The Impact of Challenges of 'Keyword-Based Search System' or KWIC Indexing on Retrievals of Search Engines

Sandip Ghosh

Assistant Librarian

Future Institute of Engineering and Management

E-mail: ghoshsandip83@gmail.com, Ph.: 9163633153

Abstract: *Keyword-based search system is the mainstay of all current search engines. Keyword indexing makes it easier to index documents alphabetically and quickly retrieve them by keyword search. However, search engines cannot retrieve documents according to user intent due to increased problems like uninformative keywords, scattered documents, scattered related term etc. Therefore, in order to increase the search facility of the user and save time, it is essential to explain the problems and do an impact analysis which will help to keep the existence of the search engines.*

Keywords: KWIC; Keyword-Based Search System; Challenges; Search Engine; Retrieval

Introduction: The use of 'Keyword-Based Search System' or KWIC indexing in search engines is undeniable in the current era of information exploration. Because, all the requirements of 'automatic computer indexing of titles' can be fulfilled only by keyword indexing. That is, indexing and retrieving by keywords maintains speed-automation-machine usability. So keyword indexing or KWIC indexing is the only essential resource of search engines by which search engines serve precise documents according to the user's intent and save their search time.

Thought of Topic: From long experience and usage, it is seen that with the increase in the amount of documents stored in the database, the precision is also decreasing even though the recall of keyword searches has increased naturally. That is, users are not getting documents according to their intents. With the increase in recalls, documents are being haphazardly arranged which is forcing people to search in elaborate. As a result, users are losing their search intent and their search time is getting longer. Also, keyword searches do not serve documents related to other terms of the given keyword like synonyms, spelling variants etc. And vocabulary in natural languages is not built properly. As a result, accurate information is not served by search engines as a whole.

Literature Review: In the 1950s and later, computers were gradually being used as a means of data gathering. In 1961, H.P.Luhn invented the method of data gathering by keywords, whose

name is 'Keyword In-context' indexing. This is an automatic indexing system. The 'KWIC' is an indexing system created by 'Word' where every word is released, with its own list of strings. (1)

Here, we see that, how KWIC is applicable for any big database like American chemical society. It is elaborated by the sample case of KWIC that indexing of documents by keyword is so easy and speedy one than others. (2)

KWIC is quite easy to handle and automatic machine based indexing system. Its' coding is fast, readable by machine than other else. (3)

From the slow and expensive manual indexing it is the demand of new exploration information age to switch over a faster and cheaper one is called automated indexing. (4)

Therefore, it is quite easy to say that, KWIC is very much applicable for search engine. It is a easy, speedy and automatic indexing system which can be operated by machine. In the 21st century keyword indexing can only be the part of search engine.

Several merits and demerits are there in KWIC. From its origin it is quite identifiable that keyword selection is some extent a problematic area and reason of that, representation of documents will be hampered and hazardous in future. (5)

The process of keyword selection is depends on rejection of non-significant words but degree of significance of keywords is not verified in this indexing system. So malfunction can be occurred in both indexing and retrieval field. (6)

Gross and Taylor found that more than a third of records retrieved by keyword searches would be lost without subject headings. A review of the literature since then shows that numerous studies, in various disciplines, have found that a quarter to a third of records returned in a keyword search would be lost without controlled vocabulary. (7)

Therefore, it is to say that, keyword searching is most needed one for search engines in modern era to speed up the process of indexing and retrieval but still it is required to modify some area due to its drawbacks.

Objectives: The main point of this paper is to highlight the various barriers in the way search engines serve information. And to be aware of their origins, scope of work, and to analyze how these problems are having an adverse effect on search engine optimization. Above all, find a solution to these problems.

1) Highlights the challenges of 'Keyword-based search system' or KWIC indexing system.
2) Analysis the impact of challenges and discover the way of recovery.

Scope: Determining how to get rid of problems by actually analyzing the source and activity of document representation problems of search engine that will become the only tool for search engine optimization and survival.

Methodology: The first is to thoroughly analyze, understand, and define the various challenges to identify key issues with search engine optimization. Second, to analyze and judgment some future steps to solve the problem as a whole, which can be used to improve search engines.

Details of Paper: Search engines are constantly being modernized to provide better service to users. Various mechanisms are used like no. of hits, ranking etc. to track the user's search intent. Again, various updated forms of traditional KWIC indexing are used like KWAC, KWOC, KLIC etc. But there are some obstacles in the way of search engines running this huge activity that are causing problems in the way of modernization of search engines. The source, activity, and scope of these problems are described below.

1) Keyword Extraction - Keywords are usually collected from the title of the document. If the title is not properly informative and meaningful, then the keyword indexing system will also become ineffective. That is, the keyword-content relationship may not be properly expressed.
2) Scattering of related documents - The alphabetical arrangement of keywords causes the related document to be completely scattered. As a result, searching for a topic requires the use of severe keywords which takes more search time. Again, as a result of searching over various pages, the user loses his search intent.
3) Scattering of related term - This method does not have a cross-reference system, making it difficult to find documents with synonyms for a particular keyword. Adjacent documents to spelling variants are also difficult to find. All these documents are searched using different keywords.
4) Less Exhaustive - Not all documents are found together by a single keyword in a large collection. That is, not all related documents can be found together under one umbrella, such as relative indexing.
5) High recall and low precision - Since the user's intent for search cannot be 100% identified. So the precision tends to decrease. On the other hand documents are searched using different keywords, and the number of collections increases day by day, resulting in increased recall.
6) Natural Language Vocabulary - Since most of the terms used in science and technology are by standard, so vocabulary is not constructed properly. Thus, this vocabulary makes it difficult to retrieve documents in the social sciences and humanities field.
7) Fixed Field Length - KWIC is normally produced by computer and the format is of fixed field length. Due to this fixed field length, part of the title is truncated when the title is lengthy. This truncation leads to funny results far away from context.

Impact: Due to these problems search engines are going through various difficulties and retrievals search engines are affected in the following cases.

1) Search Ability - The main problem with keyword search is a high recall, so users have to search deep and elaborate. As a result, the user gradually diverts from his search intent. So, the question begins to arise about the search ability of search engines.
2) Acceptability - Another problem with keyword search is low precision. This means that the user does not get the right document quickly according to his intent. As a result, search engines often fail to fulfill user requirements. This raises the question mark over the acceptability of search engines.
3) User Satisfaction – The main problem with search engines is that related documents are haphazardly represented. The result is difficult to find documents. Again documents related to related terms like synonyms, spelling variant etc. are not found by a single keyword search. As a result, the absence of all documents serially and simultaneously under one umbrella reduces the user's satisfaction level.
4) Time Consuming - Deep & elaborate search, several keyword search as a whole increases the search time indeed. As a result, users feel confused and bored and refrain from searching.
5) User Friendly - Search engines are designed to be user friendly. But in the case of retrieval, the obligation is to move search engines away from users, making them inactive. The result is a crisis of their existence.

Solution: Below is a possible solution to the problem of keyword indexing (KWIC) that is being created day by day and adversely affecting the retrieval of search engines.

1) Modify Keyword Extraction System - Keyword selection can be made more informative and content related by adopting the Significant Word and Weighted Approach instead of the Non-Significant Word approach. This will increase the precision.
2) Hierarchical Approach - Instead of a haphazard arrangement, if a sequence can be provided according to certain rules and the documents are arranged in such a way, then a direction can be given in the search for the document. This will make it easier to search.
3) Relative Approach - If the Relative Approach can be included in the KWIC Indexing, then the Relevant Term related documents can be found at a time and at a same place. This will give the benefit of relative indexing.

Conclusion: There are always some problems with any search system or indexing system. Again new problems arise in terms of long use. But that can never make the system inactive. Needless to say, we have to move forward in a new way by providing new solutions. Recovery of these indexing system issues is required to make search engines more users friendly. That is how search engines will survive in reality.

Reference

1) *Fischer, Marguerite, "The KWIC index concept: A retrospective view", Journal of the association for information science and technology, April 1966, https://doi.org/10.1002/asi.5090170203 (Last checked at 04-09-2019)*
2) *Luhn, H.P. "Keyword-in-contect index for technical literature (KWIC Index)", presented at american chemical society, Division of chemical literature at 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).*
3) *Luhn, H.P. "The automatic derivation of information retrieval en-codements from machine-readable texts," International business machines corp., Yorktown heights, N.Y. 1959, 9p. Also,min A. Kent, "Information retrieval and machine translation", Pt II, 1961, pp. 1021-1028.*
4) *Obaseki, Tony I., "Automated Indexing: The Key to Information Retrieval in the 21st Century" (2010). Library Philosophy and Practice (e-journal). pp. 338.*
5) *Sedano, John Michael, "Keyword-in-context (KWIC) indexing: Background, statistical evaluation, pros and cons, and applications", university of pittsburgh, 1964*
6) *Helbich, Jan, "Direct selection of keywords for the KWIC index", Information Storage and Retrieval, vol. 5,no. 3, pp. 123-128, Oct 1969 .*
7) *Tina Gross, Arlene G. Taylor & Daniel N. Joudrey (2015). "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching". Cataloging & Classification Quarterly, 53:1, 1-39, DOI: 10.1080/01639374.2014.917447*
8) *Kraft, Donald H., "A comparison of keyword-in-context (KWIC) indexing of titles with a subject heading classification system", Journal of the association for information science and technology, Jan. 1964, https://doi.org/10.1002/asi.5090250209.(Last checked at 04-09-2019)*
9) *J.D. Anderson, J. Perez-Carballo. "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort". Information Processing and Management 37 (2001), pp. 255-277*
10) *Beall, Jeffrey. "The weaknesses of full-text searching". The Journal of Academic Librarianship, Volume 34, Number 5, pages 438-444*