ELINVO (Electronics, Informatics, and Vocational Education), Mei 2023; vol 8(1): 113-120
ISSN 2580-6424 (printed), ISSN 2477-2399 (online) DOI: https://doi.org/10.21831/elinvo.v8i1.55851

Cross-Language Tourism News Retrieval System Using Google Translate API on SEBI Search Engine

Husni[1*], Arif Muntasa[2], Sigit Susanto Putro[3], Zulfi Osman[4]
[1,2,3,4]Informatics Engineering Department, Faculty of Engineering, University of Trunojoyo Madura, Indonesia
[*]E-mail: husni@trunojoyo.ac.id

ABSTRACT

Cross-Language Information Retrieval (CLIR) is responsible for retrieving information stored in a language different from the language of the query provided by the user. Some translation methods commonly used in CLIR are Dictionary, Parallel corpora, Comparable corpora, Machine translator, Ontology, and Transitive-based. The query must be translated to the target language, followed by preprocessing and calculating the similarity between the query and all documents in the corpus. The problem is the time and accuracy of query translation. Moreover, the queries are not written as complete sentences according to certain language rules. Stemming, for example, every language has its own method. Indonesian has basic words and affixes in the form of prefixes, suffixes, infixes, and confixes, while English only has suffixes. Stemming takes a long time in text processing. In the Indonesian search engine (SEBI), the provision of cross-language tourism news retrieval is realized using the Google Translate API, which translates the Query and all documents into English, Porter's stemming technique to convert each term to its general form, and cosine similarity to calculate similarity. This approach can deliver cross-language tourism news instantly while increasing the precision and efficiency of the SEBI search engine, although some improvements are needed to provide a more precise and efficient similarity computation.

Keywords*:* Cross-language retrieval, Cross-Language Information Retrieval, google translate API, tourism news, search engine

INTRODUCTION

Cross-Language Information Retrieval (CLIR) allows users to obtain information stored in a language different from the user's Query language. Users can access documents in English even though the Query is entered in Indonesian, and vice versa [1]. Various translation techniques are used in CLIR, such as Dictionary-based, Parallel corpora-based, Comparable corpora-based, Machine translator-based, Ontology-based, and Transitive based [2][3]. This translation must be carried out on all documents before preprocessing, including tokenization, case-folding, stop word removal, stemming, and weighting.

In Indonesian – English CLIR, the main problem is the length of time required for preprocessing, especially stemming and accuracy of Query translation. Stemming of Indonesian documents takes longer than English for the same method. This happens because Indonesian has various forms of affixes. There are variations of affixes, including prefixes (affixes that are placed at the beginning of the root word), suffixes (affixes that are located at the end of the root word), infixes (affixes that are inserted in the middle of the root word), and confixes (affixes that are located at the beginning and end of the root word at the same time).). Languages only have suffixes [4]–[6].

The complexity of stemming increases if we look at the morphological structure in Indonesian, which has a higher level of complexity than in English. For example, we have difficulty in distinguishing a word that contains affixes, both prefixes, and suffixes, from a basic word in which one syllable is part of the affix, so that requires a basic word dictionary which, of course, requires memory which causes the process to be longer [7]. Some studies, such as [8], use the Porter Stemmer when simplifying terms in Indonesian. Porter's approach accuracy is only 61%. Another study uses the Nazief-Adriani method, which bases all computations on thousands of basic words, and its accuracy is still not satisfactory.

This article tries to report the results of research that has been carried out where every document contained in the corpus is translated into English using the Google Translate API. This API quickly translates documents with a maximum length of 5000 words and queries from the Information retrieval System for tourism news. The translation results will be treated according to standard Information Retrieval, starting from pre-processing, including Stemming with the Porter technique, which has become the de-facto for English documents, weighting with TF-IDF, to the calculation of similarity using the Cosine Similarity method. Weighting and similarity calculations are widely used in Information Retrieval research because they are proven to be efficient, easy to implement, and provide high precision [9]–[14]. This study wants to find out whether the transfer of pre-processing and similarity calculations into English will provide better precision in an Information Retrieval system such as SEBI. Table 1 shows some of the CLIR studies.

Table 1. Some of CLIR's research and document translation techniques.

Authors	Methods	Result
Saravanan [15]	Query Translation Word Sense Disambiguation (WSD) and Word by Word Translation (WBWT) using VSM	precision using WSD 0,97 and WBWT 0,72
Bhattacharya [16]	Query translation using Multilingual Word Cluster	Low accuracy, MAP < 38%
Agrawal [17]	Selective Document and Query Expansion	MAP = 0.4605 and MAP = 0.4756 after Query Expansion
Prasath [11]	Query Translation Monolingual Method using two-level disambiguation model with Analyser	MAP of Monolingual 0.518, Two levels disambiguation 0.412, Two level disambiguation with analyzer 0.453.
Litschko [18]	Query Translation Cross-Lingual Embeddings Method from Comparable Documents (CL-CD)	MAP Cross-Lingual Embeddings of Comparable Documents (CL-CD) EN-NL = 0.125, EN-IT = 0.106, EN-FI = 0.176.

METHODS

The approach used in this research is to translate all documents (news related to tourism destinations on Madura Island, Indonesia) and queries on the SEBI search engine into English. This resulted in all subsequent stages in SEBI using various appropriate methods for English documents, no longer using an Indonesian-based approach, starting from tokenization, case-folding, stopword removal, and especially stemming using Porter stemmer. This series of processes is known as Preprocessing. Figure 1 shows the search engine architecture, which at least consists of at least two processing streams, namely (1) processing of the corpus, which in this case is a collection of tourism news, and (2) Query in the form of text, which represents the information needs of the user. The same preprocessing must be applied to the two streams so that the same words from different streams will give the same output from these stages. This implementation aims to increase the possibility of relevance (similarity) between the Query and the collection of news in the corpus.

Apart from being in Indonesian, it is very likely that tourism news downloaded by web crawling (part of SEBI, which is responsible for compiling corpus) is already in English. If so, the document is not translated but immediately enters the preprocessing stage. Only documents in the Indonesian language must go through the stages of translation, which are carried out using the Google Translate API. How does SEBI know if a document is already in Indonesian or English? Several approaches can be taken, at least by looking at the <html lang="id-ID"> HTML tag, which indicates the document is in Indonesian, or <html lang="en-US"> if the document is in English. This job is done right after the document has been downloaded by the Web Crawler to generate a clean document labeled with language and a list of news links which can then be downloaded. Some of the more sophisticated approaches can be seen in [19].

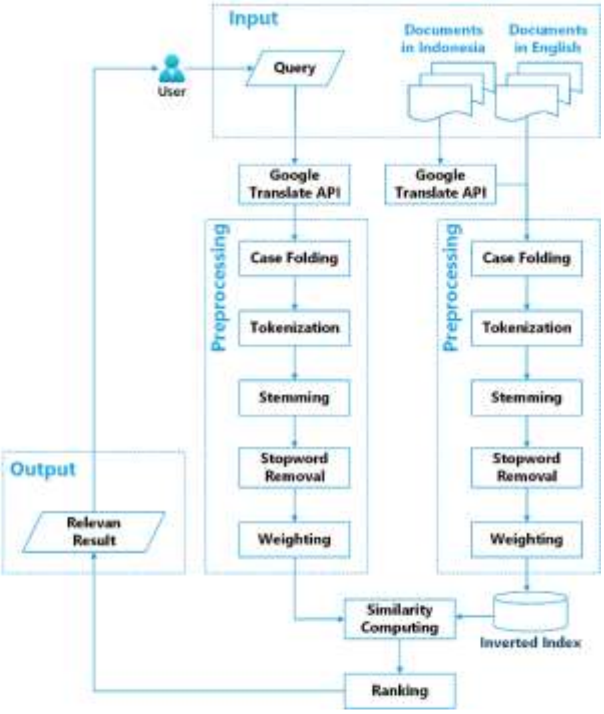

Figure 1. Search Engine architecture involving Google Translate API to standardize all documents in English

Figure 1 shows that all tourism news in the collection (corpus) and Query are switched to English. The translation of tourism news is done in the background one by one until all documents have an English version. This version will then be stored, indexed, weighted, and called a corpus. The translation of the Query is done online as soon as the user enters the Query. Details of the Query handling process in SEBI, after translation, are as follows (Figure 2):

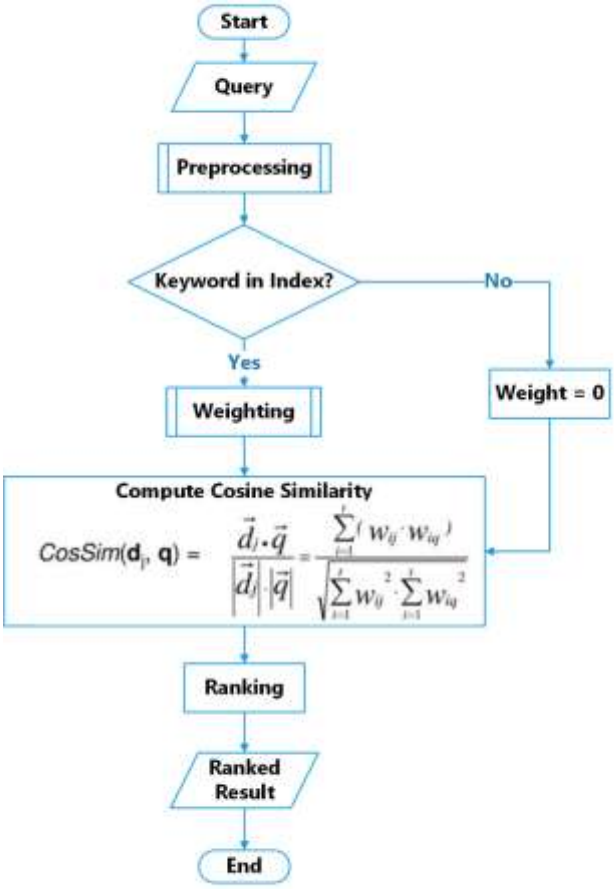
Start
Query
Preprocessing
Keyword in Index?
No
Yes
Weighting
Weight = 0
Compute Cosine Similarity
$CosSim(\mathbf{d}_j, \mathbf{q}) = \frac{\vec{d_j}\cdot\vec{q}}{|\vec{d_j}|\,|\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij}\cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$
Ranking
Ranked Result
End

Figure 2. The process of calculating the similarity between Query and Corpus. One of the advantages of this approach is that it examines the intersection of terms in a Query with a tourism news collection (corpus)

1. The user enters the query, and it is translated into English using Google Translate API.
2. Preprocessing the Query up to the Stemming stage is carried out so that query keywords are obtained that are ready to use.
3. The keyword query is matched with the keyword list in the corpus. If there is a match, then the weighting process will be carried out. If there is no match, then the process is complete (there are no documents relevant to the user's query). This step is believed to speed up the Query handling process. Similarity calculations are only

performed if the Query has an intersection with the documents in the corpus.

4. Then calculate the similarity between the keywords from the query and the keywords from the documents whose terms intersect using the cosine similarity method.
5. The ten documents with the highest similarity scores are sorted (ranked)
6. The ten documents are returned to the user and are considered tourism news relevant to the user's information needs.

The calculation of the similarity between the query and the list of documents in the corpus is the essence of a search engine. After all documents (including queries) have passed the preprocessing stage, each word or term gets a weight that is calculated, one of which is the TF.IDF formula, which is the multiplication between the number of occurrences of words in a document multiplied by the log of the total number of documents divided by the number of documents containing those terms. If we are going to calculate the similarity between Query q and document d, then every word that intersects in q and d is taken to the vector model space. The measure of the angle between q and d shows the similarity between the two. The smaller the angle, the closer the two are. If the angle between q and d is zero, then q and d are 100% alike. If the angle between q and d is 90 degrees, then q and d are not similar at all. The cosine value of the angle represents such an explanation. The cosine value of 0 degrees is one (100%), and the cosine value of 90 degrees is zero (not at all similar). An example of calculating the similarity between q and d can be seen in [18].

The process of translating documents in corpus and queries into English uses a Python library called google trans, where the steps are quite easy, namely: (1) Initialize the function for the translation process from Indonesian to English; (2) Initialize variables for the Python translator library; (3) Get the translated text.

The google translate API() function below shows how to use the Google Translate API to translate Indonesian text into English:

```
from googletrans import Translator

def googleTranslateAPI(text):
    translator = Translator()
    translated =
      translator.translate(text,
      dest='en', src='id')
    return translated.text
```

RESULT AND DISCUSSION

The proposed approach is tested using the following scenarios: (1) Scenario I: Use 10 queries consisting of two words; (2) Scenario II: Use 10 queries consisting of three words; (3) Scenario III: Use as many as 10 queries consisting of four words; (4) The query used in this test must be related to the news of Madura Island tourism destinations and by the correct Indonesian or English language rules; (5) Search results are limited to 10 documents based on the highest similarity value (calculated using the Cosine Similarity method); (6) To speed up the Query computing process and increase the similarity value, each query will only be compared with a collection of tourism news containing the same word from the Query; (7) The test records the average precision.

Testing I with a query consisting of 2 words. 10 users enter a query consisting of 2 words in Indonesian and English. The similarity between Query and 200 tourism news is calculated with and without involving stemming. When stemming is applied, Search Engine CLIR provides an average precision value of 0.41 with an average search time of 5.91 seconds. In the non-stemming approach, the average precision value is 0.4, with an average search time of 5.721 seconds. The details of this test are shown in Figure 3.

Testing II with a query consisting of 3 words. 10 users entered a query of 3 words in Indonesian and English. This test includes stemming and shows an average precision value of 0.41 with an average search time of 10.11 seconds, while when not applying stemming, the average precision value was 0.43 with an average search time of 9.24 seconds. Details of the results of this test are shown in Figure 4.

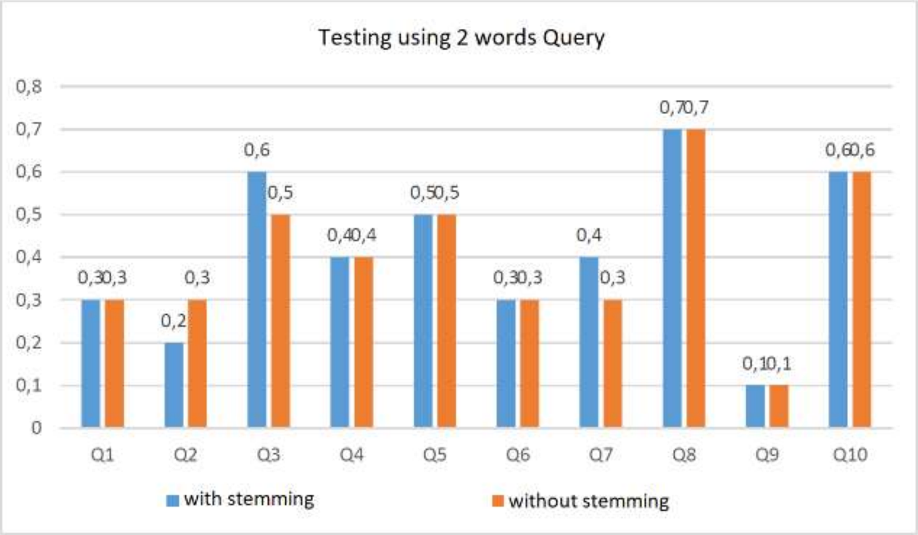

Figure 3. Testing using a 2-word query with and without stemming

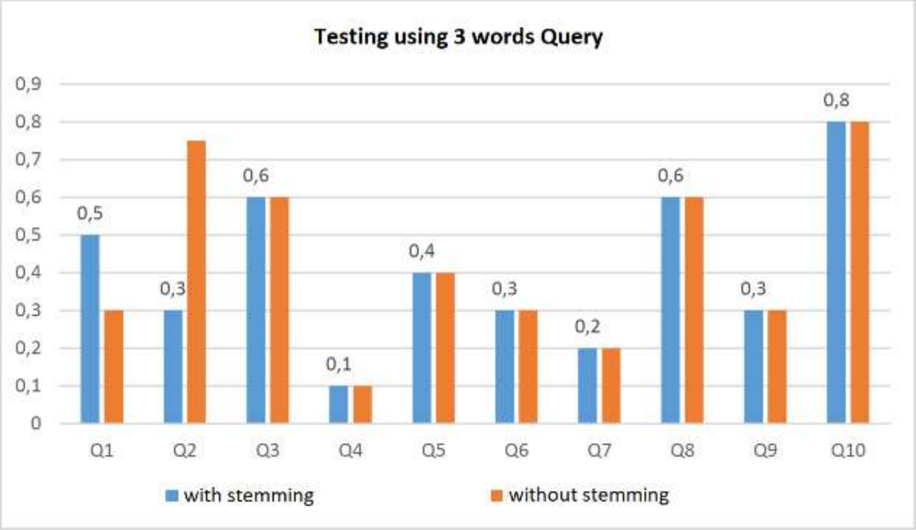

Figure 4. Testing using a 3-word query with and without stemming

Testing III with a query consisting of 4 words. 10 users entered a query of 4 words in Indonesian and English. The test results show that the application of stem gives an average precision value of 0.35 with an average search time of 11.89 seconds, while when it does not involve stemming, the average precision value is 0.27 with an average search time of 11.12 seconds. Details of the results of this test are shown in Figure 5.

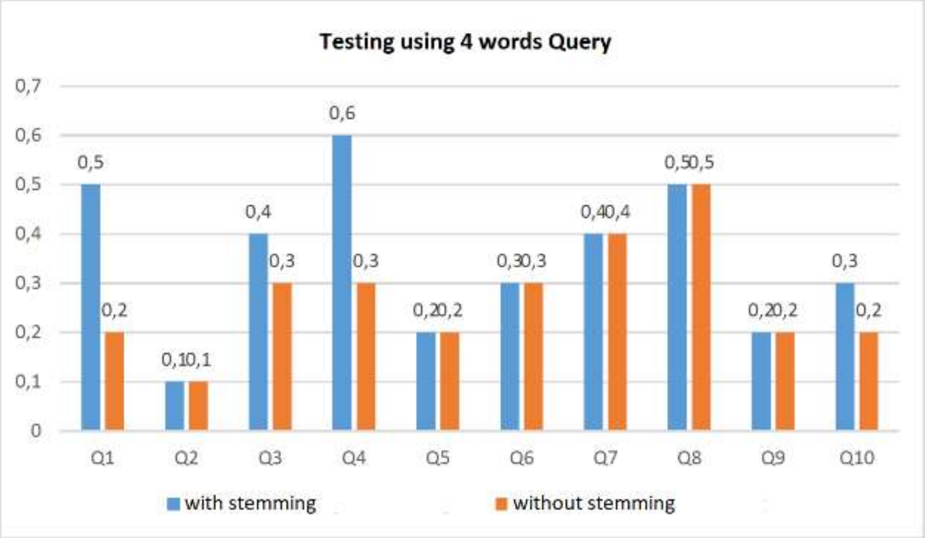

Figure 5. Testing using a 4-word query with and without stemming

Some of the tests above show that the precision values for searches involving stemming and without stemming are not far apart, and it can be said that this precision value is quite low for a production information retrieval system. In the case of this Madura tourism news corpus, several reasons make the precision value low. First, the query entered by the user is too general and less unique. Second, only 200 news collections are stored in the corpus, which is taken on the portalmadura.com website and is not specifically only about Madura tourism, resulting in a low precision average. Third, the limited number in the news itself so that the variation of articles in matching queries with documents becomes monotonous.

Google Translate API testing. This test was conducted to determine the value of the accuracy of Indonesian to English translation. The results of the expert assessment stated that the average value of accuracy was 67% which could be categorized as quite good if judged in general because most of the news had not been written using good grammar rules.

The time required for searching or calculating the similarity between queries and documents and then ranking the results depends on the number of documents in the corpus and the number of documents that need to be involved in computing cosine similarity. The longer the query, the more likely there will be more documents that intersect with the words in the query, meaning that the computation time for a 4-word query is longer than for a 3-word and 2-word query. The tests carried out took almost 6 seconds for computing a 2-word query, 10 seconds for a 3-word query, and 12 seconds for a 4-word query. It looks like there is a proportional increase in time. These figures seem quite large for a total of 200 tourism news documents. This happens because all these text documents are stored in the MySQL database server, where each read (select) operation requests access to the database server. This time value can be minimized by making some improvements to the search engine system, such as putting most of the database records in memory (RAM) so that it works as a database in memory. Another approach is to apply the concept of a cache where the Query that has been entered by the user will not experience computation but will directly retrieve the results of the previous computation. The next approach is to compare queries with queries, where queries from users do not actually experience similarity calculations with document collections in real-time. Similar queries will get the same response or answer from the SEBI system, and the new query will be used for behind-the-scenes computation and then put into the cache.

The tests carried out on SEBI above only involved 10 users who entered 2, 3, or 4 words as a representation of the Query. This number is less representative of testing a real search engine. The number of news documents that are only 200 can also be considered a weakness. Over time, SEBI will continue to compile tourism news and will involve many users in its testing. This search engine is intended for research and teaching, where every component and process in it is studied, imitated, and further developed.

We have seen that translating corpus and queries into English can provide a significant increase in precision, so this approach is very important in a search engine system like SEBI. Even though SEBI is devoted to handling Indonesian documents, it is almost difficult to deny the presence of English document that discusses Indonesia. On the other hand, various sophisticated tools, libraries, and algorithms in natural language technology focus more on English documents. Modern approaches such as Transformer and BERT have been intended for English documents from the start. Several libraries have been present to handle Indonesian language documents but are still unable to provide better accuracy. Research related to Indonesian language technology must continue to be carried out while using the library for English for study material and application development in a shorter time.

Research on SEBI is still ongoing. Several improvements and comparison methods are continuously implemented. Text translation into English or other languages can use approaches other than the Google Translate API. Language detection certainly needs to be done so as not to translate documents that have been written in English. Other approaches, such as spelling correction, classification of news and reviews based on sentiment, recommendations, and caching, are also of concern. In 2022 and 2023, SEBI will focus on document classification based on sentiment, where only documents with positive sentiment will be given to users, while negative documents will be used as input to tourism destination managers or local governments. There are many methods for sentiment-based classification, but so far, translating documents into English has proven to be able to increase the accuracy of the classification with a shorter system development time.

CONCLUSION

The use of the Google Translate API to present CLIR on the SEBI Search Engine can run well. Some tests show that the presence of stem is not able to significantly increase the precision of the similarity between Query and tourism news. SEBI precision when using stemming is 39%, and without stemming is 36%. This API-based document translation from Google can achieve 67% accuracy. This deficiency is caused by the lack of documents in the corpus, the distribution of tourism news that does not focus on the Madura destination, queries from users who are not domain-specific, and news writing that does not follow grammatical rules. The large amount of time required in computing is mainly because the documents are stored on the database server, and no caching technique or similar has been implemented.

ACKNOWLEDGMENTS

To the Institute for Research and Community Service (LPPM) University of Trunojoyo Madura (UTM), which has funded this tourism news information retrieval research through an Independent Research grant for the 2022 Research Group category.

REFERENCES

[1] K. Kayode and E. Ayetiran, “Survey on cross-lingual information retrieval,” *Int. J. Sci. Eng. Res*, vol. 9, pp. 484–491, 2018.

[2] S. Vaishnavi, “Survey on Variants of Cross-Language Information Retrieval,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 6, no. 1, pp. 167–170, 2018.

[3] P. Bajpai, P. Verma, and S. Q. Abbas, “English-Hindi Cross Language Information Retrieval System: Query Perspective.,” *J.*

Comput. Sci., vol. 14, no. 5, pp. 705–713, 2018.

[4] J. A. Hugh, E. Williams, and S. M. M. Tahaghoghi, “Stemming Indonesian language,” in *28th Australasian Computer Science Conference(ACSC2005), Conferences in Research and Practice in Information Technology*, 2005, vol. 38, pp. 1–8.

[5] P. M. Prihatini, I. K. G. D. Putra, I. A. D. Giriantari, and M. Sudarma, “Stemming Algorithm for Indonesian Digital News Text Processing,” *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 2, pp. 1–7, 2018.

[6] R. K. Hapsari and Y. J. Santoso, “Stemming Artikel Berbahasa Indonesia Dengan Pendekatan Confix-Stripping,” in *Prosiding Seminar Nasional Manajemen Teknologi XXII*, 2015, pp. 1–8.

[7] D. O. Baskoro, H. Malik, and M. H. Anshari, “Porter Stemmer Information Retrieval,” *Comput. Sci. Gadjah Mada Univ.*, 2012.

[8] M. Alif, F. Solihin, and H. Husni, “Perbandingan Metode Enhanced Confix Stripping dan Porter Stemmer Untuk Stemming Konten Bahasa Indonesia,” 2014.

[9] R. Melita, “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: Hadits Shahih Bukhari-Muslim),” Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, 2018.

[10] A. A. Maarif, “Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah,” *Dok. Karya Ilmiah| Tugas Akhir| Progr. Stud. Tek. Inform. Fak. Ilmu Komputer| Univ. Dian Nuswantoro Semarang*, vol. 5, no. 4, 2015.

[11] R. Prasath and S. Sarkar, “Cross-Language Information Retrieval with Incorrect Query Translations,” *Polibits*, no. 54, pp. 33–42, 2016.

[12] S. Napitupulu, “Analyzing Indonesian-English abstracts translation in view of translation errors by Google Translate,” *Int. J. English Lang. Linguist. Res.*, vol. 5, no. 2, pp. 15–23, 2017.

[13] H. Husni, I. O. Suzanti, Y. D. Pramudita, S. S. Putro, and L. Heryawan, “Web Service for Search Engine Bahasa Indonesia (SEBI),” in *Journal of Physics: Conference Series*, 2020, vol. 1569, no. 2, p. 22087.

[14] H. W. A. Kesuma and F. S. Pribadi, “Penerapan Cosine Similarity dalam Aplikasi Kitab Undang-Undang Hukum Dagang (Wetboek Van Koophandle Voor Indonesia),” *J. Tek. Elektro*, vol. 8, no. 1, pp. 18–20, 2016.

[15] M. Saravanan and K. Sathish, “Tamil to English Cross Lingual Information Retrieval System for Agricultural Domain Using VSM.”

[16] P. Bhattacharya, P. Goyal, and S. Sarkar, “Query translation for cross-language information retrieval using multilingual word clusters,” in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, 2016, pp. 152–162.

[17] A. J. Agrawal, “Cross Language Information Retrieval using Selective Documents Technique and Query Expansion,” 2018.

[18] J. Vembunarayanan, “Tf-idf and cosine similarity.” 2013.

[19] Y. Rajanak, R. Patil, Y.P. Singh, "Language Detection Using Natural Language Processing" in 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 2023