Query expansion using pseudo relevance feedback based on the bahasa version of the wikipedia dataset

Cite as: AIP Conference Proceedings **2679**, 020008 (2023); https://doi.org/10.1063/5.0111273
Published Online: 04 January 2023

Husni, Yeni Kustiyahningsih, Fika Hastarita Rachman, et al.

Query Expansion Using Pseudo Relevance Feedback Based on the Bahasa Version of the Wikipedia Dataset

Husni[1, a)], Yeni Kustiyahningsih[1,b)], Fika Hastarita Rachman[1,c)], Eka Mala Sari Rochman[1,d)], Hadi Yulian[1)]

[1]*Department of Informatics, Faculty of Engineering, University of Trunojoyo, Madura, Indonesia*

[a)] *Corresponding author: husni@trunojoyo.ac.id*
[b)] *ykustiyahningsih@trunojoyo.ac.id,* [c)] *fika.rachman@trunojoyo.ac.id,* [d)] *em_sari@trunojoyo.ac.id*

Abstract. The work of finding documents that are relevant to a user's query on an information retrieval system (IRS) is a very interesting study. The relevance of the list of documents returned by the IRS is influenced by the accuracy of the method of calculating the similarity between documents and the determination of the keywords. Many users are difficult to describe their information needs in words. Sometimes the user enters only one or two words that do not reflect the domain of information required. This results in a list of documents were very less relevant to the user's needs. The approach to improve the list of words in the user's query to make it more representative is called Query Expansion. One technique that can be used to expand a query is Pseudo Relevance Feedback. This paper describes the results of research that has been carried out to expand Query using Pseudo Relevance Feedback on an IRS based on the Indonesian version of the Wikipedia dataset, totaling about 450 thousand documents. Calculation of the similarity between the query and the list of tourism news documents uses the cosine similarity, while the weighting scheme for each term uses TF-IDF. The test results show that the pseudo-relevance feedback decreases the precision of the IRS up to 30%. This is due to the failure of the chosen approach to finding the right words to expand the original query. The abstract of articles in Wikipedia is general and is not limited to the tourism domain. The selection of the expansion base dataset is greatly determined by the new query quality and datasets from the same domain are recommended. It is highly recommended that the QE reference dataset is domain specific and filtered before being used as a QE basis.

INTRODUCTION

The purpose of using the Information Retrieval System (IRS) is to find a list of ranked documents that are relevant to the user's query in a short time [1][2]. There are several weaknesses of this Query-based search system, including the perception built by the user in generating queries. Many users find it difficult to formulate a list of keywords that represent their information needs. Some users make short queries consisting of one or two words that are not quite right. The large number of results were provided by Search Engines or the IRS but most of them do not meet the information needs of these users [3].

One solution that can be applied to correct the short Query above is Query Expansion (QE). The QE approach tries to extend the query by adjusting the query or adding a few new words so that a new obtained query is considered more representative of the user's information needs [4]. QE allows the elimination of ambiguity in queries and expresses the concept of detailed information. The addition of several new terms into the initial query is believed to increase the number of identified documents and the possibility of finding relevant documents [3]. However, in some cases, the QE approach does not need to be applied to certain queries.

This paper reports the results of the research of QE on the Search Engine Bahasa Indonesia (SEBI), which currently contains Sports and Tourism news. The method used pseudo-relevant feedback. The dataset in this application was

The First International Conference on Neuroscience and Learning Technology (ICONSATIN 2021)
AIP Conf. Proc. 2679, 020008-1–020008-6; https://doi.org/10.1063/5.0111273
Published by AIP Publishing. 978-0-7354-4300-6/$30.00

the Indonesian Wikipedia document where the number of records reaches 450 thousand. In this approach, the Wikipedia articles are mostly related to the Query which are identified to generate a new Query [5]. The closeness between the Query and the list of documents (including Wikipedia articles) were calculated using the cosine similarity formula. The terms contained in the Query or other documents, after going through a series of preprocessing, were weighted using TF-IDF. Cosine similarity can provide a high accuracy value where the main advantage was not affected by the short length of the document. TF-IDF weighting has proven to be effective because it is widely applied to various IRS projects. Both of these methods are quite widely used in the IRS because they are efficient, easy, and accurate [6], [7]. An explanation of QE, its types, and its role in the field of information retrieval studies can be seen in [8],[9].

The next section of this paper described the method to generate a new query based on the similarity of the query to the Wikipedia article. Analysis of the results has been discussed next part. This paper closes with a Conclusion.

METHODS

System Architecture

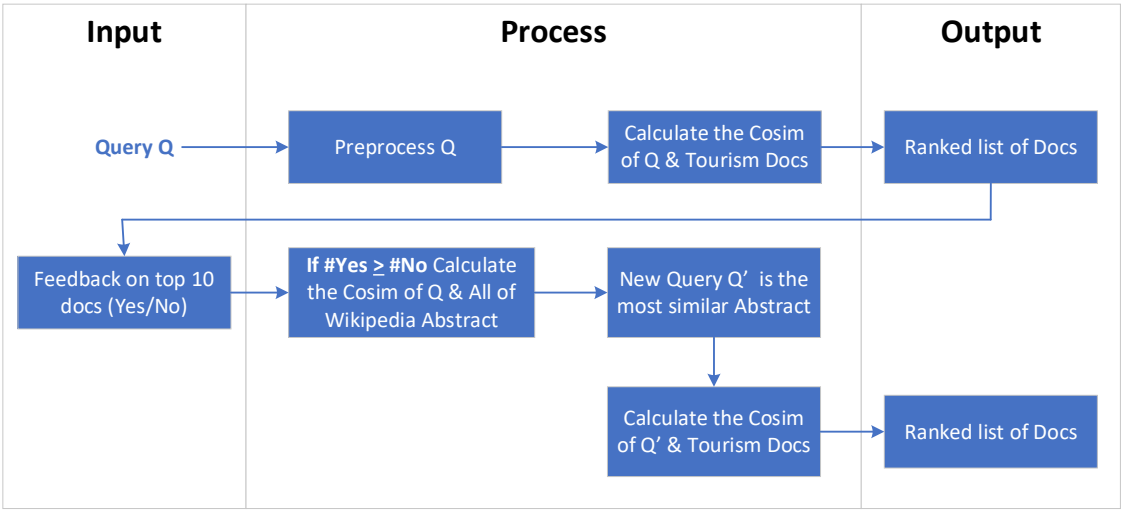

FIGURE 1. The architecture of the Indonesian Wikipedia-based QE System for the tourism news domain

The system flow of the built QE application is as follows (Figure 1):

1. Receiving Query Q from the user
2. Preprocessing Q
3. Calculating the cosine similarity (CoSim) between Q and all documents in the Search Engine corpus (tourism news).
4. Providing a list of ranked documents to users (10 best first)
5. Users are asked to provide feedback for each document received as a result by saying *Yes/No* to the suitability of the results with their information needs.
6. If the number of "No" is less than or equal to the number of "Yes" then the process is DONE.
7. The system implements QE. In this QE, each abstract from the Indonesian Wikipedia is compared with Query Q using cosine similarity. The Wikipedia abstract most similar to Q will be the new Query Q'. This approach is similar to the discussion in [10], while on how pseudo-relevance feedback works can be seen in [11], [12], [13], and [14]. How to get a dataset from Wikipedia? Please access http://download.wikipedia.org [15].
8. Calculating the similarity of each tourism news document with the New Query Q' using cosine similarity.
9. Giving the result to the user. FINISHED.

Dataset

The dataset in this study was the abstract of articles contained in the Indonesian version of Wikipedia. This dataset can be downloaded from *https://dumps.wikimedia.org/idwiki/latest/idwiki-latest-abstract.xml.gz*. The number of abstracts are 450,513. This collection of abstracts is used as the basis for searching for new Query Q' by applying query expansion pseudo relevance feedback to Query Q. Other collections of documents were 100 Indonesian language tourism news taken from various news sites on the Internet, including *detik.com, circumferencejatim.com, portalmadura.com, islandmadura.com, radarmadura.jawapos.com, traveloka.com, and liputan6.com*. The part of the news taken is the title, URL, date of posting, and the content of the article in text form (other media types are ignored). The system calculates the similarity between Query Q and the tourism news collection. If the precision given by the system is less than 50%, a query expansion can be carried out based on the Wikipedia abstracts. Other data that is also used is a collection of basic words and an Indonesian stoplist which can be obtained from URL *http://hikaruyuuki.lecture.ub.ac.id/kamus-kata-dasar-dan-stopword-list-Bahasa-indonesia/*. The number of basic words available is 28,526 words and there is a stoplist of 758 words. This data is also contained in SEBI which is an experimental search engine that is used as the basis of this research [16]. Another studies that use wikipedia articles as a reference for query expansion can be seen at [17].

Preprocessing

Preprocessing is the work of preparing raw data before the core processes of text mining such as calculating the similarity between documents is carried out. Broadly speaking, preprocessing is applied by eliminating inappropriate term or converting it into a form that is easier to process by the system [18].

The following are the preprocessing stages carried out in the proposed architecture:

1. Case Folding: change uppercase to lowercase [19].
2. Tokenizing: breaking the documents into collections of words or terms. Tokenization can be done by removing punctuation marks and separating them per space [19].
3. Stopwords removal: remove stopwords or terms that are not discriminators in the processed document. Stopword removal is used to reduce the dimensions of the document [7].
4. Stemming: removing all affixes and producing root words that match the Indonesian morphological structure [6] [8].

TF-IDF

The Term Frequency-Inverse Document Frequency (TFIDF) is a method of determining the weight of each term involved in computing in the vector space model. TF determines the importance of a term in a document, while IDF shows the importance of a term based on the number of documents containing that term. The combination of TF and IDF determines the importance of a term in the corpus or collection of documents. The weight of the term becomes important when it appears more often in a document but rarely appears in many documents. A term that is present in many documents, does not become a discriminator; so the level of importance decreases. The term *and* and *the* are certainly present in many documents so that the weight becomes zero.

The following is the formula for calculating the TFIDF of each term in the corpus [6]:

$$w_{dt} = tf_{dt} * idf_t$$

where:

w_{dt}= the weight of term *t* in document *d*

tf_{dt} = the frequency of occurrence of term *t* in document *d*

idf_t = *Inversed Document Frequency* (*log (N/df)*) where *N* = total document total and *df* = frequency or number of documents containing term *t*.

Cosine Similarity

In the vector space model, the proximity or similarity between documents can be determined by calculating the cosine value of the angle between two document vectors, so this approach is called Cosine Similarity (CoSim) [20]. The formula used for this purpose is

$$sim\ (d_j,q) = \frac{d_j \cdot q}{|d_j| \cdot |q|} = \frac{\sum_i w_{ij} \cdot w_q}{\sqrt{\sum_i w_{ij}^2} \cdot \sqrt{\sum_i w_q^2}}$$

where:

w_{ij} = the weight of the i^{th} term in the j^{th} document

w_q= the weight of query Q

RESULT

The system was built according to the architecture in Figure 1 and by applying the calculation of cosine similarity which has been tested with the following scenarios:

1. The test involved 10 users where each user was asked to enter 2 queries (two tests per user)
2. In the first test, the user entered a query consisting of only 1 word about tourism.
3. In the second test, the user entered a query that is more than 1 word long about tourism
4. For each query to be entered, the user must perform tests on two tourism news collections, containing 100 and 75 news articles, respectively.
5. After the search results were displayed, the user can provide feedback values according to a large number of documents relevant to the query. This process results in a precision value.
6. For the precision value, before query expansion is <50%, the system performed query expansion using the pseudo relevance feedback method with the Indonesian Wikipedia abstract.
7. The search results were given a feedback value by the user, which s limited to 10 documents with the highest similarity level.
8. Users have different characters so that they have their conclusions in determining the documents that are considered relevant. First, was the query entered the main discussion in the article? Second, was the title of the article following the query and describes the main discussion? Finally, the user considered the article relevant to the query if the article discussed matters related to the query.
9. The process of matching queries with documents used the cosine similarity method. Meanwhile, query expansion used pseudo-relevance feedback and cosine similarity methods.

The result was the average precision of each scenario. Table 1 shows the recapitulation of test results without QE and with QE.

TABLE 1. Recapitulation of Test Results Query Expansion Pseudo Relevance Feedback

Query Length	Without QE		With QE	
	100 Articles	**75 Articles**	**100 Articles**	**75 Articles**
One word	88%	82%	0%	40%
More than one word	63%	59%	36,67%	23,33%
Average	**75,5%**	**70,5%**	**36,67%**	**31,67%**

The comparison graph of precision without QE and with QE can be seen in Figure 2.

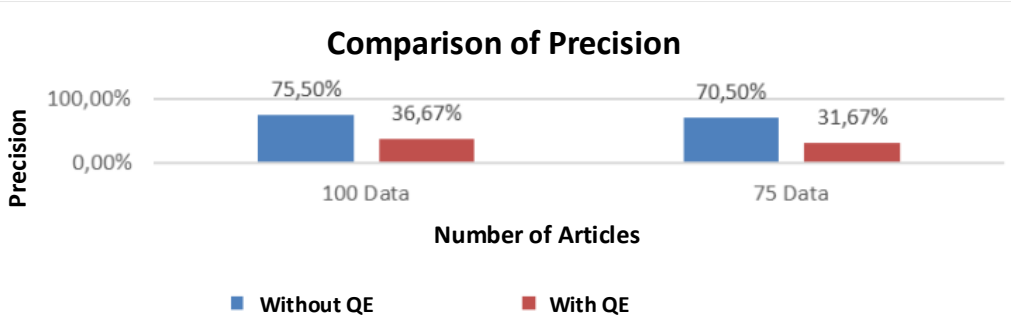

FIGURE 2. Graph of Test Results Recapitulation

Figure 2 shows that precision without QE is better than using QE. The precision of the test results without QE for 100 articles is 75.50%, while for 75 articles it is 36.67%. When QE was used in testing 100 articles, the precision was 70.50% and for 75 articles it was only 31.67%. It can be said that the number of articles and the number of words in the query also affected the precision. There were several reasons why the precision with QE is quite low. First, the queries entered by users were generally too general and less unique (not specific to tourism). So the opportunity to get a precision without QE > 50% was very possible and makes only a few queries expanded. The second was the use of Indonesian Wikipedia which was universal and did not focus on tourism. This results in the list of words were obtained after the query were expanded to not always leading to the topic of tourism. This was why articles that previously did not match the initial query were then entered into one of the top 10 search results. Finally, the limited number in the collection of articles was used. So, there was a lack of article variations in document matching.

CONCLUSION

This study shows that there is no guarantee that an information retrieval system (IRS) that implements Query Expansion (QE) will provide better performance than IRS without QE. Such results were also found by Agiyola [21]. It is highly recommended that the QE reference dataset is domain specific and filtered before being used as a QE basis.

ACKNOWLEDGMENTS

The researchers thank the University of Trunojoyo Madura for providing the opportunity and funding the 2021 research group on the topic “Development of distributed Search Engine Bahasa Indonesian (SEBI)”

REFERENCES

1. L. Afuan, A. Ashari, and Y. Suyanto, “Query Expansion in Information Retrieval using Frequent Pattern (FP) Growth Algorithm for Frequent Itemset Search and Association Rules Mining,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 263–267, 2019.
2. Q. Aini, U. Rahardja, A. Moeins, and A. M. Wardani, “Penerapan Data Market Query (DMQ) pada Sistem Penilaian Berbasis YII Framework,” *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 1, pp. 26–31, 2018.
3. J. Ooi, X. Ma, H. Qin, and S. C. Liew, “A Survey of Query Expansion, Query Suggestion, and Query Refinement Techniques,” in *4th International Conference on Software Engineering and Computer Systems, ICSECS 2015: Virtuous Software Solutions for Big Data*, 2015, pp. 112–117.
4. N. C. Wirawan and P. P. Adikara, “Analisis Sentimen Dengan Query Expansion Pada Review Aplikasi M-Banking Menggunakan Metode Fuzzy K-Nearest Neighbor (Fuzzy k-NN),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 1, pp. 362–368, 2018.
5. A. Keikha, F. Ensan, and E. Bagheri, “Query Expansion using Pseudo Relevance Feedback on Wikipedia,” *J. Intell. Inf. Syst.*, vol. 50, no. 3, pp. 455–478, 2018.
6. R. Melita, V. Amrizal, H. B. Suseno, and T. Dirjam, “Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan Cosine Similarity pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah

Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam)," *JTI (Jurnal Tek. Inform. UIN Syarif Hidayatullah*, vol. 11, no. 2, pp. 149–164, 2018.
7. A. H. Silem, H. Taktak, and F. Moussa, "A new Query Reformulation Approach using Web Result Clustering and User Profile", *Procedia Computer Science* 192, pp. 1180–1189, 2021.
8. F. P. Wardani and B. Rahayudi, "Query Expansion Pada Sistem Temu Kembali Informasi Dokumen Jurnal Berbahasa Indonesia Menggunakan Metode BM25," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2619–2625, 2019.
9. M. Y. Dahab, S. Alnofaie, and M. Kamel, "A Tutorial on Information Retrieval Using Query Expansion," in *Studies in Computational Intelligence*, vol. 740, 2018, pp. 0–16.
10. F. Colace, M. De Santo, L. Greco, and P. Napoletano, "A Query Expansion Method based on a Weighted Word Pairs Approach," in *CEUR Workshop Proceedings*, 2013, vol. 964, no. February, pp. 17–28.
11. L. Lee, "Implicit Relevance Feedback & Clickthrough Data Clickthrough Data," *CS 6740: Advanced Language Technologies*, pp. 1–10, 2010.
12. D. Kelly and J. Teevan, "Implicit Feedback for Inferring User Preference," *ACM SIGIR Forum*, vol. 37, no. 2, pp. 18–28, 2007.
13. R. Mandala, "Evaluasi Efektifitas Metode Machine-Leraning pada Search-Engine," *Seminar Nasional Aplikasi Teknologi Informasi 2006 (SNATI 2006)*, vol. 2006, no. Snati, 2006.
14. Z. Y. Pamungkas, Indriati, and A. Ridok, "Query Expansion pada Sistem Temu Kembali Informasi Dokumen Berbahasa Indonesia Menggunakan Pseudo Relevance Feedback Studi kasus: Perpustakaan Universitas Brawijaya," *Repos. J. Mhs. PTIIK UB*, vol. 6, no. 3, 2015.
15. G. J. F. Jones and B. Wang, "Query Dependent Pseudo-Relevance Feedback based on Wikipedia Categories and Subject Descriptors," *32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 59–66, 2009.
16. Husni, "Web Service for Search Engine Bahasa Indonesia (SEBI)", Journal of Physics: Conference Series 1569, 2020
17. M. Maryamah, A. Z. Arifin, R. Sarno, Y. Morimoto, "Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents", *International Journal of Intelligent Engineering and Systems*, Vol.12, No.5, 2019
18. S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Seminar Nasional Teknologi Informasi dan Komunikasi*, vol. 2016, no. Sentika. pp. 2089–9815, 2016.
19. R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro Univ. Negeri Semarang*, vol. 9, no. 1, pp. 18–23, 2017.
20. H. Wijaya, A. Kesuma, and S. Pribadi, "Penerapan Cosine Similarity dalam Aplikasi Kitab Undang-Undang Hukum Dagang (Wetboek Van Koophandle Voor Indonesia)," *Sci. J. Informatics UNNES*, vol. 8, no. 1, pp. 1–3, 2016.
21. T. Agiyola, I. Indriati, B. Rahayudi, "Relevance Feedback Pada Sistem Temu Kembali Informasi Dokumen E-Book Berbahasa Indonesia Menggunakan Metode BM25", *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* Vol. 3, No. 5, pp. 4613-4621, 2019.