Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6789044, 11 pages
https://doi.org/10.1155/2022/6789044

Research Article

Relevance Feedback and Deep Neural Network-Based Semantic Method for Query Expansion

Abhishek Kumar Shukla,[1] **Sujoy Das**,[1] **Pushpendra Kumar**,[2] **and Afroj Alam**[3]

[1]*Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India*
[2]*Department of Computer Science and Technology, Central University of Jharkhand, Ranchi, India*
[3]*Department of Computer Science, Bakhtar University, Kabul, Afghanistan*

Correspondence should be addressed to Afroj Alam; aalam@bakhtar.edu.af

Received 5 May 2022; Accepted 22 June 2022; Published 18 July 2022

Academic Editor: Kuruva Lakshmanna

Machine learning techniques have been widely used in almost every area of arts, science, and technology for the last two decades. Document analysis and query expansion also use machine learning techniques at a broad scale for information retrieval tasks. The state-of-the-art models like the Bo1 model, Bo2 model, KL divergence model, and chi-square model are probabilistic, and they work on DFR-based retrieval models. These models are much focused on term frequency and do not care about the semantic relationship among the terms. The proposed model applies the semantic method to find the semantic similarity among the terms to expand the query. The proposed method uses the relevance feedback method that selects a user-assisted most relevant document from top "*k*" initially retrieved documents and then applies deep neural network technique to select the most informative terms related to original query terms. The results are evaluated at FIRE 2011 ad hoc English test collection. The mean average precision of the proposed method is 0.3568. The proposed method also compares the state-of-the-art models. The proposed model observed 19.77% and 8.05% improvement on the mean average precision (MAP) parameter with respect to the original query and Bo1 model, respectively.

1. Introduction

The most tedious task of the retrieval system is to retrieve the exact documents that are relevant to the user query. The user starts searching documents by putting the keywords in the form of a query. But generally, queries contain very limited terms to express his/her information need. Therefore, it is impossible to retrieve all relevant documents from a large document collection. Query expansion is a technique that selects the most informative terms that are related to a user query and expands the original query. Global and local methods are two major classes of query expansion. The global method selects expansion terms from some external datasets. WordNet [1] and Word2Vec [2] are the two most well-known external datasets that are commonly used to select semantically related query terms for expansion tasks. On other hand, the local method of query expansion selects the expansion terms from initially retrieved documents. The local method is mainly classified into two categories, pseudorelevance feedback method and relevance feedback method. The pseudorelevance feedback method selects the expansion terms from top "*k*" initially retrieved documents. On the other hand relevance, the feedback method selects the expansion terms from the subset of top "*k*" ranked documents that the user thinks are relevant to the user query.

The machine learning [3] technique is broadly used to extract useful information by training a dataset from a large collection of datasets. The machine learning technique has been applied by [4, 5] to identify the liver patients from a liver patient dataset and to find the positive and negative tweets from a social dataset. The neural network technique is widely used in data mining and text mining for classification tasks. The general architecture of a neural network contains an input layer, output layer, and an activation function that works between the input layer and output layer as a

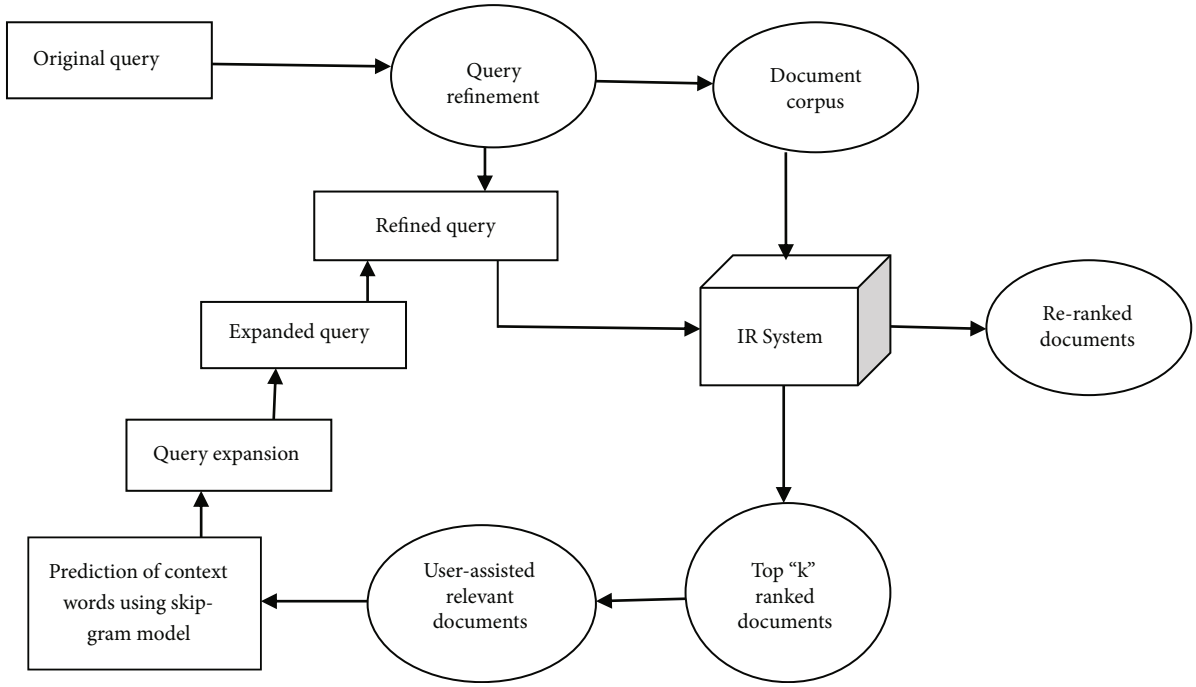

Figure 1: Architecture of the proposed model.

hidden layer. Weights are assigned initially and updated repeatedly unless some target value is not reached. A deep neural network is an extended form of neural network architecture that contains several hidden layers between the input layer and output layer. In a deep neural network, architecture weights are propagated from the input layer to the hidden layer and the hidden layer to the output layer, and vice versa. For document retrieval, there are two neural network-based approaches:

(1) Continuous bag of word model (CBOW)

(2) Skip-gram model

Continuous bag of word model predicts the center word for a given context word. The skip-gram model does the opposite of what the CBOW model does; i.e., it predicts the context word for a given center word.

The proposed model uses skip-gram architecture to train the corpus words. The proposed model initially selects the user-assisted relevant documents from top "k" ranked documents. Here "k" value is set to 30. For each query term, vocabulary terms within user-assisted relevant documents are trained to extract the context term and are then merged. The merged terms are then considered as the most informative terms for a given query and are treated as expansion terms. At the input layer, query terms are represented by one-hot encodings. In one-hot encoding representation, a vector of vocabulary size is initialized that contains all entry zero except the index where the query term appears. The index where the query term appears is initialized to 1. Then, the context words are predicted by successively updating the weight matrix between the input and the hidden layer and the hidden to the output layer. The detailed discussion is presented in Section 3. The proposed model has also been compared with state-of-the-art models [6]. The proposed model architecture is shown in Figure 1.

2. Related Work

Vector space [7] was the oldest model for the retrieval task. Every retrieval model suffers from the most common shortcoming that the keyword submitted by a user is not focused on the main topic and also, the user is unaware of what he/she is looking for [8]. This leads to retrieval of irrelevant documents. Query expansion plays a key role to maximize the retrieval of relevant documents. Maron and Kuhns [9] applied query expansion technique to improve the performance of the retrieval system. Allan [10] used relevance feedback model for selecting the most informative terms for query expansion.

Relevance-based probabilistic language model [11] over the term collection has been developed to select the most informative terms for query expansion. M. Bendersky et al. [12] used external resources and probability language-based modeling for weighting the terms for query expansion task. Miao et al. [13] proposed a proximity-based query expansion model that emphasizes the proximity of terms rather than their positional information. Lv and Zhai [14] presented a positional relevance model that focuses on the position of query terms in the relevance feedback documents. Metzler and Croft [15] used Markov random field-based term dependency model for retrieving the most informative terms for query expansion. Dalton and Dietz [16] proposed relevance feedback-based neighborhood relevance model entity linking across the document and query collections for expanding the query. The expectation-

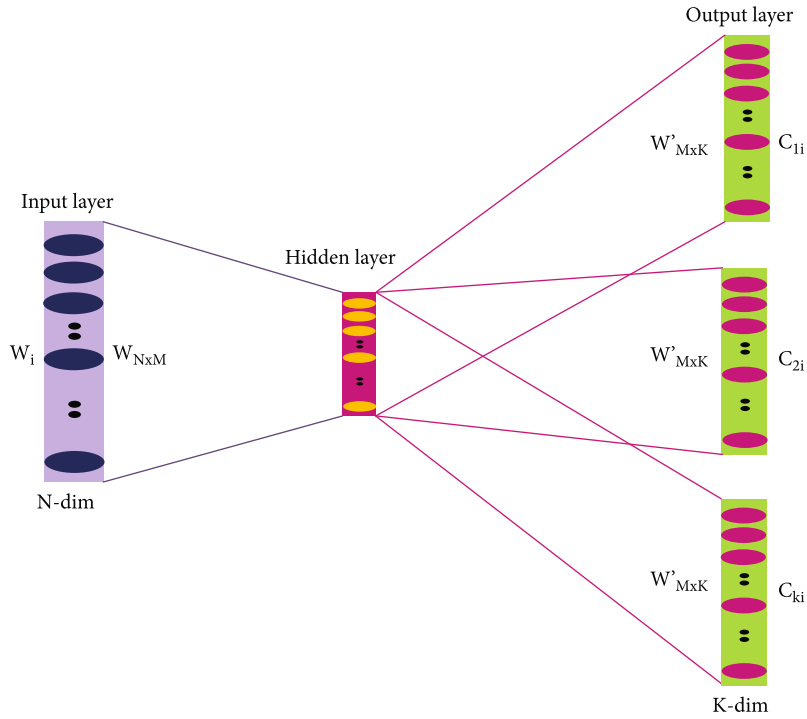

Figure 2: Skip-gram model [34].

maximization algorithm [17] is used to maximize the likelihood of relevant documents from top-ranked documents and then to retrieve the most informative terms from these relevant documents for query expansion tasks. Qiu and Frei [18] expanded the query using query log by calculating the correlation between query terms and document terms. Baeza-Yates and Tiberi [19] proposed a model where search engine query log is represented by a bipartite graph and edge connected between query node and URL node by click through. They observed an improvement of 10% on the mean average precision value. Deerwester et al. [20] proposed singular value decomposition-based Latent Semantic Indexing (LSI) technique to extract the most informative terms for query expansion. Singh and Sharan [21] proposed a fuzzy logic-based technique on top "k" ranked documents for query expansion tasks. Word embedding-based neural network technique using Word2Vec for the query expansion task was proposed by [22]. They extracted similar terms to the original query terms using the K-nearest neighbor method. Kuzi et al. and Mikolov et al. [23, 24] also used the word embedding technique for expanding the original query. Diaz et al. [25] used locally trained word embedding techniques such as Word2Vec or Glove for query expansion tasks. Imani et al. [26] proposed a Word2Vec-based continuous bag of word model for selecting the most informative terms for query expansion. Thesaurus-based query expansion model using Wikipedia was proposed by [27]. Xu et al. [28] used Wikipedia to categorize the query into three different categories broader queries, ambiguous queries, and entity queries. They selected the expansion terms using term distribution and structure of Wikipedia documents. Crouch and Yang [29] built a statistical thesaurus by clustering the whole document collection using a complete link clustering algorithm. Shukla and Das [30] proposed a pseudorelevance feedback and deep neural network-based method to expand the query. A hybrid model-based [31, 32] method was proposed to expand the query to remove word mismatch between corpus words and query terms. The article has been organized into 4 different sections. Section 3 describes the mathematical notations and formulations for the proposed model. Section 4 elaborates query expansion using the proposed model. Section 5 discusses the experimental efficiency of the proposed model. Section 6 discusses the overall performance of the proposed model. Section 7 concludes the result and the future work of the proposed model.

3. Mathematical Formulation of the Proposed Model

The proposed model is based on the deep neural network-based skip-gram model. The skip-gram model is used to predict the context words for a given center word positioned at "c." This model defines a fixed window size "n" that assumes the words at position $c-n$ to $c+n$ as context words. For example, for a sentence "Abhishek is fond of learning Mathematics, Physics and Computer," if stop words are applied, the sentence became "Abhishek fond learning Mathematics Physics computer." If the center word is "learning," then a window of size 2 context word will be {Abhishek, fond, Mathematics, Physics}. In this architecture, we create a one-hot vector representation of vocabulary size |N| for the given center word. In one-hot vector representation, we set 1 at the position where the center word has occurred and 0

1. *for each* q *in* Q, *do*
   1.1. Retrieve top "k" documents d_k from initial retrieved documents.
2. *for each* d *in* d_k, *do*
   2.1. Retrieve user assisted relevant document d_u and store them in D_u.
3. *for each* t *in* q, *do*
   3.1. Create a hot vector x form dataset D_u.
4. iter←epoch
5. Initialize weight matrix w and w' with random weights.
6. *for i←1 to* itr, *do*
   6.1. Compute $h \leftarrow w^T.x$
   6.2. Compute $v \leftarrow w'^T.h$
   6.3. Update $w'_{new} \leftarrow w' - e'$
   6.4. Update $w_{new} \leftarrow w - e$
7. mer←[]
8. *for each* t *in* q, *do*
   8.1. Retrieve top 15 context words w_{cons} for w_t such that y is maximum.
   8.2. mer←mer+ w_{cons}
9. qexp←q+mer
10. *return* qexp

Algorithm 1: Skip-gram and relevance feedback-based query expansion RESKQ(Q[],D[]).

Table 1: Original query.

Number of queries	50
Relevant	2761
Relevant retrieved	2322
MAP	0.2979
R precision	0.3188
P@5	0.4480
P@10	0.4280
P@20	0.3900
P@50	0.3152

Table 2: Query expansion using the Bo1 model.

Number of queries	50
Relevant	2761
Relevant retrieved	2424
MAP	0.3302
R precision	0.3500
P@5	0.4640
P@10	0.4580
P@20	0.3880
P@50	0.3420

on the rest positions. The sparse one-hot vector is then transformed into a lower-dimensional dense representation. The skip-gram model consists of the following notations for the mathematical modeling.

Input layer: one-hot vector "x_c" of size ($|N|$,1) with 1 at the index where center word occurred and 0 at the rest of the indexes.

Target output layer: $2n$ one-hot vector "y_c" of size ($|N|$,1) with 1 at the index where center word occurred and 0 at the rest of the indexes.

Hidden layer: hidden layer of size (M,1).

Predicted output: probability vector $\hat{y}$ of size ($|N|$,1).

Random weight matrix: two random weight matrix of w and w' of each size (N, M) between the input and predicted output layers.

Now, using the following notations, we have

$$h_c = w^T \cdot x_c, \tag{1}$$

where $(.)^T$ represents the transpose of the given matrix. Now, the value obtained is mapped to weight matrix w'. Thus, we have

$$v_c = w'^T.h_c. \tag{2}$$

Now, the predicted output is computed using the following relations:

$$\hat{y} = \text{softmax}(v_c) = P(z = z_1 = v_c) = \frac{e^{z_1}}{\sum_{i=1}^{n} e^{z_i}}. \tag{3}$$

The loss function is basically the sum of the mutual crossentropies between predicted output values and target value computed. Mathematically,

$$L\left(v_c, t_{c-n}, \cdots, t_{c-1}, t_{c+1}, \cdots, t_{c+n}, w'\right) = -\sum_{j=-n, j\neq 0}^{n} \log\left(\hat{y}_{c+j}\right), \tag{4}$$

where $t_{c-n}, \cdots, t_{c-1}, t_{c+1}, \cdots, t_{c+n}$ are the $2n$ context word for the center word t_c, Now, using gradient decent method,

Table 3: Query expansion using the Bo2 model.

Number of queries	50
Relevant	2761
Relevant retrieved	2464
MAP	0.3301
R precision	0.3514
P@5	0.4880
P@10	0.4500
P@20	0.3970
P@50	0.3560

Table 4: Query expansion using the chi-square model.

Number of queries	50
Relevant	2761
Relevant retrieved	2381
MAP	0.2912
R precision	0.3083
P@5	0.4160
P@10	0.3880
P@20	0.3650
P@50	0.3096

Table 5: Query expansion using the KL divergence model.

Number of queries	50
Relevant	2761
Relevant retrieved	2427
MAP	0.3317
R precision	0.3504
P@5	0.4560
P@10	0.4700
P@20	0.3920
P@50	0.3452

Table 6: Query expansion using the proposed model.

Number of queries	50
Relevant	2761
Relevant retrieved	2334
MAP	0.3568
R precision	0.3627
P@5	0.6880
P@10	0.5580
P@20	0.4770
P@50	0.3588

Table 7: BM25 model.

No. of queries	50
Relevant	2761
Relevant retrieved	2325
MAP	0.2970
R precision	0.3230
P@5	0.4400
P@10	0.4240
P@20	0.3860
P@50	0.3168

Table 8: MAP comparison of the proposed model with other models.

Model	MAP
Original query	0.2779
Query expansion using the Bo1 model	0.3302
Query expansion using the Bo1 model	0.3301
Query expansion using the chi-square model	0.2912
Query expansion using the KL divergence model	0.3317
Query expansion using the proposed model	0.3568

$$\frac{\partial L}{\partial w} = \left[0 \cdots ..0 \ \frac{\partial L}{\partial h_c} \ 0 \cdots 0\right],$$
$$\frac{\partial L}{\partial h_c} = 2n\left[w'\hat{y} - \frac{1}{2n}\sum_{j=-n, j\neq 0}^{n} w_{c+j}\right], \tag{5}$$

where w_{c+j} is the $(c+j)^{\text{th}}$ column of the weight matrix w. Also,

$$\frac{\partial L}{\partial w'} = \left[\frac{\partial L}{\partial w_1} \ \frac{\partial L}{\partial w_2} \ \frac{\partial L}{\partial w_3} \cdots \frac{\partial L}{\partial w_{|N|}}\right], \tag{6}$$

where

$$\frac{\partial L}{\partial w_{c+j}} = 2n\hat{y}_{c+j}h_c - h_c, \quad \text{for all} - n \le j \le n, j \neq 0,$$
$$\frac{\partial L}{\partial w_{c+j}} = 2n\hat{y}_{c+j}h_c, \quad \text{for } j = 0. \tag{7}$$

Now, our objective is to minimize the loss function; therefore, the weights are updated as follows:

$$w_{\text{update}} = w - \alpha.\frac{\partial L}{\partial w},$$
$$w'_{\text{update}} = w' - \alpha.\frac{\partial L}{\partial w'}, \tag{8}$$

where α represents the learning rate of the deep neural network architecture.

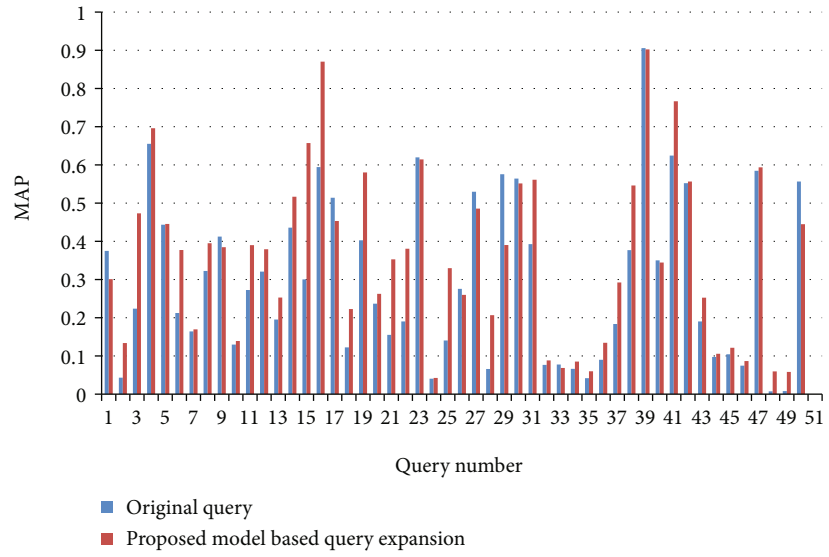

Figure 3: Performance of the proposed model vs. original query.

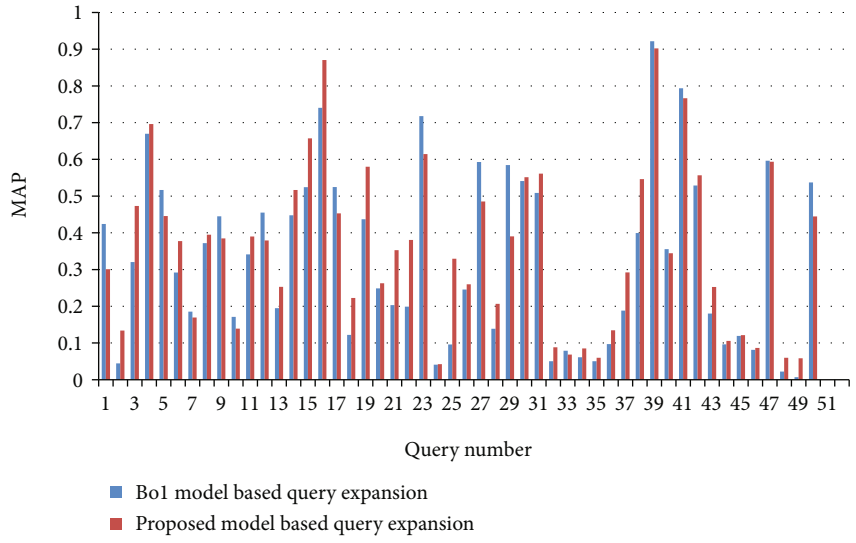

Figure 4: Performance of the proposed model vs. the Bo1 model.

4. Query Expansion Using the Proposed Method

The proposed method comprises a user-assisted relevance feedback model to select the relevant document for training the query terms using the skip-gram model. Rocchio [33] was the first who used the relevance feedback method for query expansion. The user initially selects the relevant documents from the top "k" retrieved documents. These relevant documents are then considered as a dataset to extract the most informative terms using the skip-gram model. The unique terms containing the relevant documents are considered vocabulary terms. Query terms are encoded using the one-hot encoding that creates a hot vector for each query term. The size of the hot vector is the size of the vocabulary. Each entry of the hot vector is 0 except the query term which is set as 1. In the skip-gram model, n-gram refers to the number of skip words. For example, for query "tiger conservation in India," a 2-gram representation will be "tiger conservation," "conservation in," and "India." These grams are trained to predict the context words. Skip refers to the number of times query term is presented throughout the relevant documents. The skip-gram architecture consists three layers, the input layer, hidden layer, and output layer. At the input layer, a one-hot vector is supplied to predict the context words. At the output layer, n -gram terms are trained to predict the context words. The hidden layer is the dense representation of the hot vector. A random weight matrix is assigned between the input and hidden layers and between the hidden and

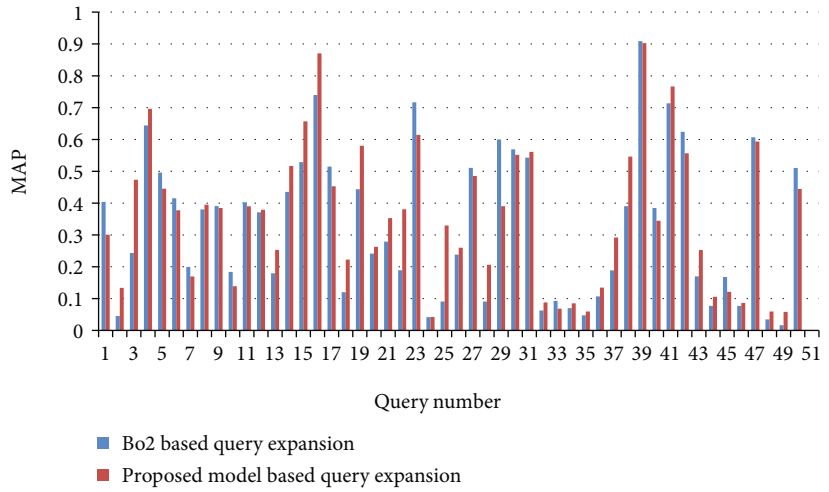

Figure 5: Performance of the proposed model vs. the Bo2 model.

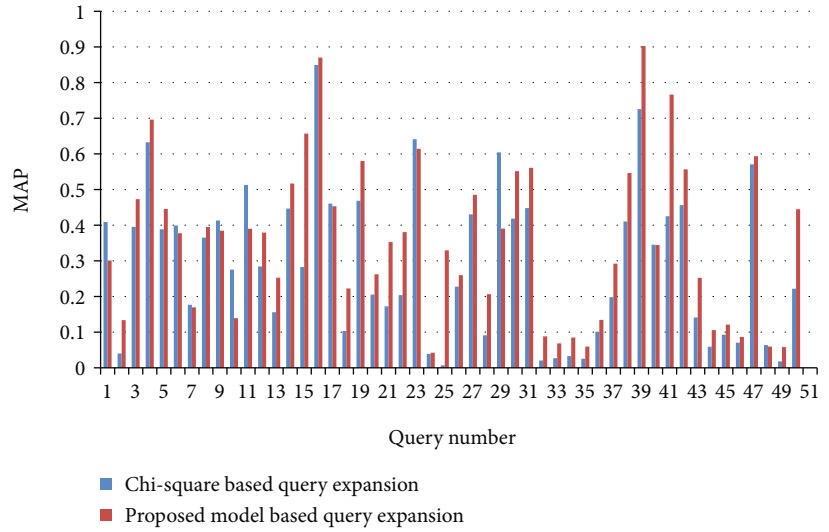

Figure 6: Performance of the proposed model vs. the chi-square model.

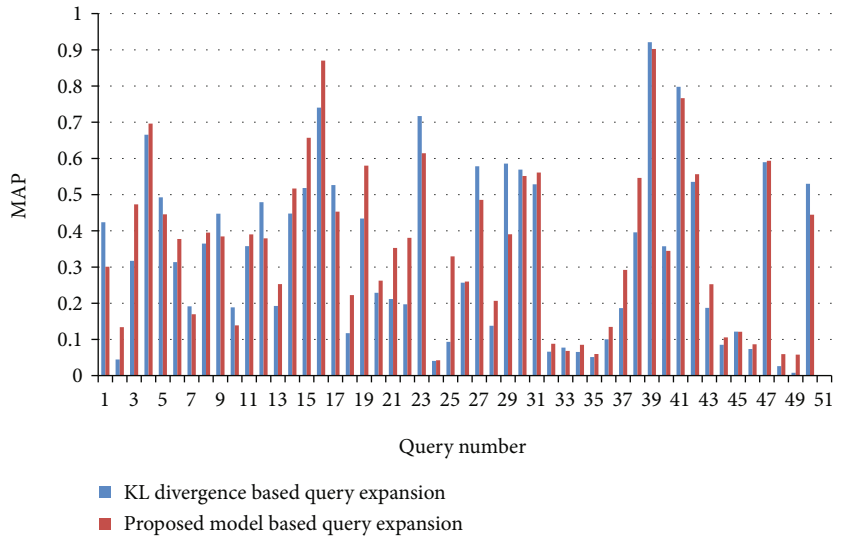

Figure 7: Performance of the proposed model vs. the KL divergence model.

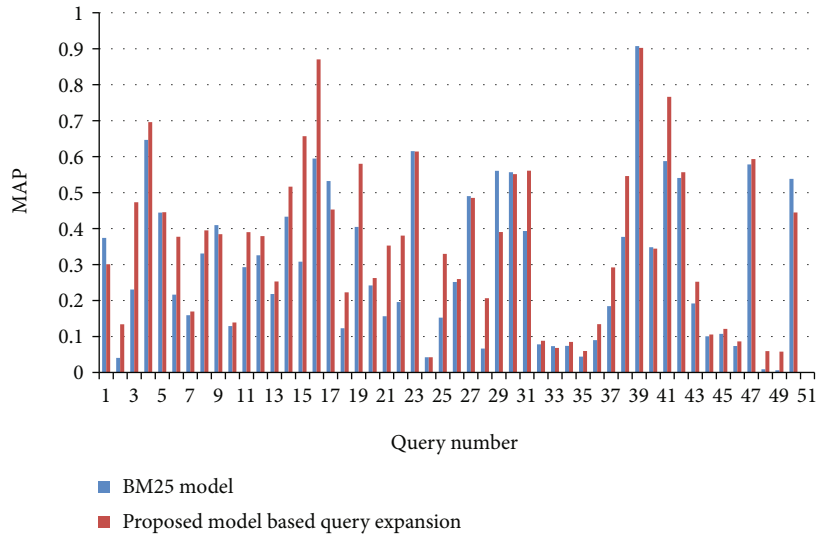

Figure 8: Performance of the proposed model vs. the BM25 model.

Table 9: Improvement of the proposed model vs. original query.

MAP	19.77%
R precision	13.77%
P@5	53.57%
P@10	30.37%
P@20	22.30%
P@50	13.83%

Table 10: Improvement of the proposed model vs. the Bo1 model.

MAP	8.05%
R precision	3.62%
P@5	48.27%
P@10	21.83%
P@20	22.93%
P@50	48.26%

Table 11: Improvement of the proposed model vs. the Bo2 model.

MAP	8.08%
R precision	3.21%
P@5	40.98%
P@10	24%
P@20	20.15%
P@50	0.7%

output layers. The structure of the skip-gram model is shown in Figure 2. The hot vector is mapped to a lower dimension representation into the hidden layer by applying the dot product between the hot vector and random weight matrix. If "x" represents the one-hot vector for a query term and "w" represent the random weight matrix, the hidden layer "h" hot vector can be represented by

Table 12: Improvement of the proposed model vs. the chi-square model.

MAP	22.52%
R precision	17.64%
P@5	65.38%
P@10	43.81%
P@20	30.68%
P@50	15.89%

Table 13: Improvement of the proposed model vs. the KL divergence model.

MAP	7.56%
R precision	3.51%
P@5	50.87%
P@10	18.72%
P@20	21.68%
P@50	3.93%

$$h = w^T \cdot x. \tag{9}$$

The hidden layer is propagated to the next layer by computing the dot product between the hidden layer and random weight matrix w' as

$$v = w'^T h. \tag{10}$$

Weights are updated at each iteration by the following computations:

$$w_{\text{new}} = w - e, \quad \text{where } e = \alpha . \frac{\partial L}{\partial w}, \tag{11}$$

Table 14: Sample queries and their expansion terms.

Query No.	Original query	Expansion terms	No. of RF relevant documents
Q130	Price hike of petroleum products	Litre, oil, rs, petrol, prices, increase, crude, rise, minister, fuel	17
Q132	Barack Obama's victory	Day, celebration, president, Americans, public, bond, mr, Tuesdays, hindrance, small, democrat, African-American, chanted, elect	7
Q139	Vanquishing the Somali pirates	Coast, strengthen, forces, three, send, seized, holding, tribal, hijacked, Aden, arrest, several, vessel, vessels	13
Q150	Bill Gates' philanthropic endeavours	Dropped, billions, friends, India, international, hundreds, Melinda, Buffett, corporation, retire, next, foundation	4
Q160	Iraq War 2003	International, likely, took, fees, early, visit, claims, plunged, people, believe, 1991, within, opposition, badly, measures, Maliki	4

$$w'_{\text{new}} = w' - e', \quad \text{where } e' = \alpha . \frac{\partial L}{\partial w'}. \tag{12}$$

The most probable context words for each query term are predicted in an unsupervised manner by successively updating the weights and calculating the probability at the softmax layer and then merged. The merged component is appended to the original query term to expand the query. The algorithm of the proposed model is shown in Algorithm 1.

5. Experimental Result

For the experimental analysis of the proposed method, we have used FIRE 2011 [35] ad hoc English test collections. Terrier 3.5 search engine has been used to evaluate the result of proposed method on the underlying dataset. Stop word and *PorterStemmer* are used to remove the stop word and to stem the root word, respectively. The experiment has been performed on 50 queries ranging from Q126 to Q175. The documents are retrieved using the InL2.0 model. The mean average precision (MAP) value of the proposed method is observed as 0.3568. To achieve this result, we have performed 70 epochs. The proposed model compares the result with other query expansion models. An improvement of 19.77% and 8.05% is observed with respect to original query and Bo1 model, respectively, on MAP parameter. The performance of the original query, query expansion using Bo1 model, query expansion using Bo2 model, query expansion using chi-square model, query expansion using KL divergence, query expansion using proposed model, and BM25 model is shown in Tables 1–7, respectively.

A Z-statistics is defined as

$$Z = \frac{(\bar{x} - \mu_0) * \sqrt{n}}{\sigma}, \tag{13}$$

where

(i) $\bar{x}$ is the sample mean

(ii) μ_0 is the mean populated in null hypothesis H_0

(iii) n is the sample size

(iv) σ is the population standard deviation

Now, sample mean $\bar{x} = 0.3568$; the population mean $\mu = 0.3302$ standard deviation of the population is $\sigma = 0.2349$; and sample size $n = 50$. Therefore,

$$Z = \frac{(0.3568 - 0.3302)}{0.2349/\sqrt{50}} = 0.801. \tag{14}$$

Since it is observed that $|z| = 0.801 \leq z_c = 1.96$, it is concluded that the null hypothesis is not rejected. Therefore, there is no evidence to claim that population mean $\boldsymbol{\mu}$ is different than 0.3302, at the 0.05 significance level. Therefore, improvement is minor.

The MAP comparison of the proposed model with other models is shown in Table 8. The query by query analysis of the proposed model vs. original query, query expansion using the Bo1 model, query expansion using the Bo2 model, query expansion using the chi-square model, query expansion using the KL divergence model, and BM25 model is shown in Figures 3–8, respectively. In the following figures, the x-axis represents the query number and the y-axis represents the MAP value.

6. Discussion

Tables 9–13 show that the proposed model has significant improvement over the original query, query expansion using the Bo1 model, query expansion using the Bo2 model, query expansion using the chi-square model, and query expansion using the KL divergence model, respectively. The proposed model has an improvement of 19.77%, 8.05%, 8.08%, 22.52%, and 7.56% in comparison to the original query, query expansion using the Bo1 model, query expansion using the Bo2 model, query expansion using the chi-square model, and query expansion using the KL divergence model, respectively. From Figures 3 to 7, it is clear that out of fifty queries, the proposed model performs well on 38, 34, 32, 40, and 34 queries with respect to the original query, query expansion using the Bo1 model, query expansion using the Bo2 model, query expansion using the chi-square model, and query expansion using the KL divergence model,

respectively. The sample query and its expansion terms are shown in Table 14.

7. Conclusion

Relevance feedback-based semantic model using skip-gram has been developed to improve the performance of the retrieval system. The proposed model has been compared with other state-of-the-art models, and experimental result shows that the proposed model has significant improvement over state-of-the-art models. The mean average precision of the proposed model is 0.3568 which is an improvement of 19.77%, 8.05%, 8.08%, 22.52%, and 7.56% compared to the original query, Bo1 model, Bo2 model, chi-square model, and KL divergence model, respectively. Out of fifty queries, the proposed model beats on 38, 34, 32, 40, and 34 queries compared to the original query, Bo1 model, Bo2 model, chi-square model, and KL divergence model, respectively. The proposed model also retrieves 12 additional relevant documents compared to the original query. In the near future, we will try to further improve the result by using other neural network architecture.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

One of the authors is pursuing a full-time Ph.D. from the Department of Mathematics, Bio-informatics, and Computer Applications, Maulana Azad National Institute of Technology (MANIT) Bhopal (MP), India. He expresses sincere thanks to the Institute for providing an opportunity for him to pursue his Ph.D. work. The author also thanks the Forum of Information Retrieval and Evaluation (FIRE) for providing a dataset to perform his experimental work.

References

[1] G. A. Miller, "WordNet," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[3] V. V. Kolisetty and D. S. Rajput, "A review on the significance of machine learning for data analysis in big data," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 6, no. 1, pp. 155–171, 2020.

[4] P. Kumar and R. S. Thakur, "Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16515–16535, 2021.

[5] S. M. Basha and D. S. Rajput, "A supervised aspect level sentiment model to predict overall sentiment on tweeter documents," *International Journal of Metadata, Semantics and Ontologies*, vol. 13, no. 1, pp. 33–41, 2018.

[6] C. Macdonald, R. McCreadie, R. L. Santos, and I. Ounis, *From puppy to maturity: experiences in developing Terrier*, OSIR@ SIGIR, 2012.

[7] G. Salton, *Modern Information Retrieval*, mcgraw-hill, 1983.

[8] A. Spink, D. Wolfram, M. B. Jansen, and T. Saracevic, "Searching the web: the public and their queries," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 3, pp. 226–234, 2001.

[9] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM (JACM)*, vol. 7, no. 3, pp. 216–244, 1960.

[10] J. Allan, "Incremental relevance feedback for information filtering," in *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 270–278, Zurich Switzerland, 1996, August.

[11] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *ACM SIGIR Forum (Vol. 51, No. 2, pp. 260-267)*, ACM, New York, NY, USA, 2017.

[12] M. Bendersky, D. Metzler, and W. B. Croft, "Parameterized concept weighting in verbose queries," in *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*, pp. 605–614, Beijing China, 2011, July.

[13] J. Miao, J. X. Huang, and Z. Ye, "Proximity-based Rocchio's model for pseudo relevance," in *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, pp. 535–544, Portland Oregon USA, 2012, August.

[14] Y. Lv and C. Zhai, "Positional relevance model for pseudo-relevance feedback," in *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pp. 579–586, Geneva Switzerland, 2010, July.

[15] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 311–318, Netherlands, 2007, July.

[16] J. Dalton and L. Dietz, "A neighborhood relevance model for entity linking," in *Proceedings of the 10th conference on open research areas in information retrieval*, pp. 149–156, Lisbon Portugal, 2013, May.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[18] Y. Qiu and H. P. Frei, "Concept based query expansion," in *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 160–169, Pennsylvania USA, 1993, July.

[19] R. Baeza-Yates and A. Tiberi, "Extracting semantic relations from query logs," in *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 76–85, San Jose California USA, 2007, August.

[20] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[21] J. Singh and A. Sharan, "A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach," *Neural Computing and Applications*, vol. 28, no. 9, pp. 2557–2580, 2017.

[22] D. Roy, D. Paul, M. Mitra, and U. Garain, "Using word embeddings for automatic query expansion," 2016, https://arxiv.org/abs/1606.07608.

[23] S. Kuzi, A. Shtok, and O. Kurland, "Query expansion using word embeddings," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, pp. 1929–1932, Indianapolis Indiana USA, 2016, October.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, https://arxiv.org/abs/1301.3781.

[25] F. Diaz, B. Mitra, and N. Craswell, "Query expansion with locally-trained word embeddings," 2016, https://arxiv.org/abs/1605.07891.

[26] A. Imani, A. Vakili, A. Montazer, and A. Shakery, *Deep neural networks for query expansion using word embeddings, European conference on information retrieval*, Springer, Cham, 2019.

[27] J. Dalton, L. Dietz, and J. Allan, "Entity query feature expansion using knowledge base links," in *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval*, pp. 365–374, Queensland Australia, 2014, July.

[28] Y. Xu, G. J. Jones, and B. Wang, "Query dependent pseudo-relevance feedback based on Wikipedia," in *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*, pp. 59–66, Boston MA USA, 2009, July.

[29] C. J. Crouch and B. Yang, "Experiments in automatic statistical thesaurus construction," in *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 77–88, Copenhagen Denmark, 1992, June.

[30] A. K. Shukla and S. Das, "Deep neural network and pseudo relevance feedback based query expansion," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3557–3570, 2022.

[31] A. K. Shukla and S. Das, "A hybrid model of query expansion using Word2Vec," in *2021 IEEE international conference on technology, research, and innovation for betterment of society (TRIBES)*, pp. 1–6, Raipur, India, 2021, December.

[32] A. K. Shukla, S. Das, and P. Kumar, "WordNet based hybrid model for query expansion," in *2021 IEEE international conference on technology, research, and innovation for betterment of society (TRIBES)*, pp. 1–6, Raipur, India, 2021, December.

[33] J. Rocchio, "Relevance feedback in information retrieval," in *The Smart retrieval system-experiments in automatic document processing*, Prentice Hall, 1971.

[34] https://www.researchgate.net/figure/The-architecture-of-Skip-gram-model-20_fig1_322905432.

[35] S. Palchowdhury, P. Majumder, D. Pal, A. Bandyopadhyay, and M. Mitra, *Overview of FIRE 2011*, In Multilingual Information Access in South Asian Languages (pp. 1-12), Springer, Berlin, Heidelberg, 2013.