Query Expansion in Information Retrieval using Frequent Pattern (FP) Growth Algorithm for Frequent Itemset Search and Association Rules Mining

Lasmedi Afuan[1], Ahmad Ashari[*2], Yohanes Suyanto[3]
Department of Informatics, Universitas Jenderal Soedirman, Purwokerto, Central Java, Indonesia[1]
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia[2, 3]

***Abstract*—Documents on the Internet have increased in number exponentially; this has resulted in users having difficulty finding documents or information needed. Special techniques are needed to retrieve documents that are relevant to user queries. One technique that can be used is Information Retrieval (IR). IR is the process of finding data (generally documents) in the form of text that matches the information needed from a collection of documents stored on a computer. Problems that often appear on IRs are incorrect user queries; this is caused by user limitations in representing their needs in the query. Researchers have proposed various solutions to overcome these limitations, one of which is to use the Expansion Query (QE). Various methods that have been applied to QE include Ontology, Latent Semantic Indexing (LSI), Local Co-Occurrence, Relevance Feedback, Concept Based, WordNet / Synonym Mapping. However, these methods still have limitations, one of them in terms of displaying the connection or relevance of the appearance of words or phrases in the document collection. To overcome this limitation, in this study we have proposed an approach to QE using the FP-Growth algorithm for the search for frequent itemset and Association Rules (AR) on QE. In this study, we applied the use of AR to QE to display the relevance of the appearance of a word or term with another word or term in the collection of documents, where the term produced is used to perform QE on user queries. The main contribution in this study is the use of Association rules with FP-Growth in the collection of documents to look for the connection of the emergence of words, which is then used to expand the original query of users on IR. For the evaluation of QE performance, we use recall, precision, and f-measure. Based on the research that has been done, it can be concluded that the use of AR on QE can improve the relevance of the documents produced. This is indicated by the average recall, precision, and f-measure values produced at 94.44%, 89.98%, and 92.07%. After comparing the IR process without QE with IR using QE, an increase in recall value was 25.65%, precision was 1.93%, and F-Measure was 15.78%.**

Keywords*—IR; query expansion; association rules; support; confidence; recall; precision

I. INTRODUCTION

The growth of the number of documents on the Internet creates problems for users, where users often find problems that are relevant to their needs. Special techniques are needed to retrieve documents that are relevant to user queries. One technique that can be used is Information Retrieval (IR). Generally the documents of finding data (generally documents) in the form of text that match documents are stored on a computer [1]. IR provides information about the subjects needed. Data includes text, audio, videos, and other documents. IR aims to produce documents that are relevant to the queries that users enter in a short and precise time.

The current IR research appears to be important developments, namely how to index documents and how to retrieve documents that are relevant to user queries [2]. IR research has been carried out at different levels but with the same goal of increasing the relevance of the documents taken. The IR research that has been carried out generally uses keywords in searching document content; often users are less able to represent the information needs needed in the form of queries. So, documents produced by IR are not relevant to the user's wishes. In fact, the number of relevant documents produced is very dependent on the query entered by the user. Vocabulary queries for users who mismatch with documents also cause no documents to be retrieved [3].

A good IR must be able to bridge the potential distance between documents and user queries [3]; to overcome this, research in IR proposes many solutions, one of which is QE [4]. QE is believed to be able to overcome problems related to user query representation. This approach is used to overcome problems in the ineffectiveness of document retrieval by expanding queries to improve the accuracy of user queries, which are believed that queries that are less accurate are the main problems related to the relevance of documents to IR. [5]. The various methods used in QE include Ontology, LSI, Local Co-Occurrence, Relevance Feedback, Concept Based, WordNet/Synonym Mapping. However, these methods still have limitations, one of them in terms of displaying the connection or relevance of the appearance of words or phrases in the document collection. To overcome this problem, this study applies AR to QE. AR, in general, is used to find the relevance of purchasing items, by analyzing the appearance of items in the transaction of daily goods sales. In this study, we apply AR to display the relevance of the appearance of a word/term with other words or terms in the document collection. So that the query used by the user can be expanded according to the relevance of the query to the word or term contained in the document.

The main contribution in this study is the use of Association rules with FP-Growth in the collection of documents to look for the connection of the emergence of

*Corresponding author's email ID: ashari@ugm.ac.id words, which is then used to expand the original query of users on IR.

The remainder of this paper is organized as follows. In Section 2 overview of related work in QE. In Section 3, we explain about Association Rules mining and FP-Growth concept. Methodology that we used in this research is explained in Section 4. Discussion and result showed in Section 5. A small summary and further study of this area will be concluded in Section 6.

II. Related Work

Research on QE has been carried out by several researchers. Based on the literature review that has been carried out, there are several methods that are often used to query expansion, among others, the LSI, Local Co-Occurrence, Relevance Feedback, Concept Based, WordNet/Synonym Mapping methods. Research [6] proposed the use of the Latent Semantic Indexing (LSI) method. This method is very powerful, implemented in two types of algorithms, namely the Singular Value Decomposition (SVD) and Probabilistic LSI. LSI builds semantic space, maps each term into the space and groups it automatically based on the meaning of the term. It is just that with the LSI method, it is difficult to control the degree of query expansion and it could be that expanded queries contain many irrelevant terms. To overcome this, [7] proposed expansion with the local co-occurrence approach, expansion was carried out based on the frequency of occurrence of words in the document collection. This method can increase the effectiveness of IR in the range of 6 to 13%. It is just that this method has not been able to display the connectedness and meaning of the word. Research conducted by[8] propose Query Expansion in Information Retrieval for Urdu. However, using query expansion with the Kullback-Leibler model only increases the MAP value by about 22-24%.

Other research conducted by [9] propose query expansion through term selection on the relevance feedback process based on the Rocchio formula on meeting the XML document information. This approach is able to overcome the two main problems in meeting the XML document information, namely the problem of overlapping the elements are taken and the problem of retrieving irrelevant elements. It is just that the use of relevance feedback is very dependent on the user's judgment, whether the resulting document is relevant or not. So, if the document is considered relevant but actually not, then the results of IR are less relevant. As with the LSI approach, relevance feedback has not been able to display the connectedness and meaning of words. Research conducted [10] proposed QE using ULMS Metathesaurus. User words or queries are mapped into UMLS CUIs by using Meta Map, and then the MRCONSO Metathesaurus table identifies the synonyms of the words and those words used for expansion queries. It is just that using Metathesaurus on several user queries; queries that are expanded actually reduce the performance of IR. Other research conducted by [11] also use WordNet to find synonyms for words entered by the user. The process of query expansion is done by identifying Part of Speech (POS) of each word preprocessing has been done using POS Tagger. Then after that, synonyms are identified for each word to expand the query using WorldNet. The results of the study showed an increase in precision and recall of around 40% and 24% compared to not being carried out by expansion queries.

Research [12] similar to research conducted [10], query expansion is done by mapping words and searching for synonyms of words entered by the user. The query entered by the user is expanded, the relevant word is searched for and reweighting is done. While research conducted by [13] propose two stages in the QE method that is carried out, namely reducing over weighting by grouping terms on queries based on semantic relationships, then using the recursive structure of the Hopfield word network that is most related to other words chosen. For the extraction of candidates the word uses WordNet. The evaluation results using the CACM and CERC collections showed an increase of 4% - 12% using MAP. It is just that the use of WordNet/Metathesaurus on some user queries, the expanded query actually decreases the performance of the IR, besides it is also less able to display connectivity between words.

III. Association Rule Mining

Association Rule is one technique that is in data mining that aims to get the rules of association or relationship between a set of items. Association rules can be obtained from various data sources, including those derived from transactional databases, data warehouses, as well as from other information storage areas. In general, the processed data is homogeneous [14] The first study of the search for association rules is obtained from itemset which often appear together [15]. One algorithm that is often used to search association rules is Apriori [16]. The importance of an association rule can be known by two parameters, namely Support and Confidence. Support is the percentage of the occurrence of a combination of items or support count of the number of items that appear in a set of transactions, and confidence is the strong relationship between items in the association rules. Association analysis is defined as a process for finding all associative rules that meet minimum support requirements, and minimum confidence requirements.

In general, the association rules are obtained as follows: For example, there are I={i1, i2, i3,...in} which is a set of items, while D is a set of transactions, where each transaction T has a set of items where $T \subseteq I$. Every transaction will have a unique TID (Transaction Identifier). Each transaction is said to contain X, a collection of items in I, if $X \subseteq T$. An association rule is formulated with form $X \rightarrow Y$, where $X \subseteq I$; $Y \subseteq I$; and $X \cap Y = \phi$. The $X \rightarrow Y$ rule has support s in transaction D if s% or the number of s in the transaction in D contains $X \cup Y$. Or in other words, support from a rule is the probability of occurrence of X and Y together or the number of events X and Y together. The $X \rightarrow Y$ rule has confidence value c if c% of transaction D contains X also contains Y. Or in other words, the confidence of a rule is consequently a conditional probability Y is true, if X is the antecedent.

A. Support

Support is the probability of an item or set of items in a transactional database as in (1).

$$Support(X) = \frac{n(X)}{n} \quad (1)$$

With n is the total number of transactions in the database, while n (X) is the number of transactions containing X itemset, or support count which is the number of items contained in the transaction.

B. Confidence

Confidence is a conditional probability, for association rules X→Y defined in (2)

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (2)$$

C. FP-Growth

The FP-Growth uses the concept of building trees in item search, not using generate candidates like the Priori Algorithm. This is what causes the FP-Growth Algorithm to be faster than the Apriori Algorithm. The FP-Growth algorithm is divided into three main steps, namely:

1) Phase of generating conditional pattern base: Conditional pattern base is a sub database that contains a prefix path and suffix pattern. Generation of conditional pattern base is obtained through FP-Tree that has been built before.

2) Phase of generating conditional FP-Tree: At this phase, support count of each item in each conditional pattern base is added up, and then each item that has a number of support counts greater than the minimum support count will be generated with a conditional FP-Tree.

3) Phase of searching frequent itemsets: If Conditional FP-Tree is a single path, and then frequent itemsets are obtained by combining items for each conditional FP-Tree. If it is not a single trajectory, then recursive generation of FP-Growth is carried out.

IV. Proposed Method

This research proposes expansion query using Association rules by utilizing the FP-Growth algorithm in frequent itemset search. In general, the architecture of the proposed model is shown in Fig. 1, there are three main processes, among others: the IR process, the process of frequent itemset search with FP-Growth, and the QE process using the Association Rules.

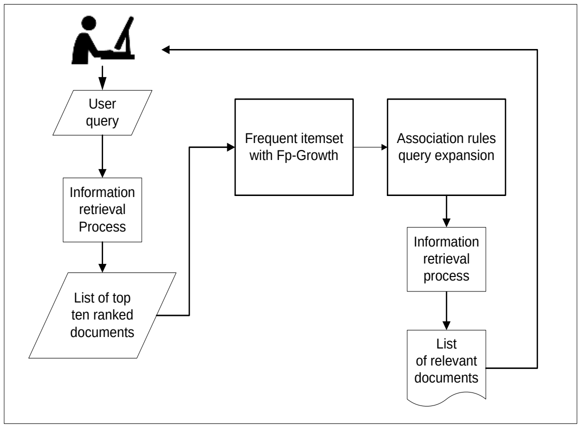
User query
Information retrieval Process
List of top ten ranked documents
Frequent itemset with Fp-Growth
Association rules query expansion
Information retrieval process
List of relevant documents

Fig. 1. Architecture of Proposed Model.

A. User Query

The initial input from the system is the user query. The user enters a query and then pre-processing. There are two things that are done on queries in pre-processing, namely stopword removal and tokenization. For stopword removal, the query entered by the user is deleted by the word/term that often appears but does not have meaning in the query, such as the appearance of the word and, or, while, and is. Queries that have been done stopword removal are then tokenized, which is separating each word into a word or phrase. In pre-processing, we do not stem the query; it aims to not eliminate the meaning of the word from the query. Queries that have been pre-processed will be used as input at the IR process.

1) IR process: The next step The IR process consists of four main processes, namely querying, indexing, searching, and ranking. In the IR process, we adapted the Vector Space Model (VSM). At the IR process, the initial document is produced; the documents will be used as input for the next process.

2) Process of frequent itemset search using FP-growth: The top documents generated from the IR process, furthermore frequent itemset search will be carried out. To search for frequent itemset by using the steps described in section 3.3. On frequent itemset search, we use Rapid Miner software. From this process, a list of the frequent itemset is generated.

3) Process of QE use association rules: The Frequent Itemset was generated. Furthermore, it used to conduct the association rule searches. We use Rapid Miner for generating association rules.

V. Experimental Setup

A. Dataset

In this study, we used a dataset of 100 Indonesian document collections which included lecture material, practicum modules, lecturer presentations, proceedings articles, and journals. These documents are material in the field of Informatics and Computer Science.

B. Testing

The proposed model testing is done by preparing a test scenario using four Indonesian language queries as initial queries to be expanded, as shown in TABLE I. We place the appropriate queries used in the evaluation. Then, calculate the correct number of documents that must be taken, for each query. Furthermore, we run the query and retrieve the top ten documents generated from the retrieval process. The resulting document will be carried out the frequent itemset search process. The frequent itemset is generated. Furthermore, the association rules mining are done. The rules produced are calculated by the value of support and confidence. Rules that have high values will be used as terms used to expand the user's initial query.

To calculate the performance of QE, we use Precision, Recall, and F-Measure evaluation metrics. Evaluation using recall and precision values is done to determine the level of relevance and accuracy of the system in searching for information requested by users. In evaluating the level of relevance, the recall value (R) is a value that shows the rate of return of results returned by a system. This value is obtained by comparing the number of relevant items returned by the system with the total number of relevant items in the system collection as in (3). The greater the recall value cannot show a good system or not. The highest recall value is 1, which means that all the documents in the collection have been found which means that all documents in the collection were found.

Recall

$$R = \frac{TP}{TP+FN} \tag{3}$$

The value of precision (P) shows the level of accuracy of a system to return relevant information to the user. This value is obtained by comparing the number of relevant items returned with the total number of items returned as in (4). The greater the precision value of a system, the system can be said to be good. The highest precision value is 1, which means that all documents found are relevant:

Precision

$$P = \frac{TP}{TP+FP} \tag{4}$$

F-Measure is a combination of recall and precision that takes the harmonic weight of the mean. F-Measure value will be high if recall and high precision, to calculate F-Measure used Eq. (5).

F-Measure

$$R = 2.\frac{PR}{P+R} \tag{5}$$

C. *Experimental Results*

Based on the information retrieval process using the query in Table 1, some initial documents are obtained as shown in Table 2.

Table 2 is the number of initial documents produced in the IR process, then frequent itemset searches are performed using FP-Growth based on the steps in Section 3.3. After successfully obtaining the next frequent itemset, the association rules are searched using the Association rules Algorithm described in Section 3. In the search for association rules, we set the minimum support = 0.1 and minimum confidence = 1. From this process many rules of association are generated, we do pruning of these rules by selecting Right Hand Side (RHS) or consequent from association rules that produce words that match the query in Table 1, the results of pruning association rules are presented in Table 3. Calculation of the value of support and confidence is using (1), (2).

Based on the association rules generated in Table 3, the next step is to expand the query. For example for Q1 queries, the expansion will be "database terminology" or "database concept". The query generated in the query expansion process is then put back into the IR process. Furthermore, precision, recall, and f-measure calculations are using (3), (4), and (5). The results of the comparison between IR without QE and IR with QE are shown in Table 4.

TABLE I. ORIGINAL QUERY

Queries	**Label/caption**
Q1	Database
Q2	Network
Q3	Protocol
Q4	Topology

TABLE II. RESULT OF INFORMATION RETRIEVAL PROCESS

Queries	**Total Document retrieved by system**
Q1	9
Q2	10
Q3	10
Q4	10

TABLE III. LIST OF ASSOCIATION RULES

No	**Queries**	**Association Rules**
1	Q1	R1 terminology → database R2 concept → database
2	Q2	R1 domination → jaringan R2 century → jaringan R3 computer → network
3	Q3	R1 http → protocol R2 computer → protocol R3 network → protocol
4	Q4	R1 http → topology R2 computer→ topology R3 network → topology R4 ring → topology R5 star → topology R6 node → topology

TABLE IV. ANALYSIS RESULTS (RECALL, PRECISION, F-MEASURE)

	IR Without Query Expansion			**IR With Query Expansion (Association Rules)**		
Queries	***Recall (%)***	***Precision (%)***	***F-Measure (%)***	***Recall (%)***	***Precision (%)***	***F-Measure (%)***
Q1	66.67	88.89	76.19	90.00	90.00	90.00
Q2	43.48	83.33	57.14	98.36	84.51	90.91
Q3	75.00	90.00	81.82	95.65	91.67	93.62
Q4	90.00	90.00	90.00	93.75	93.75	93.75
Average	68.79	88.06	76.29	94.44	89.98	92.07

Fig. 2, 3 and 4 show graphics a comparison between recall, precision, and f-measure from IR in the original query and IR with QE.

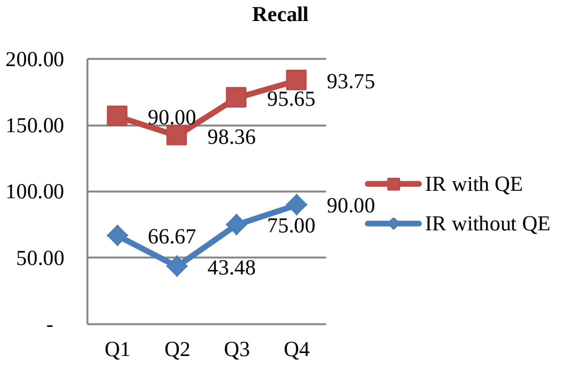

Fig. 2. Recall.

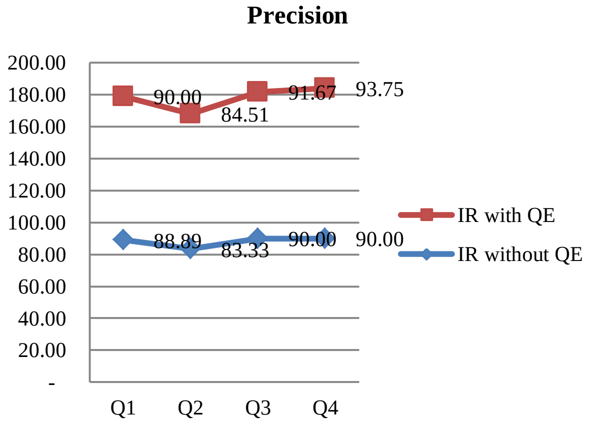

Fig. 3. Precision.

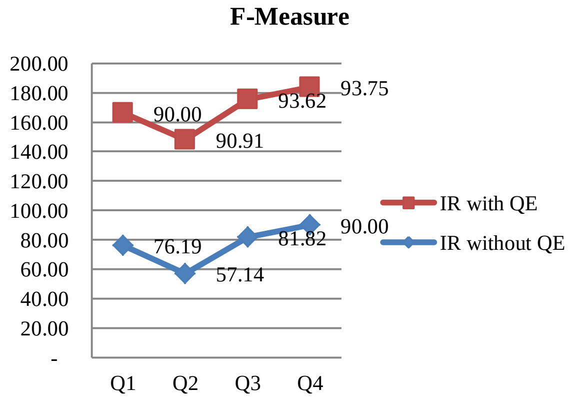

Fig. 4. F-Measure.

VI. Conclusion

In this study, we have applied the use of AR to QE. AR is used to perform rule search related to the appearance of the word/term in the document simultaneously. Based on the research that has been done, it can be concluded that the use of AR on QE can improve the relevance of the documents produced. This is indicated by the average recall, precision, and f-measure values produced at 94.44%, 89.98%, and 92.07%. After comparing the IR process without QE with IR using QE, the recall value increased by an average of 25.65%, precision 1.93%, and F-Measure by 15.78%.

For further research, we are trying to integration between AR and ontology on QE.

Acknowledgments

The authors would like to thank the reviewers for their detailed, accurate and helpful comments. Also, thank my colleagues for support us this research. Also thank Ministry of research, technology, and higher education for funding support to this study.

References

[1] C. D. Manning, P. Raghavan, and H. Schutze, "An Introduction to Information Retrieval," Online, no. c, p. 569, 2009.

[2] B. M. Sanderson and W. B. Croft, "The History of Information Retrieval Research," in IEEE, 2012, vol. 100, pp. 1444–1451.

[3] D. Pal, M. Mitra, and S. Bhattacharya, "Exploring Query Categorisation for Query Expansion : A Study," CoRR, pp. 1–34, 2015.

[4] A. Abbache, F. Meziane, G. Belalem, and F. Z. Belkredim, "Arabic Query Expansion Using WordNet and Association Rules," Int. J. Intell. Inf. Technol., vol. 12, no. 3, 2016.

[5] J. Ooi and H. Qin, "A Survey of Query Expansion , Query Suggestion and Query Refinement Techniques," Int. Conf. Softw. Eng. Comput. Syst., pp. 112–117, 2015.

[6] S. Deerwester, S. T. Dumais, and R. Harshman, "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.

[7] M. Mitra, C. Buckley, and F. Park, "Improving Automatic Query Expansion," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

[8] I. Rasheed, "Query Expansion in Information Retrieval for Urdu Language," 2018 Fourth Int. Conf. Inf. Retr. Knowl. Manag., pp. 1–6, 2018.

[9] M. Mataoui, F. Sebbak, F. Benhammadi, and K. B. Bey, "Query Expansion in XML Information Retrieval A new Approach for terms selection M'hamed," in Modeling, Simulation, and Applied Optimization (ICMSAO), 2015, pp. 4–7.

[10] M. R. A. Nawab, M. Stevenson, and P. Clough, "An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE," J. Comput. Biol. Bioinforma., vol. 5963, no. APRIL 2015, pp. 1–9, 2015.

[11] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, "Query Expansion via Wordnet for Effective Code Search," IEEE, pp. 545–549, 2015.

[12] A. Babu and S. L, "An Information Retrieval System for Malayalam Using Query Expansion Technique," Int. Conf. Adv. Comput. Commun. Informatics, pp. 1559–1564, 2015.

[13] A. Noroozi and R. Malekzadeh, "Integration of Recursive Structure of Hopfield and Ontologies for Query Expansion," Int. Symp. Artif. Intell. Signal Process., 2015.

[14] R. Gunawan and K. Mustofa, "Pencarian Aturan Asosiasi Semantic Web Untuk Obat Tradisional Indonesia," JNTETI, vol. 5, no. 3, pp. 192–200, 2016.

[15] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association in Large Databases," Proc. 1993 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '93, pp. 207–216, 1993.

[16] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceeding VLDB '94 Proc. 20th Int. Conf. Very Large Data Bases, vol. 1215, pp. 487–499, 1994.