International Journal of Computer Applications Technology and Research
Volume 7–Issue 10, 398-406, 2018, ISSN:-2319–8656

Text Mining in Digital Libraries using OKAPI BM25 Model

Gesare Asnath Tinega[1]
Student SCIT,
JKUAT
Nairobi, Kenya

Prof. Waweru Mwangi[2]
Associate Professor SCIT,
JKUAT
Nairobi, Kenya

Dr. Richard Rimiru[3],
Senior Lecturer SCIT, JKUAT
Nairobi, Kenya

Abstract: The emergence of the internet has made vast amounts of information available and easily accessible online. As a result, most libraries have digitized their content in order to remain relevant to their users and to keep pace with the advancement of the internet. However, these digital libraries have been criticized for using inefficient information retrieval models that do not perform relevance ranking to the retrieved results. This paper proposed the use of OKAPI BM25 model in text mining so as means of improving relevance ranking of digital libraries. Okapi BM25 model was selected because it is a probability-based relevance ranking algorithm. A case study research was conducted and the model design was based on information retrieval processes. The performance of Boolean, vector space, and Okapi BM25 models was compared for data retrieval. Relevant ranked documents were retrieved and displayed at the OPAC framework search page. The results revealed that Okapi BM 25 outperformed Boolean model and Vector Space model. Therefore, this paper proposes the use of Okapi BM25 model to reward terms according to their relative frequencies in a document so as to improve the performance of text mining in digital libraries.
Keywords: Online Public Access Catalogs, Relevance Ranking, Digital Libraries, Okapi BM25 Model, Text Mining, Information Retrieval Models

1. INTRODUCTION

The internet and information technology evolution has drastically transformed information development and access, especially in the library sector thus disrupting the functionality of libraries. As a result, majority of the libraries have digitized their content in order to remain relevant and exist in distributed networks [11]; [7]. Users are now using Public Access Catalogs (OPAC) to search and retrieve information from the digital library's database [5]. Khiste, Deshmukh & Awate [8] defined digital libraries as huge collection of electronic information that can be accessed by distributed users from different locations. In their study Dwivedi; Sharma & Patel, defined OPAC as a library catalog that displays a large collection of materials held by a database in which users search to access the desired documents available at a library by using in search terms such as the author, title, subject/keyword, or date of publications of the material [5]; [17].

However, studies reveal that digital libraries are still losing to other online search engines such as Amazon despite the efforts to transform library catalogs from traditional card cataloging to digital cataloging using Open Public Access Catalogs (OPACs). This is so because the results retrieved at the library's OPAC catalog does not satisfy the users need. Kumar & Vohra [9] explains that the majority of OPACs requires exact search terms to perform relevancy ranking otherwise they will display the 'no output/null retrieval in the results section. Others simply rank the results using last in/first out. The most cataloged items will show up ending up not meeting the expectations of the user. The digital libraries' OPAC use the Boolean model for information retrieval which retrieves too many or too little of the documents. These causes havoc to users when searching relevant results. It is therefore in the interest of the researcher, to establish how to improve search capabilities in the digital libraries by implementing the Okapi BM25 algorithm in order to improve relevance ranking in the online public access catalogs (OPACs) before the results are displayed to the user. The Okapi BM25 model is based on the term frequency, length normalization to improve the relevance performance of the digital libraries especially during retrieval.

2. LITERATURE REVIEW

2.1 Digitization

Information and communication technology (ICT) in libraries and many organizations has led to the increase of soft data and digitization of materials [10]. Materials are digitalized to improve their online accessibility, sorting, transmission and retrieval. Digitization refers to the process of converting print media to the digital content for electronic storage, access, and distribution among users [3]. The digitization process has facilitated storage and enhanced ease manipulation of the traditionally digitized content by researchers [25]. The process has further decentralized information storage therefore making information in the digital libraries readily accessible from anywhere anytime around the globe.

2.1.1 OPAC catalog

Online public access catalog is one of the most important tools that contain all the bibliographic collection of documents stored in the digital library database [19]. The frequent use of the internet among the researchers has slowed the usage of the library catalogs since they lack most of web 2.0 features such as relevancy ranking [12]. The huge unstructured and amorphous data available in the digital library databases has on the other hand made it difficult for developers to come up with algorithms for enhancing successful information retrieval that matches the user queries [3]. In their study Kumar & Vohra [9] established that 12.5 % of the library users at Guru Nanak Dev University found the OPAC catalogue to be slow and complicated to use thus they needed help from librarians. Current generations of library users are not satisfied with the results that the catalog retrieves because they display either too many or too little documents in a given search. The recent developments of the newer catalogs by organizations outside of libraries have resulted in vocal criticisms about the capability of digital libraries especially on relevance ranking [1].

2.3 Text mining

This paper adopts the definition of Talib et al [21] that defines text mining as a type of indexing which aims at extracting structured text data from unstructured text data. Text mining process involves gathering, preprocessing, and text analysis of document from various sources. These processes are carried out to ensure user satisfaction when accessing structured data from unstructured databases. Text mining techniques such as information retrieval, classification, clustering and categorization are thereafter used to ensure that data is analyzed and generated correctly [27]. This paper will however focus on the information retrieval (IR) approach since it aims at retrieving relevant data to users from a large library database.

2.4 Information Retrieval Process

The main objective of the OPAC catalog is to retrieve relevant documents from a large library database so as to satisfy the user information need. Information retrieval models are used to perform the matching process between the library database and the user query for retrieval. The three basic processes involved in information retrieval include indexing, query formulation and matching [13]. Indexing refers to the document representation process. Query formulation also known as indexing is done to by unique terms expressed by a user while query evaluation also known as matching process is done to estimate the level of relevance of a document to a given query [4].

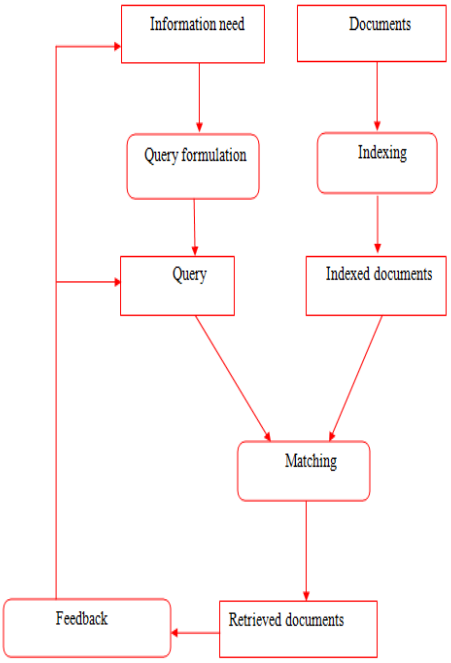

Figure 1 Information retrieval process

2.5 Information retrieval models

2.5.1 Boolean Model

It is an information retrieval model grounded on set theory to determine the prospect of document retrieval. Boolean model is an example of exact match model whereby the fate of the documents retrieval is determined based on the type of information stored in the database [14]. The model uses the logical AND, OR, and Not operators to perform document search in the library databases [23]. The AND operator retrieves results that include all the keywords linked with the operator while OR operator produces results that contain either one or all the keywords used in the user query. The NOT operator retrieves results that excludes the keyword from the user query. The Boolean model is however criticized of lack of relevance ranking when used in retrieval systems such as the OPAC catalog. Boolean model also does not support length normalization of the documents since it does not use term weight such as term frequency and inverse document frequency when retrieving documents from the library database [2].

2.5.2 Vector Space Model (VSM)

This model was introduced to overcome the limitations of Boolean model by assigning weights to term for better matching. VSM presents text documents as vectors to find the similarity between the documents stored in the database and the user query using cosine similarity. Moreover, the model is also used to find exact results with relevance ranking [17]. VSM obtains relevance ranking and information retrieval using document indexing, weighting of the indexed terms using the TF-IDF and finally ranking the documents archives as per the query comparability value [6]. The cosine similarity of the VSM is calculated using the equation 1 below.

$$sim\left(d_j, q\right) = \frac{d_j.q}{\|d_j\|\|q\|} = \frac{\sum_{i=1}^{N} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{N} w_{ij}^2} \sqrt{\sum_{i=1}^{N} w_{i,}^2}} \quad (1)$$

Where: d_j represents the total collection of documents, q signifies the user query, $W_{i,j}$ is the i^{th} term of a vector for document j, Wi,q= is the i^{th} term of a vector for query q, and N= is the total number of keywords in a given data set. The model, however, faces some major drawbacks such as poor representation of long documents which is as a result of repetitive use of terms. Moreover, Jain, et al [29], established that the model has low sensitivity to semantics. For instance the word “car” and “automobile” will not give the same match if both words are found in same document. A study by Yulianto et al [2], also revealed that VSM is hard to understand and takes a lot of time to search and match documents before retrieval.

2.5.3 Okapi BM 25 model

The Okapi Best Match 25 (BM25) model is a non-binary model that was developed as part of the Okapi Basic Search System in the TREC Conferences. Okapi BM25 is a probabilistic model that is based on the probabilistic theory. The model is a well-performed term weighting scheme that retrieves its relevant results by incorporating the use of weight term using TF-IDF, and length normalization of a given document [22]. BM25 is a bag-of-words retrieval function that ranks documents according to their relevant results. Okapi BM25 not only considers the frequency of the query terms but also the whole the length of the document under evaluation [26].

2.5.3.1 TF-IDF Weighting of Okapi BM25 Model

In Okapi BM25, term frequency also termed as document frequency shows the frequency of a query term in a document for it to be considered to be relevant. Inverse Document Frequency (IDF), on the other hand, is used to differentiate between common words and uncommon words within a

document. The simplest score for document d can be illustrated in the equation 2.

$$RSV_d = \sum_{t \in q} log \frac{N}{df_t} \qquad (2)$$

Where: N is the total number of documents in a given corpus; df_t is the document frequency of a term.

$t \in q$ is an element of a query.

TF-IDF considers short documents to have more weight than long documents therefore; Okapi BM25 model outperforms TF-IDF and vector space model by taking the average length of each document separately using tuning parameters. Tuning refers to the process by which one or more parameters are adjusted upwards or downwards to achieve an improved or specified result. The values of the tuning parameters are determined empirically using a test collection of documents, queries, and relevance judgments. K_1 is set to 1.2 to control term-frequency saturation since low values result in quicker saturation while high values results in slower saturation. The tuning parameter b is set to 0.75 to control field-length normalization of a document. The Okapi BM25 model calculates the retrieval status value of a given document in order to determine the relevance of a document as shown in equation 3.

$$RSV_d = \sum_{t \in q} log \left[\frac{N}{df_t}\right] \cdot \frac{(k_1+1)tf_{td}}{k_1((1-b)+b\times(L_d/L_{ave}))+tf_{td}} \qquad (3)$$

Where:

Retrieval Status Value: relevancy scores of a document.

N: represents documents in a given collection.

df_t-the frequency of a query term in a document.

$t \in q$ - t is an element of query q.

t- term

q- query

tf_{td} : signifies the frequency of a term in document d

***L_d* (Lave):** used to calculate the average document length in the whole collection

***k_1*:** tuning parameter set to 1.2

***b*:** tuning parameter set to 0.75

K3 tuning parameter is set to 2 in case the retrieval involves long documents as shown in equation 4.

$$RSV_d = \sum_{t \in q} \left[log \frac{N}{df_t}\right] \cdot \frac{(k_1+1)tf_{td}}{K_1((1-b)+b\times\frac{L_d}{L_{ave}}))+tf_{td}k_3+tf_{tq}} \cdot \frac{(k_3+1)tf_{tq}}{k_3+tf_{tq}} \qquad (4)$$

2.5.3.2 Example of OKAPI BM25 Model

Example query: "president lincoln"

$tf_{president,q}$= $tf_{lincoln,q}$= 1

No relevance information: R= ri= 0

"president" is in 40,000 documents in the collection: $df_{president}$= 40,000

"lincoln" is in 300 documents in the collection: $df_{lincoln}$=300

The document length is 90% of the average length: dl/avg(dl) = 0.9

We pick k1= 1.2, k2= 100, b= 0.75. Hence using the Okapi BM formula illustrated at equation 2.13 the RSV of the query is shown in table 1 below.

Table 1 Retrieval status values of Okapi BM25

$tf_{president,d}$	$tf_{lincoln,d}$	BM25
15	25	20.66
15	1	12.74
15	0	5.00
1	25	18.2
0	25	15.66

The low df term plays a bigger role.

3. METHODOLOGY

3.1 Research Design

This paper used a case study research design to generate solutions for improving information retrieval in JKUAT library. Experimental research was also used to manipulate variables and determine their effect on the dependent variable**.** This study involves manipulation of text mining technique such as information retrieval to improve the OPAC catalog used in digital library.

3.2 Model Design

A prototype was used to develop this model. Prototype model was selected because it allows development, verification in terms of performance, and reworking on the framework until an acceptable prototype is finally achieved. The prototype processes help to complete a given framework in the area of study. The figure 2 below illustrates the OPAC model design that was used for the development of the model.

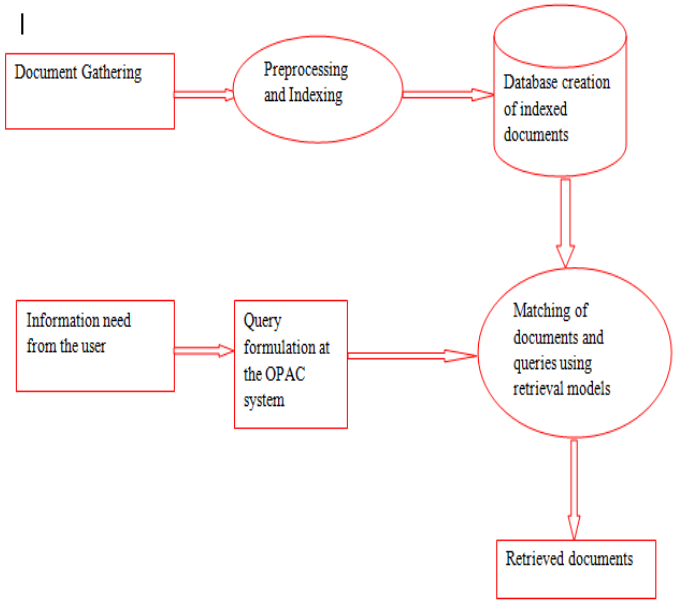

Figure 2 Opac Model design

3.3 OPAC framework Requirements

The front end of the proposed OPAC catalog was implemented using HyperText Markup Language (HTML), Cascading Style Sheets (CSS), Bootstrap, Laravel framework, and JavaScript language. MySQL was used to develop the database while server side programming of the OPAC system was done by Hypertext Preprocessor (PHP).

3.4 Document Gathering

The study utilized secondary data from the Google search engine and other online journals. The collected document were pre-processed using Google to remove inconsistencies such as tokenization, stop words and stemming before the documents were downloaded to be populated to the database. Different search queries were used to collect all the 300 documents that were used to create the database from online journals such as strategic journal of business and change management, scientific research an academic publisher, and International Journal of Computer Science and Engineering Survey among others. For instance, the query "*Text mining and digital library*" was used as a user query using the search engine and resulted in 10 articles were displayed on the first page of the search engine. Seven documents that were found in Portable Document Format were collected uploaded to the database for further analysis shown in table 2. This process was repeated until the collection of 300 documents was achieved.

Table 2 Document gathering

Doc. No	Document Title	Doc. Type	Size	Time (s)
1	Text mining in a digital library	PDF	323.2Kb	(0.47 seconds)
2	Integrating Data and Text Mining Processes for Digital Library Applications	PDF	297 kb	(0.47 seconds)
3	Application of Data Mining Technology in Digital Library	PDF	321 kb	(0.47 seconds)
4	Opportunities and Challenges of Text Mining Hathi Trust Digital Library	PDF	2.0 mb	(0.47 seconds)
5	Library Support for Text and Data Mining	PDF	197kb	(0.47 seconds)
6	Data Mining –A Librarian Overview	PDF	56.9 mb	(0.47 seconds)
7	Integrating Data and Text Mining Processes for Digital Library Applications	PDF	348 kb	(0.47 seconds)

3.5 Entity Relationship Diagram.

The database_item was populated with the 300 documents collected as shown in figure 3. An Entity Relation Diagram (ERD) that was used to create the database.

The database is made up of two entities namely administrator and books. The entities use one to many relationships and therefore, one administrator or a user can add many books to the OPAC system. The user queries the database using either of the attributes of the book entity.

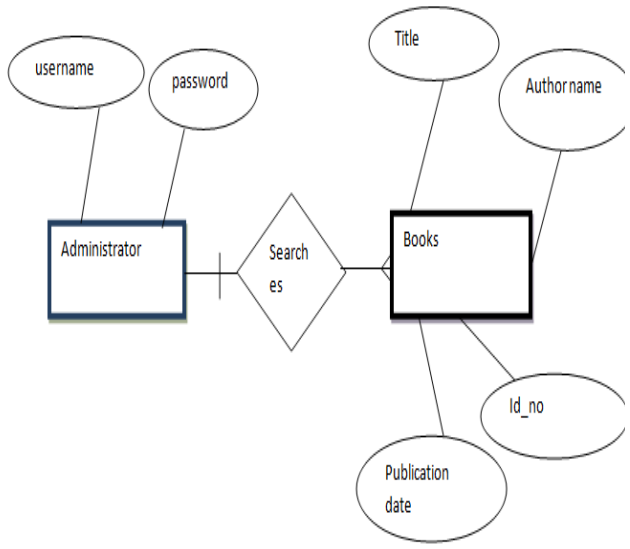

Figure 3 ERD of the Opac database

3.6 Stemming Process

Stemming seeks to reduce different grammatical forms of a word like its noun, adjective, verb, adverb among others and remove various suffixes from a word to get its common origin [24]. This is done to the retrieval models so as to save time and memory space. For example, the words user, usage, using, and usability can be rooted in the word use. The process of stemming helps a retrieval model to have exact matching stems and increase their performance level especially in document retrieval. This can be illustrated in figure 4.

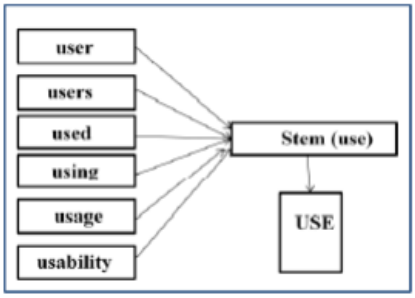

Figure 4 Stemming output

3.7 Routing

All the OPAC framework routes are registered within the *app/routes.php* file. This file tells the php framework (laravel) the URIs it should respond to and the associated controller that will give it a particular call.

3.8 Search and matching process

The user uses the search box that was created at the front end to query the database for the results to be processed by the information retrieval models. Tokenization of the documents is done to remove inconsistencies such as commas, full stops among others. Matching is done before the results are displayed to the user. It seeks to compare the user query against the indexed documents. This result in a ranked list of documents that will be used by the users in search of the information they need. The Boolean Model, Vector Space Model and the Okapi BM25 were utilized for this study. Retrieved results were displayed on the performance basis of each model. The fact that VSM and Okapi BM25 rank their results qualified them to be effective ranking models as compared to Boolean model. The following code was used for search query and matching process

```
//boolean search
foreach($books as $book){
        $keywords = explode(',', $book->title);
foreach($phrases as $phrase){
if(stripos(json_encode($keywords), $phrase) !== false){
if(in_array($book, $items)){
  } else {
array_push($items, $book);
 }
}
}
}
//vector space
$vectoritems = array();
      $books = Book::all();
foreach($books as $book){
$keywords = explode(',', $book->abstract);
foreach($phrases as $phrase){
if(stripos(json_encode($keywords), $phrase) !== false){
if(in_array($book, $vectoritems)){
   } else {
array_push($vectoritems, $book);
              } }   }
//okapi
         $okapiitems = array();
          $books = Book::all();
foreach($books as $book){
$keywords = explode(',', $book->keywords);
foreach($phrases as $phrase){
if(stripos(json_encode($keywords), $phrase) !== false){
if(in_array($book, $okapiitems)){
  } else {
array_push($okapiitems, $book);
              }
  }
}
}
```

3.9 Database connection Code

The front end and the back end of the OPAC system were connected to produce the results through the database connection. The following PHP code was used to connect MySQL and select item-database

```
<?php
$mysqli= new ,mysql (“localhost”, “username”, “ password”, “dbname”);
?>
```

When the above code connects MySQL and selects item database the user queries can now be used at the search page to display the results.

3.10 Retrieved Relevant Documents

The improved OPAC catalog is then used to retrieve relevant documents from the database. The retrieved relevant documents are then displayed at the catalog for the users to view, use and compare the performance of each information retrieval models used.

4. PERFORMANCE AND EVALUATION OF RESULTS

4.1 OPAC Results

Once the user has installed PHP software (Xampp) in the computer Apache and MySQL module are turned on as shown in figure 5 below.

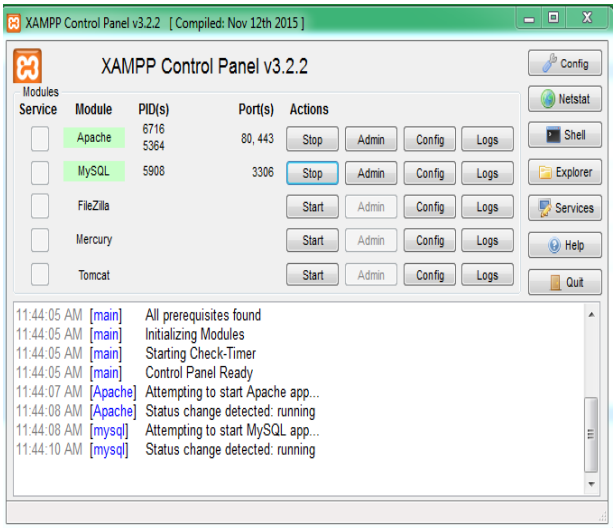

Figure 5 Xampp Control Panel

The user opens any browser and enters the url:http.localhost/opac/public/users/login The following screen will appear for the user to enter his or her email address and a password to access the system.

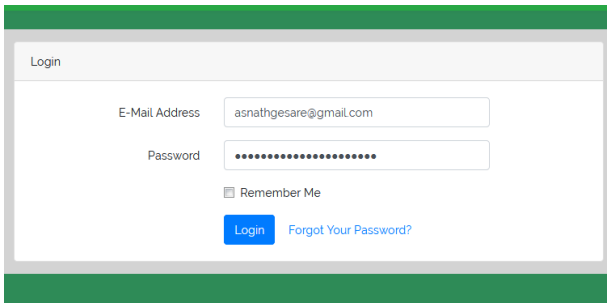

Figure 6 Opac framework

The OPAC framework displays the figure below once the user logs in the details. This can be illustrated in figure 7 below.

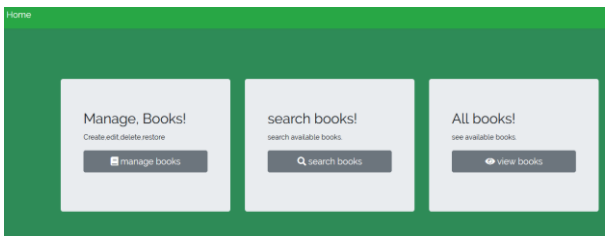

Figure 7 Opac framework search display

When the user hits the search button the following results are observed as illustrated in the figure 8 below

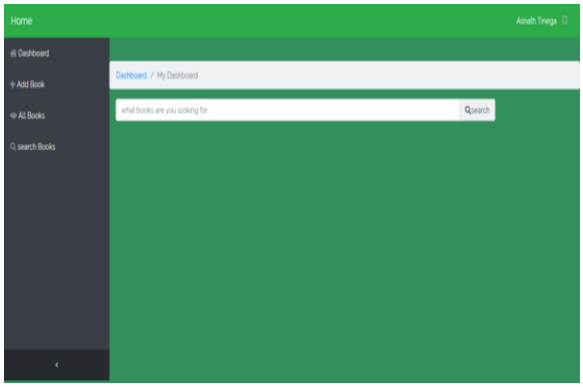

Figure 8 Search Display

When the user searches for example the query “Information System” the three models displayed the following results as shown in figure 9.

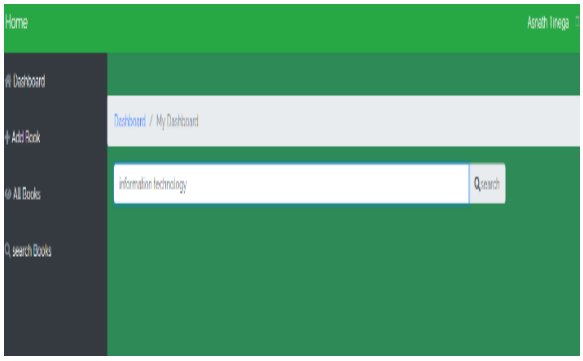

Figure 9 users enters a query" information technology

The results of the search entered by the user in figure 9 results to the retrieval of relevant documents from each model. The first page of the retrieved was screenshot as shown in figure 10. The Okapi BM 25 model retrieved documents by calculating the retrieval status value of each relevant document. Vector Space Model calculates the cosine similarity of each document that was found to match the user query was calculated while Boolean model retrieved just the book title and the author name only.

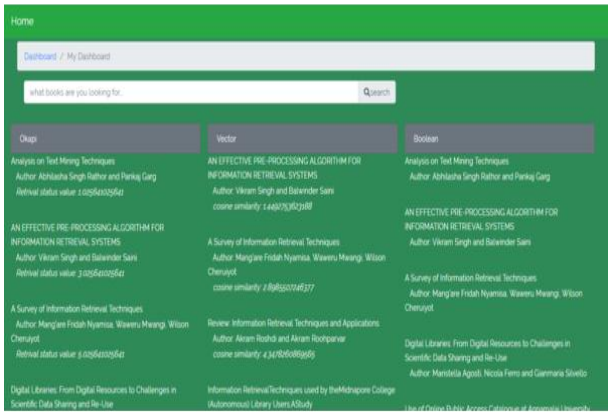

Figure 10 Retrieved documents

4.2 Tests for Performance

Evaluation of the OPAC information retrieval framework performance was done and tested using precision and recall. The Boolean model was left out because it does not retrieve relevant results to the user. Vector space model and Okapi BM25 were tested to proclaim the best model among the two since their retrieval was based on relevancy. This was done using the sample of three queries that was implied on the two models at the same time. The improved digital library's OPAC catalog allowed the users to search the catalog and sort the results by relevance ranking using the three models where the most relevant results are displayed at the top of the page. Precision is the fraction of relevant results retrieved from the total number of documents stored in the library database to meet the information need of the user. Zuva & Zuka [28], pointed out that poor performance of the models displays low values while high performance of the models results with high values. This can be calculated as shown in equation 5

$$Precision = \frac{(relevant\ items\ retrieved)}{(retrieved\ items)} \quad (5)$$

Recall denotes the fraction of the relevant documents in the collection returned by the system for use. This can be calculated using the recall formula as shown in equation 6

$$Recall = \frac{(relevant\ items\ retrieved)}{(relevant\ items)} \quad (6)$$

Precision and recall calculation for query 1: Information Systems

Table 3 Query 1- Information technology

Name of the model	Retrieved documents	Relevant documents	Precision results (%)	Recall results (%)
Okapi BM25 Model	15	11	62.2%	78.57%
Vector space model	8	3	34.4%	21.43%

Precision and recall for query 2: data mining in the digital libraries today

Table 4 Query 2- data mining in the digital libraries today

Name of the model	Retrieved documents	Relevant retrieved documents	Precision results (%)	Recall results (%)
Okapi BM model	9	6	64.29%	66.67%
Vector space model	5	3	35.71%	33.33%

Precision and recall calculation for query 3: Challenges facing the digital libraries especially in information retrieval

Table 5 Query 3- Challenges facing the digital libraries especially in information retrieval

Name of the model	Retrieved documents	Relevant retrieved documents	Precision results (%)	Recall results (%)
Okapi BM model	12	9	54.55%	56.25%
Vector space model	10	7	45.55%	43.75%

5. CONCLUSIONS

This paper's literature review exposes a vocal dissent on the use of OPAC in many digital libraries, especially with its complex search mechanisms. Although recent developments of the search capability of the OPAC have been enhanced, still OPAC is criticized for lack of relevance ranking in its search capability [16]. This paper concludes that Okapi BM 25 model can be used in information retrieval in the digital library's OPAC catalogue. A term with a high relative frequency within a document is more representative and relevant in the document characterization and ranking. Based on this research and analysis, the Okapi BM25 model is proposed to reward terms according to their relative frequencies in a document. From the results obtained, it is clear that the Okapi BM25 model which is integrated with relative term frequency information, document length normalization and tuning parameters significantly outperforms the Boolean Model and Vector Space Model on most of the representative data collections. It is a novel approach to combine the concept of relative term frequency with fundamental weighting functions in probabilistic information retrieval systems to increase performance of the model for retrieval results in the OPAC. The OPAC framework is accurate and applicable according to specified requirements.

6. REFERENCES

1. Antelman, K., Lynema, E., & Pace, A. K. (2006). Toward a Twenty-First Century Catalog. *INFORMATION TECHNOLOGY AND LIBRARIES*, *25*(3), 128-139.

2. Yulianto, B., Budiharto, W., & Kartowisastro, H, I. (2017). The Performance of Boolean Retrieval and Vector Space Model in Textual Information Retrieval. *Communication & Information Technology*, 11(1), 33-39

3. Aruleba, K. D., Akomolafe, D. T., & Afeni, B. (2016). A Full Text Retrieval System in a Digital Library Environment. *Scientific Research Publishing*, *8*(1), 1-8.

4. Boubekeur, F., & Azzoug, W. (2013). CONCEPT-BASED INDEXING IN TEXT INFORMATION RETRIEVAL. *International Journal of Computer Science & Information Technology (IJCSIT)*, *5*(1), 119-136.

5. Brahaj, A., Razum, M., & Hoxha, J. (2013). Defining Digital Library. In T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, & C. J. Farrugia, *Research and Advanced Technology for Digital Libraries.* (pp. 23-28). Berlin: Springer, Berlin, Heidelberg.

6. Dwivedi, S. J. (2014). Comparative Analysis of IDF Methods to Determine Word Relevance in Web Document. *International Journal of Computer Science Issues*, *11*, 59-65.

7. Ibba, S., & Pani, F. E. (2016, May 10). Digital Libraries: The Challenge of Integrating Instagram with a Taxonomy for Content Management. *Future Internet*, pp. 1-15.

8. Khiste, G. P., Deshmukh, R. K., & Awate, A. P. (2018, Feb 24). LITERATURE AUDIT OF 'DIGITAL LIBRARY': AN OVERVIEW. *Research gate*, pp. 403-411.

9. Kumar, S., & Vohra, R. (2013). "User perception and use of OPAC: a comparison of three universities in the Punjab region of India". *The Electronic Library*, *31*(1), 36-54.

10. Mishra, R. K. (2016). DIGITAL LIBRARIES: DEFINITIONS, ISSUES, AND CHALLENGES. *Innovare Journal of Education*, *4*(3), 1-3.

11. O'Connell, J. (2008). Information Literacy meets Library 2.0. In P. J, & G. P, *School library 2.0: new skills, new knowledge, new futures* (pp. 51-62). London: Facet Publishing.

12. Ogbole, J. O., & Morayo, A. (2017). Factors Affecting Online Public Access Catalogue Provision And Sustainable Use By Undergraduates In Two Selected University Libraries In Ogun And Oyo States, Nigeria. *Journal of Research & Method in Education*, *7*(4), 14-25.

13. Roshdi, A., & Roohparvar, A. (2015). Review: Information Retrieval Techniques and Applications. *International Journal of Computer Networks and Communications Security*, *3*(9), 373-377.

14. Ruban, s., Sam, S. B., Serrao, L. V., & Harshitha. (2015). A Study and Analysis of Information Retrieval Models. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(7), 230-236.

15. Rybchak, Z., & Basystiuk, O. (2017). Analysis of methods and means of text mining. *ECONTECHMOD. AN INTERNATIONAL QUARTERLY JOURNAL*, *6*(2), 73-78.

16. Sankari, R. L., Chinnasamy, K., Balasubramanian, P., & Muthuraj, R. (2013). A STUDY ON THE USE OF ONLINE PUBLIC ACCESS CATALOGUE (OPAC) BY STUDENTS AND FACULTY MEMBERS OF UNNAMALAI INSTITUTE OF TECHNOLOGY IN KOVILPATTI (TAMIL NADU). *International Journal of Library and Information Studies*, *3*(1), 17-26.

17. Sharma, M., & Patel, R. (2013). "A survey on information retrieval models, techniques and applications," *International Journal of Emerging Technology and Advanced Engineering*, 3(11), 542–545.

18. Shiva Kanaujia, S., & Parveen, B. (2016). Marketing and Building Relations in Digital Academic Library: Overview of Central Library, Jawaharlal Nehru University, New Delhi. *DESIDOC Journal of Library & Information Technology*, *36*(3), 143-147.

19. Sundari, G. J., & Sundar, D. (2017). A Study of Various Text Mining Techniques. *International Journal of Advanced Networking & Applications (IJANA*, *08*(05), 82-85.

20. Swaminathan, K. S. (2017). Use and Awareness of Online Public Access Catalogue (OPAC) by Students and Faculty members of Anna University Regional Campus, Coimbatore, Tamil Nadu – A Case Study. *International Journal of Scientific Research and Management*, *5*(5), 5345-5349.

21. Talib, R., Hanif, K. M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414-418.

22. Garcia, E. (2016, November 11). *A Tutorial on the BM25F Model - Minerazzi.* Retrieved June 18, 2017, from minerazzi.com: https://www.researchgate.net/publication/308991534_A_Tutorial_on_the_BM25F_Model

23. Muhammad, A. B. (2017). Efficiency of Boolean Search strings for Information Retrieval. *American Journal of Engineering Research (AJER.* 6(11), 216-222

24. Singh, V. K., & Singh, V. K. (2015). VECTOR SPACE MODEL: AN INFORMATION

RETRIEVAL. *International Journal of Advanced Engineering Research and Studies*, 141-143.

25. Tuna, G., Zogo, R., & Dermirelli., B. (2013). An Introduction to Digitization Projects Conducted by Public Libraries: Digitization and Optimization Techniques *Journal of Balkan Libraries Union*, 1(1), 28-30.

26. Zhu, R. (2016, June 5). GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY. *IMPROVEMENT IN PROBABILISTIC INFORMATION RETRIEVAL MODEL REWARDING TERMS WITH HIGH RELATIVE TERM FREQUENCY*, pp. 1-95.

27. Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications, 85*(17), 42-45.

28. Zuva, K.., & Zuva, T. (2012). Evaluation of Information Retieval Systems. *International Journal of Computer Science & Information Technology*, 4(3), 35-43.

29. Jain, A., Jain, A., Chauhan, N., Singh, V., & Thakur, N. (2017). Information Retrieval using Cosine and Jaccard Similarity Measures in Vector Space Model. *International Journal of Computer Applications , 164* (6), 28-30.