International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 10 (2020) pp.1009-1014
 https://dx.doi.org/10.37622/IJAER/15.10.2020.1009-1014

# The Impact of Relative Keyword Indexing System Instead of KWIC Indexing System on Retrieval

**Sandip Ghosh**

*Assistant Librarian, Future Institute of Engineering and Management, India.*

**Abstract**

An analysis of the effectiveness of the new "Relative Keyword Indexing" system obtained by KWWIICC (modified KWIC) is essential to solve the core problem i.e. 'High Recall and Low Precision' of all conventional search engines. This is because the use of KWIC is failing to meet the requirements of "Automatic Computer Indexing of Titles" in the current era of information exploration. So, it is needless to say, the main consideration of the new indexing system is to give importance to the user's intent in serving the document and to provide the user with the facility of document search and save search time.

**Keyword:** KWIC; KWWIICC; Relative Keyword Indexing; Search Engine; Retrieval

## INTRODUCTION

In the current era of digital information, the use of automated-computerized methods has become more important in the field of document storage and retrieval. For machine oriented storage and dissemination, all databases or search engines use 'keyword based search system' or KWIC indexing system. Because, it is the requirement for 'Automatic Computer Indexing of Titles', KWIC is one and only the essential indexing system. But long-term use experience has shown that the main problem of this method is "High Recall and Low Precision". To solve this problem, a newly introduced "Relative Keyword Indexing System" is used in an experimentally built database or search engine, and the impact on retrieval is being analyzed.

**Thought of Topic:** Any database or search engine is built on its document storage and document dissemination with an emphasis on easy automatic mechanisms and precise documents, respectively. Although KWIC indexing system is built with the above features, its quality is declining with the increase in the number of documents i.e. as the amount of recall increases, the precision decreases. Using a variety of additional mechanisms e.g. no. of hits, ranking etc. does not improve the quality indeed. As a result, it needs to be modified or replaced. Therefore, it is important to analyze the advantages and disadvantages of storage and retrieval of any database when using the 'Relative Keyword Indexing System' obtained by KWWIICC (modified KWIC) to monitor the replacement. Because, it is necessary to judge whether the indexing of documents by keywords is going to be easier than the indexing made by the KWIC and whether the precision is increasing by correctly identifying the user intent in case of retrieval. Otherwise the KWIC cannot be properly replaced by the "Relative Keyword Indexing" (RKI) System.

## LITERATURE REVIEW

KWWIICC (Modified KWIC) added some additional attributes to KWIC indexing system i.e. keywords must be Weighted and representation of documents must be Hierarchical and with APUPA relation. It tends to be a new indexing system named "Relative Keyword Indexing System" (1)

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. (2)

Solved the main problem i.e. 'high recall and low precision' of 'keyword-based search system' (made by KWIC indexing) by qualifying the keywords with weights and to increase precision by reducing no. of keyword (core element for recall). (3)

'Keyword based search system' has some problem to identify the exact user intent documents by using KWIC indexing system. These are- non informative keywords, scattering of related terms, and scattering of related documents. So, naturally precision is low in respect of high recall. (4)

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. (5)

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. (6)

Therefore, it is derived that KWIC has some problem in both indexing and retrieval field. It will take necessary to modify the KWIC for requirement of 'automatic computer indexing of titles'.

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. (7)

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. (8)

Semantic web technologies are a crucial role to retrieve meaningful information intelligently. These are called generically search engine. (9)

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. (10)

Therefore, it can be said that, every search engine should follow the basic requirement of retrieval i.e. easy acceptability to user for searching their intended document and with saving the time.

## OBJECTIVE

The main issue in this paper is the analysis of the effect of using a Relative Keyword Indexing System instead of the KWIC Indexing System in any database on document retrieval.

- Analysis of the impact of Relative Keyword Indexing System (RKI) on Retrieval.

## SCOPE

Introducing new approaches by examining how the use of the new 'Relative Keyword Indexing System' developed by KWWIICC (Modified KWIC), based on the weaknesses of the KWIC, can improve and universalize the retrieval of documents as a whole, which will be able to replace the old one and will be useful for all databases by establishing the user intent.

## METHODOLOGY

This paper will discuss two aspects; first of all, the explanations of what are the retrieval problems by using KWIC Indexing system, and secondly, the analysis of the effect of using new 'Relative Keyword Indexing System' on retrieval of any database in view of the above mentioned problems.

## DETAILS OF PAPER

In the current era of keyword-based search system, KWIC indexing is considered a necessary one. It is the main resource of Automatic and Computerized database. Speed or quickness is a key issue in the KWIC Indexing System. That is why the method of insignificant word selection has been adopted in determining the keywords i.e., insignificant words are ignored through the stop list and all other words are identified as keywords. So, it can be assume that, all the keywords here are given an equal weight. As a result, first of all, the represented documents are arranged haphazardly, that is, there is no difference between the first and the last document in terms of representation. They can change places at any time. On the other hand, arising of scattering of related documents by difficult to find spelling variant and inflections in the index because of non-existence of cross references and also arises scattering of related terms by difficult to find synonyms because of alphabetical arrangement. As a result, users have trouble finding the document and it takes more time. Thus, users lose their search intent during document search. So demand for precise documents is increases by the users. Secondly, since it is accepted that 'Natural Intelligence' can never be fully identified by 'Artificial Intelligence', exact document representation is never possible in terms of user intent, on the other hand the abundance of documents increases the amount of recall of certain databases, so, Precision is declining overall. As a result, 'High Recall Low Precision' is arises as a main problem indeed. Trying to solve this problem by introducing some new techniques i.e. No. of hits, Ranking etc. However, these are not quite enough to recover the problem. So, KWIC indexing is increasingly unable to provide user-intended documents day by day.

The new indexing named 'Relative Keyword Indexing System', created by KWWIICC (modified KWIC), first focuses on increasing precision. That's why the way to identifying the significant word has been followed which was shown by Baxendale. At the same time, in the way shown by Swanson, the importance of the significant words has been judged and weight has been imposed on them. So that on the one hand, the keyword can be selected in a scientific manner and a scientific method (weight) can be followed in indexing documents by keyword. Here is the order of keyword arrangement is below: By the example,

I am sitting on the deck of a fine ship'

**Table 1:** Keyword Arrangement by RKI

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

On the other hand, total recall documents, represented by keyword can also be classified and hierarchically arranged by following the scientific method (weight), which will help the user to search according to a fixed hierarchy. Again, an attempt has been made to increase the exclusive recall by applying some additional rules i.e. APUPA relation, which will allow the users to take advantage of relative indexing. Here is the order of document representation is below:

**Table 2:** Order of Document Representation

| 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. |

Therefore, the new 'Relative Keyword Indexing System' is trying to indirectly increase precision by providing a fixed hierarchy or classified order of document representation. Because it is real that Natural Intelligence can never be absolutely identifiable by Artificial Intelligence. So, it is unreasonable to go down the path of increasing precision by reducing recall of document. Again, in the current era of information exploration, the decline in recall of document indicates a negative approach to any search engine. Therefore, the creation of weighted, hierarchical and exclusive recalls has indirectly led to the increase of inclusive precision by facilitating the users to search their intended documents easily and saves search time indeed.

## COMPARISON

Here the presentation of two different search results are shown from two different search engine - one from existing search

engine followed by KWIC indexing system and other from exemplifying search engine followed by 'Relative Keyword Indexing System', as bellow.

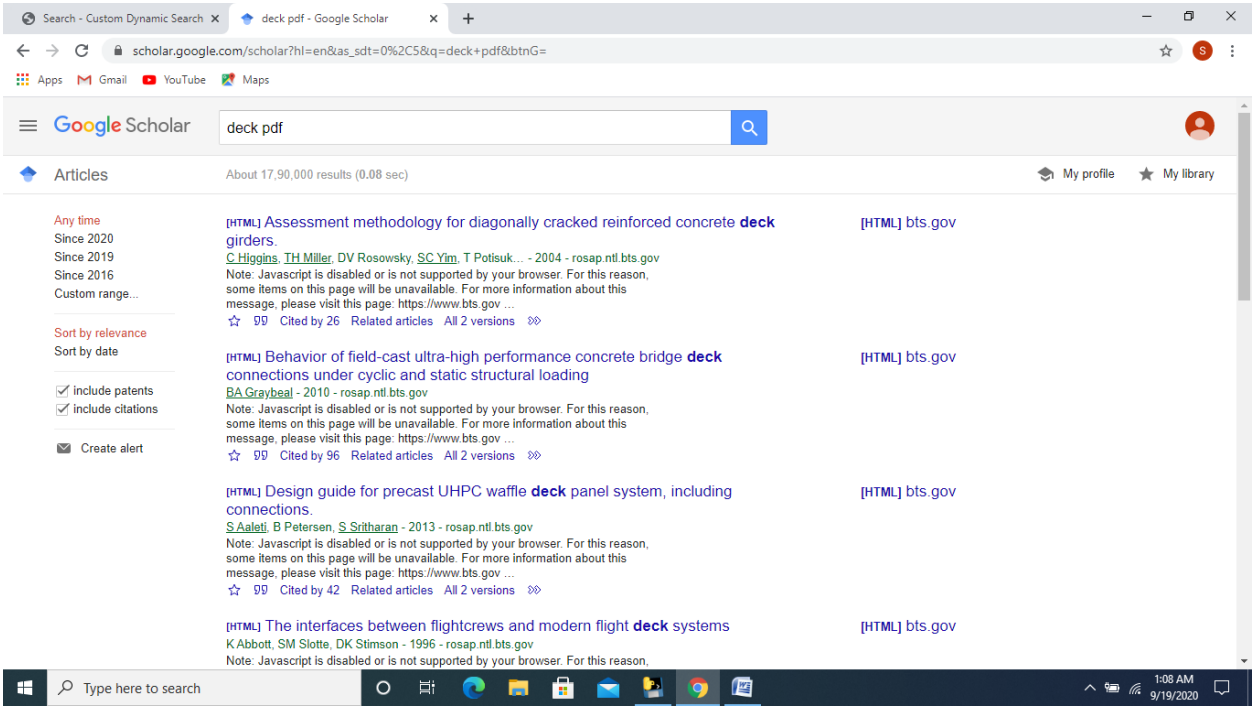


**Fig 1:** Search from KWIC based search engine

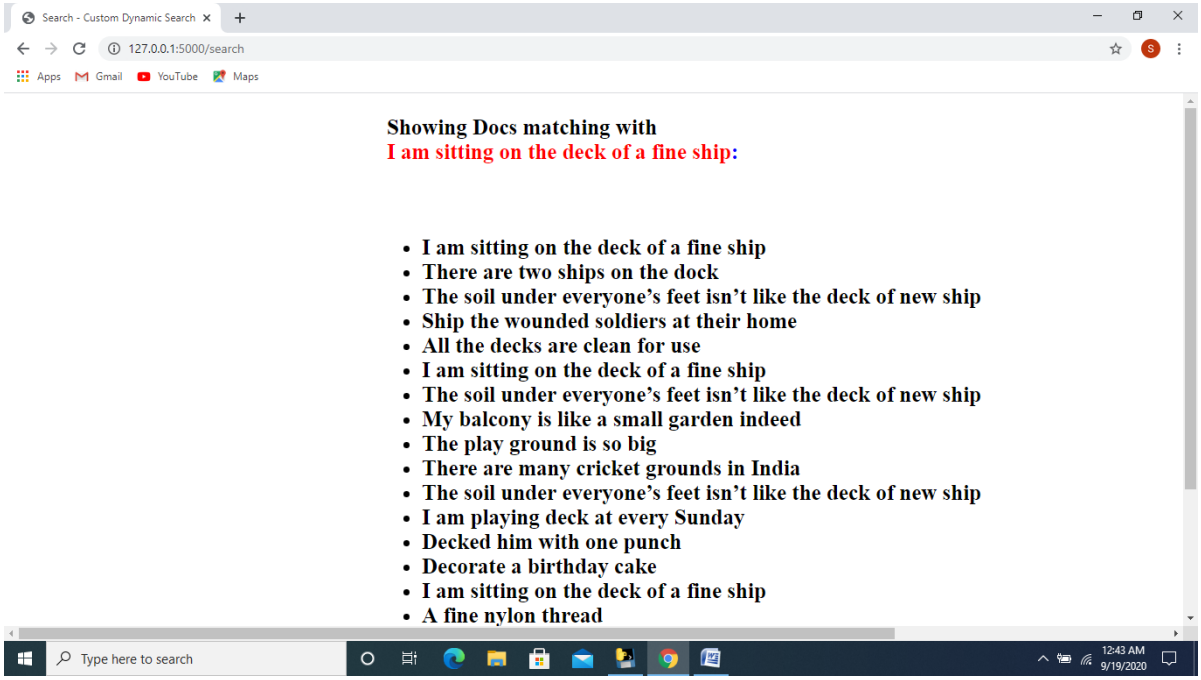


**Fig 2:** Search from RKI based search engine

From the above two different search results, it can be compare in the view of scientific approach and concept of up-datedness as follows:

1) 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 (Relative Keyword Indexing) not only reflect the context but also judge the keyword-content relation which helps

to increase the precision of the retrieve document.

2) Documents served in the search engine followed by KWIC are haphazardly arranged, 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 of instability of indexing method. But documents served by the search engines followed by KWWIICC (RKI) are classified by their own weight (imposed by the grammatical form or keyword-content relation) which is almost stable or fixed (limited to arising of new keyword) 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.

3) The search engines followed by KWIC only serves documents related to the given keywords. But search engines followed by KWWIICC (RKI) 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.

## IMPACT

The following is how the new indexing system affects or enhances the document retrieval of the keyword-based search system.

1) Keywords can be sorted in a scientific order by weighting them according to their own significant. As a result, the precision of the documents served by the keywords increases.

2) By serving the document according to the weight gained by the keyword, the user gets fixed classified, hierarchical and serially arranged document, so there is no difficulty in finding the document and time is saved which indirectly marks the increase in precision.

3) Keyword search not only yields keyword-related results, but also related documents by creating APUPA relationships. That is, KWIC indexing is accompanied by relative indexing. As a result, the increase in exclusive recall is also indirectly helping to increase the inclusive precision.

4) The addition of APUPA Relation increases the recall compared to the past but the availability of Umbra, Penumbra and Alien documents succeeded by keyword-related document increases the search capability and at the same time getting all the related documents makes the search much easier, acceptable and user friendly.

5) Although direct precision does not increase because the exact document is not served according to the user intent, however, all related documents (APUPA relation with keyword) are served according to a fixed hierarchy that gives the user a specific direction to find their intended document. This indirectly increases the inclusive precision indeed.

So, "Relative Keyword Indexing" is the well featured indexing system as it act user intent wise and solve all the problems compare to KWIC indexing system. Every search engine first wants to mitigate users' requirements then to build their own sophistication. Relative Keyword Indexing System is an easy approach for any search engines in indexing and retrieval purpose than others.

## CONCLUSION

User intent is very important when it comes to serving documents. Every database or search engine tries to follow the user intent and provide specific documents so that the users do not have difficulty in searching and the search time is less. But it is not at all possible to fully identify Natural Intelligence through Artificial Intelligence by machine or computer. Therefore, the 'Relative Keyword Indexing' system obtained by KWWIICC (modified KWIC) has been used to indirectly increase (inclusive) Precision by doing the Recall more weighted, hierarchical and exclusive in nature. This allows users to search for documents according to their own search intent in a specific redirected way and saves time. So, it is needless to say, this would be the main consideration for any search engine indeed in the current era of information exploration and the new "Relative Keyword Indexing" (RKI) system will be the replacement of KWIC indexing system to fulfill the requirement of 'Automatic Computer indexing of Titles' in the future.

## REFERENCE

[1] Ghosh, Sandip. Implications of KWWIICC (Modified KWIC) indexing system to justify retrievals of search engines. Journal of the social sciences. 48(4). Oct'2020.

[2] Ghosh, Sandip. Utilization and application of weighted keyword in retrieval. International of innovative technology and exploring engineering. 9(5). pp. 1730-1734. Mar'2020.

[3] Ghosh, Sandip. Modification of keyword selection process to get least list with weighted keywords by using essence of both 'Baxendale' and 'Swanson' experiment. International journal of computer applications. 77(12), pp. 29-35. Oct.'2019

[4] Ghosh, Sandip. The impact of challenges of 'Keyword-based search system' or KWIC indexing on retrievals of search engines. Journal of information and computational science. 10(10). Oct'2020. pp. 214-218.

[5] 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).

[6] 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).

[7] Sedano, John Michael, "Keyword-in-context (KWIC) indexing: Background, statistical evaluation, pros and cons, and applications", university of pittsburgh, 1964.

[8] 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.

[9] 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.

[10] 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, HongKong, ACM 1-58113-348-0/01/0005.

[11] Bachchhav, Kiran Prakash, "Information retrieval: search process, techniques, and strategies", IJNGLT, Vol. 2, No. 1, pp. 1-10, Feb. 2016.