation of the terms' probabilities more effectively. We can write the DS model [27] as

$$p(q_i|D) = \frac{f_{qi}, D + \mu\, p(qi|c)}{|D| + \mu} \qquad (6)$$

3.3 Third Stage

This section illustrates the QE method employed in Stage 3 of the proposed IRF. Since this stage focuses primarily on solving the problem of short query in DROs semantically, the main steps used in the QE method of the proposed IRF are as defined in the work of ALMasri, Berrut and Chevallet [24] and Almasri [28] which are

i. All Wikipedia articles related to the query are collected.
ii. By considering the title of each article as a term, the probability is estimated between the entire query and the title of each article.
iii. Next, all titles with a probability greater than zero are grouped into one set called the enrichment set.
iv. The similarity between each query term is calculated within the enrichment set.

4. EXPERIMENTS AND RESULTS

A few experiments were performed to benchmark the proposed IRF against other approaches. Traditional IR systems were developed and the CHiC2013 collection was used. All the documents need to be preprocessed before tokenization is performed so that it is easier to perform the experiment. In the beginning, all the stopwords need to be removed based on a set of comprising 571 words, both from the documents and queries. Stopwords include preposition and conjunctions e.g., "a", "the" and "him" which could not fulfil any needs from the user's point of view. Next, all non-characters in both the documents and queries such as "#" and "$" are removed. Finally, Porter algorithm which is a stemming algorithm is applied both on the documents and queries so that words which are not root words are converted into a root word such as from "plays" to "play". By doing so, the number of random variables referred to by the query and document is reduced. After this stage, the language model is applied as a retrieval model.

Two standard measures are used. The first is Mean Average Precision (MAP) and the second is precision of top ten documents (p@10) [29]. To check whether the differences between performances of the framework and the benchmarks are statistically significant, two-tailed paired t-test was used. English Wikipedia was used as an external resource to generate related articles. The benchmarks used were the CHiC2013 collection using the DE method, the CHiC2013 collection with no expansion and the CHiC2013 collection using query expansion. For retrieval, we used Dirichlet smoothing model (DS) [30] (LM). Table 1 and Table 2 present the statistics of the test collection and the LM setting respectively.

Table 1: Statistics of the test collection

Parameter Name	Value
Number of documents	1107
Number of testing queries	22
Average number of query terms	1.6

Table 2: Language model setting

Models	LM model setting
N-gram	Unigram model
Smoothing model	Dirichlet smoothing model
Smoothing parameter	μ =2000
Estimation model	Query likelihood retrieval $p(Q\|D) = \prod_{i=1}^{n} p(q_i\|D)$

The experiment were performed to discover the impact of gathering the two expansion methods and language model in a single IRF. CHiC2013_DE, CHiC2013_QE, CHiC2013_DS and CHiC2013_IRF collections were used in the experiment, and they are expanded by the DE method, enhanced by the QE method, retrieved by the EDS model and expanded by the the proposed framework respectively. The results presented in Table 3. From the table, it can be seen that MAP improves by 20.2% for CHiC2013_IRF compared with CHiC2013_ED, and MAP improves by 22.8% and 15.2% for CHiC2013_IRF compared with CHiC2013_DS and CHiC2013_QE respectively.

Furthermore, it is necessary to highlight that CHiC2013_IRF improves by 14.8%, 17.4% and 9.8% based on P@10 compared with CHiC2013_DE, CHiC2013_DS and CHiC2013_QE respectively. Additionally, the Precision-Recall curves for CHiC2013_IRF, CHiC2013_DE, CHiC2013_DS and CHiC2013_QE are given in Figure 2. It

can be seen that the precision at different recall points of CHiC2013_IRF is higher compared with those for CHiC2013_IRF, CHiC2013_DE, CHiC2013_DS and CHiC2013_QE. It can be seen that CHiC2013_IRF improves the retrieval results. Therefore, it can be concluded that putting DE, DS and QE in a single IR framework is better than performing each method and model individually in an effort to improve the effectiveness of the DRO retrieval.

Furthermore, the collaboration among three IR components of the proposed framework creates more chances of producing query based on the documents as well as enhancing the user query.

Table 3: Benchmarking results

Approach	MAP	P@10
CHiC2013_DE	0.533	0.438
CHiC2013_DS	0.501	0.412
CHiC2013_QE	0.54	0.488
CHiC2013_IRF	0.64	0.586
Improvement (IRF, CHiC2013_DE)	20.2%	14.8%
Improvement (IRF, CHiC2013_DS)	22.8%	17.4%
Improvement (IRF, CHiC2013_QE)	15.2%	9.8%

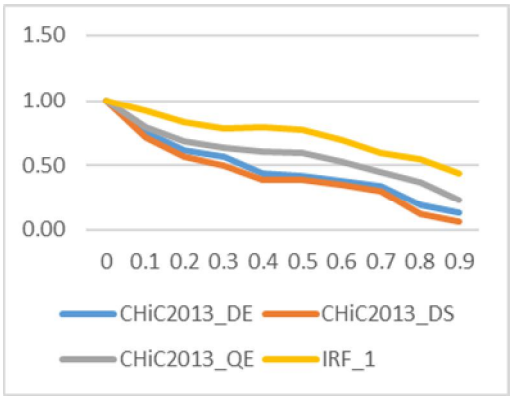

Figure 2: Comparison of the performances based on averaged 9-point precision-recall curve

5. CONCLUSION AND FUTURE WORK

An IRF for increasing the DRO retrieval performance has been presented. The proposed IRF is able to solve two fundamental issues, namely lack of quality of metadata content, and difficulty in accessing metadata content that lead to a decrease in the effectiveness of its retrieval. The principle task of IRF is to make all components of IR (DE, DS, and QE) work together to achieve the greatest benefit in improving the retrieval performance. The results show that an improvement is achieved significantly by the proposed IRF compared to other benchmarked IR approaches that use document expansion method, language model and query expansion method independently. Furthermore, the collaboration among the three IR components in the proposed framework increases the chances of producing query from DROs and enhances the query. For future work, it is possible to work on enhancing both the expansion methods (DE and QE) and the DS model in the proposed framework.

ACKNOWLEDGEMENT

We would like to thank Universiti Sains Malaysia for supporting this research under the Research University Grant - Account No. 1001/PKOMP/8014075.

REFERENCES

1. I. H. Witten, D. Bainbridge, G. Paynter and S. Boddie. **Importing documents and metadata into digital libraries: requirements analysis and an extensible architecture**, in *Research and Advanced Technology for Digital Libraries*. 2002, Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 390-405. https://doi.org/10.1007/3-540-45747-X_29
2. A. Singhal and F. Pereira, **Document expansion for speech retrieval**, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 1999, ACM, pp. 34-41. https://doi.org/10.1145/312624.312645
3. C. Carpineto and G. Romano, **A survey of automatic query expansion in information retrieval,** *ACM Computing Surveys*, 2012, Vol. 44(1), pp. 1. https://doi.org/10.1145/2071389.2071390
4. K. Darwish and D.W. Oard, **Adapting morphology for arabic information retrieval,** in *Arabic Computational Morphology*, 2007, Springer. pp. 245-262. https://doi.org/10.1007/978-1-4020-6046-5_13
5. J. Leveling and G.J. Jones, **Classifying and filtering blind feedback terms to improve information retrieval effectiveness,** in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2010, Le Centre de Hautes Etudes Internationales D'Informatique Documentaire, pp. 156-163.
6. G. Cao, J.-Y. Nie, J. Gao and S. Robertson, **Selecting good expansion terms for pseudo-relevance feedback**, in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, ACM, pp. 243-250. https://doi.org/10.1145/1390334.1390377
7. P. Mahdabi, L. Andersson, M. Keikha and F. Crestani, **Automatic refinement of patent queries using concept importance predictors,** i*n Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, ACM, pp. 505-514. https://doi.org/10.1145/2348283.2348353
8. D. Ganguly, J. Leveling and G.J. Jones. **Query expansion for language modeling using sentence**

similarities, in *Information Retrieval Facility Conference*, 2011, Springer, pp. 62-77. https://doi.org/10.1007/978-3-642-21353-3_6

9. C. Carpineto, G. Romano and V. Giannini, **Improving retrieval feedback with multiple term-ranking function combination**, *ACM Transactions on Information Systems*, 2002, Vol. 20(3), pp. 259-290. https://doi.org/10.1145/568727.568728
10. J. Bai, D. Song, P. Bruza, J.-Y. Nie, and G. Cao. **Query expansion using term relationships in language models for information retrieval.** in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, ACM, pp. 688-695. https://doi.org/10.1145/1099554.1099725
11. K. S. Lee, , W. B. Croft and J. Allan. **A cluster-based resampling method for pseudo-relevance feedback**, in *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, ACM, pp. 235-242.
12. S. Shekarpour, E. Marx, S. Auer and A.P. Sheth. **RQUERY: Rewriting Natural Language Queries on Knowledge Graphs to Alleviate the Vocabulary Mismatch Problem**, in *AAAI*. 2017. pp. 3936-3943.
13. B. He and I. Ounis. **Studying query expansion effectiveness,** in *European Conference on Information Retrieval*. 2009. Springer. pp. 611-619. https://doi.org/10.1007/978-3-642-00958-7_57
14. D. M. Christopher, R. Prabhakar, and S. Hinrich, **Introduction to information retrieval**, *An Introduction To Information Retrieval*, 2008, Vol. 151(177), pp. 5.
15. S. Büttcher, C. L. Clarke and G. V. Cormack, ***Information Retrieval: Implementing and Evaluating Search Engines,*** 2016, MIT Press.
16. J. Sanz-Cruzado, S. M. Pepa and P. Castells, **Recommending Contacts in Social Networks Using Information Retrieval Models**, in *Proceedings of the 5th Spanish Conference on Information Retrieval*, 2018, ACM, pp. 19. https://doi.org/10.1145/3230599.3230619
17. G. Bennett, F. Scholer and A. Uitdenbogerd. **A comparative study of probabilistic and language models for information retrieval**, in *Proceedings of the Nineteenth Conference on Australasian Database, Vol. 75*, 2008, Australian Computer Society Inc., pp. 65-74.
18. N. Kando and J. Adachi. **Cultural Heritage Online: Information Access across Heterogeneous Cultural Heritage in Japan**, in *Electronic Proceedings of International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society (DLKC 2004)*, 2004, pp.
19. J. Min, J. Leveling, D. Zhou and G.J.F. Jones, **Document expansion for image retrieval**, in *Adaptivity, Personalization and Fusion of Heterogeneous Information*. 2010, Le Centre de Hautes Etudes Internationales D'Informatique Documentaire: Paris, France, pp. 65-71.
20. S. Liang, Z. Ren and M. De Rijke. **The impact of semantic document expansion on cluster-based fusion for microblog search**, in *European Conference on Information Retrieval*, 2014, Springer, pp. 493-499. https://doi.org/10.1007/978-3-319-06028-6_47
21. S. Mizzaro, M. Pavan, I. Scagnetto and M. Valenti. **Short text categorization exploiting contextual enrichment and external knowledge**, in *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, 2014, ACM, pp. 57-62. https://doi.org/10.1145/2632188.2632205
22. O. Kurland and E. Krikon, **The opposite of smoothing: a language model approach to ranking query-specific document clusters**, *Journal of Artificial Intelligence Research*, 2011, Vol. 41, pp. 367-395. https://doi.org/10.1613/jair.3327
23. B. Zou *et al.*, **A Concept Language Model for Ad-hoc Retrieval**, in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, International World Wide Web Conferences Steering Committee, pp. 885-886. https://doi.org/10.1145/3041021.3054209
24. M. ALMasri, C. Berrut and J.-P. Chevallet, **Wikipedia-based semantic query enrichment**, in *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information retrieval*, 2013, ACM, pp. 5-8. https://doi.org/10.1145/2513204.2513209
25. M. Akasereh, **A quantitative evaluation of query expansion in domain specific information retrieval,** *The American Society for Information Science and Technology*, 2013, Vol. 50(1), pp. 1-7. https://doi.org/10.1002/meet.14505001046
26. M. ALMasri, K. Tan, C. Berrut, J.-P. Chevallet, and P. Mulhem, **Integrating semantic term relations into information retrieval systems based on language models,** in *Asia Information Retrieval Symposium*, 2014, Springer, pp. 136-147. https://doi.org/10.1007/978-3-319-12844-3_12
27. C. Zhai, **Risk minimization and language modeling in text retrieval**, PhD Dissertation, Carnegie Mellon University, 2002. https://doi.org/10.1145/792550.792571
28. AlMasri M., **Semantic query structuring to enhance precision of an information retrieval system: application to the medical domain**, in *CORIA*, 2013, pp. 293-298.
29. C. D. Manning, P. Raghavan, and H. Schutze, ***Introduction to Information Retrieval, Cambridge University Pres,*** *2008*, ch. 20: pp. 405-416. https://doi.org/10.1017/CBO9780511809071
30. C. Zhai and J. Lafferty, **Model-based feedback in the language modeling approach to information retrieval**, in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 2001, ACM, pp. 403-410. https://doi.org/10.1145/502585.502654