FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 3, March, 2025, pp 340 - 353
DOI: https://doi.org/10.33003/fjs-2025-0903-3242

EFFECTIVE RETRIEVAL OF RELEVANT WEB DOCUMENTS: A QUERY EXPANSION APPROACH USING FORMAL CONCEPT ANALYSIS

Abdullahi Bn Umar

Department of Computer Science, Federal University of Education, Kano, Kano State

*Corresponding authors' email: abdullahiu226@gmail.com
ORCID iD: https://orcid.org/0009-0001-5137-7831

ABSTRACT

In information retrieval, vocabulary mismatch between the search vocabulary and documents vocabulary has been a common challenge for web users, hindering information access and retrieval. This issue is often attributed to the ambiguous representation of user information needs as queries, leading to the retrieval of many irrelevant documents, particularly for non-skilled web users. This paper aims to improve user query representation for effective retrieval of relevant documents from the web. To achieve this, a query expansion strategy was employed to identify terms with similar meanings to the user's initial query term. The similarities between the expanded terms and the initial user query term were determined by calculating the cosine angle between the vector resenting document vocabulary and the vector representing query term. Thereafter, Formal Concept Analysis (FCA) was employed to analyze and present the results. Findings from the analysis revealed that concatenating similar terms with the initial user query terms resulted in a 0.16% improvement, as evident in the retrieval precision of 0.64% with the initial query and 0.80% with the expanded query terms.

Keywords: Formal concept, Precision, Query expansion, Query, Recall, Relevant document, Vector transformation

INTRODUCTION

Formal concept analysis (FCA) is a mathematical data analysis and clustering algorithm tool that represents a unit of knowledge into objects, their attributes and relationships. The product of the data analysis produced a set of object known as "concepts" (Rocco et al., 2020). FCA can be applied in different fields such as Data mining, Knowledge Discovery, Data Analysis, Software Engineering and Information Retrieval. This study focused on application of FCA in information retrieval (IR) for improve performance with query expansion (QE). IR is the system for retrieving relevant information that satisfies user's information requirement from a large collection (corpus) of information (Abdullahi and Ekuobase, 2024; Zhang et al., 2024). How users express their information needs as queries to retrieve relevant documents is crucial in IR.

A common practice by users is to represent their information needs free text. However, computers cannot comprehend free text; instead they match keywords syntactically and return documents containing the keywords with exact matches, while neglecting documents featuring different terms but having the same semantic meaning as the keywords. This lack of understanding of free text by the system results in the system keeping track of user's previous search records as search history and at times, offering them to users as suggestions. This is a case of vocabulary mismatch between the user's search vocabularies and documents' vocabularies. The complexity of this problem is rooted in the fact that a single word can have multiple meanings (Polysemy), the same way multiple words can have the same meanings (Synonymy). These limitations of common keyword-based searches have attracted researchers to identify the best approach that improves retrieval performance. One notable approach that has maintained relevance to date is query expansion (Stathopoulos et al., 2023). QE is a strategy used to generate similar terms to the original queries from the initial retrieved results to augment the original queries (Xia et al., 2024).

In Liu et al. (2011) this existing query suggestion methods based on result summarization using popular words or keyword-based cannot effectively handle the problem of query ambiguity problems resulting from 'synonymy' that is, multiple words having the same meaning and 'polysemy' that is, single word with multiple meaning (Afuan et al., 2019; Cakir, and Gurkan, 2023). To deal with this problem, QE approach was considered appropriate in resolving these problems by bridging the gaps between the user and query representation since input queries are basically expressed in words (Cakir, and Gurkan 2023; Afuan et al., 2019). QE has remain a critical task that enhance information retrieval performance (Zhang et al., 2024). Thus, numerous research studies that focus on IR effectiveness including (Afuan et al., 2019; Cakir, and Gurkan (2023) have acknowledged the relevance of this method resulting in the recent diverse approaches to query expansion using machine learning as evident in the studies of (Cakir, and Gurkan, 2023; Zhang et al., 2024). Nevertheless, machine learning approaches are heavily dependent of the availability of datasets which are grossly limited making machine learning techniques a challenging task (Cakir, and Gurkan, 2023). Since the relevance of documents in a keyword-based information retrieval system is determined by the document-to-query relationship, FCA, whose principal focus is relationship-based data analysis and visualization, was considered an appropriate method in this study. This due to its simplicity and effectiveness in relationship modelling and visualization (Boukhetta and Trabelsi, 2023; Wang et al., 2023).

The retrieval of relevant documents in response to user query has been the primary goal of information retrieval systems (Abdullahi and Ekuobase, 2024), thus, how queries are define and submitted to the search engine determines the relevance of the retrieved results. One popular strategy that minimizes and enhanced retrieval of relevant documents is QE. QE has been studied in many literature souces, such as (Afuan et al., 2019; Cakir, and Gurkan 2023; Xia et al., 2024). Messai et al. (2008) introduced an extension of FCA to deal with complex data represented as Multi-Valued (MV) contexts. They

defined a MV Galois connection based on similarity between attribute values. The basic idea was that two objects share an attribute whenever the values taken by this attribute for these objects are similar (i.e. their difference is less than a threshold). This Galois connection form the basis of their computation of MV concepts and MV concept lattices. However, the operation on data with this representation must be preceded by a transformation of MV contexts into binary contexts using an appropriate *conceptual scaling.* The choice of a scale depends on attribute interpretations making conceptual scaling user dependent task which can hardly be automated in the case of large datasets. A recent study by Xia et al (2024) utilized Large Language model (LLM) to generate new query terms to expand the initial query for improve retrieval performance. They used knowledge—aware query expansion approach. Although, their method yielded positive results, it is limited to availability large datasets. Mihai et al. (2014) describes a metric model for extracting the relevant information from large collections of Web documents, by using a new type of conceptual structures associated to the data. The t-concept lattices introduced are used as a retrieval space in which searching process is initiated. Sequences of t-concept lattices were built that can iteratively refine the set of retrieved documents, as a result of the user's query. The sequence of hierarchical structures they built provides a dynamical IR model (Mihai et al., 2014). These models do not take into account all users' preferences. The literature assumed that a combined approach, which uses the weights of terms in web pages and the rank values of these pages to define the weight function, could be a solution to the problem (Mihai et al., 2014).

El Qadi et al. (2010) proposed improving information retrieval using Formal Concepts Analysis (FCA) and related queries. To achieve this, they adopt query expansion approach to increase the quality of the search results in terms of recall, precision. However, the number of times the initial query is expanded is unspecified. Zhang & Feng (2008) used Formal Concept Analysis to group and organized search results. They generates Galois lattice from the formal context and selected the concepts similar to the query term to compute the similarity measure between the observed related concepts. Their method fails to address the generation of labels and its effectiveness measures. Poelmans et al. (2012) carried out a survey research on how to retrieve documents such as journal publications on a particular field in a given interval of time. To achieve this, they extracts keyword from the journal article's title and abstract, model this keywords into formal context and applied Formal Concept Analysis to generate concept lattice representing the document collection. Their paper only shows the clustering capability of FCA and did not explain how it can be used to improve information retrieval effectiveness. Large Language models (LLM) have been studied in (Cakir, and Gurkan, 2023; Zhang et al., 2024) to augment QE for effective information retrieval. The methods shows better retrieval performance but constrained with limited datasets that captures information in different fields. Wang et al. (2010) proposed a novel, hybrid approach to optimized query expansion source. The main idea was to analyze the documents selected by user and apply the knowledge of FCA to construct a user concept lattice; find some documents which are different from the documents mentioned above and build another concept lattice. However, their approach does not consider user's initial query for the expansion. Niu et al. (2013) proposed a reformulated query term for use in a catalog system to offer the user a suggestion with reference to user's previous search query. The system analyses user's search history using the library server to observe the effect on user's browsing experience. The real-time query reformulation to some extent can be complicated especially in the search keyword selection. Vaidyanathan et al. (2014) proposed query expansion method, by assuming that relevant information can be found within a document near the central idea. In doing so, they divide the documents into sections, paragraphs and lines. The proposed method tries to extract keywords closer to the central theme of the document. The expansion terms are obtained by equi-frequency partition of the documents obtained from pseudo relevance feedback and by using tf-idf scores. Nonetheless, this method still could not improve query ambiguities.

Kruiper et al (2023) explore a number of query expansion and documents methods to retrieve relevant information on building products using a small-dataset of user-query. Although the result show that query and document expansion can greatly improve retrieval performance, however, the approach depends heavily on the availability of dataset on the domain of interest. Making this strategy in most cases unrealistic. A similar efforts was made in (Stathopoulos et al., 2023) where a survey approach was used to evaluate several approaches to query expansion including generative AI. Cosine similarity metrics was part of the evaluation metrics used. Their findings revealed that curie-001 generative AI model outperformed other algorithm. Thus, a generative AI equally depends on availability of large dataset sufficient enough to train the AI system. Cakir, and Gurkan (2023) on the other hand proposed an alternative to QE method using a modified conditional generative adversarial network. The study recorded 10% increments in contrast to the current approaches which provides limited alternatives to QE. However, the success of this approach yet, lies in the dataset sufficient enough to train the algorithm. Overall, recent alternative approaches to existing QE strategies are based on machine learning as obvious in (Zhang et al., 2024) where a user query is semantically enriched through large language model (LLM) to produce multiple query-related document. Unfortunately, these alternatives suffers limited training data and hence, unrealistic. Thus, this study applied FCA to extracts similar terms from the initial retrieved results to augment the original query and applied cosine rule to compute the similarities between the extracted terms used to expand the original query terms in a vector space model which is known to be fast and easier for textual data analysis (Eminagaoglu, 2020; Stathopoulos et al., 2023) and hence, is the choice. As documents similarity measures are crucial components of many text analysis tasks, including information retrieval, document classification, and document clustering (Huang et al ., 2011; Stathopoulos et al., 2023). Therefore, this study aim at improving keyword-based retrieval system for better performance using query expansion with FCA.

MATERIALS AND METHODS

Process Pseudocode

The research considers a set of documents d that were returned as relevant document to the initial query q_0:

The user retrieved the first set of result as $\{d_1 \ldots\ldots d_n\}$ by submitting the first query q_0 to the Google search engine

The user identifies the top n set of relevant documents (called Relevance Feedback) present in the retrieved result from the search engine

From the top n document assumed as relevant documents the user selects a list of query terms q_1 for expansion process.

The new term obtained will be combined with the first term (query) ($q_0 + q_1$). The outcome will result in a new query q_2

which includes both the initial query and the set of terms collected from the relevant document set.
The reformulated query will be resubmitted to the search engine and the new result obtained is then compared with the previous result before reformulation.

The overall system description for the propose system for the research work is shown in Figure 1.

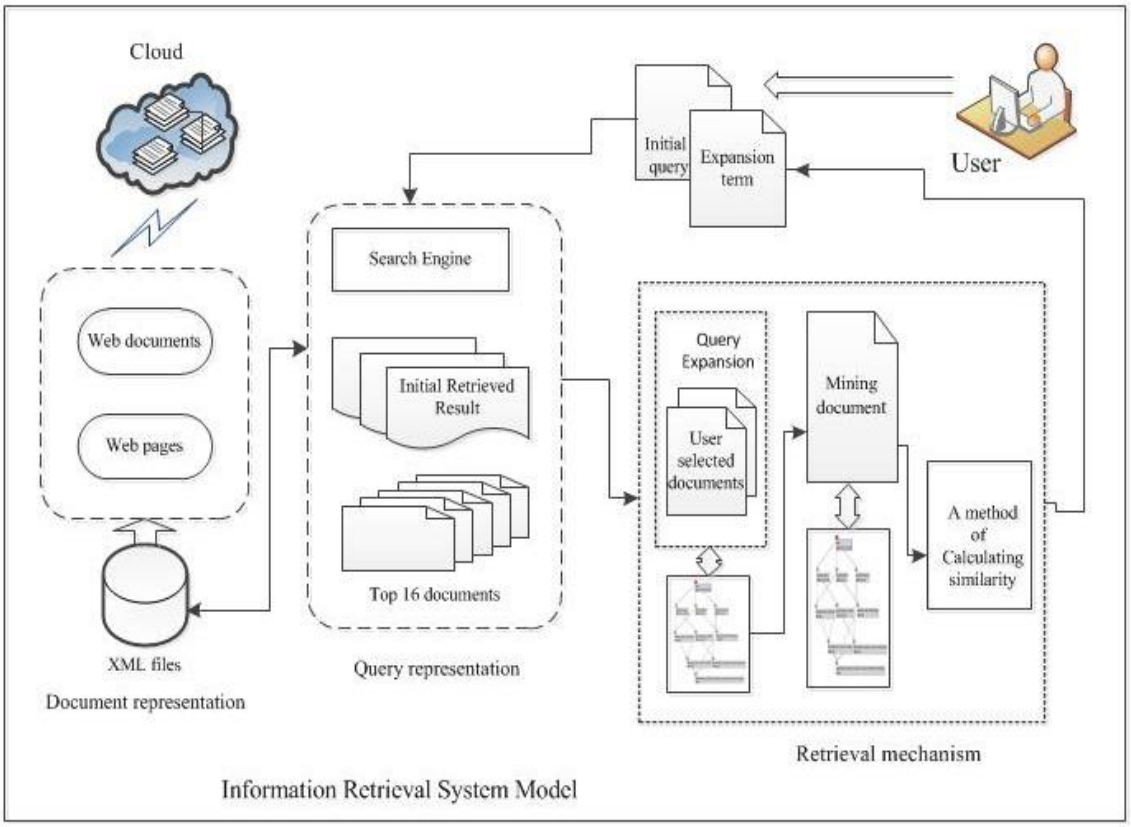

Figure 1: Propose system model

Mathematical Background of Formal Concept Analysis
This research proposed to analyze the result returned to the web users by the search engines using FCA. The choice of applying FCA amongst other data analysis tool in this research study is the efficient clustering and visual transformation of cluster relationship promoted by the FCA, its support for query refinement, and as an important tool for navigating a document corpus. The first step in the analysis is to extract index terms from the document collection.

Keyword Extraction
At this stage, the study extracts from the web document some basic words as index terms which describe a particular web document uniquely. A set of these index terms will be referred to as query term denoted by letter 'q'. After extracting the keyword from the document, the formal context between the documents and query term is constructed with O representing the n retrieved documents and A representing the query term.

Formal Context
After extracting the query terms from the initial documents retrieved, the documents are represented into document-term matrix relation called formal context. Formal context describes a binary relation between a set of object and a set of attributes. We use similar equations in Poelmans et al. (2012), El Wang et al. (2023) and Boukhetta and Trabelsi, (2023) to describe object attribute relationship. Thus, formal context Figure 2 is defined by equations;

$$K = (O, A, R) \quad (1)$$

Where O consisting of rows (i.e. the objects), A corresponds to the columns (i.e. the attributes), and crosses $R \subseteq O \times A$ (i.e. the relationships between the objects and the attributes). Mathematically this implies:

$$R \subseteq O \times A \quad (2)$$

If an object $o \in O$ has an attribute $a \in A$, the relation is given by:

$$(o, a) \in R \ (o \subseteq O, a \subseteq A), \text{ or } oRa \quad (3)$$

Contexts Family Name : Default Name

File Edit Rules Generation Algorithms Database Console

O X A

	A	B	C	D	E	F	G	H	I	J
	O X A	Cpt	Jrnal	Ntwk	Tech	Rsch	Scinc	Engrng	Math	Elctrcal
d1		X	X	0	0	0	0	0	0	0
d2		X	X	0	X	0	X	0	0	0
d3		X	X	0	0	0	X	0	0	0
d4		X	X	0	X	0	0	0	0	0
d5		X	X	0	0	0	X	0	0	0
d6		X	X	0	0	0	X	0	0	0
d7		X	X	0	0	0	0	X	0	X
d8		X	X	0	0	X	0	X	0	0
d9		X	X	0	0	0	X	0	0	0
d10		X	X	0	0	0	0	0	0	0
d11		X	X	0	0	0	0	X	0	X
d12		X	X	0	0	0	0	0	0	0
d13		X	X	0	0	0	0	X	0	0
d14		X	X	0	0	0	0	0	X	0
d15		X	X	0	0	0	0	0	0	0
d16		X	X	0	0	0	X	0	X	0
d17		X	X	0	0	0	X	0	0	0
d18		X	X	0	0	0	0	0	0	0
d19		X	X	0	0	0	X	0	0	0
d20		X	X	0	X	0	0	0	0	0
d21		X	X	0	X	0	X	X	0	0
d22		X	X	0	0	0	0	X	0	0

Figure 2: Context table for initial document

Given a formal context, Galicia-v.2.0-beta derives all concepts from this context and orders them according to a Sub-concept-Super-concept relation, resulting to the Figure 4. Galicia-v.2.0 was used to build an information display model to represent the original model more accurately, and use constructed lattices for information retrieval.

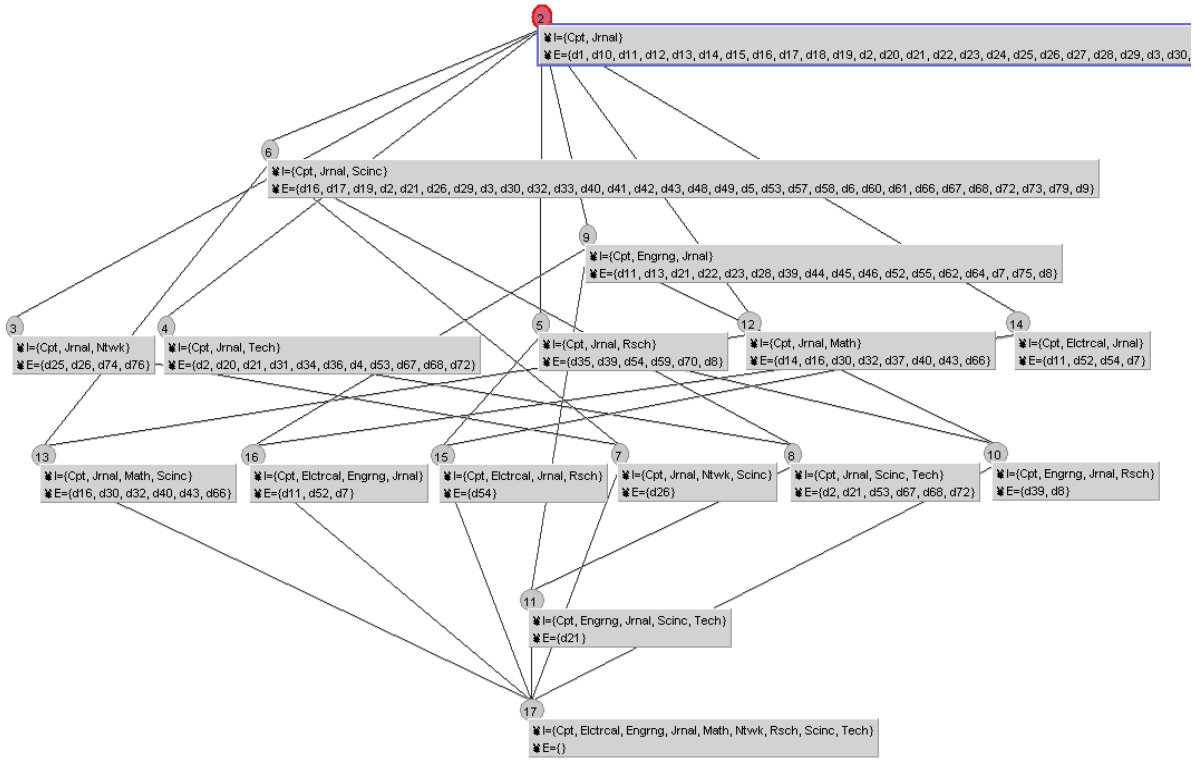

Figure 3: Document clustered into concept lattice

Concept lattice

After constructing the formal context, the information from the formal concept was used to generate concept lattice in Figure 3. A concept is made up of two components: the *Intention* and *Extension*. The extension refers to all objects belonging to the concept, while the Intention refers to all attributes shared by those objects. However, the topmost concepts and bottommost concepts in a concept lattice are considered special. The top concept has all formal objects in its extension. Its intension is often empty but does not need to be empty.

Definition:

Concept lattice is a pair that is, (O, A) of formal context (O, A, R).

C_{latt} = (O, A) (4)

Where O is referred to as the concept's *Extension* and A concept's *Intention*

For a set $A \subseteq Ob$ of objects, we define by:

$A' = \{x \in Att: X \mid x \text{ for every } x \in A\}$ (5a)

$B' = \{X \in Obj: X \mid x \text{ for every } x \in B\}$ (5b)

Equation 5a, 5b implies the set of attributes common to the objects in A and the set of objects which have all attributes in B

For a set $B \subseteq At$ of attributes, we define that;

If (A_1, B_1) and (A_2, B_2) are concepts of a given context

Then;

$A_1 \subseteq A_2$ iff $B_2 \subseteq B_1$ (6)

If $A_1 \subseteq A_2$ and $B_2 \subseteq B_1$

This implies that;

(A_1, B_1) is a sub-concept of (A_2, B_2)

(A_2, B_2) is a super-concept of (A_1, B_1);

So that;

$(A_1, B_1) \leq (A_2, B_2)$ (7)

Hence, the family of the entire formal context of K = (O, A, R) ordinate by ≤ the relation is called Galois (or concepts) lattice (Wang et al., 2023). Thus:

$O = \{d_1, d_2 \ldots d_{79}\}$ is the set of retrieved web documents, $A = \{q_1, q_2 \ldots q_{79}\}$ is the set of query terms, K= (O, A, R) is a formal context, and its incidence relation is described in Figure 1.

In a formal concept, the set of documents A referred to as the extent of the concept (A, B) and the set of term B corresponding to those documents is referred to as the intent of the concept. Thus: $(\{d_1, d_2 \ldots d_{79}\}, \{q_1, q_2 \ldots q_7\})$ is a concept of the context in Figure 2, $\{d_1, d_2 \ldots\ldots d_{79}\}$ are the concept's *extent* and $\{q_1, q_2 \ldots q_7\}$ are the concepts *intent*.

Between the concepts of a given context there exists a natural hierarchical order, the "sub-concept-super-concept" relation. In general, a concept c_1 is a sub-concept of a concept c_2 and c_2 is called a super-concept of c_1 if the extent of c is a subset of the extent of c_2 (or equivalently: if the intent of c_1 is a superset of the intent of c_2). The research uses Ganter's `Next Closure' algorithm to extract the set of all concepts of a formal context.

To calculate cosine similarity between two documents d_1 and d_2, we first represented them as vectors as shown in the Figure 4. A term represents coordinate axis in two dimension space while the number of occurrence of a term (frequency) in a document corresponds to the document length. Equation (8) is used to compute the similarity between the document and the query vector and using the size of the angle (θ) as indices for determining the similarity.

$$\text{Sim}(\vec{q}, \vec{d}) = \frac{\vec{q}.\vec{d}}{\|\vec{q}\| \|\vec{d}\|} = \frac{\vec{q}}{\|\vec{q}\|} \cdot \frac{\vec{d}}{\|\vec{d}\|} = \frac{\sum_{i=1}^{|j|} q_i d_i}{\sqrt{\sum_{i=1}^{j} q_i^2} \sqrt{\sum_{i=1}^{j} d_i^2}} \quad (8)$$

$$\text{Sim}(\vec{q}, \vec{d}) = \vec{q}.\vec{d} = \sum_{i=1}^{j} q_i d_i \quad (9)$$

$$\theta = \text{Cos}[\text{Sim}(\vec{q}, \vec{d})] \quad (10)$$

Where q_i - is the *tf-idf* weighting term i in the query and d_i is he *tf-idf* term weighting of term i in the document.

3.4 Document Weight Determination

To rank a retrieved document according to their relevance, a measure of determining document relevance is necessary. This is referred to as term frequency-inverse document frequency weighting (tf- idf) given by the equation (7)

$$\text{idf}_t = \log \frac{N}{\text{idf}_t} \quad (11)$$

Where N is the total number of documents in a collection and idf_t – is the inverse document frequency of term t.

In order to get composite weight for each term in a document collection, the term frequency and inverse document frequency were multiplied leading to equation (12).

$W_{ij} = tf_{t,d} \times idf_t$ (12)

Term frequency values will be tabulated and then normalized by applying equation (13)

$1 + \log \text{tf}_d$ (13)

After the normalization, equation 12 and 13 will be applied to show the similarities between the document vector $\vec{d}_2$ and the query vector $\vec{q}$ that is, the value of Cosθ. The similarity is the measure of positive correlation between the two vectors. The extent of the correlation will be evident in the size of Cosine of the angle that separates the two vectors. That is, the smaller the angle the closer the vectors and vice versa. The under listed are the summary of indices for determining similarity under the Cosθ similarity algorithm in Euclidian space.

If Cosθ = 1 – we have positive correlation

Cosθ = 0 – strong positive correlation

Cosθ = - 1 negative correlation

RESULTS AND DISCUSSION

This section presents the analysis, results and interpretations. Figure 4 show a screen short of user submitting a Boolean query q_0 to the search engine. After which a total of seventy nine 79 documents from fifteen (15) web pages matching the query term were retried.

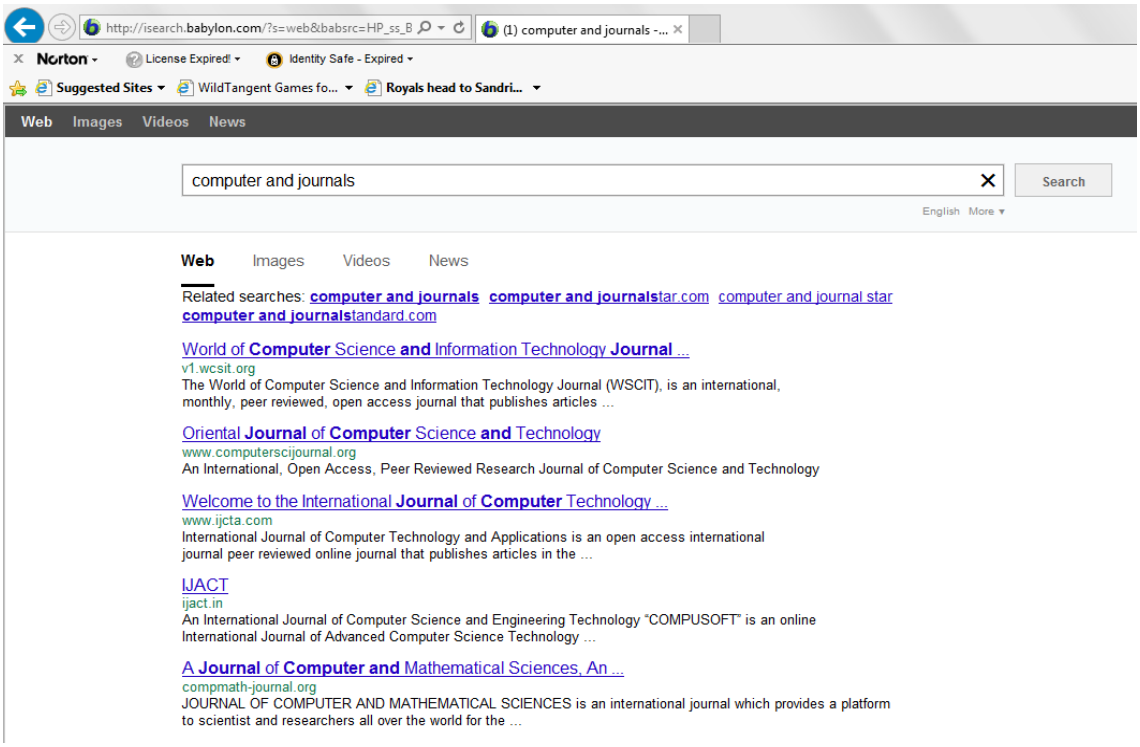

Figure 4: User initial interaction interface (Source: Google)

The retrieved result is a list of documents relevant to the query but not in a particular ranking order. It could be observed that the user's search intention (ambiguous) is not very clear since there are many journals in computer. In order to improve this result in terms of precision and recall, query q_0 was expanded. The expansion process modifies initial query q_0 by adding a term closely related to q_0. For the purpose of extracting expansion term, Figure 2. (Information source directory) was represents formal context where the key terms occurring in the individual document in the collection serves as attributes and individual documents in the document collection serve as the objects. In order to cluster the document source into related concepts (document and query term relationship) and to provide a visual interpretation of information contained in Figure 2, Ganter's Next Closure algorithm were used to interpret Figure 2 into concept lattice in Figure 3. The user then extracted the top 16 documents which contain the keywords q = {Science, Technology, Engineering, and Information} to reformulate another query q' given the binary relation in (Figure 5). Figure 6 is a concept lattice associated with Figure 5.

Contexts Family Name : Default Name

File Edit Rules Generation Algorithms Database Console

O X A

A	B	C	D	E
O X A	Technology	Science	Information	Engineering
d1	o	X	o	o
d2	o	X	o	o
d3	o	o	o	o
d4	o	X	o	o
d5	o	o	o	o
d6	o	X	o	o
d7	o	o	o	X
d8	o	o	o	X
d9	o	o	o	o
d10	o	X	o	o
d11	o	o	o	o
d12	o	o	o	o
d13	X	X	o	o
d14	o	X	X	o
d15	o	o	o	o
d16	o	X	o	o

Figure 5: Corresponding binary relations for the top 16 documents

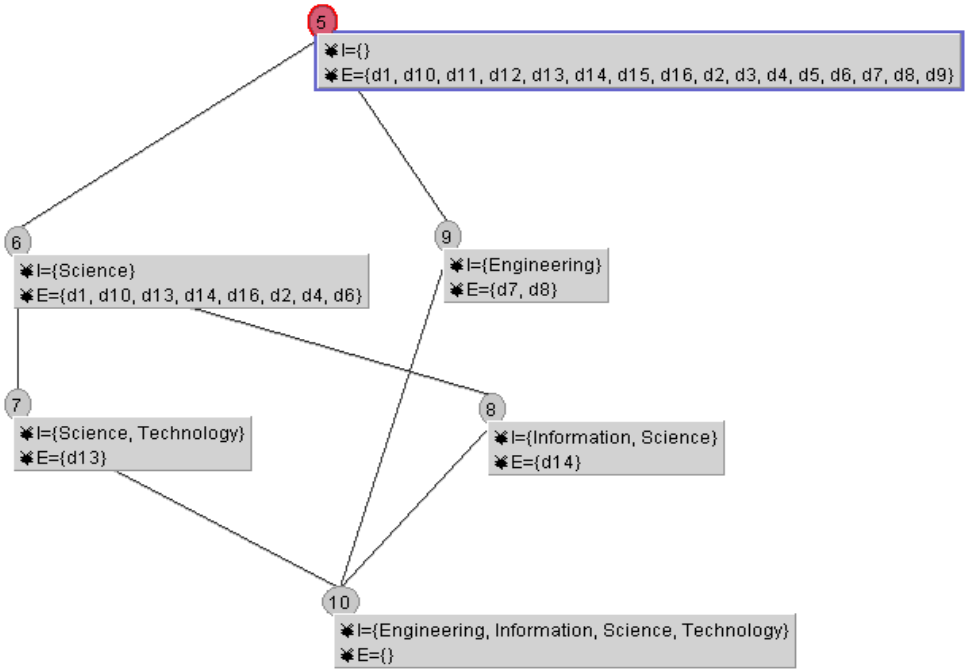

Figure 6: User concept lattice

Considering concepts C_6 = {(d_1, d_2, d_4, d_6, d_{10}, d_{13}, d_{14}), (q_4)}, C_7 = {(d_{13}), (q_2, q_4)}, C_9 = {(d_7, d_8), (q_5)}, C_8 = {(d_{14}), (q, q_4)} in Figure 5, similarity between these concept were computed. The computation algorithm used in this work is based on vector space model which allows documents and queries to be represented as vectors in two or three dimensional Euclidean spaces as shown in Figure 6. In this case, the similarity between the query vector and document vector depend on the occurrence frequencies of the keywords in the query and in the documents. Considering Figure 7a, we observed that document vector d_2 and query vector *q* is not equal in length. To bring the lengths of document d_2 and query vector q closer to each other and to transform them from text-based representation into real values, length normalization and document frequency term (*tf-idf*) weighting were computed using equation (8) and (11) respectively. Where *tf* – represents documents in terms of the frequencies of occurrence of the terms in the document and query vector, *idf* – is the inverse document frequency of the number of documents that contain a query or document term as presented in (Table 2, and Table 3) to obtain the result in Figure 7.

Vector Transformation

Table 1: Document and query terms frequency - Vector component frequency value

Query/Doc.	Science (q1)	Eng. (q5)	Tech. (q4)	Maths (q)
C_{13}	5	1	11	0
C_{16}	32	3	4	6
C_{19}	3	17	1	0
C_{24}	5	0	0	8

Table 2: Document and query terms frequency - Log Frequency values

Query/Doc.	Science (q2)	Eng. (q5)	Tech. (q4)	Math (q)
C_{13}	1.70	1.00	2.04	0.00
C_{16}	2.51	1.48	1.60	1.78
C_{19}	1.48	2.23	1.00	0.00
C_{24}	1.70	0.00	0.00	1.90

Table 1, presents frequency values of both the document and the query vectors while Table 2 is a normalized value of these components. Length normalization is necessary in vector space model so as to transform the textual representation of document and query vector into real values for easy computation and to bring the document vector length closer to the query vector as shown in (Figure 7b).The values in (Table 3) present the dot product of the vector components over their Euclidean length. Thus the components of the document vector in C_{13} were obtained using equation (9). The dot product of the individual vector components in (Table 3) were then added together component by component. Thus we have;

Table 3: Normalized vector components

Query/Document	Science (q2)	Engineering (q5)	Technology (q4)	Maths (q3)
C_{13}	0.455	0.350	0.734	0.000
C_{16}	0.664	0.518	0.576	0.684
C_{19}	0.392	0.781	0.360	0.000
C_{24}	0.455	0.000	0.000	0.936

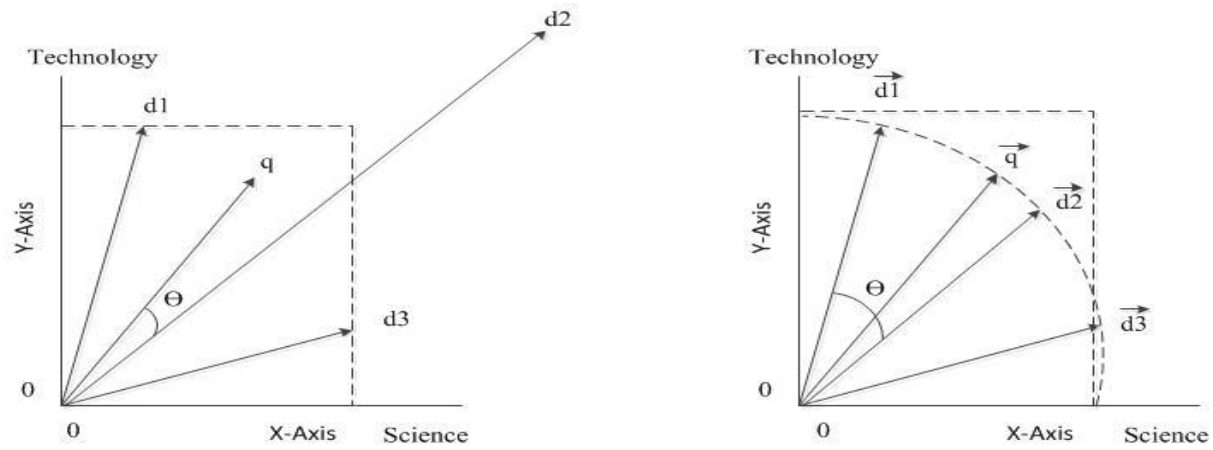

Figure 7: Documents and query representation in vector space

$$\text{Sim}(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{j} q_i d_i}{\sqrt{\sum_{i=1}^{j} q_i^2}\sqrt{\sum_{i=1}^{j} d_i^2}} = \frac{1}{\sqrt{5.622}} = 0.422$$

$$\text{Sim}(\vec{q},\vec{d}) = \sum_{i=1}^{j} q_i d_i = 0.422 \times 1.000 = 0.422$$

Hence, equation 9 computes concept C_7 as higher in similarity value of 0.422. This implies that the set of query {Science, Technology} which is an intention of concept C_7 is closer to the initial query q_0. It can also be observed that even though the intension (i.e. science) of concept C_6 appeared more frequently in a greater number of documents (d_1, d_2, d_4, d_6, d_{10}, d_{13}, d_{14}, d_{16}) compared to the frequency of number of documents in which the intension of concept C_7 (Science, Technology), yet the intention of concept C_7 is calculated as closer to the user's initial query q_0 as shown in Figure 6 Similarly, taking into account all the documents returned by the search engine containing the query term q_0, in order to illustrate the problem better, the research excluded the initial query q_0 from Figure 1 to arrive at a formal context in Figure 8. In doing so, the incident relation between the concepts C_6, C_7, C_8 and C_9 in figure 6 and the concepts C_{12}, C_{13}, C_{14}, C_{15}, C_{16}, C_{17}, C_{18}, C_{19}, C_{20}, C_{21}, C_{22}, C_{23}, C_{24}, C_{25} and C_{26} in figure 9 can be clearly visualized.

Formal Context

Contexts Family Name : Default Name

File Edit Rules Generation Algorithms Database Console

OXA OXA

A	B	C	D	E	F	G	H
OXA	Ntwrk	Technlgy	Rsrch	Scnce	Engrng	Maths	Electrcl
d1	0	0	0	0	0	0	0
d2	0	X	0	X	0	0	0
d3	0	0	0	X	0	0	0
d4	0	X	0	0	0	0	0
d5	0	0	0	X	0	0	0
d6	0	0	0	X	0	0	0
d7	0	0	0	0	X	0	X
d8	0	0	X	0	X	0	0
d9	0	0	0	X	0	0	0
d10	0	0	0	0	0	0	0
d11	0	0	0	0	X	0	X
d12	0	0	0	0	0	0	0
d13	0	0	0	X	X	0	0
d14	0	0	0	0	0	X	0
d15	0	0	0	0	0	0	0
d16	0	0	0	X	0	X	0
d17	0	0	0	X	0	0	0
d18	0	0	0	0	0	0	0
d19	0	0	0	X	0	0	0
d20	0	X	0	0	0	0	0
d21	0	X	0	X	X	0	0
d22	0	0	0	0	X	0	0

Figure 8: Mining Context

Formal Concept Lattice

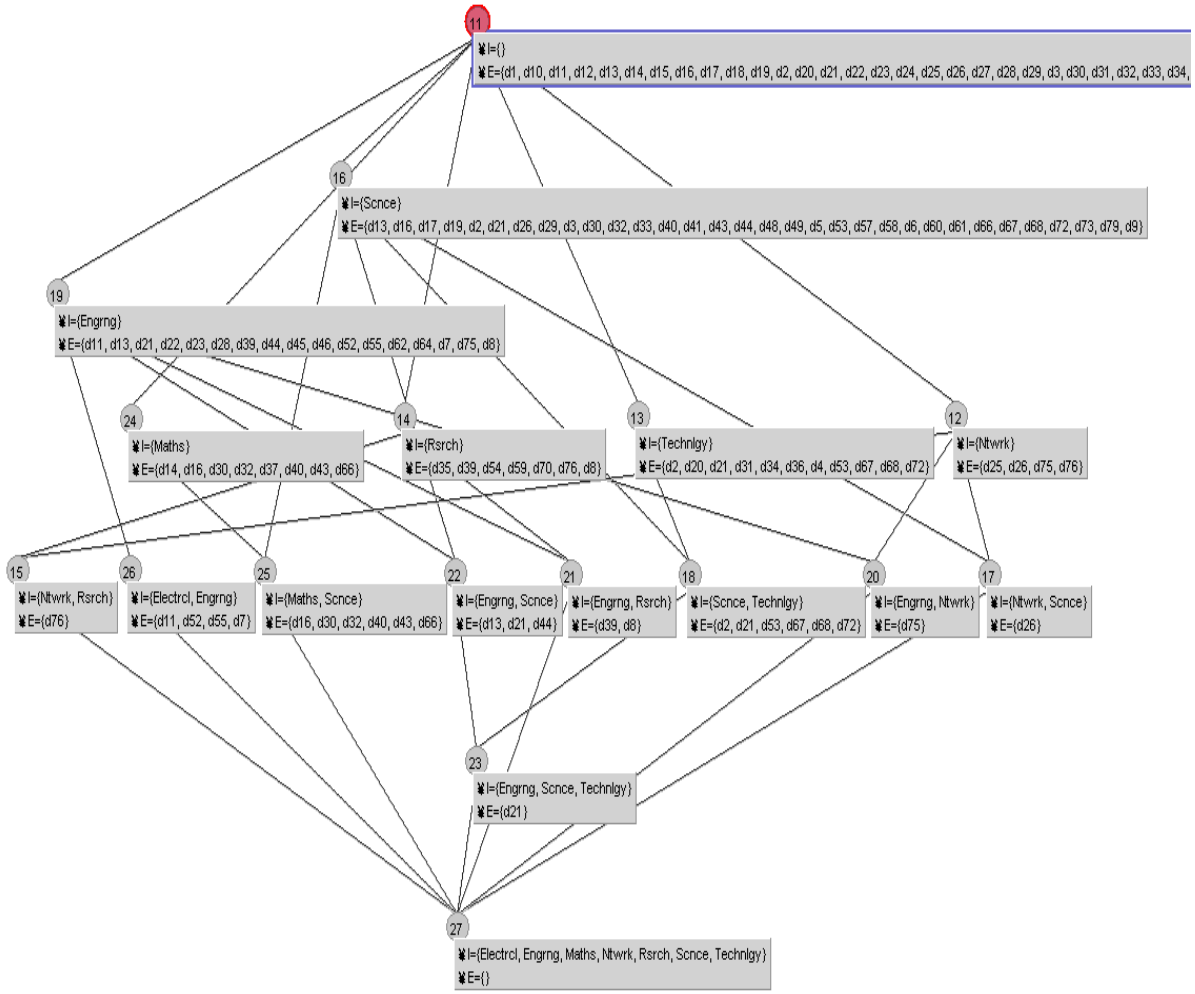

Figure 9: Associated mining lattice

On the other hand, the similarities between concepts in Figure 9 were also computed and compared with the similarity result obtained from (Figure 6). However, concepts C_5, C_{10}, C_{11}, and C_{27} in (Figure 6 and 9) were treated as special concepts and therefore were excluded in our calculation since they have no intention and extension respectively. Considering the tabulated frequencies of occurrence of terms in the documents of concepts in Figure 9, the similarity between the concepts in the mining lattice is hereby calculated.

$$\mathrm{Sim}(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{j} q_i d_i}{\sqrt{\sum_{i=1}^{j} q_i^2}\sqrt{\sum_{i=1}^{j} d_i^2}} = \frac{1.70}{\sqrt{(1.70)^2+(2.51)^2+(1.48)^2+(1.70)^2}}$$

$$= \frac{1.70}{3.74} = 0.455$$

$$\mathrm{Sim}(\vec{q},\vec{d}) = \sum_{i=1}^{j} q_i d_i$$

$\mathrm{Sim}(q_3, d_4) = 0.455 \times 0.350 + 0.664 \times 0.518 + 0.392 \times 0.781$

$\mathrm{Sim}(q_1, d_2) = 0.809$

Following the same procedure as above, the similarities results between concepts C_{13}, C_{16}, C_{19} and C_{24} were obtained as follows:

$\mathrm{Sim}(q_3, d_4) = 0.394$

$\mathrm{Sim}(q_1, d_2) = 0.809$

$\mathrm{Sim}(q_1, d_3) = 0.853$

$\mathrm{Sim}(q_2, d_3) = 0.827$

Comparing the similarities between $(q_3, d_4), (q_2, d_3), (q_1, d_2)$ and (q_1, d_3) it was observed that $\mathrm{Sim}(q_2, d_3) > \mathrm{Sim}(q_3, d_4)$ which implies that concept C_{13} with its intension{Science, Technology} is higher in similarity value compared to other concepts of the mining lattice. Similarly, in the user concept lattice in Figure 5 concept C_7 with intesion {Science, Technology} also shows higher similarity value when compared with concepts C_6, C_7, C_8 and C_9 respectively. Hence, the new expasion term will be "Science, Technology" and the expanded or reformulated query will now be q_1 = {Computer, Science, Technology, Journal}.

Performance Evaluation

In information retrieval field, the performance of any information retrieval strategy is all is subject to the retrieval of precise documents that meets user information need/requirements. This is achieved by taking into account how much document is returned from the document corpus (Recall) and from the returned documents how many documents meet user information need (Precision). That is, Recall is a fraction document returned amongst the documents in the collection, whereas the Precision on the other hand is the fraction of the relevant document contained in the document returned. To determine the performance of this work, precision and recall obtained at various stages of the analysis were computed.

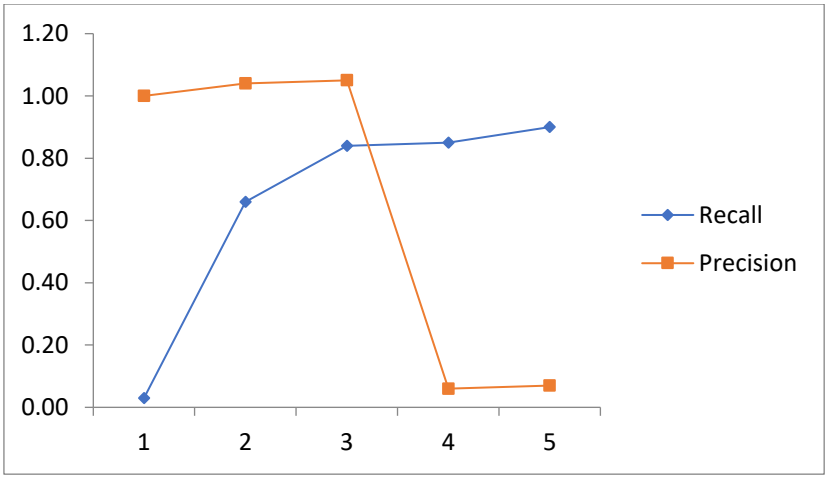

Figure 10: Precision and Recall at 102 documents before query reformulation

Table 4: precision/Recall

Recall	Precision
0.03	1.00
0.66	1.04
0.84	1.05
0.85	0.06
0.90	0.07

Figure 10 presents the precision and recall at 102 documents from the initial result obtained result after user submits query q_0 to the Google search engine. It is obvious that the precision curve drops significantly. This is because the relevant document retrieved is highly insignificant of the total recall at 102 documents. So the retrieval result in this case can be considered as poor.

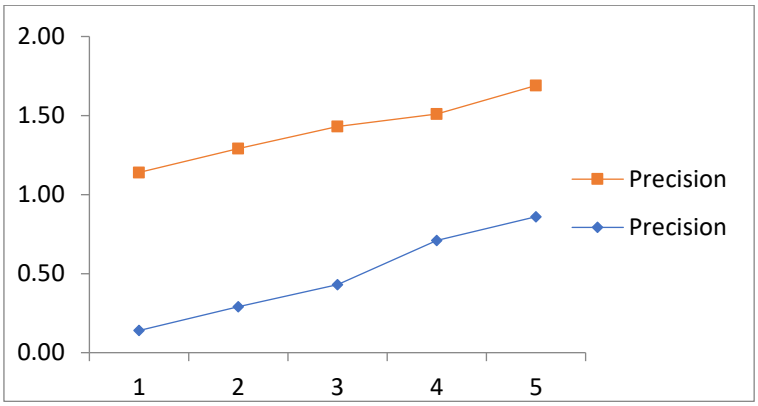

Figure 11: Precision and Recall at 7 documents for the first Page

Table 5: Precision/ Recall

Recall	Precision
0.14	1.00
0.29	1.00
0.43	1.00
0.71	0.80
0.86	0.83

Figure 11 represents precision and recall curve at seven document of the first page retrieved after a reformulated query q_1 was submitted to the search engine. The simultaneous increase in both precision and recall curves as observed in this figure implies an improvement over the initial result before new query formulation. A rise in precision curve is an indication that even within a small number of retrieved results, a significant number of documents relevant to user information need have been retrieved.

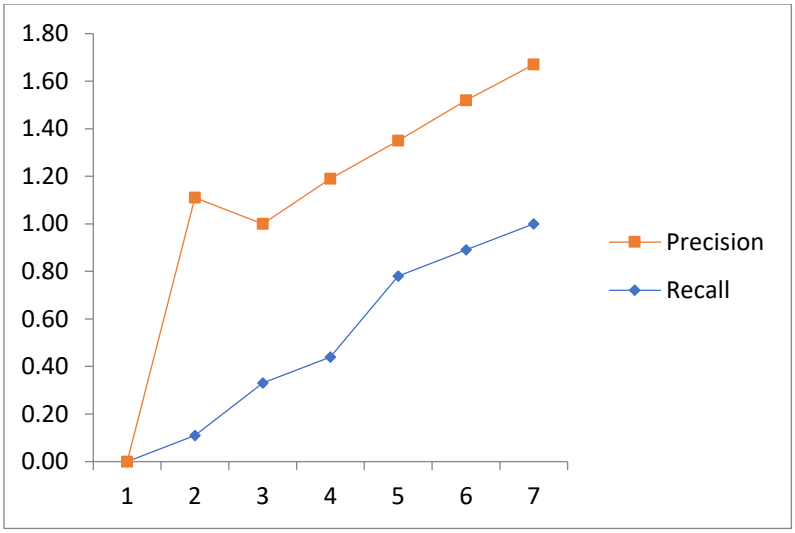

Figure 12: Precision and Recall at 9 documents for the second Page

Table 6: Precision/Recall

Recall	Precision
0.09	1.00
0.18	1.00
0.27	1.00
0.73	0.50
0.82	0.56
0.91	0.60
1.00	0.64

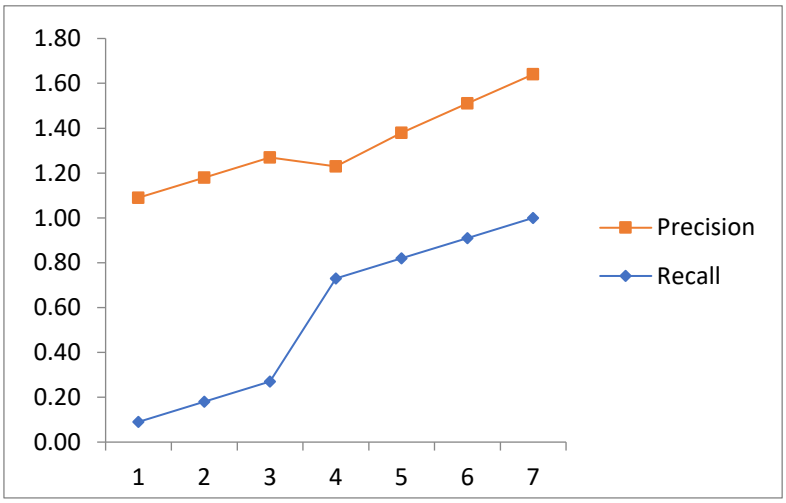

Figure 13: Precision and Recall at 11 documents from the third Page

Table 7: Precision/Recall

Recall	Precision
0.00	0.00
0.11	1.00
0.33	0.67
0.44	0.75
0.78	0.57
0.89	0.63
1.00	0.67

From figures 12 and 13 the precision curve obtained at various recall levels from second and third pages of the retrieved result after user submitted a new query confirms the expectation of the retrieval system of ranking documents in order of relevance. By this, it can be concluded that the first three pages of the result after q_1 was submitted are relevant.

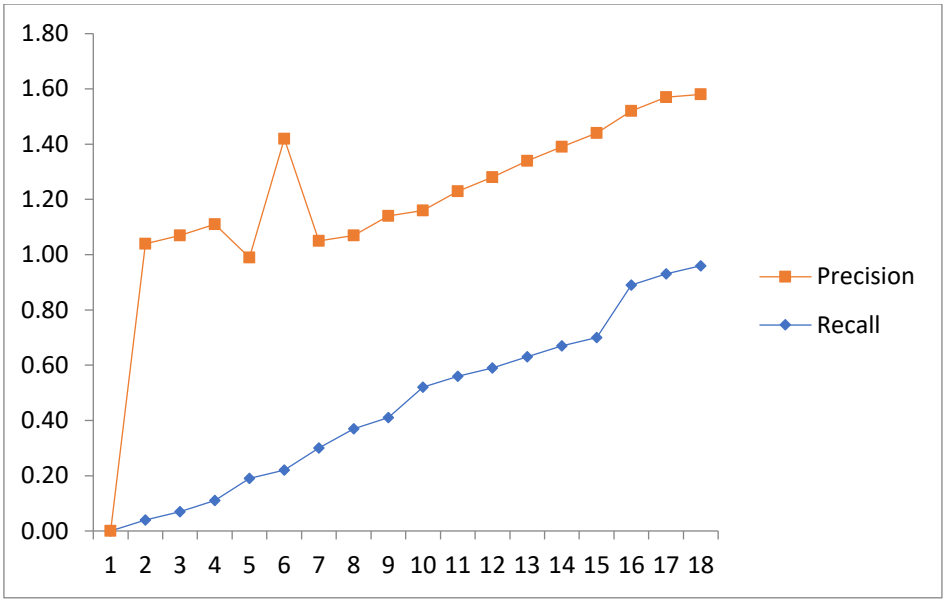

Figure 14: Precision and Recall at 27 documents from the first three Pages

Table 8: Precision/Recall

Recall	Precision
0.00	0.00
0.04	1.00
0.07	1.00
0.11	1.00
0.19	0.80
0.22	1.20
0.30	0.75
0.37	0.70
0.41	0.73
0.52	0.64
0.56	0.67
0.59	0.69
0.63	0.71
0.67	0.72
0.70	0.74
0.89	0.63
0.93	0.64
0.96	0.62

Figure 14 presents the precision and recall at the 27 documents which is the sum total of the precision and recall of the first three pages retrieved with query q_1. This is a further justification that relevant document can be found at the early pages of the retrieved documents if queries are properly formulated and that translates to improved performance of the proposed information retrieval system.

Discussion

The research observed that, the relevance of any retrieved result from the web is subject to system understands of user information requirement specified to it. User information need is usually presented as a free text (as unstructured data) to the search engine. Because free texts are unstructured, system finds it difficult to understand them. So a way of representing user information need is by systematic query formulation. In this research work, an approach to improving retrieval feedback through proper query formulation was carried out. The obtained precision and recall values from the initial retrieved result with q_0 and the subsequent result retrieved after submitting a reformulated query q_2 were compared. The average precision for the initial result is 0.64% and the average precision obtained after submitting a reformulated query q_2 is 0.80%. It is observed that the average precision of the reformulated query q_2 is higher than the average precision of the initial query q_0 by 25%.

The consequence of this is a ranked document in order of relevance as evident in the precision and recall curves at various levels of recall and document in Figures 10 – 11, and Figure 13 respectively. As noticed from the graphs, the precision and recall curves for the first, second and third retrieved pages both increases even at a low recall level. This justifies our method for improving information retrieval system using a combine approach. For further study a researcher can build its own document collection and perform a retrieval process over the document in the collection and

compared with retrieval system over unknown document collection.

In this research work attention was focused on how to improve web retrieval result by query formulation and user's browsing experience using formal concept analysis. To do this, the research first extracts the key words or query terms from the initial retrieved result using query q_0 and organized them into document/term relationship known as formal context or term-document matrix incidence relation. In this format, document collection (corpus) set were referred to as the objects while query terms serves as the attributes set these documents. Formal Concept Analysis was then used to cluster the documents into a set of objects and attributes known as formal concept or concept (a cluster). A collection of the family of all formal context (formal concepts) and their interrelationship in a pictorial representation is referred to as formal concept lattice or concept lattice. In this concept lattice, some concepts that appeared to share similarities were identified and the similarities between them were computed using cosine similarity metric measure. The process of identifying a term(s) closely (similar) related to the initial query term q_0 and adding it to q_1 is referred to as query reformulation (or expansion) and is represented as q_2. The reformulated query will then be the sum ($q_0 + q_1$) of the initial query q_0 and the query term q_1. The reformulated query q_2 was again submitted to the Google search engine as shown in Figure 6 and the retrieved results were evaluated and compared in terms of precision and recall. The outcome of the comparism were presented on a precision/recall curve in Figures 10, 11, 12, and 13 respectively

CONCLUSION

The knowledge of how web users specify their information requirements to the search engine is paramount to the successful retrieval of relevant documents from the information retrieval systems. However, it is important to note that the inability of user to retrieve relevant documents from the web is associated with three basic factors: One, query formulation, two, user's lack of knowledge of the information collection in the information repository (vocabulary mismatch) and third, inexperience of the search strategies.

REFERENCES

Abdullahi, U.B. and Ekuobase, G. O. (2024). A Lingual Agnostic Information Retrieval System. The Scientific World Journal 2024, Article ID 6949281, 37. https://doi.org./10.1155/2024/6949281.

Afuan, L; .Ashari, A. and Suyanto, Y. (2019). A study: query expansion methods in information retrieval. Journal of Physics: Conference Series 1367 (2019) 012001 https://doi.org./10.1088/1742-6596/1367/1/012001

Boissier, B.; Rychkova, I.; Le Grand, B. (2024). Using Formal Concept Analysis for Corpus Visualisation and Relevance Analysis. 16th International Conference on Knowledge Management and Information Systems, Nov 2024, Porto, Portugal. 120-129, ff10.5220/0013047800003838ff. ffhal-04808054ff

Boukhetta, E. S. and Trabelsi, M. (2023). Formal Concept Analysis for Trace Clustering in Process Mining. International Conference on Conceptual Structures, Sep. 2023. Berlin, Germany. 73-88.

Cakir, A. and Gurkan, M. (2023). Modified query expansion through generative adversarial networks for information extraction in e-commerce. Machine Learning with Applications, 14, (2023) 100509.

El Qadi A., Aboutajdine D., & Ennouary Y. (2010) Formal Concept Analysis for Information Retrieval: International Journal of Computer Science and Information Security, 7 (2), 119-125

Eminagaoglu, M. (2020). A new similarity measure for vector space models in text classification and information retrieval. Journal of Information Science, 48(4). https://doi.org/10.1177/0165551520968055

Huang, L., Milne, D., Frank, E., & Witten, I. H. (2011) Learning a Concept-based Document Similarity Measure: Retrieved on 3rd July 2015 from www.cs.waikato.ac.nz/

Jothilakshmi, R., Shanthi, N., & Babisaraswathi, R., (2013). A survey of semantic query Expansion: *Journal of Theoretical and Applied Information Technology, 57, (1), 128-138*

Kruiper, R.; Konstas, I; .Alasdair J.G.G.; Sadeghineko, F.; Watson, R. and Kumar, B. (2023). Document and Query Expansion for Information Retrieval on Building Regulations Lattices. Frontiers in Computing and Intelligent Systems, 3(3), 81-83.

Liu, Z., Natarajan, S. & Chen, Y. (2011). Query Expansion Based on Clustered Results *Proceedings of the VLDB Endowment*, 4, (6)

Messai, N., Devignes, M., Napoli, A., & Smail-Tabbone, M. (2008). Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval: *ECAI 2008 IOS, https://doi.org./10.3233/978-1-58603-891-5-127*

Mihai C.V (2014) Metric & Topological Aspect in Distributed System: Thesis Summary, Institute of Computer Science, Romanian Academy – Iasi Branch

Niu, X., & Hemminger B. M. (2011). Effective of Real-time Query Expansion; ASIST '11,

Niwattanakul S., Singthongchai J., NaenudornE. & Wanapu (2013) Using of Jaccard Coeffient for Keyword similarity: Proceedings of the International Multi-conference of Engineers and Computer Scientist 1, 13 – 15

Poelmans J., Dmetry I. I., Viaene S., Dedene G & Kuzuetsov S. (2012) Text Mining Scientific Papers: A survey on FCA – based Information Reaserch ICDM 2012, LNA 73, 77, 273 – 287

Rocco, M.C.; Hernandez-Perdomo, E. and Mun, J. (2020). Introduction to formal concept analysis and its applications in reliability engineering. Reliability Engineering and System Safety, 202, October 2020, 107002. https://doi.org/10.1016/j.ress.2020.107002.

Stathopoulos, E.A.; Karageorgiadis, A.I.; Kokkalas, A.; Diplaris, S.; Vrochidis, S.; Kompatsiaris,I. (2023) A Query Expansion Benchmark on Social Media Information Retrieval: Which Methodology Performs Best and Aligns with Semantics? Computers, 12, 119. https://doi.org/10.3390/computers12060119.

Vaidyanathan, R., Das, S., & Srivastava, M. (2014). Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval: *international Journal of Computer Applications, 105 (8), 0975 – 8887*

Wang, X., Hu, Z., Bai, R., & Mou, Y. (2010). Automatic Semantic Retrieval and Visualization Model Based on the Integrated Ontology Library: *Journal of Computational information Systems 6, (1), 139-145*

Wang, Y.; Song, Y. and. Wang, Y. (2023). A Survey of Formal Concept Analysis and Concept Lattice. Frontiers in Computing and Intelligent Systems, 3(3), 81-83, 2023

Wenjie, L. & Qiuxiang X., (2011) A method of Concept Similarity Computation Based on Semantice Distance: Journal proceedings in Control engineering and Information 15 (2011), 3854 – 3859

Xia, Y., Wu, J., Kim, S., Yu, T., Rossi, A. R., Wang, H. and McAuley, J. (2024). Knowledge-Aware Query Expansion with Large Language Models for Textual and Relational Retrieval. 1-12, arXiv: 2410.13765v1 [sc.CL] 17 Oct 2024.

Zhang Y, GuoQiang, C., & Yue, J. D. (2011). A Modified Method for Concepts Similarity Calculation: *Journal of Convergence Information Technology, 6, (1).*

Zhang, L., Wu, Y., Yang, Q. and Nie, J. (2024). Exploring the Best Practices of Query Expansion with Large Language Models. Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1872–1883. November 12-16, 2024

Zhang, W.; Liu, Z.; Wang, K. and Lian, S. (2024). Query expansion and verification with Large Language Model for information retrieval. Advance Intelligent Computing Technology and Applications. ICIC 2024, 341-351. https://doi.org/10.1007/978-981-97-5672-8_29.

Zhang, Y., & Feng, B., (2008). Clustering Search Result based on Formal Concept analysis: *Journal of Information Technology 7, (5), 746 – 753.*