

Efficiency of Boolean Search strings for Information Retrieval

Muhammad Bello Aliyu

Computer Science Unit, Department of Mathematics, UsmanuDanfodiyo University Sokoto.

Corresponding Author: aliyu.mbello@udusok.edu.ng

Abstract: The review of available literature is a foundation requirement for most research projects. The relevant literatures should be searched from multiple sources. Search engines and on-line bibliography resource sites are conventionally used to find the relevant literatures using key word search. However, with little automated help for the free text search query. In this paper, the technique of Boolean search string is explored in details along with the analysis/evaluation of the effectiveness of the technique. Searching engines such as google, google scholar and online bibliography sources such as IEEE Xplore, ACM and Science Direct were used to implement the technique. The technique was evaluated based on three (3) criteria: Number of documents retrieved, the time taken to retrieve them and the relevance of the documents to the query or research question. The analysis of this technique shows that Boolean search strings technique returns more relevant articles compared to the free text query by at least 77% and in shorter time frame. Hence, Boolean search strings are very useful for information retrieval.

Keywords: Information retrieval, search query, Boolean search string, search engines, internet searches.

Date of Submission: 06-11-2017

Date of acceptance: 25-11-2017

I. INTRODUCTION

Information retrieval from a large collection or online sources that satisfies the information need is very challenging especially for a research project. The dominant searching techniques are based on the ranked retrieval models [7]. Using these models, users specify free text queries i.e. Compose one or more words to search for information. The system decides which documents best satisfy the query [2]. The overall goal for the information retrieval is to find, within a large collection of documents in different databases, those documents which satisfy the user information need [3]. Therefore, rather than let the system makes the decision of which documents best satisfy the query, the user should take charge of this decision by using a precise choice of words with operators for building up query expression. This technique is based is a Boolean operators. The Boolean retrieval model for information generates a query which is in the form of a Boolean expression of terms. That is, terms are combined with the operators AND, OR, and NOT. The model views each document as just a set of words.

A number of tools supporting systematic review provide support for this technique. Tools such as SESRA [4] and Parsifal [5] have enabled the Boolean search string formulation using population, intervention, comparison, output and context (PICOC) components. The search string also involves the generation of the synonyms of the various components of the PICOC elements. However, none of the tools does automate the PICOC elements completion from the research questions, users have to manually break the components down.

Several online databases also incorporate this technique in the search of related articles. IEEE Xplore, ACM digital library and the host of the giant online databases support this technique. However, each of the above does require a slight specification for the way that Boolean search string is constructed. The details of the various specifications are outline in section 3.

When searching for information from online sources, related documents must be returned as much as possible and within a reasonable timeframe. These, in addition to the relevance of the documents to the query, are part of the criteria for information searches. Therefore, the usefulness of this technique (Boolean technique) is measured by these three (3) criteria.

Measuring the effect of this technique is very important for information retrieval and the review quality as a whole. However, unlike the number of the articles returned and the time taken to return them, the relevance of the papers to the research questions is not easy to measure. Notwithstanding, the frequency of the keywords in the query is used to inform the relevance of the returned papers by the search engines.

In this paper, the technique of the Boolean search string is introduced along with its computational benefits.

II. BOOLEAN SEARCH STRINGS

Searches using the free text query returns a number of records or documents. The documents returned are those which match the free query submitted by the user. As shown in the previous section, the system decides which documents match the query. The search engines takes a long time to look at hundreds of records. By putting a little effort into constructing search strings (what you type into the search box) you can save a lot of time. The database can do a lot of work for you if you know how to add a little sophistication to your search strings. If the search returns too many records, search can be narrowed by adding more search terms. To ensure all the records your search finds contain all the search terms, link them with the **AND** operator.

Table 2.0 the Boolean operators

OPERATOR	USE	EXAMPLE
AND	This concatenate the words and produce results that include all the keywords linked with AND	Java AND Android
OR	This produce the result that include either or all the keywords	Britain OR UK
NOT	Results Excludes a keyword from your search	NOT(Java)
QUATION MARKS “ “	Search for an exact phrase (Consider keywords in quotation marks as a whole word)	“Systematic review”
BRACKETS	Group multiple search strings and set priorities	Tool OR (“software engineering”)

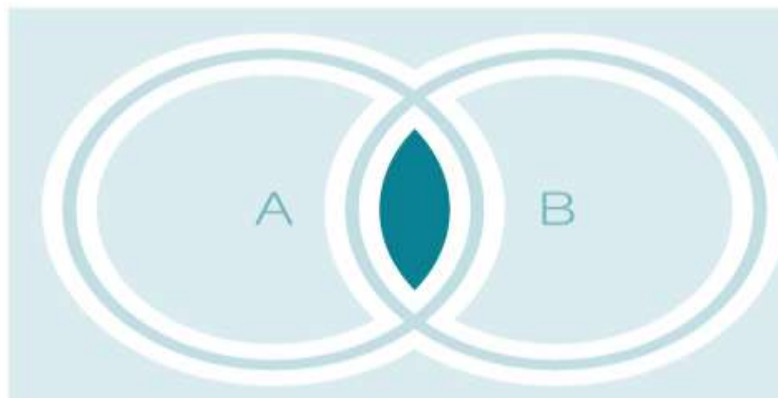


Figure 2.1(A AND B) | $A \cap B$

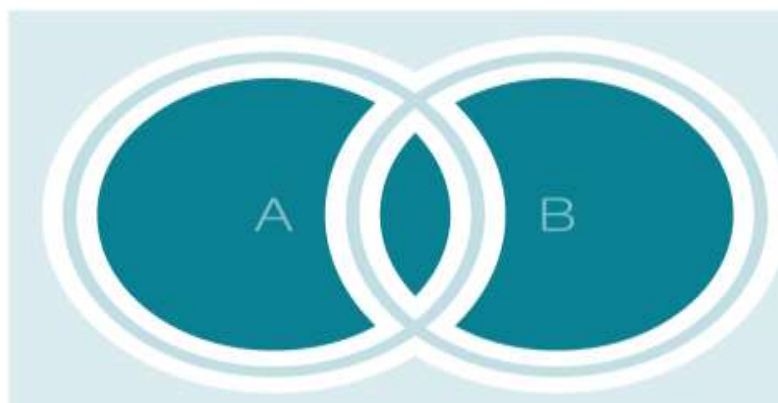


Figure 2.2(A OR B) | $A \cup B$

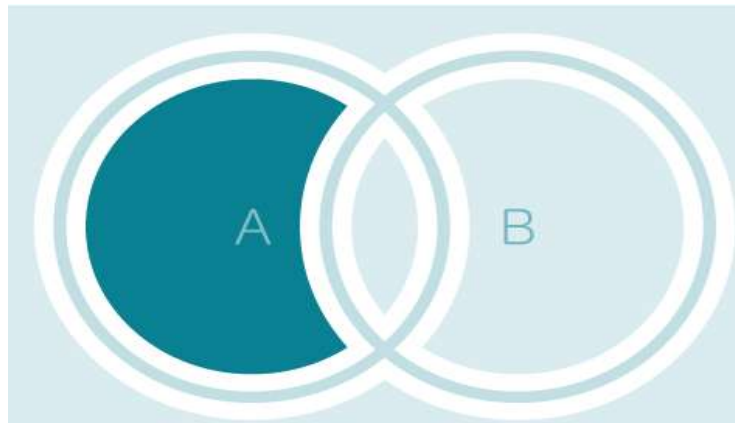


Figure 2.3 (A NOT B) | A - B

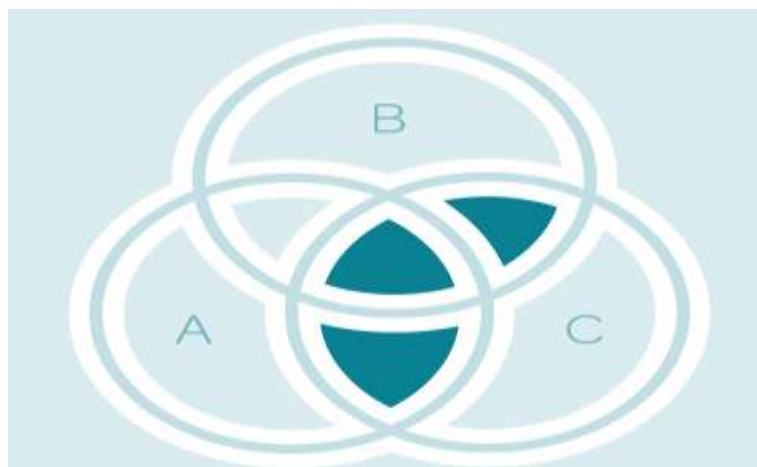


Figure 2.4(A OR B) AND C

III. Background

The Boolean search strategy is the process by which the research question is translated into the research string that can be used to retrieve relevant articles from the online sources: databases or search engines. The search string is not a complete sentence or question rather a set of key words connected by the Boolean operators highlighted above [6]. The following are the main steps for constructing the search strategy

1. Extract the main ideas - This is the identification of the key words and subjects in the research questions. To extract main ideas, it is important to understand the subject area so the key ideas can easily be generated easily and effectively [7]. The key ideas should exclude the auxiliary verbs, definite articles and the conjunctions. For example, the following question shows the keywords and subject highlighted in red.

Analysis of JAVA and PHP for enterprise applications

2. Create a list of synonyms and related words - Here, the synonyms of the highlighted subject and keywords above, are suggested. This is because, the various literatures may cover the concept in different terms which are known in the subject. Search engines are blinded by the keywords specified, hence any document not represented by the search terms could be missed.
3. Link synonyms and related words with OR – The keyword and its corresponding synonyms are linked using OR Boolean operator.
4. Connect the main ideas with AND. The AND operator is used to keywords and their respective synonyms.
5. Exclude unrelated ideas with NOT – Unwanted keyword or its synonyms can be excluded using the NOT operator.
6. Utilise phrases and truncation – for any priorities, use the parenthesis.

Boolean algebra has been used for information retrieval. [8] Presented a mathematical model of a weighted Boolean retrieval system for the evaluation of the relevance of the document retrieved from search. Similarly, [9] developed an extended model for the Boolean search retrieval. In the model, the precision of the model was calculated. Four different collections of the documents: documents in bio-medicine, library science electrical engineering and computer science were used to evaluate the model. Similar investigations were carried out using the Boolean search strings such as: [10] However, for most of the previous works, no attempt was made to measure the efficiency of this technique with respect to the search engines or other online sources. This research does that.

IV. IMPLEMENTING THE SEARCH STRATEGY

Here, the technique is put into use. A research topic is randomly selected and implemented using the technique. Consider the research topic: **“Software Cost Estimation Methods”**.

From this topic, we can search for relevant papers using the Boolean search technique as follows.

First, the keywords and concepts were identified from the topic as shown in table 4.1 below. Some key concepts may be a single word like ‘software’ or a phrase like “Cost estimation”. Unlike a single word concept, a phrase requires the quotation marks. As shown below

Table 4.1 keywords concepts

Concept	Keywords/subjects
1.	Software
2.	“Cost estimation”
3.	Methods

From the identified keywords or concepts above, the related synonyms and words is brainstormed as follows:

Table 4.1 Keywords and synonyms

Concept	Keywords/subjects
1.	software, project
2.	“Cost estimation”
4.	Method, approach

The synonyms are then linked with OR operator as follows:


Table 4.2The ‘OR’ concatenation

Concept	Keywords/subjects
1.	software OR project
2.	“Cost estimation”
4.	Method OR approach

The results of the above OR operator linkage is connected using the ‘AND’ operator. The ‘AND’ operator ensures that only results that contain all the connected terms is returned.

Table 4.3 the ‘AND’ concatenation

Concept	Keywords/subjects
1.	software OR project
2.	“Cost estimation”
4.	Method OR approach



The diagram shows two blue curved arrows pointing to the right, each labeled 'AND'. The top arrow connects the first and second rows, and the bottom arrow connects the second and fourth rows, indicating the logical AND operation between the rows.

Combining the ‘OR’ (corresponding to the rows) and the ‘AND’ constituents will generate the required Boolean search string. Ideally, the entire row (a concept and its associated synonyms and words) is taken (connected with AND operator) and combined with another row (concept and its associated synonyms). However, if all the rows appear to be much, a combination of the row is selected and run at several times. So, the following search strings were derived from the above tables. For efficiency, from the combination of two (2)

AND operators i.e. three (3) concepts before taking them as a whole. Hence, the Boolean search string for the above research question becomes:

(Software OR application OR project) AND (“cost estimation”) AND (method OR approach)

This Boolean search string can be generally applied to most search engines and bibliography sources. However, there are some restrictions imposed on some of them as to the number of Boolean elements in the search string. Let's explore the search strings specification for the various online databases.

4.1 Running the Search Query

Using the research question as the free text query, we first run the query on the four online sources: google scholar, IEEE Xplore, Science Direct and the ACM Digital Library. Later, the Boolean search string is also run. The output from the two searches were collected are analysed. This can be seen in the result section.

V. Result

The table 1 and table 2 below show the number of documents returned by the various bibliographic databases and search engine when both the free text query and the Boolean search string are run; and the time taken for the retrieval. The google scholar and the Science Direct display the time taken by the query to fetch the results while the IEEE Xplore and the ACM Digital Library do not display the

Table 4.1 Free text query result

Source	Number of documents	Time (sec)
Googlescholar	3220,000	0.20
IEEE Xplore	1377	
ACM Library	213549	
ScienceDirect	615	0.34

Table 4.2 Boolean search strings result

S/N	Number of documents	Time (sec)
Googlescholar	84000	0.16
IEEE Xplore	305	
ACM	94	
ScienceDirect	48	0.12

From the tables above, it can be seen that irrelevant documents from search engines can be screened out by Boolean search strings at least 97% in Google scholar, 77% in IEEE Xplore, 99% in ACM and 92% in Science Direct. The figure below shows the computation. It also saves time, since the search engines will not bother about the irrelevant ones. Here is how the figures are arrived at. It is based on the percentage of the free text query result and the Boolean search string results displayed in the tables 1 and 2 above.

$$\begin{aligned}
 \text{Google scholar} &= \frac{(3220\ 000 - 84,000)}{3220000} \\
 &= 97.4\% \\
 \text{Ieeexplore} &= \frac{(1377 - 305)}{1377} \\
 &= 77.9\% \\
 \text{ACM} &= \frac{(213549 - 94)}{213549} \\
 &= 99.9\% \\
 \text{Science Direct} &= \frac{(615 - 48)}{615} \\
 &= 92.2\%
 \end{aligned}$$

The number of returned documents for both the free text query and the Boolean search strings are also displayed in charts for easy glance. The difference can also be seen easily between the two results

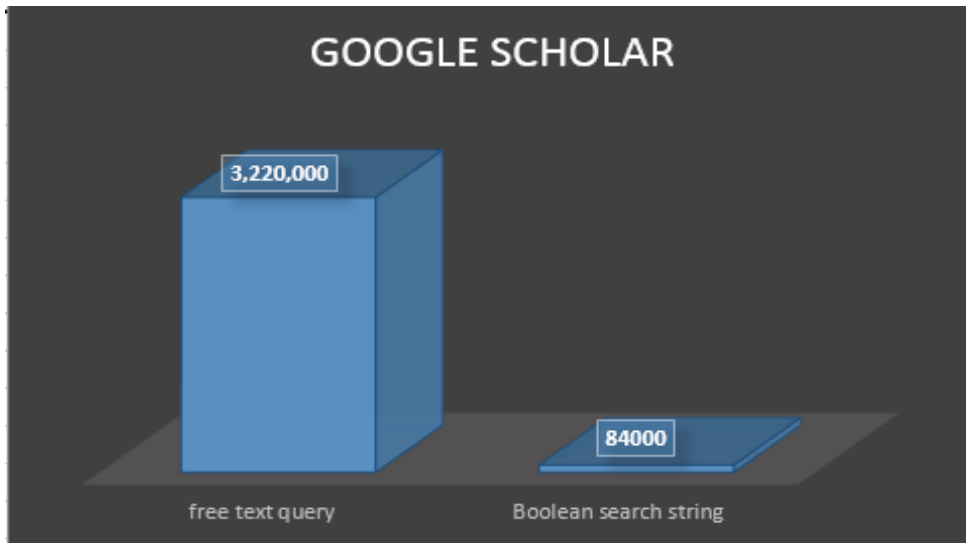


Fig 1.0 the google scholar results

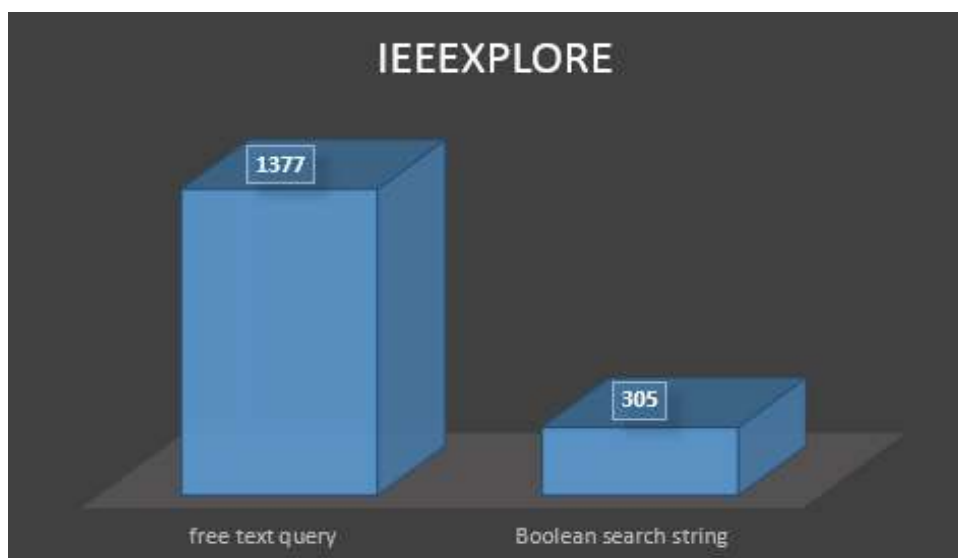


Fig 2.0 the IEEE Xplore results

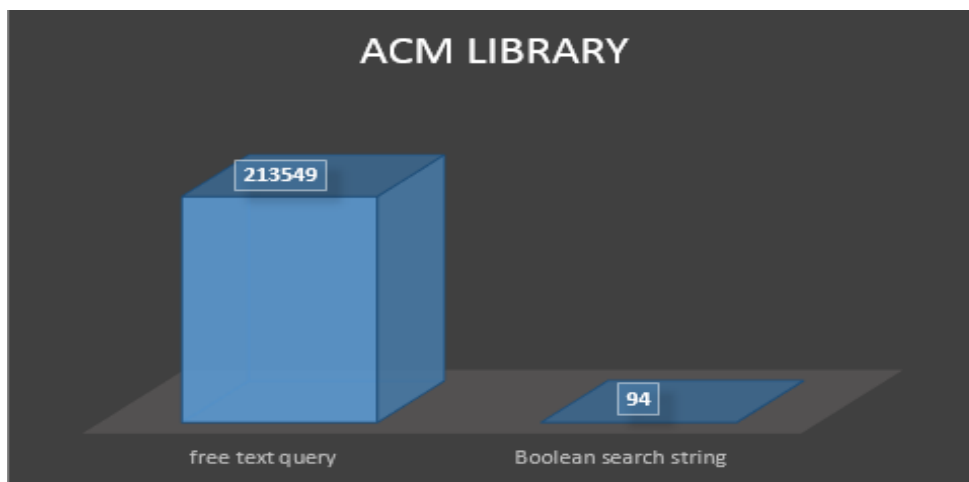


Fig 3.0 the ACM Library results

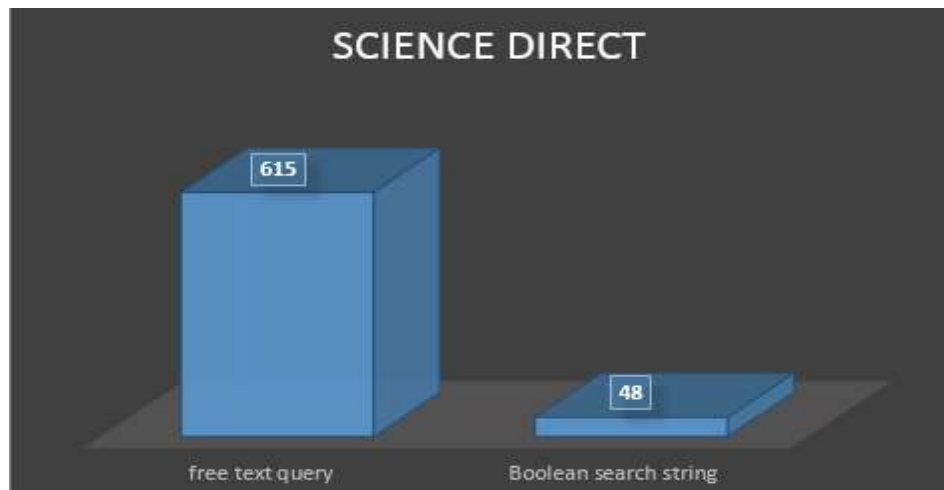


Fig 4.0 the Science Direct results

VI. CONCLUSION

The information retrieval is essential for the researchers. In the digital age, most information is stored electronically. Harnessing the power to effectively acquire the information is essential. The researchers rely on the documents they are able to get from the online sources. However, the search for the documents needs to be, not just timely but relevant. By taking over the control of which document are relevant, the search for the articles is more efficient and time economical. This is because, instead of the search engines to evaluate the large of documents (many of them not even related to the subject), and specific documents are selected, hence saving the time taken for conducting the searches. However, the big question is: are all the relevant documents captured? What if relevant documents are left by the Boolean search strings? This is carefully analysed by the two results (free text query and the Boolean search strings) and it shows that the Boolean search strings return more relevant documents (in less number) than the free text query which returned large number of documents with lots of irrelevant ones.

REERENCES

- [1]. Turtle, H. R. and Croft, W. B. (1992) 'A Comparison of Text Retrieval Models'. *The Computer Journal* 35 (3), 279-290
- [2]. Stanford, N. G. (2009) *Boolean Retrieval*. Online: Cambridge UP
- [3]. Lewis, D. D. (ed.) (2014) *Machine Learning: Proceedings of the Eighth International Workshop. 'Learning in Intelligent Information Retrieval'*
- [4]. Molléri, J. S. and Benitti, F. B. V. (eds.) (2015) *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. 'SESRA: A Web-Based Automated Tool to Support the Systematic Literature Review Process'*: ACM
- [5]. Verner, J.M., Brereton, O.P., Kitchenham, B.A., Turner, M. and Niazi, M., (2012). *Systematic literature reviews in global software development: A tertiary study*.
- [6]. Montori, V. M., Wilczynski, N. L., Morgan, D., Haynes, R. B., and Hedges Team (2005) 'Optimal Search Strategies for Retrieving Systematic Reviews from Medline: Analytical Survey'. *BMJ (Clinical Research Ed.)* 330 (7482), 68
- [7]. Biondi-Zoccai, G. G., Agostoni, P., Abbate, A., Testa, L., and Burzotta, F. (2005) 'A Simple Hint to Improve Robinson and Dickersin's Highly Sensitive PubMed Search Strategy for Controlled Clinical Trials'. *International Journal of Epidemiology* 34 (1), 224-225
- [8]. Waller, W. and Kraft, D. H. (1979) 'A Mathematical Model of a Weighted Boolean Retrieval System'. *Information Processing & Management* 15 (5), 235-245
- [9]. Salton, G., Fox, E. A., and Wu, H. (1983) 'Extended Boolean Information Retrieval'. *Communications of the ACM* 26 (11), 1022-1036
- [10]. Aromataris, E. and Riitano, D. (2014) 'Constructing a Search Strategy and Searching for Evidence. A Guide to the Literature Search for a Systematic Review'. *The American Journal of Nursing* 114 (5), 49-56

Muhammad Bello Aliyu "Efficiency of Boolean Search strings for Information Retrieval."
 American Journal of Engineering Research (AJER), vol. 6, no. 11, 2017, pp. 216-222.