International Journal of Computational Research and Development (IJCRD)
Impact Factor: 5.015, ISSN (Online): 2456 - 3137
(www.dvpublication.com) Volume 5, Issue 2, 2020

IMPLICATIONS OF KWWIICC (MODIFIED KWIC) INDEXING SYSTEM TO JUSTIFY RETRIEVALS OF SEARCH ENGINES

Sandip Ghosh
Assistant Librarian, Future Institute of Engineering and Management,
Kolkata, West Bengal

Cite This Article: Sandip Ghosh, "Implications of KWWIICC (Modified KWIC) Indexing System to Justify Retrievals of Search Engines", International Journal of Computational Research and Development, Volume 5, Issue 2, Page Number 8-13, 2020.

Abstract:

KWWIICC (modified KWIC) is an updated version to search engines to address the main problem of document Retrieval i.e. high recall and low precision. It tries to modify the definition of precision and precise document as per user intent. It establishes a new structure of indexing-retrieval mechanism which inspires search engines to serve documents in a new manner i.e. relational document and tends to be a new indexing named by "Relative Keyword Indexing".

Key Words: KWIC; KWWIICC; Relative Keyword Indexing; Search Engine; Retrieval

Introduction:

The KWWIICC indexing system is trying to give a modified version of the most popular KWIC indexing system. This will allow some changes to the 'Automatic Computer Indexing of Titles' and will be able to address the root problem of document retrieval i.e. "High Recall and Low Precision". This means that the main goal of this paper is to improve the quality of document retrieval using a new mechanism instead of the conventional retrieval mechanism used in any search engine or to retrieve the most precise document according to the user's intent.

Thought of Topic:

With the growth of documents, it is becoming more and more difficult to search for documents in search engines. Because, here the documents are arranged haphazardly, so users lose their intent for search while searching for the document and it takes more time. Thus, the demand for precise documents by users is increasing due to the difficulty of searching. That is why modern search engines are moving towards providing more and more precise documents using a variety of mechanism tools like ranking, no. of hits etc. Generally, precise document refers to the type of document that the users are looking for, that is, the user intend oriented document. In order to accurately represent a precise document, the contents of the obtained document require in-depth verification, which in the latter case will help to increase the scope of context analysis during keyword extraction. And determine the exact keywords. So,

Precise Document = User (Keywords having highly relation along with content) Document

All types of search engines and the retrieval mechanisms used in them retrieve documents following the above formula. But there are different degrees of precise document or user documents according to giving importance (choice) to different intended documents by the users. Therefore, weight can be applied to user documents according to user intent or use value (Importance). Again, the A to Z document given to search engines in terms of specific keywords cannot be said to be equal weighted, that is, it can never be A = Z. So there can be no spatial interplay between them. Thus, there needs to be an order (hierarchy) in terms of a certain set of rules (Weight) for representation of documents. As a result, the importance of each document is realized. Here, if the proven decision or rule (weight), i.e., the keyword of the noun form represents the most content, then the objective form and the verb form, respectively, are followed, then the served documents will be organized and it will be convenient to search. Again, if observed closely, the retrieved documents are arranged haphazardly. That is, it is not serially served according to the degree of keyword-content relation. As a result, the user has to go to the minute and elaborate search for the desired document. So, the search time is getting longer as there is no order (hierarchy) and rule (weight) of serving the document. Thus, the definition of 'Precise' needs to be defined before determining the precise document.

In view of the above explanation, it can be said that 'Precise' refers to 'a particular order and rule' i.e., the order and rules served according to the degree of the keyword-content relationship. So, 'precise document' is weighted, hierarchically represented user document.

Precise Document= Weight + Hierarchy + User Document

Again, defining 'Precise' by inclusiveness shows that, only given keyword related documents do not serve the user document. Synonyms of given keyword, used meaning of given keyword and synonymous word of used meaning of given keyword are also indirectly indicate the user document indeed. The hierarchy or order of document representation is below:

Order	Document Representation
1	Given Keyword (Own Used Meaning)
1a	Document related to given keyword

1b	Document related to other word-formation of given keyword e.g. es, ed (limited to Own Used Meaning)
1c	Document related to double meaning of given keyword. (limited to Own Used Meaning)
1d	Document related to synonyms of given keyword.
1e	Document related to double meaning of synonyms of given keyword. (limited to Own Used Meaning)
1f	Document related to synonyms of used meaning of given keyword.
1g	Document related to double meaning of synonyms of used meaning of given keyword. (limited to Own Used Meaning)
1h	Document related to other word-formation of synonyms of used meaning of given keyword.
2	Given Keyword (Other Used Meaning)
2a	Document related to other used meaning of given keyword (including formal meaning).
2b	Document related to other word-formation of other used meaning of given keyword e.g. es, ed (limited to Other Own Used Meaning).
2c	Document related to double meaning of other used meaning of given keyword. (Limited to Other Own Used Meaning).
2d	Document related to synonyms of other used meaning of given keyword.
2e	Document related to double meaning of synonyms of other used meaning of given keyword. (Limited to Other Own Used Meaning).
2f	Document related to synonyms of used meaning of other used meaning of given keyword.
2g	Document related to double meaning of synonyms of used meaning of other used meaning of given keyword. (Limited to Other Own Used Meaning).
2h	Document related to other word-formation of synonyms of used meaning of other used meaning of given keyword.
3	[For multiple keyword] Next Keyword (Own Used Meaning)
4	Next Keyword (Other Used Meaning)
5	So On.

Table 1: Hierarchy or Order of Document Representation

Thus, the documents obtained by following a 'APUPA' relation to the given keyword also indicate the user document. So 'Precise document' is weighted, hierarchically represented user document with APUPA relation.

Precise Document= Weight + Hierarchy + User Document ~ APUPA Relation

Here, it is quite necessary to say that, natural intelligence can never be fully identified by artificial intelligence. Thus, even after enriching artificial intelligence with various tools, it is not possible for any retrieval mechanism to serve the exact user intent document or achieve 100% precision. Thus, the direct method of increasing precision by reducing recall is almost impossible. But if the documents can be properly classified and arranged serially, then it will be easier for the users to find the documents and it will take less time which will indirectly provide precision for the users to get the intended documents. Thus, users can get the benefit of obtaining documents by indirectly giving precision instead of going 100% precision directly which will be much more reliable and user friendly. So, instead of reducing the recall for high precision, one can go towards creating a weighted, hierarchical and exclusive recall, which will indirectly increase the inclusive precision. Therefore, in the present context, an attempt has been made to justify the search results by making implications of the new KWWIICC (modified KWIC) indexing system for the use of the above formula of precise document or precision in retrieval.

Literature Review:

Modified the KWIC indexing system by imposing weights on keywords to achieve a new method or system named 'key-word weighted-in-intra contextual-content' (KWWIICC) and on the other hand, draft a mechanism of an advance search engine to provide most relevant and precise documents as per user intent. (1)

Solved the main problem i.e. 'high recall and low precision' of 'keyword-based search system' by qualifying the keywords with weights and to increase precision by reducing no. of keyword (core element for recall). (2)

By the 1950 computer began to use for data storage. In 1961 automatic indexing system named KWIC was invented by Luhn. Here it was tried to index the documents by words where each word have its own list of strings. (3)

Luhn formatted automatic indexing system for the purpose of machine storage. Here, it is mentioned the background, need, structure of KWIC indexing system in details. And also application of KWIC in technical literature has shown elaborately. (4)

So, for using computer as data storage device updated mechanism is so much necessary since beginning stage by the 1950. Though, Luhn invented a very famous indexing system but it also need to update

due to exploration of information. Due to the need of the users precision is become crucial point for each and every search engine established by the KWIC indexing system.

For overcoming the problems of traditional search engines it is introducing intelligent semantic search engines to provide accurate information by save search time of the users. (5)

The context guided information retrieval process is extraction of semantic keyword and clustering automatically generation of new, augmented queries. The result is semantically ranked, again, using context. (6)

Semantic web technologies are a crucial role to retrieve meaningful information intelligently. These are called generically search engine. (7)

Traditional Keyword-Based Search system is lacking of semantic. To overcome this issue it will be considered the context (concept) using semantic search terms to index the search engine. (8)

Therefore, modern search engines are working on the new mechanism by which it can serve precise document as per user intent and also save the search time. It can only be done by modifying the existing indexing system and or by introducing new indexing system having updated characteristics.

Objective:

The main point of this paper is the presentation of precise document in case of retrieval. This requires the introduction of new formulas, the invention of new mechanisms, and the comparative analysis of search results. So, the main objectives are.

- Identification of Precise documents
- Hierarchical representation of user document with APUPA relation
- Justify search results

Scope:

The implications of KWWIICC will make it possible to remove the conventional haphazardness of document retrieval and provide a universally accepted 'order and rule'. As a result, the document can be retrieved in an inclusive and classified manner and the search time will be less.

Methodology:

A comparison method is being used in this paper. The search results obtained from the search engines used by KWWIICC in a specific database are compared with the search results obtained from the conventional search engines by observing a specific database.

Details of Paper:

The KWWIICC (modified KWIC) indexing system is based on the “weighted keywords” obtained by applying the new ‘Grammatical-Hierarchical Logic’ method which is the essence of both the analysis and interpretation of the grammatical form of the keyword by Baxendale and the experiment of using weights to determine the significant and non-significant keywords by Swanson. Here the mechanism is constructed by arranging the order according to the weight imposed on the grammatical form of the keywords extracted from the various documents. That is, the ‘Noun-Adjective-Verb’, respectively, is preserved in this order. Thus, the keywords are getting a scientific arrangement instead of the alphabetical arrangement as before to index the document. For example:

‘I am sitting on the deck of a fine ship’

Keyword	Grammatical Form	Weight
Ship	Noun	3
Deck	Noun	3
Fine	Adjective	2
Sitting	Verb	1

Table 2: Keyword Arrangement to Index

Here, both pre-coordinate and post-coordinate processes are used when retrieving documents. In case of single keyword search, document retrieval is done by following pre-coordinate process i.e. according to the indexing order of the keyword. But in the case of multiple keyword searches, the document is retrieved in the declining order of the weights determined by the grammatical form of the keywords used which marks the process in post-coordinates. So, on the one hand, documents with high to low weight in terms of both grammatical form of keywords and keyword-content relationship (both are proven equal in nature) are being retrieved hierarchically in respect of user intent and on the other hand, penumbra (other used meaning related), alien (synonymous word of used meaning related) documents are being succeeded with umbra document by establishing APUPA relationship (exclusive recall of document and this includes the inclusion of the feature of relative indexing in KWIC, giving KWIC the benefit of relative indexing) which as a whole increase inclusive precision. Therefore, retrieval is being composite and weighted which provide an easy search mechanism and time saver for users.

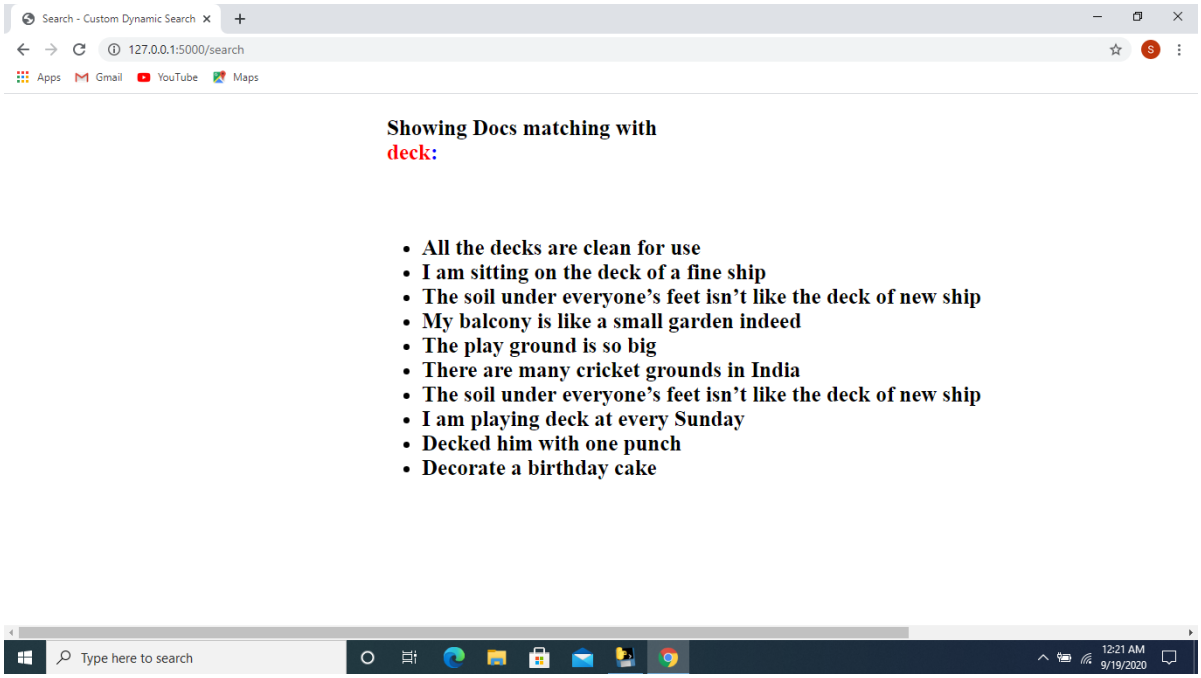

Figure 1: Single Keyword Search

Like other search engine it is a keyword based search (relating content but limited to context) but retrieving all kinds of related documents (umbra, penumbra, and alien) altogether with classified in nature. Though it increase the recall of documents quite enough but weighted, hierarchical and exclusive recall indirectly increase the inclusive precision of document indeed by providing right direction to easy and time consumable search. So, this kind of indexing and retrieving mechanism which have all features (both KWIC indexing and Relative indexing) to solve the main problem i.e. 'high recall and low precision' of "automatic computer indexing of titles" can be awarded by the name "Relative Keyword Indexing System" and can be used in future.

Justification:

Here the two different search results from two different search engine made by KWWIICC and KWIC indexing system respective as follows:

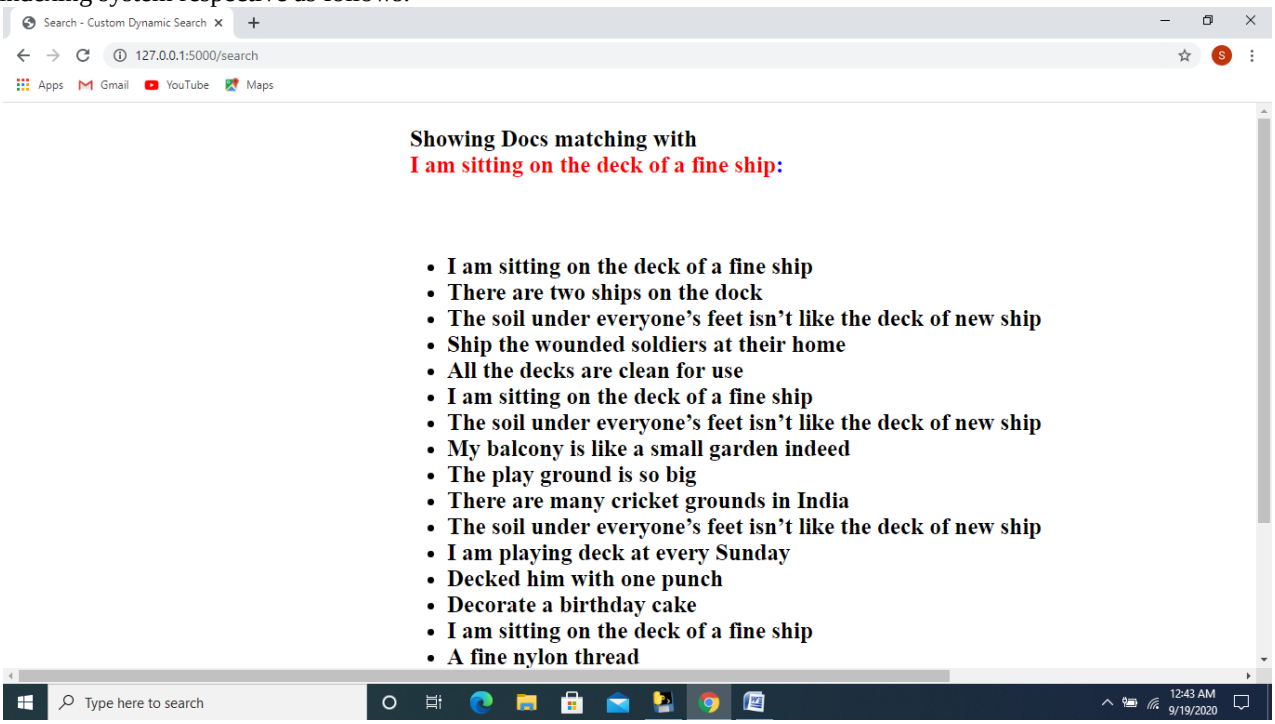

Figure 2: Multiple Keyword Search (KWWIICC)

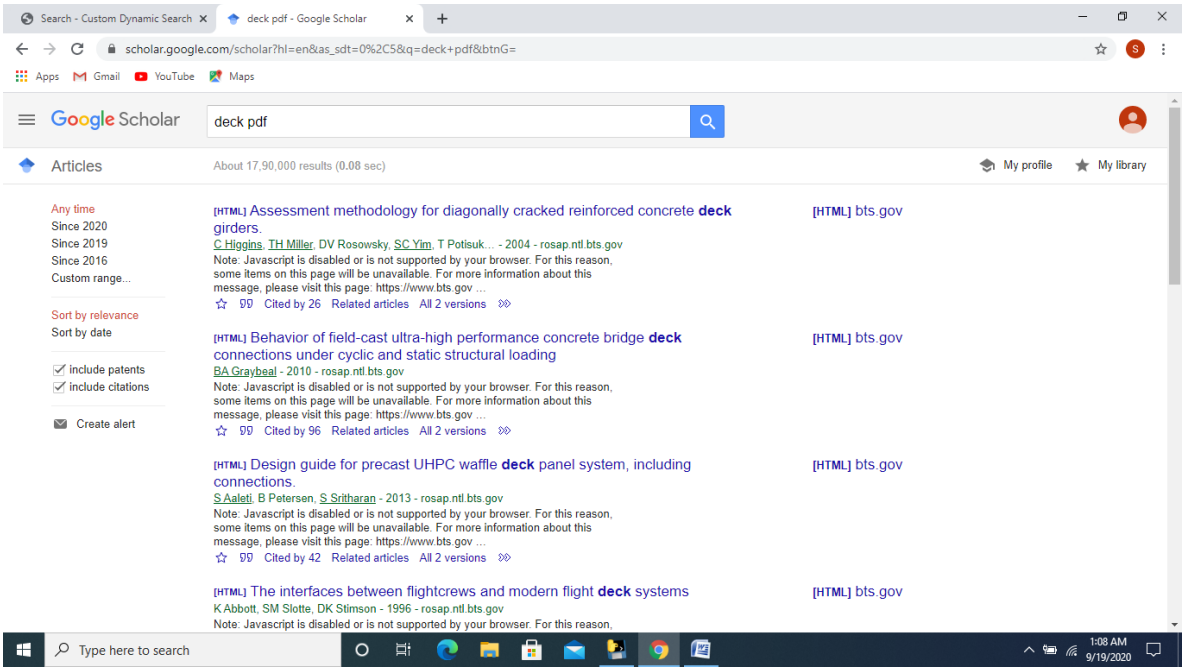

Figure 3: Keyword Search (KWIC)

After observing the search results of two different search engines, it can be said,

- Documents served by keywords in search engines followed by KWIC reflect only the context, but documents served by weighted (which is parallel to the grammatical form of the keyword and easily imposed) keywords in search engines followed by KWWIICC not only reflect the context but also judge the keyword-content relation which helps to increase the precision of the retrieve document.
- Documents served in the search engine followed by KWIC are haphazardly halved, because here the weight of all documents is assumed to be equal. That is why documents are repeatedly rearranged to serve user intended documents using a variety of methods which indicates the nature instability of the indexing method. But documents served by the search engines followed by KWWIICC are classified by their own weight (imposed by the grammatical form or keyword-content relation) which is almost stable or fixed arrangement. As a result, it is easier for the user to search for a document in a specific format or direction, which indirectly increases the precision of retrieve document.
- The search engines followed by KWIC only serves documents related to the given keywords. But search engines followed by KWWIICC also serves other documents having APUPA relation succeeded by documents related to the given keywords. So, all kinds of relational documents are available here at the same time which helps to get the benefit of relative indexing. As a result, on the one hand, the exclusive recall increases and on the other hand, the related searching facility indirectly increases the inclusive precision.

Merit:

The benefits of using this method for storage and retirement are described below.

- Keywords receive weighted arrangement instead of alphabetical arrangement, thus establishing a scientific base.
- Returning a document in a classified way gives users a fixed direction to search for a document. In other words, in the case of document representation, scatterings of documents are eliminated, resulting in less search time.
- All kinds of relational documents (APUPA relation) related to specific keywords are served together which increases the search ability of search engines.

Demerit:

Disadvantages of using this method are described below.

- The recall increases exponentially, causing users to become bored while searching.
- Vocabulary control is required.

These difficulties are easily solvable. Users' boredom is overcome by search engine's find ability and the need for vocabulary control can be easily solved by the grammatical form of the keyword or Grammatical-Hierarchical Logic.

Conclusion:

Artificial Intelligence can never completely identify Natural Intelligence. For example, Sugar= Sucrose (Food) and Drug (Narcotic). Now for single keyword search, sucrose and drug can never be identified separately

by looking at the 'sugar' keyword according to individual user intent. Again, in the case of multiple keyword searches, individual identification is possible through content verification, but documents adjacent to multiple keywords are arranged haphazardly among themselves. This makes it difficult for users to search for intent oriented documents and takes more time. In other words, users lose their search intent when searching. So the core problem of search engines formed by KWIC i.e. 'High Recall and Low Precision' is not solved. However, KWWIICC can be used to find a single solution to both AI deficiency and haphazard representation of documents by providing a fixed order of document representation according to the weight of the keywords. As a result, users will get a specific direction in the search and search time will be less. So, it is needless to say that even if the search engine exclusive recall is high, the user is not lost from his search intent, which indirectly increases the inclusive precision. And this modified approach (KWWIICC) creates a new automatic computer indexing by title named "Relative Keyword Indexing System".

Reference:

1. Ghosh, Sandip. Utilization and application of weighted keyword in retrieval. International of innovative technology and exploring engineering. 9(5). pp. 1730-1734. Mar'2020.
2. Ghosh, Sandip. Modification of keyword selection process to get least list with weighted keywords by using essence of both 'Baxendale' and 'Swanson' experiment. 77(12), pp. 29-35. Oct.'2019
3. Fischer, Marguerite. The KWIC index concept: A retrospective view. Journal of the association for information science and technology. Apr.'1966. https:// doi.org/10.1002/asi.5090170203 (Last checked at 10-09-2020).
4. Luhn, H.P. Keyword-in-context index for technical literature (KWIC index). Presentated at American chemical society. Division of chemical literature Atlantic City. N.J. 14 Sept.'1959. Rept. no. RC 127, International business machines corp. York-town heights. N.Y. 1959. 16p. Also in Amer. Documentation 11,288-295 (1960).
5. Sedano, John Michael, "Keyword-in-context (KWIC) indexing: Background, statistical evaluation, pros and cons, and applications", university of pittsburgh, 1964.
6. Roshdi, Akram, Roohparvar, Akram, "Review: information retrieval techniques and applications", International journal of computer networks and communications security, Vol. 3, No. 9, pp. 373-377, Sept. 2015.
7. Dinesh, Jagtap, Nilesh Argade, Shivaji Date, Sainath Hole, Mahendra Salunke, "Implementation of intelligent semantic web search engine", International journal of engineering research and technology, Vol. 4, No.4, pp. 114-117. Apr. 2015.
8. Finkelstein, Lev, Gabrilovich, Evgeniy, Matias, Yossi, Rivlin, Ehud, Solan, Zach, Wolfman, Gadi, Ruppin, Eytan, "Placing search in context: the concept revisited", WWW 10, May 2-5, 2001, Hong Kong, ACM 1-58113-348-0/01/0005.
9. Madhu, G., Govardhan, A., Rajinikanth, T.V., "Intelligent semantic web search engines: a brief survey", International journal of web & semantic technology", Vol.2, No. 1, pp 34-42, Jan. 2011.
10. Abdullah, K-K.A., Robert, A.B.C., Adeyemo, A.B., "Semantic indexing techniques on information retrieval of web content", International journal of advanced research in computer and communication engineering, Vol. 5, No. 8, pp. 347-352, Aug. 2016.
11. Malve, Ankita, Chawan, P.M., "A comparative study of keyword and semantic based search engine", International journal of innovative research in science engineering and technology, vol. 4, No. 11, pp. 11156-11161, Nov. 2015.
12. Raju, K. Butchi, vani, Velde, "An empirical techniques of information retrieval system in searching", International journal of engineering research and development, Vol. 13, No. 9, pp. 37-42, Sep, 2017.
13. Shah, Vidhi, Shah, Akshat, deulkar, Khushali, "comparative study of semantic search engine", International journal of engineering and computer science, Vol. 4, No. 11, pp. 14969-14972, Nov. 2015.
14. Chitre, Nikhil, "Semantic web search engine", International journal of advance research in computer science and management studies, Vol. 4, No. 7, pp. 47-52, July 2016.
15. Bachchhav, Kiran Prakash, "Information retrieval: search process, techniques, and strategies", IJNGLT, Vol. 2, No. 1, pp. 1-10, Feb. 2016.