Journal of Engineering Science and Technology
Vol. 18, No. 1 (2023) 534 - 553

USING OF THESAURUS IN QUERY EXPANSION ON INFORMATION RETRIEVAL AS VALUE CREATION STRATEGY THROUGH BIG DATA ANALYTICS

GRAHA NOVIANA[1,*], AGUS RAHAYU[1], RATIH HURRIYATI[1], LILI ADI WIBOWO[1], ERWIN YULIANTO[2], RIZKY JUMANSYAH[3]

[1]Department of Business and Management, Universitas Pendidikan Indonesia, Bandung, Indonesia
[2]Department of Informatics, Langlangbuana University, Bandung, Indonesia
[3]Department of Information System, Universitas Komputer Indonesia, Indonesia
*Corresponding Author: graha.noviana@gmail.com

Abstract

Research still does not frequently use the concept of value creation through the application of big data analytics. Additionally, the internet serves as a limitless supply of knowledge that comes from different people all over the world and is delivered in a variety of ways, including text documents, photos, audio files, and video. Search engine is a tool used to find different types of information on the internet. The issue that frequently occurs when looking for information online is that the search results for the intended subject are frequently unrelated to the terms entered. In order to increase the relevance of the data obtained from the search results, the goal of this study is to implement the query expansion approach, which involves reformulating the initial query in order to add several phrases utilizing a thesaurus. The waterfall model was used in conjunction with the Software Development Life Cycle (SDLC) in the research. The results showed that the creation of a document search engine based on an information retrieval system was able to enhance data relevance and the standard of document search outcomes in big data analysis. The recall results are higher due to the usage of query expansion, which allowed for the retrieval of more pertinent documents. It is clear that by adopting this technique, the relevancy of the data in the search system can be improved.

Keywords: Big data analytic, Information retrieval, Query expansion, Thesaurus, Value creation strategy

1. Introduction and Research Background

Along with the development of information and communication technology that is increasingly advanced, and globalization is increasingly widespread, the world has become a global village that cannot be separated from the need for knowledge. Science and technology need information but also 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. Search engine is a medium used for searching various information on the internet [1].

The world of business that is full of competition makes businessmen must always think about breakthrough strategies that can guarantee the continuity of their business. Information search is a business strategy that is very helpful in decision making [1]. 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 for the desired topic are often irrelevant to the keywords entered. For example, when foreign terms are translated directly into Indonesian and entered as keywords, the results obtained are often very irrelevant. This is quite common when doing searches on topics that contain a lot of foreign terms whose meanings are not yet known and greatly hinders in searching for certain topics.

For example, one of the terms that is increasingly popular to be searched for research titles in Indonesia is the use of Information Communication and Technology (ICT). ICT research is more dominant in education and subsequently in management [2]. Likewise, with 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 behaviors. Studies resulting from the various searches above are more likely to give another aspect of how technology affects consumer behavior [3-5].

The construct of value creation through the implementation of big data analytic is still not widely found in research. Along with the role of the internet as an unlimited source of information originating from various people around the world with various 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 anyplace. Value creation from implementing big data is the key to victory in competing with increased business efficiency [5].

According to the research results [6-10], the use of the latest technologies such as artificial intelligence and big data in the industrial sector can increase efficiency, accuracy, fraud detection, help fulfil with regulation, offer innovative services, provide a better customer experience, reduce costs and increase revenue. The company's ability to increase value creation to consumers can increase company profits [11].

Information retrieval system is a system, method, procedure used to retrieve stored information 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 [11]. 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 [12, 13]. Based on the research background above, the formulation of the problems to be studied and analysed are:

(i) How to build a reliable document search engine system in big data analysis? and

(ii) How to increase the level of relevance of the data obtained if there are keywords using foreign languages as added value in big data analysis?

Based on the problems and background, the research aims to create an information retrieval system capable of increasing data relevance and document searching result quality in big data analysis. In addition, this research also aims to implement query expansion by expanding queries on thesaurus to expand additional terms, phrases or words associated with foreign synonyms as added value in big data analysis.

2. Literature Review

2.1. Value creation in big data analytic

Value creation is the ability of a company's ability to make a strong difference for the company. Companies must continuously provide value creation to customers so that the company can have a strong position in the market in the long term. Strategic orientation reflects a company's philosophy on how to run a business and guides its efforts to achieve superior performance [14].

Value creation can be achieved through the environmental scanning process at the company. Environmental scanning is the process of monitoring, evaluating, and disseminating information from the external and internal environment to important people within the company. It aims to identify external and internal strategic factors that will assist in the analysis of corporate strategic decisions [15]. The ability to discern attractive opportunities and having the competence to succeed in these opportunities are two different things. Thus, every business needs to periodically evaluate internal strengths and weaknesses in marketing, finance, manufacturing, and company competence [16].

The priority of relationship marketing is to improve the customer experience, by leveraging big data analysis. Big data can be defined as a system that integrates the real world, humans and cyberspace [17]. Big data can be classified into two categories, namely data originating from the real world where this data is obtained through technology that is able to capture all types of data circulating in the real world such as natural data, climate, maps, biology and others [17]. Second, big data comes from human society, where this data is obtained through the internet of things such as social media, the internet, and other technologies [18].

Considering that big data is related to data and technology, data in the context of big data can be categorized into 5V, namely Volume, Velocity, Variety, Veracity, and Value [19]. The entire business case from big data analysis can be seen in Fig. 1.

Big data technologies can provide data in many formats beyond simple numeric data. The data includes numeric data, text, audio, and video files that are increasingly numerous and interconnected. By analysing existing big data,

insights can be developed that can be used to identify and forecast trends in consumer behavior. Ultimately, this will lead to maximizing value creation and organizational value.

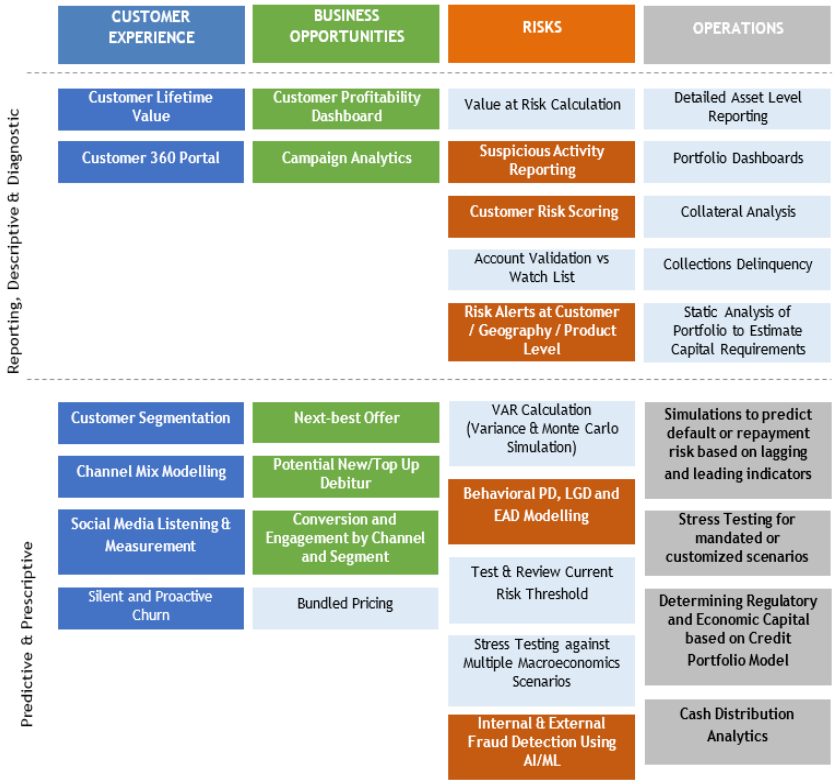

Fig. 1. Business case on big data analysis.

Big data analytics has been considered as a game changer that allows increased efficiency and effectiveness in business because of the potential for significant improvements in business and operational areas. The emerging literature on big data analytics has identified a positive relationship between the application of data analytics to support decision-making as value creation and firm performance [20].

2.2. Information retrieval

ISO 2382/1 published in 1984 defines “information retrieval” as actions, methods and procedures to recover stored data, then provide information on the required subject. Based on this standard, these actions include text indexing, inquiry analysis, and also relevance analysis. Data includes text, tables, images, speech, and video, information including related knowledge needed to support problem solving and knowledge acquisition [21].

The goal of information retrieval is to meet the information needs of users by recovering all relevant documents and at the same time obtaining as little as possible irrelevant documents. This system uses a heuristic function to obtain documents that are relevant to the keywords entered by the user. A good

information retrieval system allows the user to quickly and accurately determine whether the contents of the document are satisfactory. Only relevant documents should be returned based on user queries. For better document representation, documents with similar topics are grouped together [22].

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, conformity measurement, ranking, feedback relevance and others. In the context of evaluating information retrieval, the methods used are recall and precision. Recall is a comparison of the number of relevant documents retrieved according to the given query with the total collection of documents relevant to the query [22]. Precision is the ratio of the number of documents relevant to the query with the number of documents retrieved from the search results [23]. Precision can be interpreted as the accuracy or compatibility between a request for information and an answer to that request, while the term recall relates to the ability to retrieve information that has been stored [24].

Recall is stated as part of the relevant documents in the documents found, or the total number of relevant documents. Furthermore, precision is expressed as part of the relevant documents found, or the total number of documents found. Both describe the performance of the information retrieval system by calculating the number of relevant documents in the search results. This measurement of recall and precision is a calculation performed on the 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 recall rate and precision are high, then the information retrieval system can be considered good.

2.3. Query expansion using thesaurus

Query expansion that is applied in this case is using query expansion in thesaurus. In this information retrieval system, there is a pre-processing stage beforehand, where a collection of documents with the extension *.pdf will be converted into a text file (*.txt). Next, the text operation process is carried out using the stemming algorithm (see https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm/, retrieved on 16 March 2020). After going through text operations, document data and terms will be indexed in the database using the linked list order method [25].

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 they have similarities or other terms with these words. If it has similarities, the system will execute simultaneously between the keywords that the user typed and the keyword equations.

In the search process the system will use a hash table structure with an inverted index data structure. Then before being displayed again as results, there is a final step where the data or documents are sorted based on the ranking of the most relevant documents based on the queries or keywords entered. The system testing process uses the recall and precision methods. Figure 2 shows a flowchart of the query expansion that will be examined.

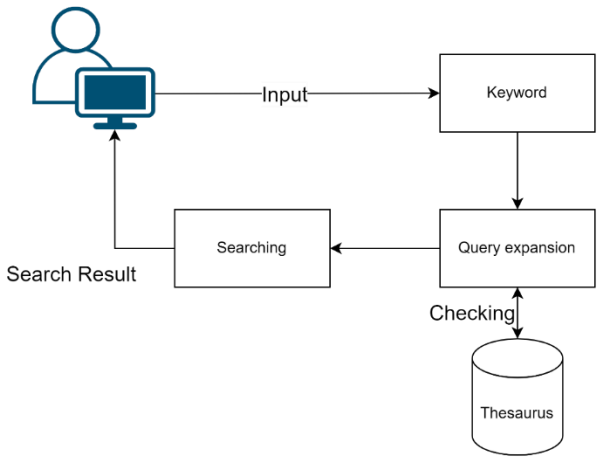

Fig. 2. Query expansion flowchart.

3. Methodology

The research method is the stages that is carried out in the research. With systematic activity systematics, 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 waterfall model is most appropriate are:

(i) Requirements are very well documented, clear and fixed.

(ii) Product definition is stable.

(iii) Technology is not dynamic and understood.

(iv) There are no ambiguous requirements.

(v) Ample resources with required expertise are available to support the product.

(vi) The project is short.

Figure 3 illustrates a representation of the different phases of the waterfall model.

The sequential phases in waterfall model are:

(i) Requirement analysis, all possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.

(ii) System design, the requirement specifications from first phase are studied in this phase and the system design is prepared. This system design helps in specifying hardware and system requirements and helps in defining the overall system architecture.

(iii) Implementation, with inputs from the system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality, which is referred to as unit testing.

(iv) Testing, all the units developed in the implementation phase are integrated into a system after testing of each unit.

(v) Deployment, once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.

(vi) Maintenance, there are some issues which come up in the client environment. To fix those issues, patches are released. Maintenance is done to deliver these changes in the customer environment.

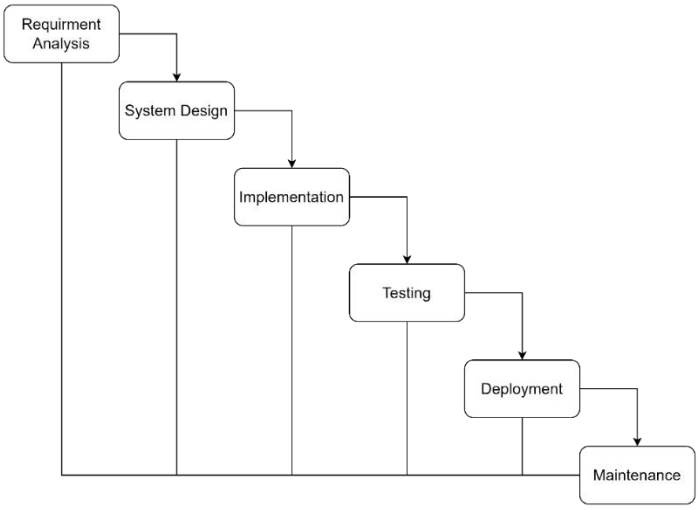

Fig. 3. Waterfall model [26].

4. Results and Discussion

4.1. System flowchart

To obtain information, a system is needed that can facilitate users in terms of searching. Although in general an information retrieval system equipped with a text operation process is able to provide adequate results, it will not be optimal when the keywords entered are in the form of short text or foreign terms. Based on this, an additional process was designed for information retrieval in the form of a query expansion process to maximize search results, especially at the level of relevance of existing data searches.

The analysis carried out in making the information retrieval system architecture in this study uses a flowchart as shown in Fig. 4.

The function used during the searching process is AND - OR. To search for more than 1 word, an AND will be added to each word, for example the keyword "change design" becomes "change AND design". When going through the query expansion process, each of these words will first look for similarities or related terms, then combined with the AND function added with the OR function, for example the word "change" earlier after checking in the thesaurus found the word "edit" and for "design" found the word "draft". The results after going through the query expansion process become "change design", "edit design", "edit design", so that when the search process uses the AND - OR function it becomes (change AND design) OR (change AND design) OR (edit AND design) OR (edit plan).

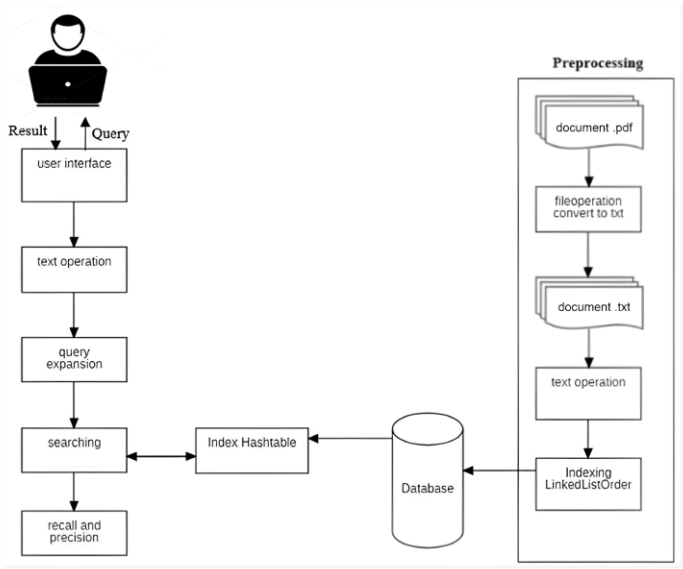

Fig. 4. Information retrieval system flowchart.

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. Table 1 describes the roles of each actor.

Table 1. Actors' role.

Actor	Role
Admin	Login
	Add documents
	Add term expansion
User	Document searching
	Document download

The interactions between actors and the information retrieval system that has been developed is shown in Fig. 5.

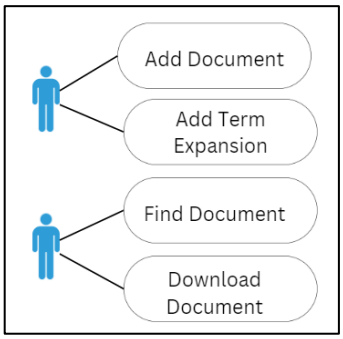

Fig. 5. Information retrieval system use case diagram.

4.2.2. Class diagram

The design of the class diagram on the information retrieval system with query expansion will be divided into 4 major classes namely text operations, indexing, text formulation and searching. The process of text operations aims to reduce the complexity of representing raw documents and processing data into terms that are ready to be indexed [27]. Text operations are performed on the keywords entered by the user and on documents that have been uploaded. Class text operation as shown in Fig. 6 consists of 3 classes namely stopword, stemming and tokenizer.

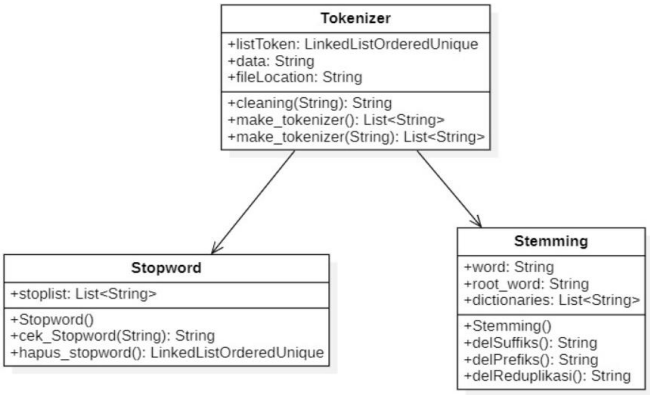

Fig. 6. Text operation class diagram.

The indexing process is used for managing text that has gone through a text operation process to be stored and weighted in a database. Figure 7 shows a class diagram design for the indexing process.

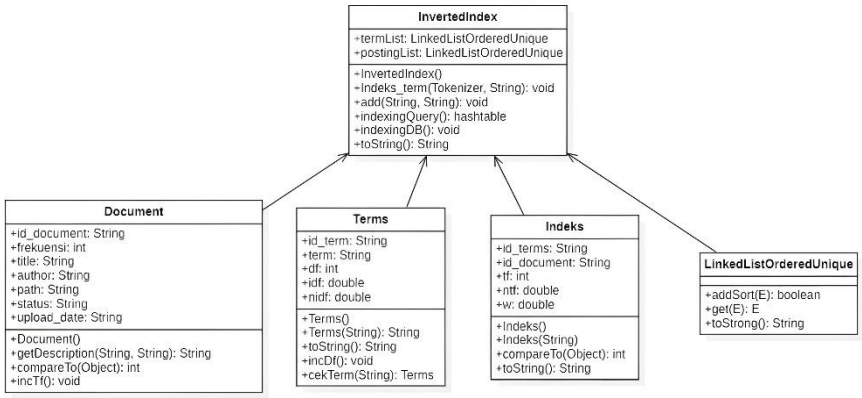

Fig. 7. Indexing class diagram.

Class text formulation as shown in Fig. 8 is in charge of managing the query expansion function or expansion of the query. Here, in Fig. 8 several of the attributes are written in Bahasa Indonesia, such as *persamaan* which means ‘equation’, *kombinasiKata* which means ‘word combination’, and *proses_QE* which means ‘QE process’.

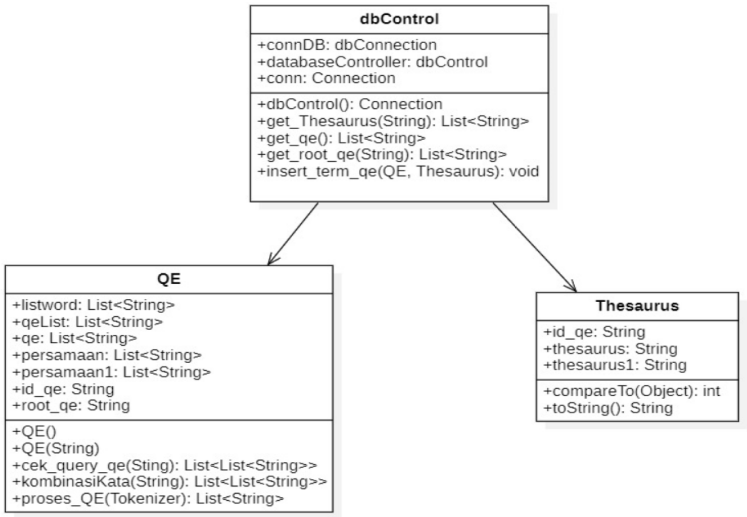

Fig. 8. Text formulation class diagram.

Class searching as shown in Fig. 9 functions to manage the document. Search process for each word or keyword entered. In Fig. 9, several of the attributes are written in Bahasa Indonesia, such as *potongKata* which means 'word removal', *listHasil* which means 'results list', and *tampungKata* which means 'word collection'.

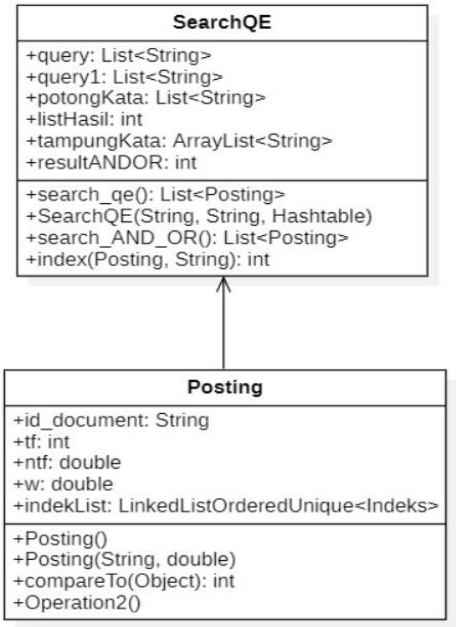

Fig. 9. Searching class diagram.

4.3. Database structure design

The database design describes the structure of the fields required by the system. The following Table 2 of the database structure on 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
Index	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 or document preparation before indexing which serves to reduce the complexity of representing raw documents and processing data into terms that are ready to be indexed [27]. This stage includes the selection of words in queries and documents (term selection) in the transformation of documents or queries into terms index (index of words). Three algorithms are used in the text operation which are tokenization, stopwords removal, and stemming.

4.4.1. Tokenization (word separation)

This process works by separating a row of words in a sentence, paragraph or page into a single word chunk or termed word. This stage also removes special characters such as punctuation marks and changes all words or tokens to lower case. An example of the process of cutting words can be seen in Fig. 10.

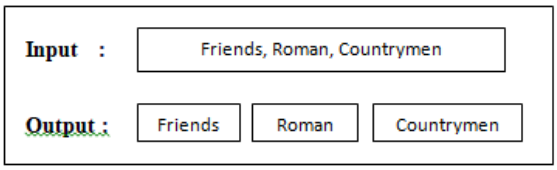

Fig. 10. Tokenization.

4.4.2. Stopwords removal (general word removal)

Stopword is defined as a term that is irrelevant to the main subject of the database even though the word is often present in the document. Examples of Indonesian stopwords are *juga*, *dari*, *dia*, *kami*, *kamu*, *aku*, *saya*, *ini*, *itu*, *atau*, *dan*, *pada*, *dengan*, *adalah*, *contohnya*, *kepada*, *tidak*, *engga*, *dalam*, *di*, *jika*, *lalu*, *disana*, *lainnya*, *hanya*, *cuma*, *namun*, *seperti*, etc. [28]. Stopwords removal is a process to remove common words from a document.

4.4.3. Stemming

A process in the information retrieval system that transforms the words contained in a document into root words using certain rules. For example, in Bahasa

Indonesia, the words of "*Bersama,*" "*kebersamaan*", and "*kesamaan*" will be stamped into the root word, namely "*sama*". The stemming process in Indonesian texts is different from stemming in English texts. In English language texts, the process that is needed is only the process of removing the suffix. Whereas in Indonesian language texts, apart from suffixes, confixes must also be removed.

In this stemming process, the algorithm used by the author is the algorithm [24]. The steps of the algorithm are:

a. Look up the word to be stylized in the dictionary. If found, it is assumed that the word is the root word. Then the algorithm stops.
b. Inflection suffixes ("-*lah*," "-*kah*," "-*ku*", "-*mu*", or "-*nya*") are omitted. If it is in the form of particles ("-*lah*", "-*kah*", "-*tah*" or "-*pun*") then this step is repeated again to remove any possesive pronouns ("-*ku*", "-*mu*", or "-*nya*").
c. Remove derivation suffixes ("-*i*," "-*an*" or "-*kan*"). If the word is found in the dictionary, the algorithm stops. If not then go to step c(1).
   (i) If the "-*an*" has been removed and the last letter of the word is "-*k*," then the "-*k*" is also removed. If the word is found in the dictionary, the algorithm stops. If not found then do step c(2).
   (ii) The deleted suffix ("-*i*," "-*an*" or "-*kan*") is returned, go to step d
d. Remove derivation prefixes. If in step c any suffixes were removed then go to step d(1), otherwise go to step d(2).
   (i) Check for disallowed prefix-suffix combinations. If found then the algorithm stops, if not go to step d(2).
   (ii) For i = 1 to 3, specify the prefix type then remove the prefix. If the word root has not been found, do step e, if so, the algorithm stops.
e. Recoding.
f. If all steps have been completed but not successful then the initial word is assumed to be the root word.
g. Process complete.

4.5. Query formulation

The query formulation process is a set of techniques for modifying queries with the aim of fulfilling an information need. One method of expanding the query is using query expansion [29].

Query expansion or query expansion is the process of reformulating the original 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 extending search queries to find documents that match the query [30]. The method used in the expansion is to look for the meaning of foreign terms in the unstemmed-term form of the query. As for the query expansion method itself is divided into three types, namely:

a. Manual Query Expansion (MQE) where the user modifies the query manually. The system provides no assistance to the user at all.
b. Automatic Query Expansion (AQE) where the system will modify queries automatically without the need for control assistance from the user. Several techniques commonly used include [27]:

(i) Global analysis operates by examining all documents in the collection to build a structure similar to thesaurus. The query is expanded with terms closely related to the query terms in the scope of the collection. Thesaurus provide information about synonyms, semantically related words and phrases.

(ii) Local analysis, the system returns documents with an initial query, selects and checks a number of top-ranking documents, assumes that the top documents are relevant, and then generates a new query.

c. Interactive Query Expansion (IQE) includes methods in which the user interacts with the system in expanding the query. The techniques included in it are relevance feedback. Relevance feedback is a widely accepted method for increasing the effectiveness of interactive feedback. An initial search is performed by the system using a query provided by the user. The query is run to find back a more relevant set of documents. This process can be repeated until the user feels his information needs are met [31].

4.6. Document indexing

Indexing process is the process of storing documents in an orderly manner. Document storage is carried out so that it can be processed again through the document search process. Building an index from a collection of documents is the main task at the preprocessing stage in the information retrieval system. The document index is a set or collection of terms that indicate the contents or topics in the document. The index will distinguish a document from other documents in the collection. A large index makes it possible to find many relevant documents but at the same time can increase the number of irrelevant documents and reduce search speed.

Inverted file index is a mechanism for indexing words from text collections that are used to speed up the search process [27]. The inverted index as shown in Fig. 11 consists of two parts, namely the dictionary and posting list. The dictionary contains a list of terms and a post list containing document ids related to the term [32].

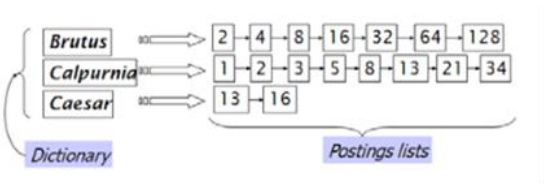

Fig. 11. Inverted index.

The representation of the inverted index data structure in Fig. 12 shows that the dictionary contains a collection of terms that have been sorted alphabetically and each term has a post list which contains a collection of sorted document ids [32].

4.7. Document searching

Search subsystem (matching) is the process of finding back information or documents that are relevant to a given query. Documents retrieved by the system are documents that are in accordance with the wishes of the user. Figure 12 shows the relationship between relevant documents, documents retrieved by the system and relevant documents retrieved by the system.

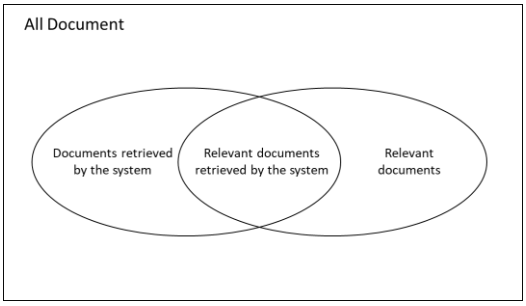

Fig. 12. Document searching Venn diagram.

In conducting a search, the order of documents that will be displayed on the information retrieval system is based on weight. The Term Frequency-Inverse Document Frequency (TF-IDF) method is a way of giving weight to the relationship of a term to a document. For a single document, each sentence is considered as a document. This method combines two concepts for calculating weights, namely Term Frequency (TF) is the frequency of occurrence of the word (t) in sentence (d). Document Frequency (DF) is the number of sentences where a word (t) appears. The frequency with which a word appears in a given document indicates how important that word is in that document. The frequency of documents containing the word indicates how common the word is. The weight of the word is greater if it appears frequently in a document and is smaller if it appears in many documents [32].

4.8. Interface implementation

The interface display is very important for a user. so that the information retrieval system implements a lightweight and attractive interface according to the design. The following is an explanation of some of the main menus developed:

(i) Searching page. There are 2 search page menus namely information retrieval search without query expansion and search with query expansion, as seen in Figs. 13 and 14. Users must enter keywords and press the search button to get the document they are looking for. The document can be downloaded by clicking on the document title.

(ii) Admin page. On the admin menu page, as shown in Fig. 15, a form is available for entering document data and new term expansions.

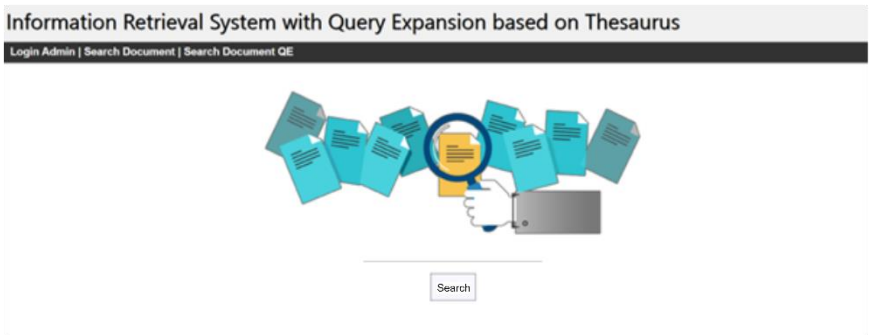

Fig. 13. Searching page with query expansion.

Fig. 14. Searching page without query expansion.

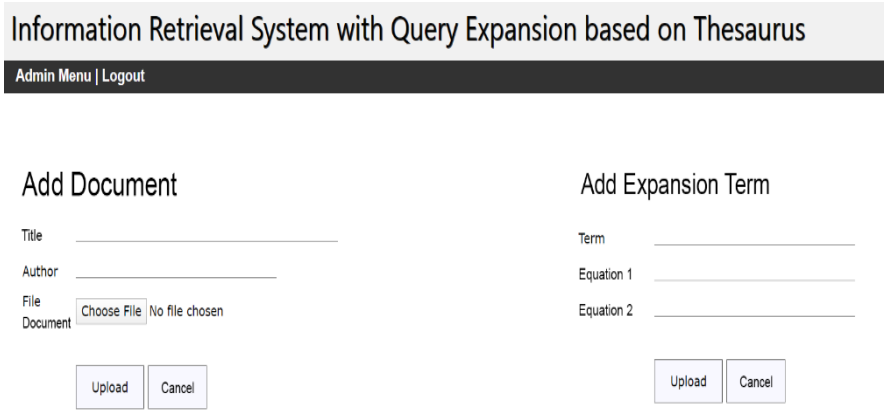

Fig. 15. Admin page.

4.9. System testing

This section will explain the analysis of the results and the systematic testing carried out on searches that do not use query expansion and searches using query expansion to determine the level of relevance of the results, then an analysis will also be carried out on the quality of access time from the two methods.

Testing is carried out by entering keywords (for this test, the keywords used are data warehouse) in both search menus, which will then determine the level of relevance using the recall and precision methods.

4.9.1. Search without using query expansion.

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

Table 3. Recall precision test 1.

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 make Table 4 for recall interpolation points and search precision without query expansion.

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 query expansion

The search results found 28 documents with 11 relevant documents and 11 relevant documents in the collection. In Table 5, recall and precision can be calculated after knowing that 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 make Table 6 for recall interpolation points and search precision using query expansion.

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%

Comparison of the two search tests can be seen in Fig. 16.

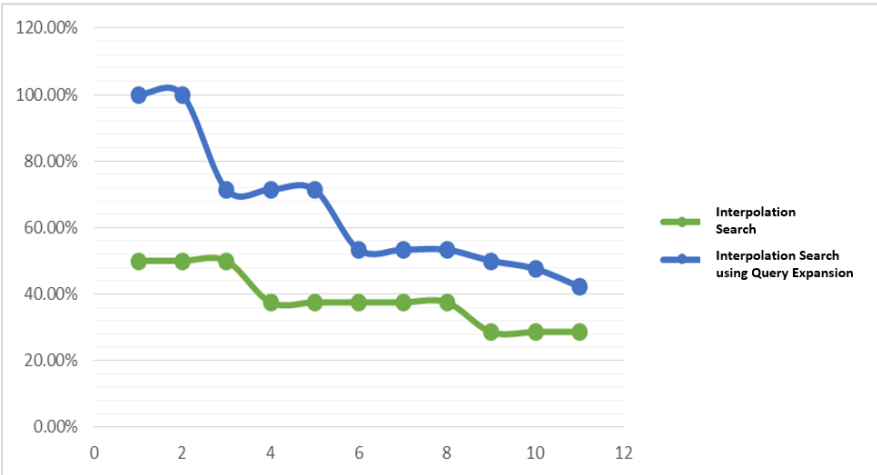

Fig. 16. Interpolation testing graphic.

Furthermore, it can be compared that searching with query expansion is much better than without query expansion. In searches without query expansion, it has never reached a precision value of 100%, then it can be seen again in the difference between the best values, the search precision without query expansion has almost reached half or 50%.

Searching with query expansion is better because the query expansion method is used in thesaurus. With the keyword "data warehouse", a document about "data warehouse" will also appear, so it is likely that the document needed by the user is from the keyword "data warehouse".

5. Conclusion

The conclusion that can be drawn is that development of an information retrieval system-based document search engine is capable of increasing data relevance and document searching result quality in big data analysis. Based on the test results using the recall and precision methods, a graph is obtained that shows the results of the relevance of the query expansion method experiencing improvements in the level of data relevance. Because using query expansion the recall results are higher which allows for more relevant documents to be retrieved. Unfortunately, in terms of query execution, the search time becomes longer than searches without query expansion due to the process of checking the results of query expansion in thesaurus so that the more similarities in words or terms found, the more time is used. Implementation of query expansion by expanding queries on thesaurus to expand additional terms, phrases, or words related to foreign synonyms to make added value in big data analysis.

References

1. Suryana, A.; and Yulianto, E. (2019). Application of data mining with association rules to review relationship between insured, products selection and customer behavior. *Universal Journal of Electrical and Electronic Engineering*, 6(3A), 45-61.
2. Rusdi, J.F.; Salam, S.; Abu, N.A.; Baktina, T.G.; Hadiningrat, R.G.; Sunaryo, B.; Rusmartiana, A.; Nashihuddin, W.; Fannya, P.; Laurenty, F.; Shanono, N.M.; and Hardi, R. (2019). ICT research in Indonesia. *SciTech Framework, Journal of Science and Technology*, 1(1), 1-23.
3. Smith, A.; de Salas, K.; Lewis, I.; and Schüz, B. (2017). Developing smartphone apps for behavioural studies: The alcorisk app case study. *Journal of Biomedical Informatics*, 72, 108-119.
4. Winskel, H.; Kim, T.-H.; Kardash, L.; and Belic, I. (2019). Smartphone use and study behavior: A Korean and Australian comparison. *Heliyon*, 5(7), e02158.
5. Wamba, S.F.; Gunasekaran, A.; Akter, S.; Ren, S.J.F.; Dubey, R.; and Childe, S.J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356-365.
6. Vijai, C. (2019). Artificial intelligence in Indian banking sector: Challenges and opportunities. *International Journal of Advanced Research*, 7(5), 1581-1587.
7. Jewandah, S. (2018). How artificial intelligence is changing the banking sector– A case study of top four commercial Indian banks. *International Journal of Management, Technology and Engineering*, 8(7), 525-530.

8. Kumar, K.N.; and Balaramachandran, P.R. (2018). Robotic process automation - A study of the impact on customer experience in retail banking industry. *Journal of Internet Banking and Commerce*, 23(3), 1-27.
9. Vedapradha, R.; and Ravi, H. (2018). Application of artificial intelligence in investment banks. *Review of Economic and Business Studies*, 11(2), 131-136.
10. Eng, T.-Y. (2004). Does customer portfolio analysis relate to customer performance? An empirical analysis of alternative strategic perspective. *Journal of Business and Industrial Marketing*, 19(1), 49-67.
11. Kanai, H.; and Kumazawa, A. (2021). An information sharing system for multi-professional collaboration in the community-based integrated healthcare system. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, 2(1), 1-14.
12. Imran, H.; and Sharan, A. (2009). Thesaurus and query expansion. *International Journal of Computer Science and Information Technology (IJCSIT)*, 1(2), 89-97.
13. Zhou, K.Z.; Yim, C.K.; and Tse, D.K. (2005). The effects of strategic orientations on technology- and market-based breakthrough innovations. *Journal of Marketing*, 69(2), 42-60.
14. Singgih, I.K. (2020). Air quality prediction in smart city's information system. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, 1(1), 35-46.
15. Goonjur, M.K.; Sumitra, I.D.; and Supatmi, S. (2020). Enhanced the weighted centroid localization algorithm based on received strength signal in indoor wireless sensor network. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, 1(1), 13-22.
16. Chen, C.L.P.; and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275, 314-347.
17. Paculaba, A.M.C.; Bathan, M.A.S.; and Niego, E.L. (2022). Log monitoring system using quick response (QR) code: A state university’s covid–19 contact tracing system. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, 3(2), 65-74.
18. Al Husaeni, D.; and Nandiyanto, A. (2022). Mapping visualization analysis of computer science research data in 2017-2021 on the google scholar database with VOSviewer. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, 3(1), 1-18.
19. Germann, F.; Lilien, G.L.; Fiedler, L.; and Kraus, M. (2014). Do retailers benefit from deploying customer analytics? *Journal of Retailing*, 90(4), 587-593.
20. Rizaldi, A.; Margareta, F.; Simehate, K.; Hikmah, S.N.; Albar, C.N.; and Rafdhi, A.A. (2021). Digital marketing as a marketing communication strategy. *International Journal of Research and Applied Technology (INJURATECH)*, 1(1), 61-69.
21. Murad, M.A.A.; and Martin, T. (2007). Word similarity for document grouping using soft computing. *IJCSNS International Journal of Computer Science and Network Security*, 7(8), 20-27.
22. Kurniawan, D. (2012). Evaluasi sistem temu kembali informasi model ruang vektor dengan pendekatan user judgement. *Jurnal Sains MIPA Universitas Lampung*, 16(3), 155-162.

23. Kurniasih, D.; and Akbar, F.M. (2021). E-commerce pandemic covid-19 home industries and SMEs. *International Journal of Research and Applied Technology (INJURATECH)*, 1(1), 70-75.
24. Adriani, M.; Asian, J.; Nazief, B.; Tahaghoghi, S.M.; and Williams, H.E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4), 1-33.
25. Mariyandi, D.D.; Sakti, A.W.; and Wulandary, V. (2021). Reading skill of elementary school students and relationship to foreign language (German and Japanese) contained in the text. *International Journal of Research and Applied Technology (INJURATECH)*, 1(1), 84-89.
26. Tamara, Y.; Sakti, A.W.; and Wulandary, V. (2021). Analysis of the level of interest of junior high school students in learning basic Japanese language. *International Journal of Research and Applied Technology (INJURATECH)*, 1(1), 109-114.
27. Rahayu, S. (2022). Implementation of blockchain in minimizing tax avoidance of cryptocurrency transaction in Indonesia. *International Journal of Research and Applied Technology (INJURATECH)*, 2(1), 30-43.
28. Uriawan, W. (2020). SWOT analysis of lending platform from blockchain technology perspectives. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, 4(1), 103-116.
29. Fitriawati, M.; and Lestari, R. (2022). Designing information systems for general administration management in playgroups in North Cimahi district. *International Journal of Research and Applied Technology (INJURATECH)*, 2(1), 54-60.
30. Soegoto, E.; Alifia, N.; Salsabila, T.; and Mardika, C. (2022). The effect of using applications to facilitate medicine purchase amid the covid-19 pandemic. *International Journal of Research and Applied Technology (INJURATECH)*, 2(1), 71-81.
31. HDP, M.Z.Y.; Zidane, G.A.; and Hartaman, R.P. (2022). Application of population data collection in Padasuka village, Bandung regency, based on website. *International Journal of Research and Applied Technology (INJURATECH)*, 2(2), 19-23.
32. Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503-520.