Proceedings of the Federated Conference on Computer Science and Information Systems pp. 161–164
DOI: 10.15439/2018F260
ISSN 2300-5963 ACSIS, Vol. 15

A New Subject-based Document Retrieval from Digital Libraries Using Vector Space Model

Sayed Mahmood Bakhshayesh*, Azadeh Mohebi†, Abbas Ahmadi‡, and Amir Badamchi§
*Amirkabir University of Technology, Tehran, Iran
Email: s.ma.bakhshayesh@aut.ac.ir
†Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran
Email: mohebi@irandoc.ac.ir
‡Amirkabir University of Technology, Tehran, Iran
Email: abbas.ahmadi@aut.ac.ir
§Amirkabir University of Technology, Tehran, Iran
Email: badamchi@aut.ac.ir

***Abstract*—Document retrieval from digital libraries based on user's query is highly affected by the terms appeared in the query. In many cases, there are some documents in the digital libraries that do not share exactly the same terms with the query, but they are related to the user's need. We address this challenge in this paper by introducing a new subject-based retrieval approach in which, apart from ranking documents based on the terms in the query, a new subject-based scoring scheme is defined between the query and a document. We define this score by introducing a new vector space model in which a vectorized subject-based representation is defined for each document and its keywords, and the terms in the query, as well. We have tested the new subject-based scoring scheme on a database of scientific papers obtained from Web of Science. Our Experimental results show that in 83% of times users prefer the proposed scoring scheme with respect to the classic scoring ones.**

I. Introduction

NOWADAYS a considerable amount of information is spread over billions of various documents saved in digital libraries. Although, various retrieval tools and algorithms have been developed to address accessing such information easily, in many cases these algorithms and tools are limited by the user's query. Many of the retrieval methods try to go beyond the exact terms in user's query. In other words, instead of only relying on Bag-of-Words (BoW) representation of the query, new approaches have been developed such as interactive query refinement, relevance feedback from user, word sense disambiguation, and clustering search results [1], [2], [3], [4], [5] to guid the user in his/her *journey of information retrieval*. More specifically, some of these methods are based on improving user involvement (implicitly or explicitly) in the retrieval process by receiving relevance feedback or providing interactive search tools. Some other methods rely on expanding user's query using query expansion techniques [6].

In this paper, we address these challenges by introducing a new *subject-based* document retrieval approach. Instead of applying query expansion techniques or using semantic relations between words and terms based on ontologies, we introduce a new subject-based representation for each document in the digital library, using vector space model. By the use of the proposed approach, we can measure how much a document and a given query are similar and share same subjects, even if they do not share same terms. The proposed approach is applicable in specialized, scientific digital libraries in which in addition to a set of keywords/tags usually assigned to each document, a set of predefined disciplines/subjects are also available and each document usually falls into a specific discipline, subject or category. In such specialized libraries, documents are usually indexed based on subjects and keywords assigned to them, to improve the indexing, retrieval and archiving tasks.

In the proposed method, first, a new subject-based vectorized representation for each keyword is introduced by relying on the knowledge obtained from all documents that have been already indexed in the digital library. Then, a probabilistic, vectorized subject-based representation for each document is estimated. Each element of this vector shows how much each document belongs to a specific subject. This consideration is based on the assumption that each document might belongs to more than one subject/category. This is a valid assumption that is usually considered in well-known retrieval/indexing approach such as topic modeling. Then we use these vectors in order to calculate subject-based similarity between a given query and documents.

After a brief literature review in section two, we describe our method in details in section three. Then, in section four, a series of experiments are presented to show the effectiveness of our approach, and The experimental results are analyzed. Finally, we present our conclusions in section five.

II. Literature Review

Different types of research have been done in order to improve the performance of retrieval algorithms, by considering semantic relationship between the query and documents. Tai et al. used supervised learning to improve vector space information retrieval model [7] by using matrices with 1s and 0s to show the relevance of queries and documents. Hofmann presented a statistical model based on Latent Semantic Analysis (LSA) leading to probabilistic latent Semantic Analysis

(PLSA) [8]. Maitah et al. investigated the use of an adaptive algorithm under vector space model, extended Boolean model, and language model in information retrieval [9]. Wang et al. presented a new document retrieval framework that learns a probabilistic knowledge model for improving document retrieval [10]. The model was represented by a network of association among concepts defining key domain entities and is extracted from a corpus of documents or from a domain knowledge base. Campos et al. proposed a probabilistic model based on Bayesian network for document retrieval [11] and used the network to compute posterior probabilities for the relevance of the documents. Mohebi et al. proposed a new subject-based retrieval method to retrieve all documents from a scientific digital library related to that subject. Their proposed method does not rely on user's query, rather the user specifies a specific topic or subject, and all related scientific documents related to this subject are retrieved [12]. Siddiqui proposed a hybrid IR model with two stages: first, the document collection is downsized using vector model based on a given query, second a conceptual graph based representation is used to rank the documents [13].

Sometimes retrieving the relevant text is hard because the query and the document may use different vocabularies. Mitra et al. trained a word2vec embedding model to improve the ranking of retrieved documents. In their model they map the query words into the input space and the document words into output space, and compute a relevance score by aggregating the cosine similarities across all word pairs [14].

Relevant document may be clustered together with other relevant items that may not contain query terms and could be retrieved through a clustered search [15].

Most of the methods in the literature rely completely or partially on the terms presented in the user's query. However, when a document does not contain any of the terms in the query, but is related to the query, then that document has a low chance to appear in the top retrieved documents. We address this challenge in this research by proposing a new method based on Vector Space Model. In this model a new subject-based representation for each document and the query is defined, that is independent of the query terms. Subject-based mapping of all documents in this method is a pre-processing activity that should be done once for all documents in the data-base.

III. Proposed subject-based scoring scheme

The proposed scoring scheme can be applied on a basic retrieval model such as BM25, in order to re-order the ranking of a set of retrieved documents. The proposed scheme calculates a new subject-based distance between a document and a query. This distance is a semantic-based one which calculates the relationship between a query and a given document apart from their joint terms. For this purpose, we assume that $\mathcal{D}$ is the document collection, with N documents, while every document has a set of keywords and a set of subjects associated with it. We aggregate all subjects and all keywords of all documents in set $\mathcal{S}$ and $\mathcal{K}$, respectively, i.e.:

$$\mathcal{D} = \{d_1, \ldots, d_N\}, \mathcal{S} = \{s_1, \ldots, s_M\}, \mathcal{K} = \{k_1, \ldots, k_L\}. \quad (1)$$

Our ultimate goal is to define a vector space model in order to represent each document as a subject-based vector. Consequently, the subject-based vector for each document can be compared with the subject-based vector of a given query to compute their relationship. In order to do so, we rely on the keywords for each document. In other words, we introduce a method to represent each keyword as a subject-based vector, with the size of M, to reflect how much the keyword is related to every subject. For a keyword k_l, this vector is defined as:

$$\mathbf{vk}_l = \Big(p_l(s_1),\ p_l(s_2), \ldots,\ p_l(s_M)\Big), \quad (2)$$

where $p_l(s_m)$ shows how much keyword k_l is related to subject s_m. In other words, $p_l(s_m)$ can be considered as the conditional probability that a given keyword belongs to a specific subject, defined by:

$$p_l(s_m) = P(s_m|k_l) = \frac{P(s_m, k_l)}{P(k_l)}. \quad (3)$$

We estimate this probability based on the data available in $\mathcal{D}$, as follows:

$$\hat{P}(s_m|k_l) = \frac{\sum_{d_i \in D_l} ds^i_m}{|D_l|}, \quad (4)$$

where D_l is the set of all documents with keyword k_l, and ds^i_m denotes the number of documents in D_l containing subject s_m. Finally, for a document d_i with L_i keywords, we represent d_i as a $L_i \times M$ matrix (X_i) where each row corresponds to each keyword of d_i and each column corresponds to a subject. For the sake of simplicity, we assume that $k_1, k_2, \ldots, k_{L_i}$ are keywords of d_i, then we have:

$$\mathbf{X_i} = \begin{bmatrix} \mathbf{vk}_1 \\ \vdots \\ \mathbf{vk}_{L_i} \end{bmatrix} = \begin{bmatrix} p_1(s_1) & p_1(s_2) & \ldots & p_1(s_M) \\ \vdots & & & \\ p_{L_i}(s_1) & p_{L_i}(s_2) & \ldots & p_{L_i}(s_M) \end{bmatrix}. \quad (5)$$

Now we can define a subject-based representation vector for every document, based on the matrix in 5. We call this vector $\mathbf{vs_l}$. Each component in this vector corresponds to a subject in $\mathcal{S}$, showing how much the document is related to that subject. Thus, each document d_i is mapped to a subject-based vector:

$$\mathbf{vs_i} = \frac{\sum_{l=1}^{L_i} \mathbf{vk_l}}{L_i} = \Big(\frac{\sum_{l=1}^{L_i} p_l(s_1)}{L_i}, \ldots, \frac{\sum_{l=1}^{L_i} p_l(s_M)}{L_i}\Big). \quad (6)$$

Every query Q, can also be mapped to a subject-based, M-sized, vector too. For this purpose, the query is processed first in order to extract its distinguished terms, i.e. $q_1, q_2, \ldots, q_r$. Thus, we have:

$$\mathbf{vs_q} = \frac{\sum_{l=1}^{r} \mathbf{vk_l}}{r} = \Big(\frac{\sum_{l=1}^{r} p_l(s_1)}{r}, \ldots, \frac{\sum_{l=1}^{r} p_l(s_M)}{r}\Big). \quad (7)$$

Based on the subject-based vectors for d_i and Q, a new subject-based scoring function is defined:

$$Score_{subject}(d_i, Q) = \frac{1}{\|\mathbf{vs_q} - \mathbf{vs_i}\|}. \quad (8)$$

A. Final combined scoring scheme

The proposed subject-based scoring scheme can be combined with different basic retrieval scoring schemes such as Okapi BM25 which is based on the probabilistic retrieval framework and ranks a set of documents based on the query terms appearing in each document. Given a query Q with r distinct terms, BM25 score is:

$$Score_{BM25}(d, Q) = \sum_{j=1}^{r} IDF(q_m) \frac{freq(q, d)(c+1)}{freq(q, d) + c(1 - b + b\frac{|d_j|}{avgdl})}, \quad (9)$$

where q_i is i-th term of query, $freq(q_i, d)$ is term frequency of q_i in document d, $|d|$ is the length of d in words and $avgdl$ is the average document length in the whole collection. Parameters c and b are usually chosen as $c \in [1.2, 2.0]$ and $b = 0.75$. $IDF(q_i)$ is the inverse document frequency (IDF) weight of the query term q_i and is usually calculated as:

$$IDF(q_i) = \log \frac{(N - n(q_i) + 0.5)}{(n(q_i) + 0.5)}, \quad (10)$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing term q_i. Now, we can define a combined scoring scheme based on the subject-based and BM25 scores:

$$Score_{final}(d, Q) = \alpha\ Score_{subject}(d, Q) + (1 - \alpha)\ Score_{BM25}(d, Q), \quad (11)$$

where $\alpha \in [0, 1]$ is a weighing parameter that need to be tuned. This score is applied on a set of documents retrieved based on a basic model such as BM25, in order to represent a new ranking for the retrieved documents.

IV. Experiments and Results

In order to examine the proposed method, we have considered a collection of scientific documents (articles) extracted from Web of Science (WoS) which contains all papers published from Iran in years $2013-2017$. The collection contains 98497 documents. Each document has a title, abstract, author, keywords and subjects. The subjects are assigned for each document by WoS, based on a list of predetermined categories in WoS. The collection contains 340836 keywords and 1200 different subjects. Two domain experts have classified 1200 subjects to eight main subjects including Art, Biosciences and Natural Sciences, Basic Sciences, Empirical Sciences, Humanities Sciences, Medicine and Treatment, Engineering.

In our experiment, we choose the top 100 documents for our query. Then, the top selected documents are ranked again based on $Score_{final}$. The subject-based vectors for keywords and documents in the database are calculated once. Thus, all vectors are calculated offline, and for every query presented to the system, only the corresponding vector for the query is calculated. For instance, given the following query:
"A Robust optimization for the milkrun problem under demand and travel time uncertainty",
the subject-based vector for the query is calculated based on the vectors of each term in the query, after stop word removal, as shown in Table I. $Score_{final}$ is calculated for the selected top documents based on the query $\mathbf{vs_q}$. Table II shows the $Score_{BM25}$, $Score_{subject}$, and $Score_{final}$ for the top retrieved documents, when $\alpha = 0.3$. Thus, we calculated retrieved documents changes as following:

$$Change(\%) = n_q \sum_{i=1}^{n_q} \min |R_i - i|, \quad (12)$$

while R_i is rank of i-th result in BM25 ranking and n_q is number of queries in experiment. In Fig. 1 we show how the ranking changes based on (12) in terms of α. In the proposed scoring scheme, when α is very small, the contribution of subject-based score is small, thus BM25 plays the key role in ranking the results. Alternatively, when α is large, near 1, the subject-based scoring share the most contribution in the final score. However, in a specific range, i.e. when $\alpha \in [0.25, 0.55]$, there is a competition between BM25 ranking and subject-based ranking. In this range, we see the maximum changes in the ranking between these two ranking schemes.

In order to evaluate the proposed approach on users' opinion, we have launched our model on a server and represented users the ranking obtained based on BM25 and proposed approach for a set of queries, while α changes. Then, we have asked the users to choose the best ranking. We have observed that the users prefer more the results based on the proposed scoring scheme than BM25 scoring scheme, when $\alpha = 0.3$.

TABLE I
SUBJECT-BASED VECTOR FOR QUERY: *"Robust optimization for the milkrun problem under demand and travel time uncertainty".*

Query term	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Robust	0	0	0	0	0	0	1	0
Optimization	0	1	0	1	0	0	1	0
Milkrun	0	0	0	0	0	0	0	0
problem	0	0	0	0	0	0	0	0
demand	0	0	0	0	0	0	1	0
travel	0	0	0	0	0	0	1	0
time	0	0	0	0	0	0	0	0
uncertainty	1	1	0	0	1	1	1	0
$\vec{\mathbf{vs_q}}$	0.125	0.25	0	0.125	0.125	0.125	0.625	0

TABLE II
BM25 RANKING VERSUS PROPOSED RANKING FOR TOP 5 RETRIEVED DOCUMENTS WITH $\alpha = 0.3$

BM25 ranking	$Score_{BM25}$	$Score_{final}$	**Proposed ranking**
1	0.628	0.637	1
2	0.508	0.436	2
3	0.358	0.157	**5**
4	0.357	0.193	**3**
5	0.348	0.167	**4**

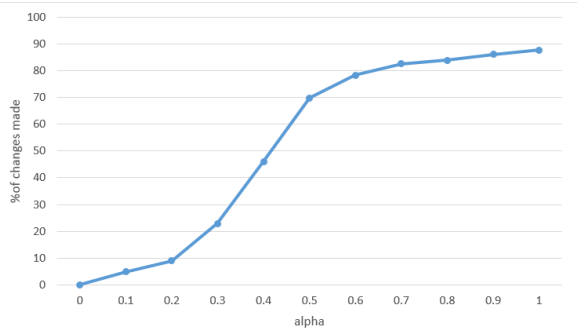

Fig. 1. How much the proposed method is able to change the ranking, when α changes. The vertical axis reflects the ranking difference between proposed method and BM25.

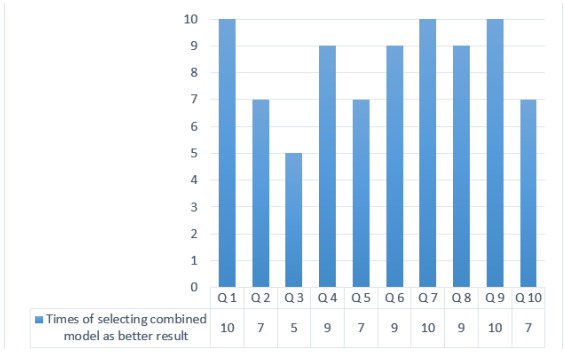

Fig. 2. How much different users prefer the proposed method versus BM25, in 100 experiments

We have also represented 10 users, 10 different queries (users and queries are independent) with both BM25 and proposed ranking scheme and ask them to choose the best results.

Based, on 100 experiments and results, we have obtained that in 83% of times, users preferred the proposed ranking scheme, as shown in Fig. 2.

V. Conclusions

This paper introduced a new vector based model for improving document retrieval, specifically when the documents are from a set of scientific databases, and each contains a set of keywords, and subjects assigned to it. A new scoring scheme is defined in which each document is represented as a vector of subjects. Based on this vector, a new subject-based scoring scheme is defined that can be combined with a basic scoring scheme such as BM25 in order to assign a new score for each document. The new scoring scheme is specifically practical when some terms in the user's query have not been appeared in the database. Thus, rather than retrieving documents based on the exact appearance of the user's term in the database, the proposed approach looks for documents related to the query conceptually, by comparing the subject-based vectorized representation. We have evaluated our proposed scoring scheme to examine how much it is able to change the results effectively, comparing with BM25. In addition we have evaluated the proposed approach based on user's satisfaction, and obtained that in 83% of times the users prefer the proposed scoring scheme than the basic frequency-based scoring scheme. For future research directions, we propose to examine other basic retrieval method rather than BM25, and combine them with the subject-based scoring scheme.

References

[1] S. Momtazi, M. Lease, and D. Klakow, "Effective term weighting for sentence retrieval," in *Research and Advanced Technology for Digital Libraries*, M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, Eds. Springer Berlin Heidelberg, 2010, pp. 482–485. [Online]. Available: https://doi.org/10.1007%2F978-3-642-15464-5_62

[2] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *ACM SIGIR Forum*, vol. 51, no. 2. ACM, 2017, pp. 176–184. [Online]. Available: https://doi.org/10.1145%2F3130348.3130365

[3] S. Acid, L. M. De Campos, J. M. Fernández-Luna, and J. F. Huete, "An information retrieval model based on simple bayesian networks," *International Journal of Intelligent Systems*, vol. 18, no. 2, pp. 251–265, 2003. [Online]. Available: https://doi.org/10.1002%2Fint.10088

[4] J. Zhang, J. Gao, M. Zhou, and J. Wang, "Improving the effectiveness of information retrieval with clustering and fusion," *Computational Linguistics and Chinese Language Processing*, vol. 6, no. 1, pp. 109–125, 2001.

[5] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. ACM, 2006, pp. 178–185. [Online]. Available: https://doi.org/10.1145%2F1148170.1148204

[6] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1:1–1:50, 2012. [Online]. Available: https://doi.org/10.1145%2F2071389.2071390

[7] X. Tai, M. Sasaki, Y. Tanaka, and K. Kita, "Improvement of vector space information retrieval model based on supervised learning," in *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*. ACM, 2000, pp. 69–74. [Online]. Available: https://doi.org/10.1145%2F355214.355224

[8] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57. [Online]. Available: https://doi.org/10.1145%2F3130348.3130370

[9] W. Maitah, M. Al-Rababaa, and G. Kannan, "Improving the effectiveness of information retrieval system using adaptive genetic algorithm," *International Journal of Computer Science & Information Technology*, vol. 5, no. 5, p. 91, 2013. [Online]. Available: https://doi.org/10.5121%2Fijcsit.2013.5506

[10] S. Wang, S. Visweswaran, and M. Hauskrecht, "Document retrieval using a probabilistic knowledge model," in *Internation Conference on Knowledge Discovery and Information retrieval*, 2009. [Online]. Available: https://doi.org/10.5220%2F0002293400260033

[11] L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete, "A layered bayesian network model for document retrieval," in *Advances in Information Retrieval*, F. Crestani, M. Girolami, and C. J. van Rijsbergen, Eds. Springer Berlin Heidelberg, 2002, pp. 169–182. [Online]. Available: https://doi.org/10.1007%2F3-540-45886-7_12

[12] A. Mohebi, M. Sedighi, and Z. Zargaran, "Subject-based retrieval of scientific documents, case study: Retrieval of information technology scientific articles," *Library Review*, vol. 66, no. 6/7, pp. 549–569, 2017. [Online]. Available: https://doi.org/10.1108%2Flr-10-2016-0090

[13] T. Siddiqui and U. Tiwary, "A hybrid model to improve relevance in document retrieval," *Journal of Digital Information Management*, vol. 4, pp. 73 – 81, 2006 2006.

[14] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana, "Improving document ranking with dual word embeddings," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International WWW Conferences Steering Committee, 2016, pp. 83–84. [Online]. Available: https://doi.org/10.1145%2F2872518.2889361

[15] Y. Kural, S. Robertson, and S. Jones, "Clustering information retrieval search outputs," in *Proceedings of the 21st Annual BCS-IRSG Conference on Information Retrieval Research*, ser. IRSG'99. Swindon, UK: BCS Learning & Development Ltd., 1999, pp. 9–9.