Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 16-02-2026; Accepted 24-05-2026; Published online 31-05-2026

Improving Information Retrieval Using Association Rule–Based Query Expansion

Fadi Yamout FADI.YAMOUT@LIU.EDU.LB
Lebanese International University,
School of Arts & Sciences,
Computer Science & Information Technology,
Beirut, Lebanon

Hussein Chawich HUSSEIN.CHAWICH@LIU.EDU.LB
Lebanese International University,
School of Arts & Sciences,
Computer Science & Information Technology,
Beirut, Lebanon

Corresponding Author: Fadi Yamout

Copyright © 2026 Fadi Yamout and Hussein Chawich. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Searching for documents using search engines is done by comparing the query submitted by the user to the documents in the collection. The submitted query contains one or more words carefully selected to retrieve as many relevant documents as possible. There have been many approaches that performed query expansion using synonyms. Retrieval effectiveness can be increased using query expansion, like the query term "PC" can be expanded with "Laptop", "Desktop", and "Computer", then documents containing these synonyms will also be retrieved. This paper aims in introducing a new model that mines association rules from the document collection itself using the Apriori algorithm and expands the query with words learned from association rule mining. Confidence values for association rules are used, where terms with high Confidence will be added to our query, which are learned from association rule mining. The query without any expansion will be our baseline query. We use a vector space model with TF-IDF weighting and cosine similarity. The proposed technique is evaluated on the Medline test collection, consisting of 1,033 documents and 30 queries. Queries are first run as-is, then expanded using associated terms, and the results are compared using Precision and Recall. This approach can complement existing query expansion methods by adding terms automatically discovered from the document collection itself, without relying on external knowledge sources.

Keywords: Information retrieval, Data mining, Association rules, Query expansion.

Citation: Fadi Yamout and Hussein Chawich. Improving Information Retrieval Using Association Rule–Based Query Expansion. **Advances in Artificial Intelligence and Machine Learning**. 2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.63309

1. INTRODUCTION

Major problems dealt with by Information Retrieval(IR) systems can be: getting hold of documents from a variety of sources, representing documents in an effective manner so that they can be retrieved easily, ranking different results obtained after a query to match relevance, and outputting these documents to the user as quickly as possible. The objective of IR systems is to retrieve the maximum number of relevant documents with the minimum number of non-relevant documents. The steps involved include indexing, modeling, querying, ranking, and evaluation [1, 2].

Query expansion techniques based on synonym relations use additional terms, closely related or synonymous to the query terms, and retrieve documents matching these expanded queries to improve recall values. Many previous techniques for query expansion use information gathered from external resources or pseudo-relevance feedback or make use of complex semantic representations [3].

1.1 Data Mining

Data mining helps discover knowledge from large databases. This technology uncovers hidden information which can predict unknown patterns [4]. It extracts implicit, previously unknown, and potentially useful relationships/patterns from data.

1.2 Association Rule Mining

Association rule mining discovers rules of the form $X \rightarrow Y$ (called an association rule) which imply a strong relationship between item X and item Y. An association rule has a confidence associated with it [5, 6]. Confidence measures how often the rule is found to be true. For example, if there is a rule {music magazine $\rightarrow$ car magazine} with a confidence of 60%. It means 60% of the users who buy a music magazine also buy a car magazine. The Apriori algorithm is used to extract these rules efficiently from large databases [6].

1.3 Main Contribution

We propose a query expansion method which aims to retrieve relevant documents efficiently. Words associated with terms in the query are mined using association rule mining from the collection of documents itself. Terms with higher Confidence are added to the query for retrieval. It is expected that only terms having a strong association with query words are considered which should increase Precision at higher recall values than that of baseline retrieval(which uses only query terms) [3, 5].

1.4 Methodology

Tests were carried out on the Medline test collection which contains a total of 1,033 documents and 30queries. The Vector Space Model with TF-IDF [7] weighting and cosine similarity measure is used for the retrieval of documents. Queries in both their original and expanded(using word

associations mined) form are run on the collection. Documents retrieved by both these queries were used to plot a precision-recall graph to compare the effectiveness of both the methods of querying directly [1, 2].

1.5 Paper Description

Section 2 consists of related work and previous methods of query expansion discussed. The proposed method has been discussed in detail in Section 3. Section 4 describes the experimental setup, the dataset used, and the implementation details of the proposed approach. Results and discussions have been depicted in Section 5. Conclusion and Future work have been presented in section 6.

2. RELATED WORK

Query expansion methods are used to increase retrieval effectiveness by considering additional terms that were not entered by the user but can be derived by query expansion algorithms. Queries are usually shorter than a user's actual information need. Due to this low Recall of queries, there is also low Precision, causing users to not receive many relevant documents as search results. As a result, query expansion techniques have been explored [8, 9].

Synonyms, semantic resources, or statistical methods based on term distribution are common methods for query expansion. However, they increase Recall, but many rely on external knowledge bases or pseudo-relevance feedback [10].

Newer methods have examined utilizing large language models (LLMs) and corpus-aware methods to retrieve expansion terms. Zhang et al. research how LLMs are able to generate terms that are semantically related to expand a query [9]. Lei et al. introduce a corpus-guided query expansion method that extracts relevant sentences from documents that are retrieved to use as query expansion terms and showed that utilizing relevant sentences from retrieved documents can better estimate query-document relevance [11].

Other approaches apply data mining techniques to enhance retrieval. Rajkumar and Vignesh integrate association rule mining with natural language processing techniques to discover term relationships directly from the document collection [12]. These methods improve Precision and Recall, but many still rely on additional semantic resources, pseudo-relevance feedback, or complex deep learning architectures.

In contrast, the proposed approach mines association rules directly from the document collection using the Apriori algorithm, adding terms to the query based on the Confidence of the rules. Retrieval is performed using a classical vector space model with TF-IDF weighting [7] and cosine similarity, providing a simple yet effective query expansion technique without external resources [10, 12]. This approach addresses a gap left by existing methods and improves Precision at higher recall levels, as highlighted in our experimental evaluation.

3. METHODOLOGY

The methodology describes the technique used to retrieve documents efficiently using a query expansion method implemented through the Association Rule mining technique. Query expansion is commonly implemented in information retrieval systems. It works by expanding the original query used for searching with related terms to improve retrieval effectiveness [1, 3]. The proposed technique aims to expand the user query by identifying relationships between terms that frequently occur together within the collection of documents. The relationship between terms is mined through association rules generated by applying the Apriori algorithm [6].

3.1 Association Rules Generation

The initial stage involves extracting association rules from the collection of documents available. Documents are considered transactions, and the terms appearing within those documents are considered items present in those transactions. Association rule mining is first applied to the collection of documents after preprocessing it by indexing and term weighting the collection using the TF-IDF model.

Terms are then filtered out from documents if they have very low TF-IDF weights. This filtration technique helps in lowering the number of candidate terms by removing non-informative words present within the document collection. Mining is then performed on the processed set of documents by applying the Apriori algorithm, which generates association rules between terms that are present within those documents [4, 6].

An association rule is of the form $X \rightarrow Y$, where X represents the antecedent term, and Y is associated with X. Rules have two values attached to them, known as support and Confidence, which are used to determine their accuracy. Assume, for example, that the query contains the words "house computer computer pen", then the set Level-1 has four words as shown in FIGURE 1:

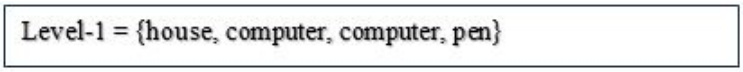

Figure 1: The set Level-1

Shown below in TABLE 1 are examples of rules mined from the collection of documents.

3.2 Query Expansion with Association Rules

The second step involves expanding the user's query using the generated rules. Words contained in the original query are called Level-1 terms because they reflect the user's expressed information need.

The database of rules is searched for words that have been associated with Level-1 terms. Only rules with support and Confidence greater than experimentally determined minimum levels are used. See TABLE 2 for rules remaining after minimum support and confidence levels are specified.

Table 1: Example of rules generated with the corresponding Confidence and support

Number	Association Rules	Confidence	support
1	house → home, roof	60%	50%
2	computer → PC	60%	50%
3	computer → mainframe	10%	10%
4	pen → book	40%	70%
5	home → country	90%	60%
6	home → Lebanon	80%	20%
7	roof → ceiling	80%	70%
8	roof → top	10%	90%
9	PC → Toshiba	10%	90%
10	PC → HP	10%	90%
. . .	. . .	. . .	. . .

Table 2: Rules remaining after specifying minimum Confidence and support

Number	Association Rules	Confidence	support
1	house → home, roof	60%	50%
2	computer → PC	60%	50%
3	**computer → mainframe**	**10%**	**10%**
4	**pen → book**	**40%**	70%
5	home → country	90%	60%
6	**home → Lebanon**	80%	**20%**
7	roof → ceiling	80%	70%
8	**roof → top**	**10%**	90%
9	**PC → Toshiba**	**10%**	90%
10	**PC → HP**	**10%**	90%
. . .	. . .	. . .	. . .

For the set Level-1 = {house, computer, computer, pen}, we find word "home" and "roof" associated with "house, the word "PC" associated with "computer", and none of the words are associated with "pen". Therefore, we label the new set of words added to the query as the set Level-2. Set Level-2 is now = {home, roof, PC, PC}; the word "PC" is repeated twice since it is associated with the word "computer" that exists twice in Level-1. The updated query now has eight words, as shown in FIGURE 2

Level-1 = {house, computer, computer, pen}

Level-2 = {home, roof, PC, PC}

Figure 2: The sets Level-1 and Level-2

The new model searches for more words that are associated with the words in set Level-2 with a specified minimum support defined by the user. For the set Level-2 = {home, roof, PC, PC}, we find the word "country" associated with "home", and the word "ceiling" associated with "roof", and none of the words are associated with "PC". Therefore, the new set of words added to the query is labeled as set Level-3. Set Level-3 is now = {country, ceiling}. The updated query now has ten words as shown in FIGURE 3:

Level-1 = {house, computer, computer, pen}
Level-2 = {home, roof, PC, PC}
Level-3 = {country, ceiling}

Figure 3: The sets Level 1, 2, and 3

We consider the words in Level-1 are more significant than the words in Level-2 since the user originally chose these words. The same applies to Level-2 and Level-3, where we consider the words in Level-2 to be more significant than the words in Level-3. Therefore, Level-2 = {home, roof, PC, PC} becomes {home, roof, PC, PC, home, roof, PC, PC} and the final query will now contain 14 words as shown in FIGURE 4:

Level-1 = {house, computer, computer, pen},
Level-2 = {home, roof, PC, PC, home, roof, PC, PC}
Level-3 = {country, ceiling}

Figure 4: Levels after doubling β

3.3 Thresholds for the Levels

We multiply each level by thresholds α for Level-1, β for Level-2, and δ for Level-3; α, β, and δ are initially equal to 1. If we double the threshold value of Level-2 by making $\beta = 2$, then we double the words in set Level-2. The syntax used to indicate the minimum Confidence, minimum support, threshold values of the three levels is shown in FIGURE 5:

| S | Support | C | Confidence | - | α | β | δ |

Figure 5: The syntax used for thresholds

For example, S5C7-121 indicates that the words selected are based on a minimum support of 5, a minimum Confidence of 7, and the words in set Level-2 were doubled. S5C7-210 would double the words in set Level-1, and remove the words from set Level-3.

3.4 The new model's pseudocode

The pseudocode for the new model is shown in FIGURE 6.

```
1: Document Preprocessing & Index Construction
FOR each document d ∈ D DO
    Remove stop words, & apply Stemming
    FOR remaining terms t in document d DO
        Compute TF-IDF(t,d)
        Store in InvertedIndex
    END FOR
END FOR
2: Association Rule Generation
Transactions ← documents as sets of terms
Generate the itemsets & FrequentItemsets
Rules ← Ø
FOR each FrequentItemsets > min_support DO
    Generate candidate rules X → Y
    Compute Confidence(X → Y)
        Add rule to Rules
    END IF
END FOR
    END FOR
```

```
3: Query Expansion
FOR each query q ∈ Q DO
    FOR each term t ∈ q DO
        Find rules (t → x) in Rules
        Add x to q
        Duplicates words based on αβδ
4: Querying & Ranking
FOR each query q (original & expanded) DO
    Represent q as a TF-IDF vector
    FOR each document d ∈ D DO
        Compute cosine between q & d
    END FOR
    Rank documents in descending similarity
END FOR
5: Evaluation
FOR each query result DO
    Compute Precision & Recall
END FOR
```

Figure 6: The pseudocode for the new model

3.5 Summary

The technique presented combines association rule mining with the basic Information Retrieval model to automatically learn how to expand user queries. Instead of using an external lexical resource or neural networks to learn associations between terms [8, 11], this method mines for patterns present in the document collection to determine expansion terms. This technique can be used alongside current query expansion methods, such as semantic [5] or machine-learning-based techniques [9] to augment queries with learned associated terms.

4. EXPERIMENTAL SETUP

4.1 Test Collection

The test collections employed in this study are based on the Medline test collection, which in turn is one of the collections from TextREtrieval Conference (TREC) benchmark collections. Collections provided by the TREC project have been extensively used in the IR community as standard test-beds to assess retrieval performance and query expansion approaches [13]. Medline is a set of journal articles which provides manually created relevance judgments done by experts on the retrieval of medical documents. In total, there are 1,033 documents and 30 queries in the Medline test collection used in this work. Each query in the test collection is accompanied by a set of relevant documents for it, allowing for the calculation of standard performance measures. TABLE 3. Characteristics of the test collection used give a brief summary of the employed test collection.

Table 3: Test collection

Name of Test Collection	Number of documents	Number of queries
Medline	1,033	30

The dataset was chosen to conduct experiments in a well-defined setting where different retrieval methods can be tested and compared according to standard relevance judgments [14].

4.2 Evaluation Metrics

Evaluation of experimental results: To analyze how well the proposed query expansion technique is working, we measure the performance with the help of Precision and Recall, two popular evaluation measures used in the Information Retrieval domain. Precision: shows the percentage of retrieved documents that are relevant. It is defined as

$$Precision = \frac{Relevant\ Documents\ Retrieved}{Retrieved\ Documents}$$

Recall: shows the percentage of relevant documents that are retrieved successfully. It is defined as

$$Recall = \frac{Relevant\ Documents\ Retrieved}{Relevant\ Documents}$$

Precision and Recall are popular evaluation measures as they not only tell how good the retrieval process is, but also help us to determine the trade-off between retrieving as many documents as possible while keeping Precision high [1]. In our experiments, we calculate these metrics for our baseline system, which uses the original queries as they are. We then calculate these metrics for the system using queries generated after applying our proposed association rule technique. We finally compare the metrics obtained from these two configurations to analyze how our proposed query expansion technique has impacted the retrieval process.

5. RESULTS

This section describes the experiments done using the new model. The new model is tested on the Medline test collections composed of 1033 documents with 30 queries. The intention from the experiments is to get better Precision than the baseline at higher recall values [15] since this will be equivalent to search engines getting more relevant documents on the first pages. Support = 6% and Confidence = 10% or 15% were chosen based on the Support-Confidence plot of the 1,120 mined rules. Support = 6% was chosen because Support values lower than this (e.g., between 3 and 5%) pruned too many rules such that rules could not be expanded upon, whereas Support values greater than or equal to 7% would not prune enough rules that had items with low frequency. Confidence values of 10% and 15% were chosen because they are below (10%) and around the average (15%) of the Confidence range, which has a median of 13.5% and a mean of 15.9%. These percentiles were chosen to reflect rules with below average and average.

5.1 Combination of threshold values

We have tried all possible combinations for the thresholds, starting from $\alpha = 0$, $\beta = 0$, and $\delta = 0$ to $\alpha = 6$, $\beta = 6$, and $\delta = 6$. We have then dropped all the threshold values that have $\alpha = 0$ since the results do not include the words of the query. It should be noted here that if we assign the same value

for all thresholds, the results will not be affected. For example, the results we get when applying $\alpha = 1$, $\beta = 1$, and $\delta = 1$ will be the same as applying $\alpha = 2$, $\beta = 2$, and $\delta = 2$, and the same as applying $\alpha = 3$, $\beta = 3$, and $\delta = 3$. In another example, the results we get when applying $\alpha = 2$, $\beta = 1$, and $\delta = 3$ will be the same as applying $\alpha = 4$, $\beta = 2$, and $\delta = 6$. The threshold values giving the best results are: $\{\alpha = 3, \beta = 1, \delta = 2\}$,$\{\alpha = 3, \beta = 2, \delta = 1\}$, $\{\alpha = 5, \beta = 2, \delta = 1\}$, $\{\alpha = 5, \beta = 3, \delta = 1\}$, $\{\alpha = 6, \beta = 3, \delta = 1\}$, $\{\alpha = 6, \beta = 4, \delta = 0\}$, $\{\alpha = 6, \beta = 4, \delta = 1\}$, and $\{\alpha = 6, \beta = 5, \delta = 1\}$.

5.2 Baseline results

The baseline results (FIGURE 7) are the same as running the new model with $\alpha = 1$, $\beta = 0$, and $\delta = 0$, since we will be using the same words in the query without expanding it with additional words, since the threshold values of Level-2 and Level-3 are both equal to zero. It shows Precision in the retrieval of relevant documents first, then degrades to retrieve the less relevant ones.

Recall	Precision Baseline
0.0	0.8008
0.1	0.7576
0.2	0.6666
0.3	0.6362
0.4	0.5858
0.5	0.5120
0.6	0.4610
0.7	0.4274
0.8	0.3932
0.9	0.3631
1.0	0.3470

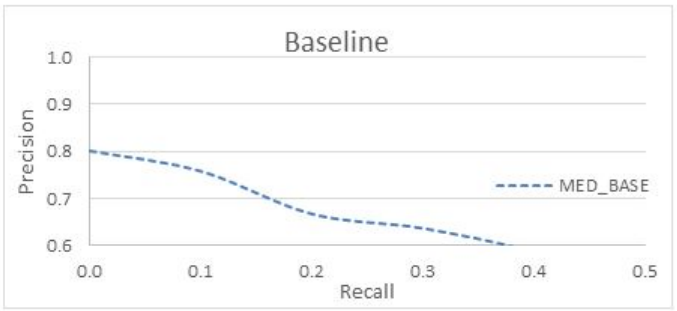

Figure 7: Results for the baseline

5.3 Results of Confidence = 10% and Support = 6%

We have run experiments with variant threshold values with Confidence = 10% and support = 6%. We limited the results to the ones that showed the improvements, as shown in TABLE 4:

In FIGURE 8, our model outperforms the baseline at low recall levels, suggesting that our enhancement returns more relevant results while the Recall is less than 0.25. Our model again demonstrates higher Precision (FIGURE 9) than the baseline at low recall levels, with a Precision of 0.8820 compared to 0.8008 at zero Recall, suggesting a consistent improvement in retrieving highly relevant documents when the result set is small.

In FIGURE 10, our improvement consistently improves result relevance across a broad range of retrieval depths. Nevertheless, Precision still falls below the baseline at higher recall levels (0.26 and beyond). The new model, in FIGURE 11, shows improved Precision over the baseline at low recall levels, with a precision of 0.8154 versus 0.8008 at zero Recall. Showing the same consistent pattern across the experiments.

Table 4: Results of Confidence = 10% and Support = 6%

		Precision							
		Baseline	S6C10-312	S6C10-321	S6C10-521	S6C10-531	S6C10-631	S6C10-640	S6C10-651
Recall	0.0	0.8008	0.8232	0.8220	0.8133	0.8154	0.8172	0.8165	0.8142
	0.1	0.7576	0.7696	0.7649	0.7657	0.7637	0.7615	0.7664	0.7555
	0.2	0.6666	0.6816	0.6768	0.6814	0.6816	0.6810	0.6686	0.6676
	0.3	0.6362	0.6194	0.6103	0.6273	0.6176	0.6225	0.6338	0.5881
	0.4	0.5858	0.5583	0.5483	0.5695	0.5594	0.5664	0.5804	0.5250
	0.5	0.5120	0.4776	0.4683	0.4950	0.4780	0.4879	0.5101	0.4478
	0.6	0.4610	0.4233	0.4203	0.4474	0.4259	0.4368	0.4651	0.3985
	0.7	0.4274	0.3755	0.3694	0.4001	0.3778	0.3900	0.4230	0.3527
	0.8	0.3932	0.3356	0.3342	0.3617	0.3388	0.3503	0.3878	0.3234
	0.9	0.3631	0.2972	0.2966	0.3204	0.3008	0.3084	0.3546	0.2862
	1.0	0.3470	0.2670	0.2650	0.2874	0.2695	0.2772	0.3257	0.2558

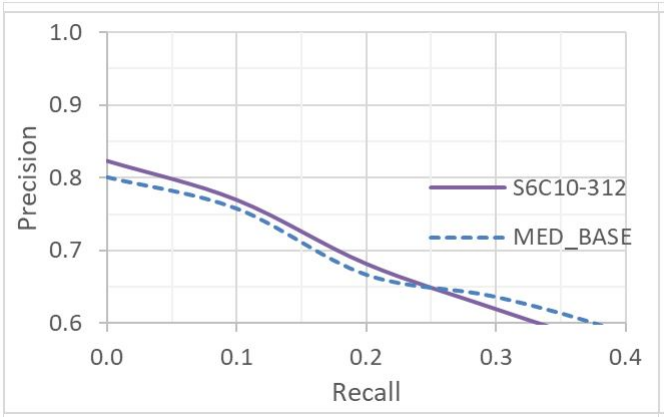

Figure 8: S6C10-312 with the baseline

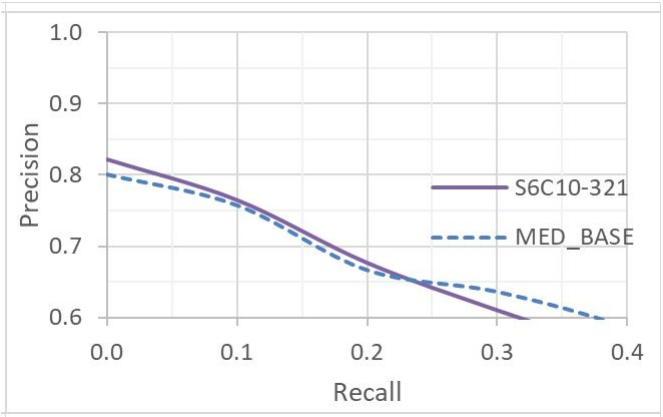

Figure 9: S6C10-321 with the baseline

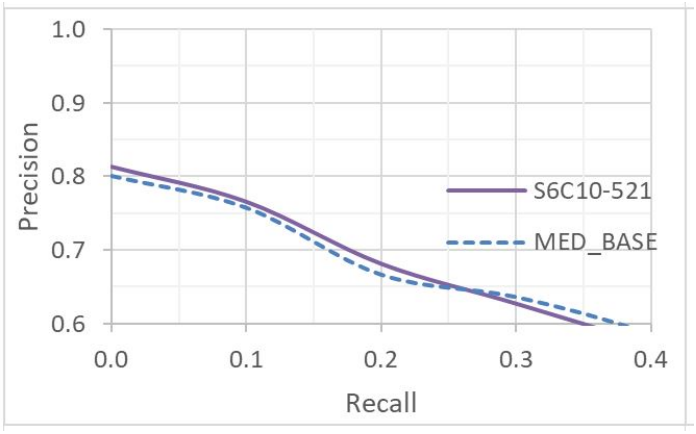

Figure 10: S6C10-521 with the baseline

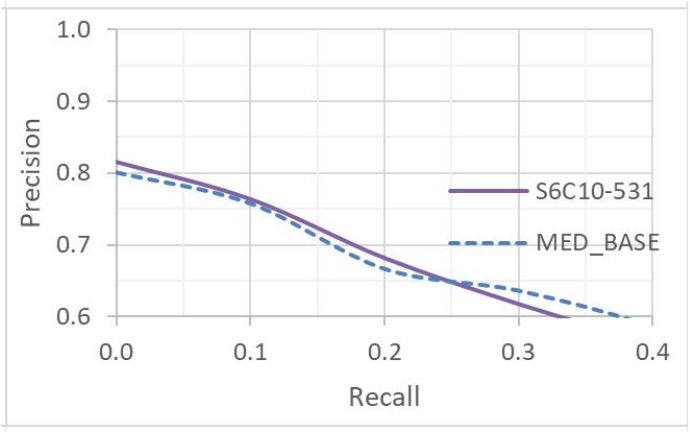

Figure 11: S6C10-531 with the baseline

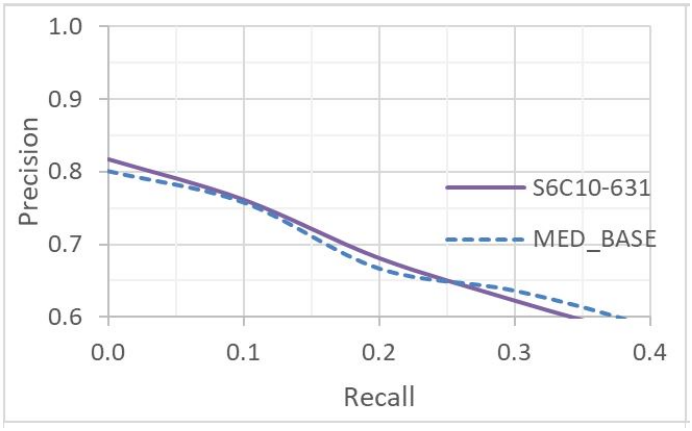

Figure 12: S6C10-631 with the baseline

The result shown in FIGURE 12 is consistent with the pattern observed across previous experiments, further confirming that our enhancement reliably improves Precision for top-ranked results. In FIGURE 13, our model demonstrates its closest performance to the baseline yet, maintaining higher Precision through the 0.2 recall point and remaining competitive with only marginal differences at mid-range recall levels.

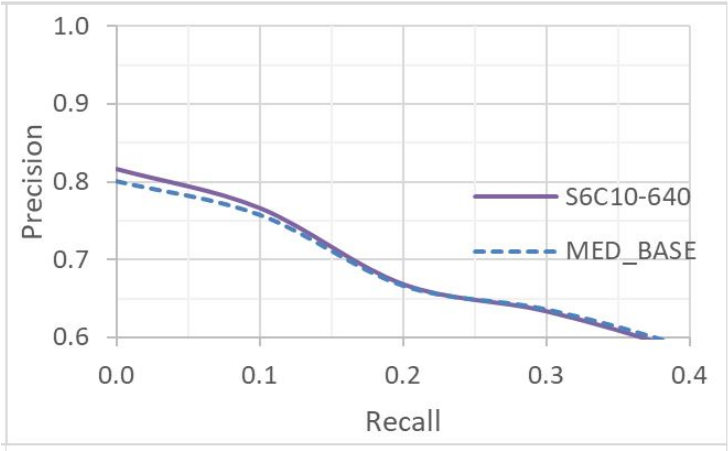

Figure 13: S6C10-640 with the baseline

Finally, FIGURE 14 shows that this configuration appears to be the weakest performer overall.

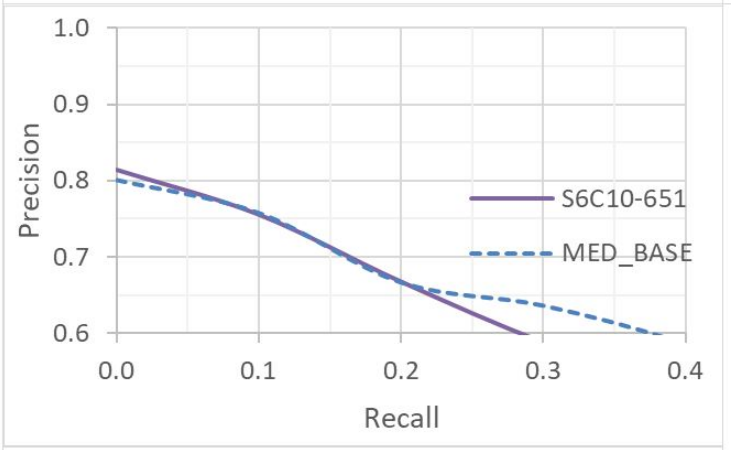

Figure 14: S6C10-651 with the baseline

5.4 Results of Confidence = 15% and Support = 6%

We have run experiments with variant threshold values with Confidence = 15% and support = 6%. We limited the results to the ones that showed the improvements, as shown in TABLE 5:

Table 5: Results of Confidence = 15% and Support = 6%

		Precision							
		Baseline	S6C10-312	S6C10-321	S6C10-521	S6C10-531	S6C10-631	S6C10-640	S6C10-651
Recall	0.0	0.8008	0.8160	0.8160	0.8163	0.8165	0.8173	0.8175	0.8113
	0.1	0.7576	0.7681	0.7681	0.7693	0.7685	0.7689	0.7672	0.7568
	0.2	0.6666	0.6823	0.6832	0.6763	0.6808	0.6820	0.6689	0.6782
	0.3	0.6362	0.6362	0.6368	0.6361	0.6374	0.6383	0.6342	0.6317
	0.4	0.5858	0.5779	0.5754	0.5800	0.5773	0.5783	0.5779	0.5666
	0.5	0.5120	0.5080	0.5054	0.5095	0.5062	0.5079	0.5094	0.5009
	0.6	0.4610	0.4641	0.4620	0.4680	0.4617	0.4645	0.4639	0.4548
	0.7	0.4274	0.4165	0.4112	0.4222	0.4126	0.4186	0.4239	0.4040
	0.8	0.3932	0.3806	0.3788	0.3864	0.3818	0.3863	0.3896	0.3696
	0.9	0.3631	0.3415	0.3349	0.3505	0.3377	0.3435	0.3535	0.3268
	1.0	0.3470	0.3173	0.3085	0.3258	0.3117	0.3194	0.3324	0.3020

FIGURE 15 shows that our model maintains higher Precision than the baseline from 0.0 through 0.3 and remains closely competitive at mid to high recall levels with only minimal differences. This consistency further reinforces that our model (FIGURE 16) reliably delivers precision gains at low Recall while maintaining a relatively small performance gap at higher recall levels, marking it as one of the more stable and balanced configurations observed across all experiments.

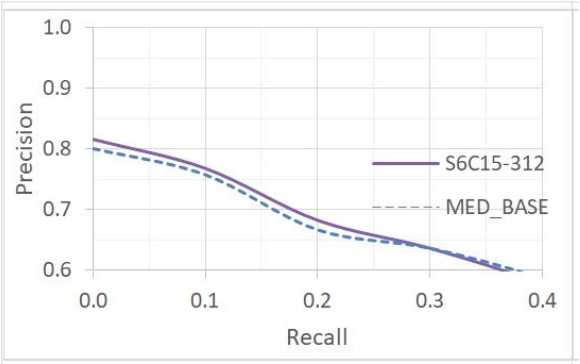

Figure 15: S6C10-312 with the baseline

In FIGURE 17, this configuration represents one of the best overall balances between early precision gains and late-recall performance among all tested settings. FIGURE 18 shows a stable pattern where our model reliably delivers early precision improvements while keeping the performance gap at higher recall levels relatively contained.

FIGURE 19 shows that it continues the trend of strong, balanced results observed in this group of experiments. FIGURE 20 shows near-baseline performance at higher recall levels, combined with early precision improvements.

In FIGURE 21, this relatively early crossover and consistent underperformance beyond the lowest recall point place our technique among the weaker results in the experimental set.

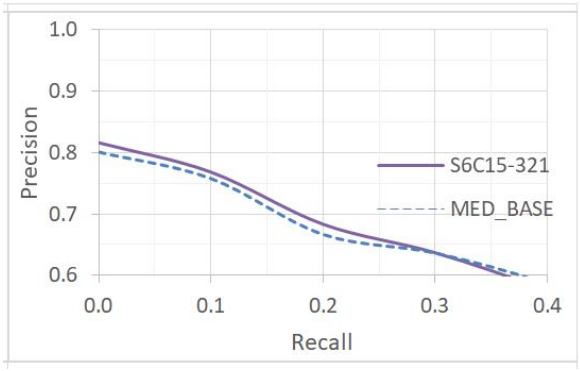

Figure 16: S6C10-321 with the baseline

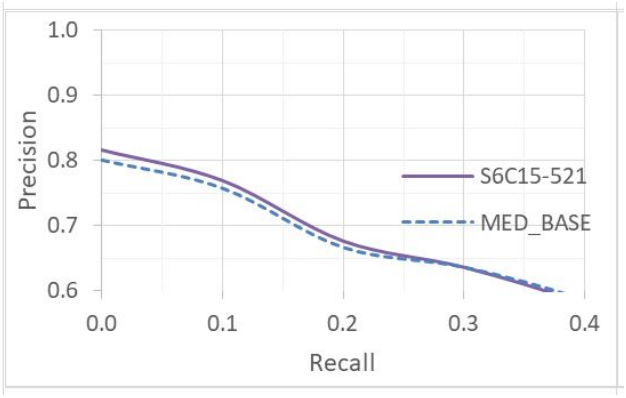

Figure 17: S6C10-521 with the baseline

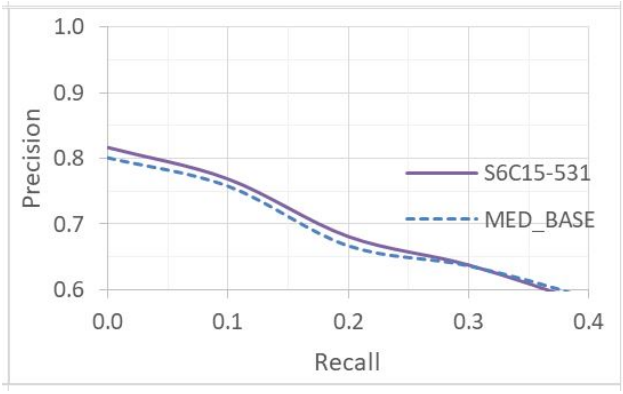

Figure 18: S6C10-531 with the baseline

6. CONCLUSION

The query expansion technique discussed in this paper uses association rule mining in order to mine relationships between terms within a document collection and retrieve documents effectively based

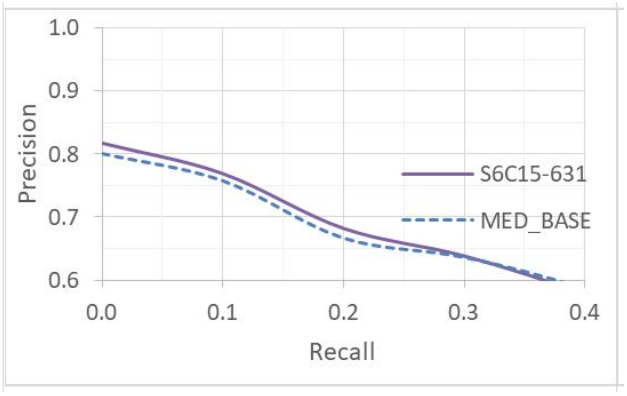

Figure 19: S6C10-631 with the baseline

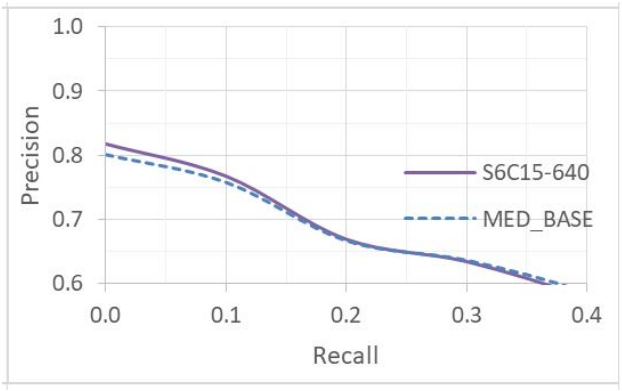

Figure 20: S6C10-640 with the baseline

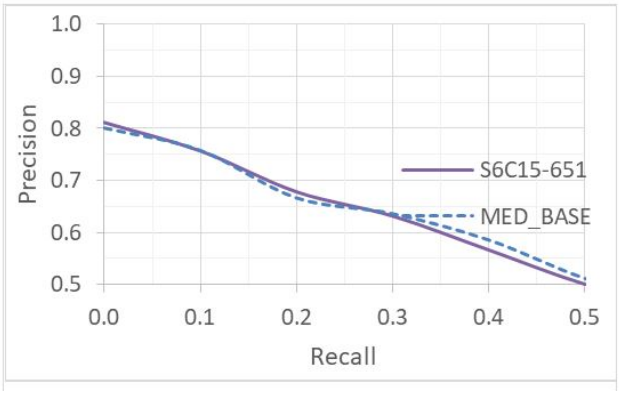

Figure 21: S6C10-651 with the baseline

on those relationships. The difference between this method and other expansion techniques is that association rule mining looks at word relationships within a given collection rather than synonyms or other lexical tools that may have been manually developed outside the document collection itself. Additionally, the query expansion method discussed here automatically expands the query by using

the Apriori algorithm to find relationships between terms in the document collection, then associates terms that meet a minimum confidence value to the query terms. Information retrieval is accomplished using the Vector Space Model with TF-IDF weighting and cosine similarity. Minimum support and confidence values were used to test the effectiveness, as well as threshold values α, β, and γ to limit query expansion to one level, two levels, or any number of levels. The method was tested on the Medline test collection, which contains 1,033 documents and 30 queries. Precision and Recall were used to calculate effectiveness. It was found that a minimum support of 6% and a minimum Confidence of either 10% or 15% produced the best results with certain threshold values. Another takeaway from the results was that added expansion (associating terms to the query) performed better than reformed expansion (replacing the query terms with associated terms).

7. FUNDING DETAILS

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

8. CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper.

9. ACKNOWLEDGMENTS

The authors would like to thank the Lebanese International University for its continuous support of research and innovation. The academic environment and resources provided by the university played an important role in enabling the successful completion of this work.

References

[1] Vechtomova O. Review of: Introduction to Information Retrieval, by Manning CD, Raghavan P, Schütze H. Comput Linguist. 2009;35(2):307-309.

[2] Yamout F, Mouallem Y. Effects of Weighting Schemes on Retrieval Precision. In 2024, the IEEE International Conference on Progress in Informatics and Computing. IEEE. 2024:152-156

[3] Allahim A, Cherif A, Imine A. Semantic Approaches for Query Expansion: Taxonomy, Challenges, and Future Research Directions. Peer J Comput Sci. 2025;11:e2664.

[4] Larose DT, Larose CD. Data mining and predictive analytics. 2nded. Hoboken, NJ: John Wiley & Sons. 2015.

[5] Rajkumar M, Vignesh S. Data-Driven Information Retrieval Using Association Rule Mining and NLP-Based Stemming Techniques. Int J Comput Sci Eng Tech. 2025;9:273-284.

[6] Agrawal R, Imieliński T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data. 1993:207-216.

[7] Yamout F, Lakkis R. Improved TFIDF Weighting Techniques in Document Retrieval. In Thirteenth International Conference on Digital Information Management (ICDIM). IEEE. 2018:69-73.

[8] Azad HK, Deepak A. Query Expansion Techniques for Information Retrieval: A Survey. 2017. ArXiv Preprint: https://arxiv.org/pdf/1708.00247

[9] Zhang L, Wu Y, Yang Q, Nie JY. Exploring the Best Practices of Query Expansion With Large Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2024. ACL. 2024:1872-1883.

[10] Li M, Lv X, Zou J, Chen T, Zhang C, et al . Query Expansion in the Age of Pre-Trained and Large Language Models: A Comprehensive Survey. 2025. arXiv preprint: https://arxiv.org/pdf/2509.07794

[11] Lei Y, Cao Y, Zhou T, Shen T, Yates A. Corpus-Steered Query Expansion With Large Language Models. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers). 2024:393-401.

[12] Brahme A, Shaikh S, Lokare S, Kulkarni S, Mundhe S, et al. Association Rule Mining and Information Retrieval Using Stemming and Text Mining Techniques. J Inf Syst Eng Manag. 2025;10.

[13] Voorhees EM, Harman DK. TREC: Experiment and Evaluation in Information Retrieval. Cambridge. MA: MIT Press. 2005.

[14] `https://www.researchgate.net/publication/2556316_Relating_the_New_Language_Models_of_Information_Retrieval_to_the_Traditional_Retrieval_Models`

[15] Zaki MJ, Meira W Jr. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. 2nd ed. Cambridge. UK: Cambridge University Press. 2020.