International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 05, Issue 02, May 2024
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

A Query Expansion Using Support Vector Machine (SVM) and Best Matching 25 (BM25)

Muhadi
Department of Master of Science and Information Technology
President University
Bekasi, Indonesia
muhadi@student.president.ac.id

Abstract - The information retrieval system needs to be upgraded constantly to improving retrieval performance. Expansion Query is one of the information retrieval methods for retrieve more relevant documents. Several technique expansion queries are synonym expansion, Thesaurus Expansion, and Word Embedding Expansion. This research uses a Support Vector Machine (SVM) to get new terms from the corpus. Some people are sometimes confused about what must to write to get the desired document or are too lazy to write a lot of words. In this research, Based on the input query data SVM will search for condition data from the existing corpus so that the data obtained will later become an expansion for the query. Adding additional terms to capture a broader range of relevant documents will retrieve more documents and improve the relevance of search results.

***Keyword**: Information Retrieval, Query Expansion, Support Vector Machine, Best Matching 25*

I. INTRODUCTION

Along with advances in internet technology today, we have an abundance of new document sources. In terms of document types, the internet also provides various choices of document types including multimedia elements. With so many documents whose data would be very difficult if document searches were done manually, users need more sophisticated tools to find their relevant information. Therefore, a number of language technologies are used in various information management applications of multilingual search engines.The main problem of this article is trying to provide an alternative document search engine through expansion queries using a Support Vector Machine. It is believed that this technique will be able to improve the relevance of search results.

Query Expansion (QE) is a process in Information Retrieval that consists of selecting and adding terms to the user's query with the goal to capture more relevant documents and thereby improving retrieval performance.

Several studies have focused on A Query Expansion Method Using a Machine Learning algorithm. Research [1] Use Multinomial Naive Bayes to get additional terms. Researh[2][3] has studied and surveyed about Expansion Query. This research using a Support Vector Machine (SVM) for get new terms and Best Matching 25 (BM25) for Document retrieval.

II. RESEARCH METHODS

1.1 Diagram

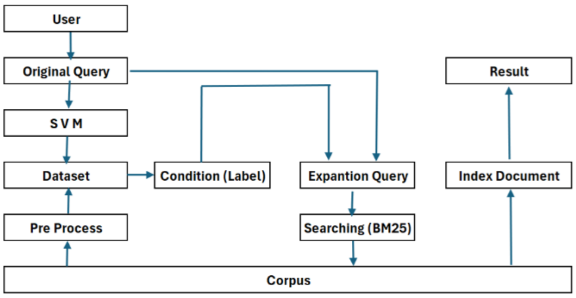
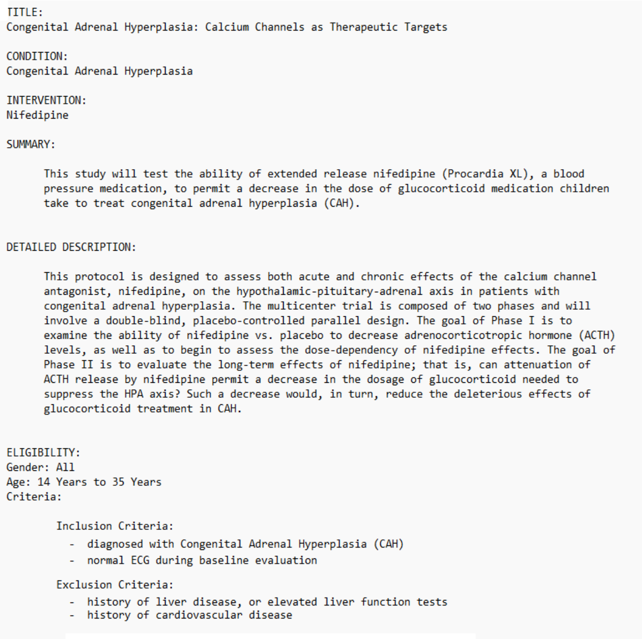


Figure 1. Diagram

2.2 Corpus

In this research, we use corpus data from http://www.trec-cds.org/2017. The corpus consists of a set of clinical trials, on text format and the document contains a title, condition, intervention, summary, details description, and eligibility (gender, age, criteria).

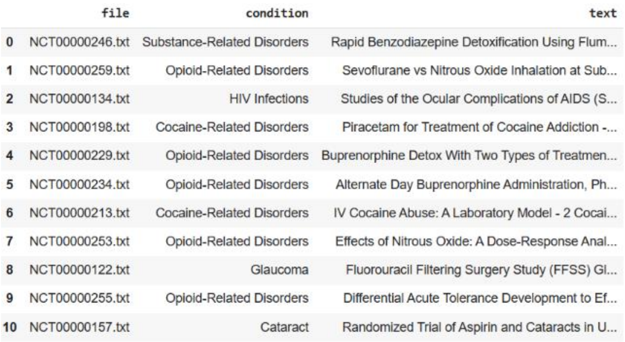


```
TITLE:
Congenital Adrenal Hyperplasia: Calcium Channels as Therapeutic Targets

CONDITION:
Congenital Adrenal Hyperplasia

INTERVENTION:
Nifedipine

SUMMARY:

      This study will test the ability of extended release nifedipine (Procardia XL), a blood
      pressure medication, to permit a decrease in the dose of glucocorticoid medication children
      take to treat congenital adrenal hyperplasia (CAH).


DETAILED DESCRIPTION:

      This protocol is designed to assess both acute and chronic effects of the calcium channel
      antagonist, nifedipine, on the hypothalamic-pituitary-adrenal axis in patients with
      congenital adrenal hyperplasia. The multicenter trial is composed of two phases and will
      involve a double-blind, placebo-controlled parallel design. The goal of Phase I is to
      examine the ability of nifedipine vs. placebo to decrease adrenocorticotropic hormone (ACTH)
      levels, as well as to begin to assess the dose-dependency of nifedipine effects. The goal of
      Phase II is to evaluate the long-term effects of nifedipine; that is, can attenuation of
      ACTH release by nifedipine permit a decrease in the dosage of glucocorticoid needed to
      suppress the HPA axis? Such a decrease would, in turn, reduce the deleterious effects of
      glucocorticoid treatment in CAH.


ELIGIBILITY:
Gender: All
Age: 14 Years to 35 Years
Criteria:

        Inclusion Criteria:
          -  diagnosed with Congenital Adrenal Hyperplasia (CAH)
          -  normal ECG during baseline evaluation

        Exclusion Criteria:
          -  history of liver disease, or elevated liver function tests
          -  history of cardiovascular disease
```

Figure 2. Corpus

2.3 Dataframe

For an easy process in machine learning, need to convert data from text file to data frame format. The data frame format will be easy for reading, filtering, and processing in machine learning applications. The data will be split only into 3 fields: file_name, condition, and text. A text field is a result of combining several sections of data in a text file: data title, condition, intervention, summary, and description. Figure 3 shows the dataset in data frame format.

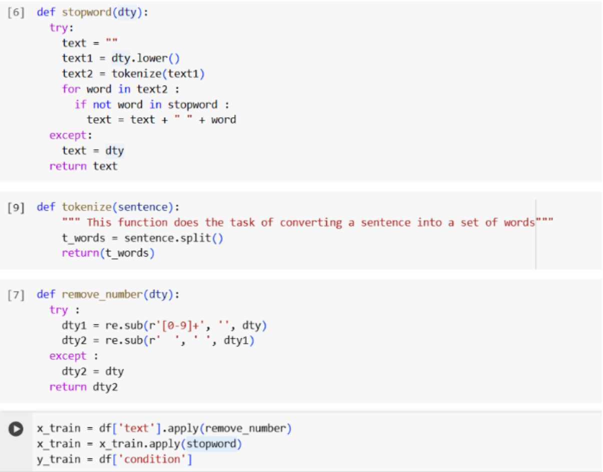


	file	condition	text
0	NCT00000246.txt	Substance-Related Disorders	Rapid Benzodiazepine Detoxification Using Flum...
1	NCT00000259.txt	Opioid-Related Disorders	Sevoflurane vs Nitrous Oxide Inhalation at Sub...
2	NCT00000134.txt	HIV Infections	Studies of the Ocular Complications of AIDS (S...
3	NCT00000198.txt	Cocaine-Related Disorders	Piracetam for Treatment of Cocaine Addiction -...
4	NCT00000229.txt	Opioid-Related Disorders	Buprenorphine Detox With Two Types of Treatmen...
5	NCT00000234.txt	Opioid-Related Disorders	Alternate Day Buprenorphine Administration, Ph...
6	NCT00000213.txt	Cocaine-Related Disorders	IV Cocaine Abuse: A Laboratory Model - 2 Cocai...
7	NCT00000253.txt	Opioid-Related Disorders	Effects of Nitrous Oxide: A Dose-Response Anal...
8	NCT00000122.txt	Glaucoma	Fluorouracil Filtering Surgery Study (FFSS) Gl...
9	NCT00000255.txt	Opioid-Related Disorders	Differential Acute Tolerance Development to Ef...
10	NCT00000157.txt	Cataract	Randomized Trial of Aspirin and Cataracts in U...

Figure 3. Data frame format

2.4 Text Pre-Processing

The characteristics of a corpus sometimes use unusual language, various abbreviations, and certain characters. so the data obtained is still in unstructured form. The system will experience difficulty analyzing data like that, pre-processing steps need to be carried out. Likewise, in this research pre-processing steps were carried out. The first is to change all letters to lowercase. Then remove stop words, hyperlinks, non-alpha characters, and other words that have no meaning.



```
[6] def stopword(dty):
      try:
        text = ""
        text1 = dty.lower()
        text2 = tokenize(text1)
        for word in text2 :
          if not word in stopword :
            text = text + " " + word
      except:
        text = dty
      return text
```

```
[9] def tokenize(sentence):
        """ This function does the task of converting a sentence into a set of words"""
        t_words = sentence.split()
        return(t_words)
```

```
[7] def remove_number(dty):
      try :
        dty1 = re.sub(r'[0-9]+', '', dty)
        dty2 = re.sub(r'  ', ' ', dty1)
      except :
        dty2 = dty
      return dty2
```

```
x_train = df['text'].apply(remove_number)
x_train = x_train.apply(stopword)
y_train = df['condition']
```

Figure 4. Text Preprocessing

2.5 Best Matching 25 (BM25) Scoring

Best Matching 25 is a ranking system that can sort the match results for a document. This ranking function is commonly used in information retrieval systems to rank documents based on their relevance to a particular query. BM25 is an extension of the TF-IDF weighting scheme.

2.5.1 Term Frequency (TF) Calculation

BM25 calculates the term frequency (TF) for each term in each document. Unlike standard TF in TF-IDF, BM25 uses a saturation function to dampen the effect of excessively high term frequencies.

The TF component in BM25 is calculated using the formula:

$$TF(t,d) =$$
$$(k+1)\times tf(t,d)k\times(1-b+b\times|d|avg\_dl)+tf(t,d)$$
$$TF(t,d) = k\times(1-b+b\times avg\_dl|d|)+tf(t,d)(k+1)\times tf(t,d)$$

$tf(t,d)$tf(t,d) is the raw term frequency of term tt in document dd.
$|d|$|d| is the length of document dd.
avg_dlavg_dl is the average document length in the corpus.
kk and bb are tuning parameters that control the scaling and normalization of term frequencies. Typically, kk is set to a value such as 1.2 and bb to a value such as 0.75.

2.5.2 Inverse Document Frequency (IDF) Calculation

Like TF-IDF, BM25 also incorporates IDF to weigh down terms that occur in many documents. However, unlike TF-IDF, BM25 uses a simplified IDF calculation without the logarithm.
The IDF component in BM25 is calculated as:

$IDF(t)=log\ \ (N-n(t)+0.5n(t)+0.5)IDF(t)=log(n(t)+0.5N-n(t)+0.5)$

NN is the total number of documents in the corpus.
$n(t)$n(t) is the number of documents containing term tt.

2.5.3 Query Term Weighting

After calculating TF and IDF for each term in each document, BM25 calculates a score for each document based on how well it matches the query. For a given query qq, the score of document dd is calculated as the sum of the weights of its terms that match the query terms.

The final score for document dd is calculated as:
$score(d,q)=\sum t\in qTF(t,d)\times IDF(t)score(d,q)=\sum t\in q\ TF(t,d)\times IDF(t)$

2.5.4 Ranking

Finally, the documents are ranked based on their scores. The documents with higher scores are considered more relevant to the query and are typically presented to the user first in the search results.

Overall, BM25 is a powerful ranking function that takes into account both term frequency and document length normalization, making it suitable for ranking documents in information retrieval systems. Adjusting the tuning parameters kk and bb allows fine-tuning of the ranking algorithm based on the characteristics of the corpus and the specific requirements of the application.

2.6 Support Vector Machine (SVM)

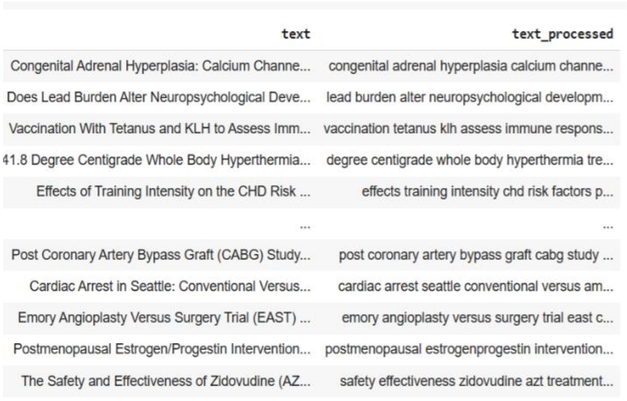
This research uses a Support Vector Machine with C parameter = 1 and linear kernel. For SVM text classification setting C=1 is common, and the result for this setting is often good. This setting represents a balanced trade-off between achieving a low training error and a low testing error.

Linear kernels are popular choices and generally faster to train compared to more complex kernels like RBF. Linear kernel can work well for text classification especially when using techniques like TF-IDF or word embeddings, the data is often high-dimensional.

III. RESULT

3.1 Text Preprocessing and SVM Accuracy

After going through text processing, some characters and words from the collection of documents will be lost, including special characters / strange characters, numbers, and words that are considered unimportant in the stopword list. Figure 5 shows the difference between the original text data and the text after text processing. You can see in the picture the comparison between the original test and the preprocessed text. If we compare, there are several characters missing, there are even words removed too. Open-close bracket characters and numbers are not visible in the preprocessed results. Irrelevant words are also lost such as 'and', 'of', and 'in'. This format provides a structured overview of the NLP processing results, making it easy to understand and interpret the findings.



text	text_processed
Congenital Adrenal Hyperplasia: Calcium Channe...	congenital adrenal hyperplasia calcium channe...
Does Lead Burden Alter Neuropsychological Deve...	lead burden alter neuropsychological developm...
Vaccination With Tetanus and KLH to Assess Imm...	vaccination tetanus klh assess immune respons...
41.8 Degree Centigrade Whole Body Hyperthermia...	degree centigrade whole body hyperthermia tre...
Effects of Training Intensity on the CHD Risk ...	effects training intensity chd risk factors p...
...	...
Post Coronary Artery Bypass Graft (CABG) Study...	post coronary artery bypass graft cabg study ...
Cardiac Arrest in Seattle: Conventional Versus...	cardiac arrest seattle conventional versus am...
Emory Angioplasty Versus Surgery Trial (EAST) ...	emory angioplasty versus surgery trial east c...
Postmenopausal Estrogen/Progestin Intervention...	postmenopausal estrogenprogestin intervention...
The Safety and Effectiveness of Zidovudine (AZ...	safety effectiveness zidovudine azt treatment...

Figure 5. Text Processed

This research uses Machine Learning with text classification Support Vector Machine algorithm, an accuracy value of 100% was obtained from the data train and 58% for the data test. This accuracy was obtained from a dataset of 274 documents.

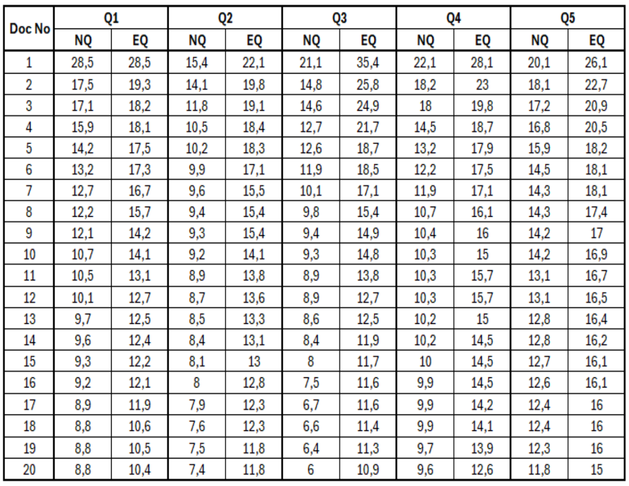
3.2 BM25 Scoring

Results of this research, the BM25 score value of the document query results has increased both in terms of the score of each document and the number of documents that have a score > 0. Table 1 shows comparison of the number of documents that have a score between normal query (NQ) and expansion query (EQ). All experiments show that the number of documents that have a score for Expansion Query has increased until 67 %.

Table 1. The number of documents have a score

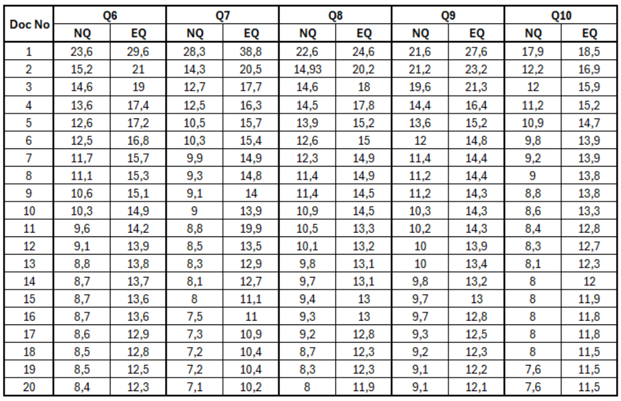
The number of documents have a score				
No	**Query**	**NQ**	**EQ**	**%**
1	'impaired glucose (carbohydrate) tolerance'	165	276	67,2
2	'reductions in central vision'	163	172	5,5
3	'risk factors for cataract'	100	182	82,0
4	'effectiveness of buprenorphine'	230	286	24,3
5	'effects of nitrous oxide in humans'	233	287	23,2
6	'effects of treatment medications'	272	294	8,1
7	'role cotinine plays in nicotine addiction'	164	210	28,0
8	'clinical protocol to detoxify patients'	193	271	40,4
9	'tablet buprenorphine formulations'	211	291	37,9
10	'evaluate the efficacy of desipramine'	227	281	23,7

Tables 2 and 3 show the top 20 score values from the query results both in normal query (NQ) and in Extension Query (EQ). Table 4 shows that in all experiments top 20 query scores increased until 50%.



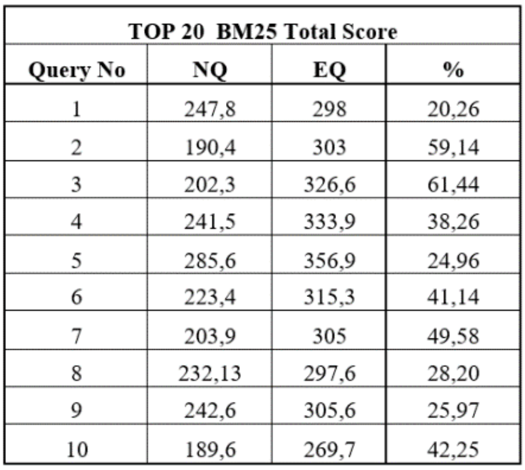
Doc No	Q1		Q2		Q3		Q4		Q5	
	NQ	EQ	NQ	EQ	NQ	EQ	NQ	EQ	NQ	EQ
1	28,5	28,5	15,4	22,1	21,1	35,4	22,1	28,1	20,1	26,1
2	17,5	19,3	14,1	19,8	14,8	25,8	18,2	23	18,1	22,7
3	17,1	18,2	11,8	19,1	14,6	24,9	18	19,8	17,2	20,9
4	15,9	18,1	10,5	18,4	12,7	21,7	14,5	18,7	16,8	20,5
5	14,2	17,5	10,2	18,3	12,6	18,7	13,2	17,9	15,9	18,2
6	13,2	17,3	9,9	17,1	11,9	18,5	12,2	17,5	14,5	18,1
7	12,7	16,7	9,6	15,5	10,1	17,1	11,9	17,1	14,3	18,1
8	12,2	15,7	9,4	15,4	9,8	15,4	10,7	16,1	14,3	17,4
9	12,1	14,2	9,3	15,4	9,4	14,9	10,4	16	14,2	17
10	10,7	14,1	9,2	14,1	9,3	14,8	10,3	15	14,2	16,9
11	10,5	13,1	8,9	13,8	8,9	13,8	10,3	15,7	13,1	16,7
12	10,1	12,7	8,7	13,6	8,9	12,7	10,3	15,7	13,1	16,5
13	9,7	12,5	8,5	13,3	8,6	12,5	10,2	15	12,8	16,4
14	9,6	12,4	8,4	13,1	8,4	11,9	10,2	14,5	12,8	16,2
15	9,3	12,2	8,1	13	8	11,7	10	14,5	12,7	16,1
16	9,2	12,1	8	12,8	7,5	11,6	9,9	14,5	12,6	16,1
17	8,9	11,9	7,9	12,3	6,7	11,6	9,9	14,2	12,4	16
18	8,8	10,6	7,6	12,3	6,6	11,4	9,9	14,1	12,4	16
19	8,8	10,5	7,5	11,8	6,4	11,3	9,7	13,9	12,3	16
20	8,8	10,4	7,4	11,8	6	10,9	9,6	12,6	11,8	15

Table 02. Top 20 BM25 Score(Query no 1 ~ 5)



Doc No	Q6		Q7		Q8		Q9		Q10	
	NQ	EQ	NQ	EQ	NQ	EQ	NQ	EQ	NQ	EQ
1	23,6	29,6	28,3	38,8	22,6	24,6	21,6	27,6	17,9	18,5
2	15,2	21	14,3	20,5	14,93	20,2	21,2	23,2	12,2	16,9
3	14,6	19	12,7	17,7	14,6	18	19,6	21,3	12	15,9
4	13,6	17,4	12,5	16,3	14,5	17,8	14,4	16,4	11,2	15,2
5	12,6	17,2	10,5	15,7	13,9	15,2	13,6	15,2	10,9	14,7
6	12,5	16,8	10,3	15,4	12,6	15	12	14,8	9,8	13,9
7	11,7	15,7	9,9	14,9	12,3	14,9	11,4	14,4	9,2	13,9
8	11,1	15,3	9,3	14,8	11,4	14,9	11,2	14,4	9	13,8
9	10,6	15,1	9,1	14	11,4	14,5	11,2	14,3	8,8	13,8
10	10,3	14,9	9	13,9	10,9	14,5	10,3	14,3	8,6	13,3
11	9,6	14,2	8,8	19,9	10,5	13,3	10,2	14,3	8,4	12,8
12	9,1	13,9	8,5	13,5	10,1	13,2	10	13,9	8,3	12,7
13	8,8	13,8	8,3	12,9	9,8	13,1	10	13,4	8,1	12,3
14	8,7	13,7	8,1	12,7	9,7	13,1	9,8	13,2	8	12
15	8,7	13,6	8	11,1	9,4	13	9,7	13	8	11,9
16	8,7	13,6	7,5	11	9,3	13	9,7	12,8	8	11,8
17	8,6	12,9	7,3	10,9	9,2	12,8	9,3	12,5	8	11,8
18	8,5	12,8	7,2	10,4	8,7	12,3	9,2	12,3	8	11,5
19	8,5	12,5	7,2	10,4	8,3	12,3	9,1	12,2	7,6	11,5
20	8,4	12,3	7,1	10,2	8	11,9	9,1	12,1	7,6	11,5

Table 3. Top 20 BM25 Score(Query no 6 ~ 10)



TOP 20 BM25 Total Score			
Query No	**NQ**	**EQ**	**%**
1	247,8	298	20,26
2	190,4	303	59,14
3	202,3	326,6	61,44
4	241,5	333,9	38,26
5	285,6	356,9	24,96
6	223,4	315,3	41,14
7	203,9	305	49,58
8	232,13	297,6	28,20
9	242,6	305,6	25,97
10	189,6	269,7	42,25

IV. CONCLUSION AND FUTURE WORK

Table 1 shows that the number of documents has increased, this means that the results obtained with expansion queries are in accordance with their objectives. Query expansion techniques aim to capture more relevant documents

Tables 2 and 3 show that the score value for each document has increased, improving the relevance of search results

The increase in the number of documents and score values shows that the method of using SVM has been successfully used for expansion queries which can improve the relevance of documents and effectiveness of search results.

This research only uses 274 documents which consist of a set of clinical trials, and need to be tried with more datasets and with other themes.

REFERENCES

[1]. Silva, Sergio, Adrián Seara Vieira, Pedro Celard, Eva Lorenzo Iglesias, and Lourdes Borrajo. 2021. "*A Query Expansion Method Using Multinomial Naive Bayes*." Applied Sciences (Switzerland)

[2]. Afuan, Lasmedi, Ahmad Ashari, and Yohanes Suyanto. 2019. "*A Study: Query Expansion Methods in Information Retrieval*." Journal of Physics: Conference Series 1367(1).

[3]. Azad, Hiteshwar Kumar, and Akshay Deepak. 2019. "*Query Expansion Techniques for Information Retrieval: A Survey*." Information Processing and Management 56(5): 1698–1735.

[4]. Blanco, Roi, and Paolo Boldi. 2012. "*Extending BM25 with Multiple Query Operators.*" SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval: 921–30.

[5]. Hou, Fenglei, and Bingxi Wang. 2001. "*Text-Independent Speaker Recognition Using Support Vector Machine*." 2001 International Conferences on Info-Tech and Info-Net: A Key to Better Life, ICII 2001 - Proceedings 3(306): 402–7.

[6]. Hwang, Young Sup. 2014. "*Wrapper-Based Feature Selection Using Support Vector Machine*." Life Science Journal 11(7): 632–36.

[7]. Maier, J., and K. Ferens. 2009. "*Classification of English Phrases and SMS Text Messages Using Bayes and Support Vector Machine Classifiers*." Canadian Conference on Electrical and Computer Engineering: 415–18.

[8]. Ramadhani, P. P., and S. Hadi. 2021. "*Text Classification on the Instagram Caption Using Support Vector Machine*." Journal of Physics: Conference Series 1722(1).

[9]. Svore, Krysta M., and Christopher J.C. Burges. 2009. "*A Machine Learning Approach for Improved BM25 Retrieval*." International Conference on Information and Knowledge Management, Proceedings: 1811–14.

[10]. Whissell, John S., and Charles L.A. Clarke. 2011. "*Improving Document Clustering Using Okapi BM25 Feature Weighting*." Information Retrieval 14(5): 466–87.