Journal of Theoretical and Applied Information Technology

15th November 2023. Vol.101. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THESAURUS-BASED QUERY EXPANSION ON INFORMATION RETRIEVAL TO IMPROVE THE QUALITY OF DOCUMENT SEARCHING RESULT

YIYI SUPENDI[1] , ERWIN YULIANTO[2] , DESHINTA ARROVA DEWI[3*] , KM SYARIF HARYANA[4]

[1, 2, 4]Senior Lecturer, Langlangbuana University, Department of Informatics, Indonesia

[1,3]Senior Lecturer, INTI International University, Faculty of Data Science and Information Technology, Malaysia

[*]Corresponding Author

E-mail: [1]yiyi.supendi@gmail.com, [2]rwinyulianto@yahoo.com, [3]deshinta.ad@newinti.edu.my, [4]kmsyarif@gmail.com

ABSTRACT

With the role of the internet as an unlimited source of information originating from various people around the world with a variety of languages and various forms of delivery such as text documents, images, audio, or video. Information can be accessed wherever we are in real-time, anywhere, and any place. A search Engine is a medium that is used in finding various information on the internet. The phenomenon of problems that arise in the search for information on the internet is the search results of the desired topic are often not relevant to the keywords entered. Information Retrieval is a system, method, and procedure used to recover information stored from a collection of information based on a query entered by the user. The performance of searches based on Information Retrieval can be improved by various methods. One of them uses the Query Expansion method, which works by reformulating the initial query by adding several terms using Thesaurus to the query entered so that it is expected to be able to increase the relevance of data obtained from search results. The contribution obtained from this research is the development of an Information Retrieval System-based document search engine that can improve data relevance and the quality of document search results. Based on the test results using the recall and precision methods, a graph was obtained showing that the relevance results with the Query Expansion method had increased in the level of data relevance. Because when using expansion queries, the recall results are higher, allowing more relevant documents to be retrieved.

Keywords: *Information Retrieval, Query Expansion, Thesaurus, Data Revelation, Process Innovation*

1. INTRODUCTION

As the development of communication and information technology is increasingly advanced and globalization is increasingly widespread, resulting in this world has become a global village that can not be separated from the need for knowledge. Science and technology need information, but also at the same time produce information. The internet is a medium that is used by many people regardless of status, age, and education and is an unlimited source of information originating from various people around the world in various languages and various forms of delivery such as text documents, images, audio or video. Many systems have been built that operate in real-time and online, which allows one to access data from anywhere and get the latest information. Information can be accessed wherever we are in real-time, anywhere, and anyplace. A search Engine is a medium that is used in finding various information on the internet. [1]

The world of business is full of competition making the perpetrators must always think about breakthrough strategies that can guarantee the continuity of their business. Information search is one business strategy that is very helpful in decision-making [2]. The problem that often arises in searching for information on the internet is when a user searches for a particular topic and the search results of the de-sired topic are often not relevant to the keywords entered. Especially when searching for topics that contain a lot of unfamiliar terms whose meaning is unknown, of course this is very inhibiting in searching for certain topics. When foreign terms

are translated directly into Indonesian and entered as keywords. Often the results obtained are highly irrelevant.

This problem is one of the challenges in Information Retrieval (IR) which can be solved through various approaches and research. A lot of research has been carried out to overcome this problem so that it is in line with the research topic that will be discussed, including the use of a thesaurus for query expansion, namely that information search systems can use a thesaurus or synonym dictionary to expand keywords so that it can help in overcoming problems when keywords are used. foreign languages are translated into Indonesian, but the results are not relevant. Using relevant synonyms from Indonesian can increase the relevance of search results. Other research studies the use of Word Embeddings for Multilingual Search, where word embeddings are vector representations of words in multidimensional space. These studies try to overcome the problem of irrelevance of search results when faced with foreign terms that are difficult to understand by using various approaches involving language processing techniques to improve the quality of search results. [2, 9, 10, 11, 12, 15, 19, 20, 21, 22, 23, 24]

For example, searching for research based on the use of information communication and technology keywords in research titles in Indonesia has increased over time. ICT research is more dominant in education and subsequently in management. [3]. Likewise, a search using the keyword "smartphone restaurant" produces articles and is not directly related to the food ordering service, most of the articles pay attention to psychological aspects of consumer behavior. Studies resulting from the various searches above are more likely to give another aspect of how technology affects consumer behavior [4],[5],[6], [7],[8].

Information Retrieval System is a system, method, and procedure used to recover information stored from a collection of information based on a query entered by the user. In the book "Information Storage and Retrieval Systems Theory and Implementation", an information retrieval system is a system capable of storing, searching, and maintaining information [9]. Unfortunately, often the search results from the Information Retrieval System are not optimal because they only compare queries with documents at the word or sentence level rather than at the semantic level [10].

Based on the background of the above research, the formulation of the problems that will be reviewed and analyzed include:

a. How to build a reliable document search engine system?
b. How do increase the level of relevance of the data obtained if keywords are using foreign languages?

The research objectives to be achieved based on the identified mask formula are:

a. Creating an Information Retrieval System that can increase the relevance of data and document searching result quality.
b. Implement Expansion Queries on Thesaurus to expand additional terms, phrases, or words related to foreign synonyms.

2. LITERATURE REVIEW

2.1 Information Retrieval

ISO 2382/1 issued in 1984 defines Information Retrieval as actions, methods, and procedures for recovering stored data, then providing information about the subjects needed. Based on this standard, these actions include text indexing, inquiry analysis, and also relevance analysis. Data includes text, tables, images, speech, and videos, and information including related knowledge needed to support problem-solving and knowledge acquisition [11].

The purpose of Information Retrieval is to meet the information needs of users by getting back all relevant documents and at the same time getting the minimum amount of irrelevant documents. This system uses the heuristic function to get documents relevant to the keywords entered by users. A good IR system allows users to determine quickly and accurately whether the contents of the document are satisfactory. Only relevant documents must be returned based on user queries. For better representation of documents, documents with similar topics are grouped [12].

This Information Retrieval consists of several steps, namely Text Operation, Query Formulation, Document Indexing, and Document Searching. In information retrieval, there are various methods used in word weighting, measurement of suitability, ranking, the relevance of feedback, and others. In the context of Information Retrieval evaluation, the methods used are recall and precision. The recall is a comparison of the number

of relevant documents taken per a given query with the total collection of documents relevant to the query [13]. Precision is a comparison of the number of documents relevant to a query with the number of documents taken from search results. Precision can be interpreted as the accuracy or compatibility between the request for information with the answer to the request, while the term recall relates to the ability to recover information that has been stored [14].

The recall is stated as part of the relevant document in the document found, as Equation 1 follows.

$$\text{Recall} = \frac{\text{the Number of Relevant Documents Found}}{\text{the Sum of All Relevant Documents}} \quad (1)$$

Furthermore, precision is stated as part of the relevant document found with formulas as in Equation 2 below.

$$\text{Precision} = \frac{\text{the Number of Relevant Documents Found}}{\text{the Total Number of Documents Found}} \quad (2)$$

Both describe the performance of the information retrieval system by calculating the number of relevant document search results. Measuring recall and precision is a calculation performed on a collection of documents search results (set-based measure) as a whole. The size of both is usually given in the form of a percentage value of 1 to 100%. If the level of recall and precision is high then the information retrieval system can be considered good.

2.2 Query Expansion Using Thesaurus

The expansion query applied in this case is using the expansion of the query in the Thesaurus. In this IRS there is a pre-processing stage first, where the document extension with *. Pdf will be converted into a Text (* .txt) file. Then the text operation process is performed using a stemming algorithm [15]. After going through the text data operation documents and terms will be indexed in the database using the Linked List Order method.

When the user performs a search process using certain keywords, the system will perform a series of text operations, then the search keywords are expanded by checking the Thesaurus data in the database whether it has similarities or other terms of the word. If you have an equation, the system will execute concurrently between the key-words that the user types and the keyword equation.

Data / Information modeling collected from business process modeling is used to define data objects needed for business [16]. In the process of searching the system will use the structure of the Hash Table with the inverted index data structure. Then before being displayed again, as a result, there is a final step where the data or documents are sorted by ranking the most relevant documents based on the query or keywords entered. The system testing process uses the Recall and Precision methods. The following Figure 1 is a flowchart of the Query Expansion that will be examined.

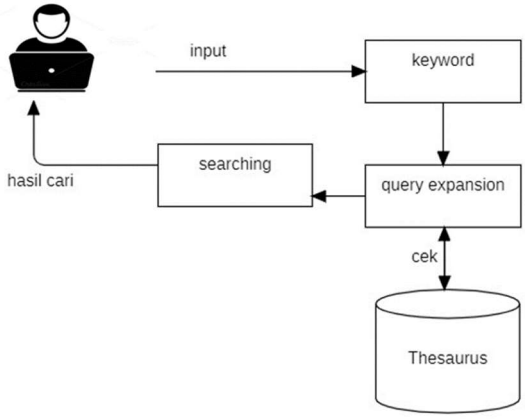

Figure 1: Query Expansion Flowchart

3. METHODOLOGY

The research method is the stages that will be carried out in the research. With systematic activity, the Software Development Life Cycle (SDLC) used in this study is the Waterfall model which is a linear-sequential life cycle model.

The Waterfall Model was the first Process Model to be introduced. In a waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases. In this waterfall model, the phases do not overlap. Some situations where the use of the Waterfall model is most appropriate are:

- ✓ Requirements are very well documented, clear, and fixed.
- ✓ Product definition is stable.
- ✓ Technology is not dynamic and understood.
- ✓ There are no ambiguous requirements.
- ✓ Ample resources with the required expertise are available to support the product.
- ✓ The project is short.

The following Figure 2 is an illustration is a representation of the different phases of the Waterfall Model.

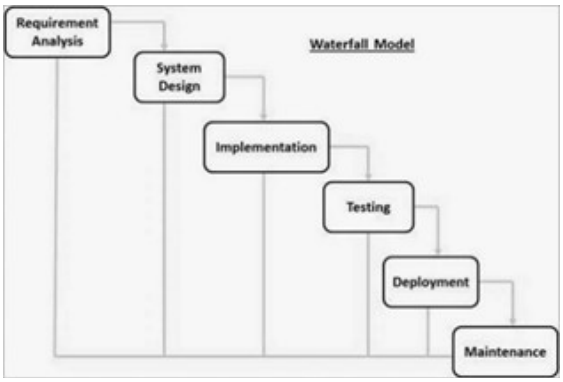

Figure 2: Waterfall Model [17]

The sequential phases in the Waterfall model are:

- ✓ Requirement Analysis, all possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
- ✓ System Design, the requirement specifications from the first phase are studied in this phase and the system design is prepared. This system design helps in specifying hard-ware and system requirements and helps in defining the overall system architecture.
- ✓ Implementation, with inputs from the system design, the system is first developed in small programs called units, which are integrated into the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.
- ✓ Testing, all the units developed in the implementation phase are integrated into a system after testing each unit.
- ✓ Deployment, once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.
- ✓ Maintenance, some issues come up in the client environment. To fix those issues, patches are released. Maintenance is done to deliver these changes in the customer environment.

4. RESULT AND DISCUSSION

4.1 System Flowchart

To obtain information, a system is needed to make it easier for users to search. Although in general the Information Retrieval System which is equipped with a text operation process can provide adequate results, it is not optimal when the keywords entered are in the form of short texts or foreign terms. Based on this, it is designed to add processes to the Information Retrieval in the form of a query expansion process to maximize search results, especially at the level of relevance of existing data search.

The analysis conducted in making the Information Retrieval System architecture in this study uses the flowchart as Figure 3 below.

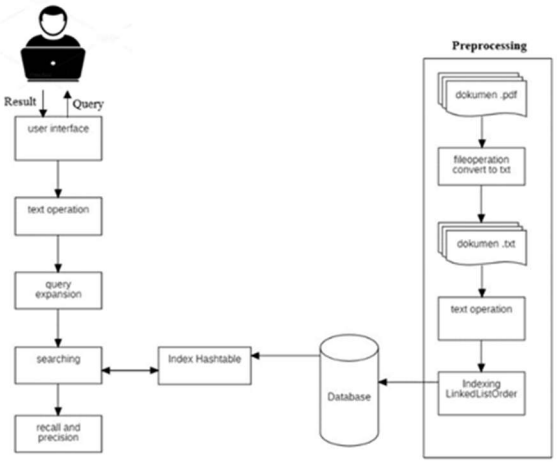

Figure 3: Information Retrieval System Flowchart

The function used during the searching process is AND - OR. For searching more than 1 word, AND for each word will be added AND for example the keyword "change design" to "change AND design". When going through the Expansion Query process, each of these words will first be searched for related equations or terms, then merged with the AND function which is added to the OR function, for example, the word "change" after checking in Thesaurus found the word "edit" and for "design "Found the word" design ". The results after going through the Expansion Query process to "change design", "edit design", "edit design", so that when the search process uses the AND - OR function (change AND design) OR (change AND design) OR (edit AND design) OR (edit design).

4.2 System Requirement Analysis

4.2.1 Use Case Diagram

In the Information Retrieval System, there are two actors involved namely the admin and the user. The following table 1 explains the roles of each actor:

Journal of Theoretical and Applied Information Technology
15th November 2023. Vol.101. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Actors Role.

Actors	Access Rights
Admin	Login Add Documents Add Term Expansion
User	Document Searching Download Document

The interaction between actors and the Information Retrieval System that will be developed can be seen in Figure 4 below.

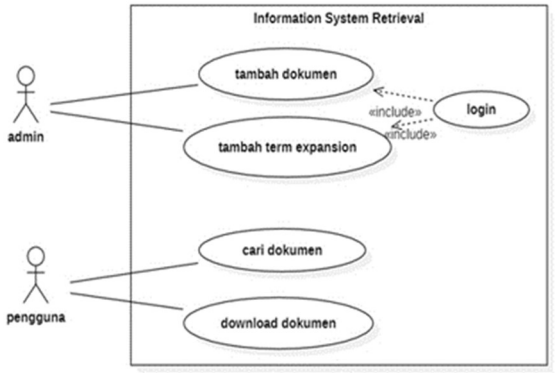

Figure 4: Use Case Diagram of Information Retrieval System

4.2.2 Class Diagram

The design of Class Diagrams on Information Retrieval System with Expansion Query will be divided into 4 big classes namely text operation, indexing, text formula-tion, and searching. The text operation process aims to reduce the complexity of the repression of raw documents and process data into index-ready terms [18]. Text opera-tion is performed on keywords entered by users and documents that have been up-loaded. Class Text Operation as shown in Figure 5 consists of 3 classes namely Stop-word, Stemming, and Tokenizer.

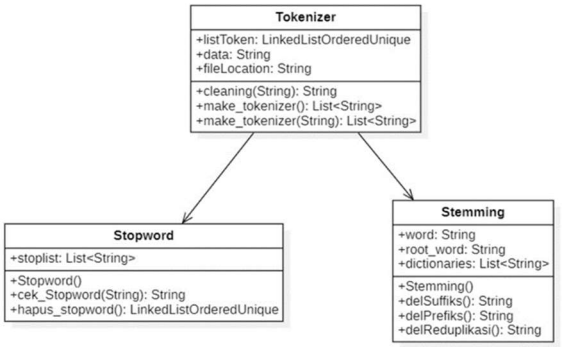

Figure 5: Text Operation Class Diagram

The indexing process is used to manage text that has been through a text opera-tion process to be stored and weighted in the database. Next figure 6 is a class diagram design for the Indexing process.

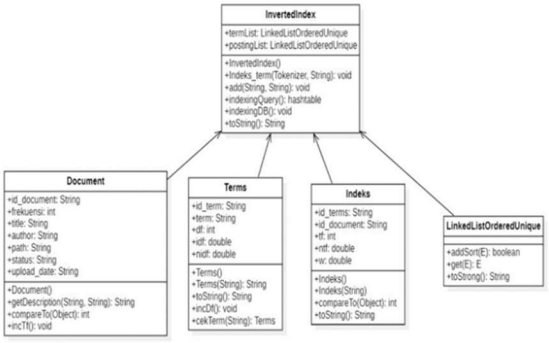

Figure 6: Indexing Class Diagram

The Text Formulation class as shown in Figure 7 is tasked with managing the Expansion Query function or the expansion of the query.

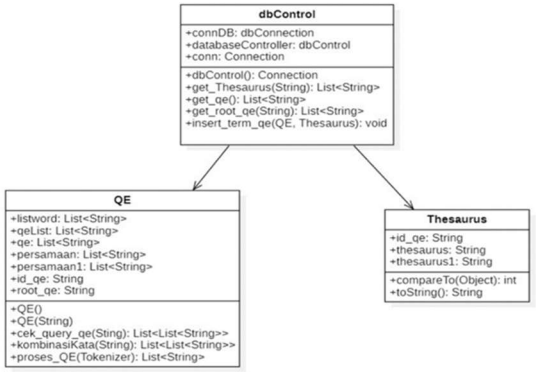

Figure 7: Text Formulation Class Diagram

Class Searching as Figure 8 functions to manage the process of searching docu-ments for each word or keyword entered.

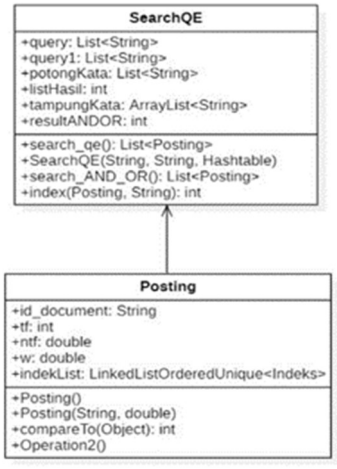

Figure 8: Searching Class Diagram

4.3 Database Structure Design

Database design explains the structure of the fields needed by the system. Following table 2 of the database structure in the Information Retrieval System.

Table 2: Information Retrieval System Database Structure

Table Name	Field Name
Documents	**ID_DOCUMENT (PK),** TITLE, PATH, STATUS, UPLOAD_DATE, AUTHOR
Indeks	**ID_DOCUMENT (PK),** ID_TERMS, TF, NTF, W
Terms	**ID_TERMS (PK),** TERMS, DF, IDF, NIDF
Term_Expansion	**ID_QE (PK),** ROOT_QE
Thesaurus	**ID_QE (PK),** EQUATION
Dictionary	**ID_DICTIONARY (PK),** ROOT_WORD
Stopwords	**ID_STOPWORD (PK),** STOPWORDS

4.4 Text Operation

Text Operation is the initial stage of preparation of documents before indexing which functions to reduce the complexity of the repression of raw documents and process data into terms that are ready in the index [18]. This stage includes the selec-tion of words in the query and document (term selection) in the transformation of documents or queries into terms index (index of words). Three algorithms are used in Text Operation:

4.4.1 Tokenization (Word Separation)

This process works by separating rows of words in a sentence, paragraph, or page into a single word or term word. This stage also removes special characters such as punctuation and changes all words or tokens to form lowercase letters (lower case). Examples of word-cutting processes can be seen in Figure 9 below.

Figure 9: Tokenization.

4.4.2 Stopwords Removal (Elimination of Common Words)

Stopword is defined as an irrelevant term with the main subject of the database even though the word is often present in the document. Examples of Indonesian stop-words are those, also, from, he, we, you, me, me, this, that, or, and, that, at, with, are, that is, to, no, no, at, at, if, then, there are, too, other, only, only, however, like, then, etc. [19]. Stopwords Removal is a process to eliminate common words that are in a document. Examples of the process of removing common words can be seen in Figure 10 below.

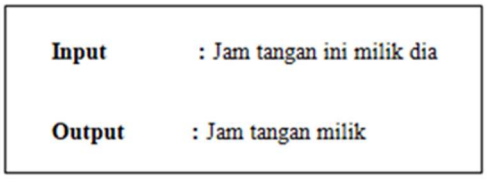

Figure 10: Stopwords Removal

4.4.3 Stemming

A process in the Information Retrieval System that transforms the words contained in a document to the root word using certain rules. For example, shared words, togetherness, equal, will distem to the root word that is "the same". The stemming process in Indonesian texts is different from stemming in English texts. In English texts, the only process required is the process of removing suffixes. Meanwhile, in the Indonesian language text, besides suffixes, confixes must also be removed.

In this Stemming process, the algorithm used by this research is an algorithm with the following stages [11]:

- ✓ Look for the word to be distem in the dictionary. If found, it is assumed that the word is the root word. Then the algorithm stops.
- ✓ Inflection Suffixes ("-lah", "-lah", "-ku", "-mu", or "his") removed. If it is in the form of particles ("-lah", "-lah", "-tah" or "-pun") then this step is repeated again to remove the Positive Pronouns ("my", "your", or "his")) , If there is.
- ✓ Remove Derivation Suffixes ("-i", "-an" or "-kan"). If the word is found in the dictionary, the algorithm stops. If not then proceed to step c (3).
  - o If "-an" has been deleted and the last letter of the word is "-k", then "-k" has also been deleted. If the word is found in the dictionary the algorithm stops. If not found then do step 3b.
  - o Deleted endings ("-i", "-an" or "-kan") are returned, proceed to step 4
- ✓ Remove the Derivation Prefix. If in step 3 a suffix was deleted then go to step 4a, if not go to step 4b.
  - o Check for combinations of prefixes that are not permitted. If found the algorithm stops, if not go to step 4b.

- For i = 1 to 3, specify the prefix type then delete the prefix. If the root word has not been found, do step 5, if it has then the algorithm stops.

✓ Perform Recoding.
✓ If all the steps have been completed but are not successful then the initial word is assumed to be the root word.
✓ The process is complete.

4.5 Query Formulation

The Query Formulation Process is a collection of techniques for modifying a query to meet an information need. One method of query expansion is to use the Expansion Query [20].

Query expansion is the process of reformulating the initial query by adding a few terms or words to the query to improve performance in the information retrieval process. In the context of web search engines, this includes evaluating user input and ex-tending search queries to get documents that match the query [21]. The method used in the expansion is to look for the meaning of foreign terms in the unstemmed-term form of the query. The Expansion Query method itself is divided into 3, namely:

✓ Manual Query Expansion (MQE). the user modifies the query manually. The system does not provide user assistance at all.
✓ Automatic Query Expansion (AQE), the system will modify the query automatically without the need for control from the user. Some techniques commonly used include : [18]
  - Global Analysis operates by examining all documents in the collection to build structures similar to Thesaurus. Queries are expanded by terms that are closely related to query terms within the scope of the collection. Thesaurus provides information about synonyms, words, and phrases that are semantically related.
  - Local Analysis, the system returns documents with an initial query, selects and checks some documents with the highest rank, assumes that the top documents are relevant, and then generates a new query.
✓ Interactive Query Expansion includes methods in which users interact with the system in expanding queries. The technique included is relevance feedback. Relevance feedback is a widely accepted method for increasing the effectiveness of interactive returns. An initial search is per-formed by the system using queries provided by the user. Queries are run to find more relevant documents. This process can be repeated until the user feels his information needs are being met [22].

4.6 Document Indexing

The indexing process is the process of storing documents in order. The document storage is done so that it can be processed again through the process of searching for documents. The construction of an index from a collection of documents is the main task at the pre-processing stage in the Information Retrieval System. A document index is a collection or collection of terms that indicate the contents or topics contained in a document. The index will distinguish a document from other documents in the collection. A large index allows many relevant documents to be found but at the same time can increase the number of irrelevant documents and decrease the search speed.

An inverted file index is a mechanism for indexing words from a collection of texts used to speed up the search process [18]. The inverted index as Figure 11 consists of two parts, namely a dictionary and a posting list. The dictionary contains a list of terms and a list of posts containing the document id associated with the term [23].

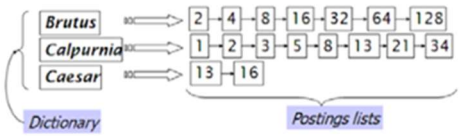

Figure 11: Inverted Index

The inverted index data structure representation in Figure 11 above shows the dictionary contains a collection of terms that have been sorted alphabetically and each term has a list of lists that contain a collection of sorted document IDs [23].

4.7 Document Searching

Search subsystem (matching) is the process of rediscovering information/documents that are relevant to a given query. Documents taken (retrieved) by the system are documents per the wishes of the user (relevant). Figure 12 below shows the relationship between relevant documents, documents taken by the system, and relevant documents taken by the system:

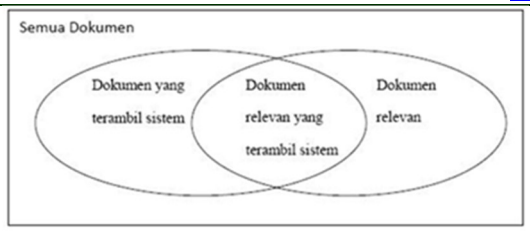

Figure 12: Document Searching Venn Diagram

In searching, the order in which documents will be displayed on the Information Retrieval System is based on weighting. The Term Frequency-Inverse Document Frequency (TF-IDF) method is a way of giving the weight of a word (term) to documents. For a single document, each sentence is considered a document. This method combines two concepts for weight calculation, namely Term Frequency (TF) is the frequency of occurrence of words (t) in sentences (d). Document Frequency (DF) is the number of sentences where a word (t) appears. The frequency with which words appear in a given document indicates how important that word is in the document. The frequency of documents containing the word indicates how common the word is. The weight of the word is greater if it often appears in a document and smaller if it appears in many documents [24]. Calculation of the Weight (W) of each document with the following Figure 13.

$W_{dt} = TF_{dt} * IDF_t$

dengan :
d = kalimat ke-d
t = kata (*term*) ke-t
TF = *term frequency*
W = bobot kalimat ke-d terhadap kata(*term*) ke-t
IDF = *inverse document frequency*

Figure 13: Term Weight Formulation.

Next is the process of sorting the cumulative value of W for each sentence. The three sentences with the largest W values are used as the result of a summary or as the output of automatic text summation. Based on the formula above, regardless of the value of tfij, if N = n, you will get 0 (zero) results for the calculation of Idf. For that you can add value 1 on the Idf side, so the weight calculation becomes as Equation 3 follows:

$$W\ ij = tfij * (logN/n)+1) \quad (3)$$

4.8 Interface Implementation

The display interface is very important for a user. so, in this Information Retrieval System, a lightweight and attractive interface is implemented following the design. Following is an explanation of some of the main menus developed:

✓ Searching Page

There are 2 search page menus namely searching for Information Retrieval with-out Expansion Query and searching with Expansion Query, as seen in Figures 14 and 15. The user must enter a keyword and press the search button to get the document sought. The document can be downloaded by clicking on the document title.

Figure 14: Searching Page with Query Expansion.

Figure 15: Searching Page without Query Expansion.

✓ Admin Menu Page.

On the admin page, as shown in Figure 16, a form is available for inputting new document data and term expansion.

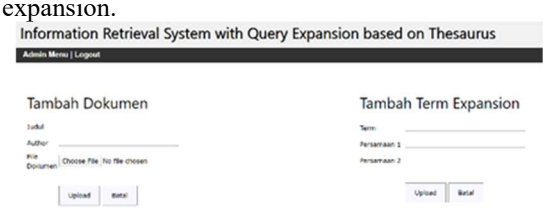

Figure 16: Admin Menu Page.

4.9 System Testing

This section will explain the results analysis and systematic testing conducted for searches that do not use Expansion Query and searches using Expansion Query to determine the relevance level of the results, and will also analyze the quality of time access of the two methods.

Testing is done by entering keywords (for this test, the keywords used are Data Warehouse) in the two search menus, which will then be determined

on the level of relevance using the recall and precision methods.

4.9.1 Search Without Using Expansion Queries

The search results found 15 documents with 4 relevant documents and 11 relevant documents in the collection. In table 3 following, recall and precision can be calculated after knowing the number of documents found is 15 and the number of relevant documents at the testing time is 4.

Table 3: Recall Precision Test I.

Doc No	Result	Recall	Precision
248		0	0
95	R	0,25	0,5
165		0,25	0,33333333
25		0,25	0,25
239		0,25	0,2
33		0,25	0,16666667
178	R	0,5	0,28571429
90	R	0,75	0,375
224		0,75	0,33333333
228		0,75	0,3
63		0,75	0,27272727
167		0,75	0,25
190		0,75	0,23076923
69	R	1	0,28571429
31		1	0,26666667

The next step is to create Table 4 for the recall and precision interpolation points of search without expansion queries.

Table 4: Interpolation Testing 1

Recall	Precision
0%	50,00%
10%	50,00%
20%	50,00%
30%	37,50%
40%	37,50%
50%	37,50%
60%	37,50%
70%	37,50%
80%	28,57%
90%	28,57%
100%	28,57%

4.9.2 Search Using Expansion Queries

The search results found 28 documents with 11 relevant documents and 11 relevant documents in the collection. In the following table, 5 recall and precision can be calculated after knowing the number of documents found is 28 and the number of relevant documents at the time of testing is 11.

Table 5: Recall Precision Test 2.

Doc No	Result	Recall	Precision
181	R	0,090909	1
139	R	0,181818	1
28		0,181818	0,66666667
248		0,181818	0,5
178	R	0,272727	0,6
171	R	0,363636	0,66666667
95	R	0,454545	0,71428571
165		0,454545	0,625
25		0,454545	0,55555556
67		0,454545	0,5
239		0,454545	0,45454545
33		0,454545	0,41666667
178	R	0,545455	0,46153846
90	R	0,636364	0,5
90	R	0,727273	0,53333333
179		0,727273	0,5
224		0,727273	0,47058824
123	R	0,818182	0,5
228		0,818182	0,47368421
63		0,818182	0,45
95	R	0,909091	0,47619048
167		0,909091	0,45454545
255		0,909091	0,43478261
190		0,909091	0,41666667
190		0,909091	0,4
69	R	1	0,42307692
172		1	0,40740741
31		1	0,39285714

The next step is to create Table 6 for the recall and precision interpolation points of search using Expansion Queries.

Table 6: Interpolation Testing 2

Recall	Precision
0%	100,00%
10%	100,00%
20%	71,43%
30%	71,43%
40%	71,43%
50%	53,33%
60%	53,33%
70%	53,33%
80%	50,00%
90%	47,62%
100%	42,31%

A comparison of the two search tests can be seen in Figure 17 below.

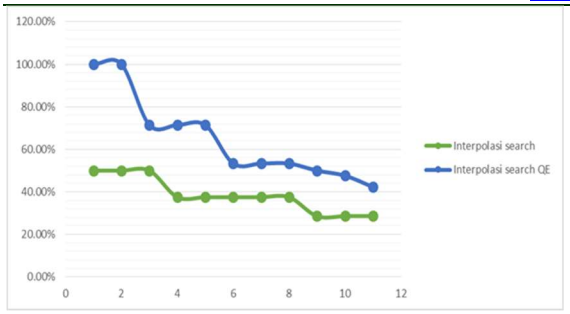

Figure 17: Interpolation Testing Graphic.

Furthermore, it can be compared that searching with Expansion Query is much better than without Expansion Query. In a search without Expansion Query has never reached the value of 100% precision, then it can be seen again in the difference be-tween the best value of precision search without Expansion Query almost reaching half or 50%.

Searching with Expansion Query is better because it uses the expansion query method in Thesaurus. With the keyword "data warehouse", there will also appear documents about "data warehouse", so most likely the documents needed by users are from the keyword "data warehouse".

4.10. Discussion Based on the Result

The relationship between "Thesaurus-Based Query Expansion on Information Retrieval to Improve the Quality of Document Searching Result" and the other topics can be explored as follows:

- ✓ Implementation of Single Sign-On Using the Concept of Method OAuth (Open Authorization) [1, 10]

  There might be a connection between the two in terms of user authentication and security. Implementing Single Sign-On can enhance the user experience in information retrieval systems, and OAuth is a relevant authentication method.
- ✓ Data Mining with Association Rules. [2, 11]

  Data mining techniques, specifically association rules, can be applied in the context of information retrieval to identify patterns and relationships in search queries and document content. This can lead to more effective query expansion strategies.
- ✓ Information Technology and Communication Research. [5, 6, 10]

  Thesaurus-based query expansion is a topic within the broader field of information technology and communication research. Research in this area contributes to advancements in information retrieval methods.
- ✓ Information Storage & Retrieval System: Theory & Implementation. [9, 10]

  This topic is directly related to the core theme of improving information retrieval. The theory and implementation of retrieval systems can benefit from techniques like thesaurus-based query expansion.
- ✓ Word Similarity for Document Grouping Using Soft Computing [10, 12]

  Word similarity measures can be integrated into thesaurus-based query expansion to enhance the selection of relevant terms, improving the quality of document search results.
- ✓ Confix-Stripping: Approach to Stemming Algorithm. [10, 15]

  Stemming algorithms are a fundamental component of information retrieval systems. The relationship here could be that stemming techniques can complement thesaurus-based query expansion in optimizing the search process.
- ✓ Inverse Document Frequency [10, 23, 24]

  Inverse Document Frequency is a key concept in information retrieval, particularly in ranking the importance of terms. It can be used alongside thesaurus-based query expansion to better understand and weigh the relevance of expanded terms in the retrieval process.

Thesaurus-Based Query Expansion on Information Retrieval is closely related to these topics, as they all contribute to the development and improvement of information retrieval systems, each from its own perspective, such as authentication, data mining, research, implementation, word similarity, stemming, and term weighting.

5. CONCLUSIONS AND SUGGESTION

5.1 Conclusions

Based on the results and discussion of the above research, the conclusions that can be drawn from the use of Thesaurus in the Expansion Query in Information Retrieval to improve the quality of document search results, namely:

- ✓ The development of a document search engine based on the Information Retrieval System can improve the relevance of data and document searching result quality.
- ✓ Based on the test results using the recall and precision method, obtained a graph that shows the results of relevance to the Query Expansion

method has improved in the level of data relevance. Because in the use of expansion queries the results of recall be-come higher which allows more relevant documents to be retrieved. Unfortunately, in the case of query execution, the search time is longer than the search without Query Expansion due to the process of checking the results of query expansion in Thesaurus so that the more words or terms found, the more time is used.

5.2 Suggestions

Suggestions recommended for the research on the development of Information Retrieval in the future can be seen as follow:

- ✓ Increased query execution time for searches using Expansion Query.
- ✓ Add the pre-processing process to the Admin menu "Add Document".

REFERENCES:

[1] Awan Setiawan, Mokh. Hendayun, Suci Fitri Yanti, "Implementation Single Sign On Using the Concept of Method OAUTH (Open Authorization) on the Web Portal", Vol. 86, No. 3, 2016, pp. 339-346.

[2] Ase Suryana, Erwin Yulianto, "Application of Data Mining with Association Rules to Review Relationship between Insured, Products Selection and Customer Behavior, Horizon Research Publishing Corporation", Vol. 6, No. 3A, 2019, pp. 45-61.

[3] J.F. Rusdi, S. Salam, N.A. Abu, S. Sahib, M. Naseer, A.A. Abdullah, "Drone Tracking Modelling Ontology for Tourist Behavior", *J. Phys. Conf. Ser.*, 1201, 2019, 012032.

[4] J.F. Rusdi, S. Salam, N.A. Abu, B. Sunaryo, R. Taufiq, L.S. Muchlis, T. Septiana, K. Hamdi, B. Ilman Arianto, F.R. Kodong Desfitriady, A.V. Vitianingsih, "Dataset Smartphone Usage of International Tourist Behavior", *Data Br.*, 27, 2019, 104610.

[5] Jack Febrian Rusdi, Sazilah Salam, Nur Azman Abu, Tedja Gurat Baktina, R Gumilar Hadiningrat, Budi Sunaryo, Arlinda Rusmartiana, Wahin Nashihuddin, Puteri Fannya, Fretycia Laurenti, N.M. Shanono, Richki Hardi, "ICT Research in Indonesia", *Journal of Science and Technology*, Vol. 1, No. 1, 2019, pp. 1-23.

[6] Jack Febrian Rusdi, Nur Azman Abu, Nova Agustina, Mohammad Kchouri, Sintia Dewi, "ICT Research in Indonesia", *Journal of Science and Technology*, Vol. 1, No. 1, 2019, pp. 24-33.

[7] A. Smith, K. de Salas, B. Schüz, "Developing smartphone apps for behavioural studies: The AlcoRisk app case study", *Journal of Biomedical Informatics*, 72, 2017, pp. 108-119. https://doi.org/10.1016/j.jbi.2017.07.007

[8] H. Winsket, T.H. Kim, L. Kardash, I. Belic, "Smartphone use and study behavior: A Korean and Australian comparison", *Heliyon*, Vol. 5, No. 7, 2019, e02158.

[9] Gerald J. Kowalski, "Information Storage & Retrieval System : Theory & Implementation".

[10] Hazra Imran, Sharan Aditi, "Thesaurus & Query Expansion", *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol. 1, No. 2, 2009, pp. 89-97.

[11] Krzysztof J. Etc. Cios, "Data Mining - A Knowledge Discovery Approach".

[12] Azmi MA Murad, Trevor Martin, "Word Similarity for Document Gouping using Soft Computing", *IJCSNS International Journal of Computer Science and Network Security*, Vol. 7, No. 8, 2007, pp. 20-27.

[13] Dikdik Kurniawan, "Evaluasi Sistem Temu Kembali Informasi Model Ruang Vektor Dengan Pendekatan User Judgement", *J. Sains MIPA*, Vol. 16, No. 3, 2010, pp. 155-162.

[14] P.L. Pendit, "Perpustakaan Digital Dari A Sampai Z".

[15] Bobby Nazief, Adriani Mirna, "Confix-Stripping : Approach to Stemming Algorithm for Bahasa Indonesia", *Faculty of Computer Science, University of Indonesia*, 2007.

[16] Awan Setiawan, Erwin Yulianto, "Implementation of Risk Control Self Assessments Using Rapid Application Development Model in Bank Operational Risk Management Processes", Vol. 97, No. 11, 2019, pp. 2957-2968.

[17] Tutorialspoint, "SDLC - Waterfall Model", 2020. https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm

[18] R. Baeza-Yates, B. Riberio-Neto, "Modern Information System", 1999.

[19] H. Etc. Schutze, "Introduction to Information Retrieval", *Cambridge University Press*, 2008.

[20] E.W. Selberg, "Information Retrieval Advances Using Relevance Feedback", *Department of Computer Science & Engineering, University of Washington Seattle*, 1997.

[21] Y. Qiu, H.P. Rfe, "Concept-Based Query Expansion", *SIGIR*, 1993, pp. 160-169.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[22] Buckley Chris, "The Effect of Adding Relevance Information in a Relevance Feedback Environment", *SIGIR*, 1994, pp. 292-300.

[23] C. Manning, "Inverted Indexing", 2009. http://home.deib.polimi.it/lbondi/data/uploads/irdm15-16/slides/07_inverted_indexing_v1. pdf

[24] Stephen Robertson, "Understanding Inverse Document Frequency: On theoretical arguments for IDF", *England: Journal of Documentation*, Vol. 60, 2005, pp. 502-520.