International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 01 Jan 2024, pp: 463-475 www.ijaem.net ISSN: 2395-5252

# A Recommender Search Engine for Improved Retrieval of Books in a Digital Library Using User's Preferences

Caroline. N. Asogwa[1], Francis .S. Bakpo[2], Emmanuel.C. Ukekwe[3], Helen.C. Ugwuishiwu[4]
*Department of computer science, university of nigeria, nsukka*

Date of Submission: 20-01-2024 Date of Acceptance: 30-01-2024

**ABSTRACT**
Most digital libraries make use of traditional search engines that work with google PageRank (PR) technology when retrieving books. However, the approach does not consider specific user preferences in its search queries. This study presents an improved recommender search engine (RSE) that takes cognisance of both search phrase and library user's book preferences/interests to recommender and retrieve the appropriate book off the database of a digital library. An empirical justification of the model was carried out using the popular “goodreads_bbe_dataset[1] from Github. A principal component analysis (PCA) carried out on the data identified publication year, cost price of books and popularity as the significant user preferences required for the RSE model. Subsequently, the K-nearest neighbour algorithm was employed to predict a library user's priority preference based on the PCA's result. An accuracy of 97% was recorded in prediction of user's preference. The model could be adopted and implemented as a search engine for book retrieval in libraries.

**Keywords:** Recommender, user preferences, digital, intelligent, search engine, PageRank

## I. INTRODUCTION

The term “Digital Library” could be referred to as electronic library, virtual library, library without walls. The definition of Digital Library depends on how we think about them and how much information we wish to store in them [1]. According to [2], Digital Libraries basically store materials in electronic format and manipulate large collections of those materials effectively. It is an ITC based library management system where users of library services can find information related to reading books or other documents in digital form [3]. The ability to store materials in electronic copy is one thing and the ability to retrieve the stored materials for the required purpose is another. Digital Libraries are known to be a repository for electronic materials especially books. No matter the conceptual definition given to it, Digital Libraries are just online libraries that provide the functionalities of a normal physical library without being limited by space and time. Digital materials in a Digital Library are usually accessible online through the Library’s host website. Host websites are designed to allow new users register and become part of a known community that makes use of the library materials. Borrowing of materials, reading and even outright purchasing can be seen as part of the Digital Library services. Some Digital Libraries may require a user to pay a token for borrowing a softcopy of their book while some lend it out for free. Digital Libraries are known to habour several research assisted materials in form of books, magazines, video and audio clips, journal articles and lots more. Hence, Digital Libraries can be classified based on the nature of materials they store. Popular known Digital Libraries include; NASA Multimedia, Smithsonian Images (database), New York Public Library Digital Gallery, Google Books, Universal Digital Library (Million book collection), Digital Video Library, GEM - Gateway to Educational Materials. Several others could be found here[2].

Basically, Digital Libraries have been found to be a major source of ideas for research,

[1] http://doi.org/10.5281/zenodo.4265096

[2] https://www.csc.lsu.edu/~wuyj/Teaching/7410/sp16/hw/ExampleDLs.html

history, culture and education in general. After all, without proper information, research breakthroughs will be difficult [4]. However, one major challenge that seems to bedevil Digital Libraries is the process of information retrieval from their storage. The authors in [5] defines Digital Library in terms of computing as an assemblage of digital computing, storage and communications machinery together with the software needed to reproduce as well as retrieve the stored information within it. The success of Digital Library depends to a large extent on the robustness of the search engine. The ability of users to retrieve desired information relating to their need is paramount. Unfortunately, most known Digital Libraries implement their search engine using the traditional PageRank (PR) algorithm employed by Google. In this traditional based algorithm, emphasis is usually on the search query entered and how many times the words occurred in the related materials within the database. According to [6], the way in which the PR algorithm is used to display web pages in a search engine is not a mystery but involves applied math and good knowledge of computing. Several approaches have been suggested as a better way of computing the PR algorithm. A new local randomized algorithm for approximating personalized PR was propounded in [7], an improved PR algorithm that computes the Page Rank values of the Web pages correctly with a consistent total sum of 1 was developed in [8], a new technique based on graph theory in [9] was also developed. These algorithms are all aimed at improving the PR for effective search and retrieval. However, they do not change the fact that users of the Digital Library need more than a listing of related books or materials based on their search query. Unfortunately, the traditional approach to search engines seems to stop at listing search results related to the search query alone.

Apart from looking for books in a given search area (which is where the traditional search engines stops), when people search for books in a Digital Library, they have a lot of other things in mind. They may be interested in the year of publication of such book, they may also be interested in the number of pages, how good the book is rated by other users with similar interest, how popular the book is among readers and even how affordable the book is assuming they have to purchase it. Foundational historical books would naturally have an older year of publication than emerging concepts. Some people prefer not to be bored with much detail, hence they prefer books with fewer pages while some others would prefer to go for books having very high rating and high popularity. The traditional PR search engine does not make provision to consider user preferences in their search. Hence there is need to ascertain these from onset and incorporate it in the search engine using a recommender algorithm.

Recommender systems have gained a lot of recognition in recent times. Recommender systems represent tools that aid the user to gain a personalized view of a situation thereby prioritizing things likely to be of interest to the user [10]. Recommender Systems are designed to generate meaningful suggestions to users for items or products that might interest them [11] According to the authors in [12], recommender systems can be applied to assist users to locate the information and products they desire so much to find. Hence, recommender systems can also be applied to find the appropriate books for users of digital libraries.

There are several book recommender systems with each having almost the same or different implementations. The authors in [13] developed a book recommender system using users genre interest. The authors in [14] developed a book recommender system by taking a deeper look at the table of contents which is expected to give the user a better reason for choosing the book. A book shopping recommender system was also developed using the choices of both similar and different users to make a wider recommendation for users [15]. Features such as content filtering, collaborative filtering and association rule mining were employed by [16],[17] to produce efficient and effective recommendations for book purchase. The authors in [18], made use of user's profile to provide highly personalized book recommendations by employing machine learning approach. [19] enumerated several recommender systems for books such as Quambo for DSpace, Matchbook for OpenAIRE, PubMedReco for PubMed and others.

However, most of the recommender systems seem to focus on recommending the right item to users using established user profile and preferences. In real life, the of a Digital Library user are dynamic. A user may be interested in a cross disciplinary research which will be a deviation from the profile norm. Hence having a fixed user profile as a determining factor upon which prediction is based may be misleading.

In this paper, a different approach is proffered. People have different motivations for seeking for a book. Some want older books that have foundational definitions, some want books with few detail and pages while some prefer books that have been rated as being of high quality. Others prefer books that are popular among readers or books that the cost is of specific range. Satisfying these motivations and interest are our main focus in this work. The system accepts the immediate preferences of a user and a recommendation is given specifically to improve the search engine results thereby allowing the user to view the ordered list of related search based on the most important user need determined by the KNN algorithm. The algorithm itself being a non-parametric algorithm is a Machine Learning algorithm which is based on Supervised Learning technique. It has been used in studying economic events and prediction of companies in distress [20]. It has also been used for predicting network data card sales between three competing network providers in [21], accuracy of product delivery in [22] and in other areas for predictive purposes. Hence, it will be used in this study to predict the most important user need upon which the search engine is to focus its result ordering.

## II. METHODS

The dataset, conceptualization, the model and other necessary concepts needed for the study are presented in this section.

### 2.1 Conceptualisation

Some characteristics of digital libraries, according to [Irsa Arma] include; use a computer to manage, use of electronic channels to connect information providers with information users, use of electronic transactions. and use of electronic facilities to store, manage, and convey information to users.

Quick and timely Information retrieval characterises a good digital library. Hence, the expectation of every library user is that a search query is typed and a search is initiated across the library database for similarities in the typed word, sentence or phrase. Information retrieval in a digital library is heavily supported and dependent on the computer skill as well as hardware support. The discourse here assumes that every information retrieval will be done across online platforms using a computer device. According to the authors in [23], the huge volume of information within the library database necessitates the use of reliable information retrieval methods in which computers play a major role. It is even recommended that for strategic retrieval of library information, the application of computer knowledge skills is required [24]. Computers are therefore essential in organisation and retrieval of information in a digital library. According to the author in [25] , a library user can search for a book via a computer using any or all of the following attribues:

i. By Author
ii. By Title
iii. By Subject
iv. By edition
v. By character and so on.

In this case, the term "attributes" will refer to the library user preferences. Using these annotations, the International Federation of Library Associations and Institutions (IFLA) modified the objective of a good information retrieval method to:

i. To find entities that corresponds to the user’s search criteria
ii. To identify an entity
iii. To select an entity that is appropriate to the user’s preferences
iv. To enquire or obtain access to the entity described.

Hence, it has become imperative to put into cognisance user's preferences and interest while retrieving library information queries. The following represents typical user preferences for a book search in a digital library.

i. Book title
ii. Year of publication
iii. Number of pages
iv. Cost/subscription of hardcopy
v. Book Rating
vi. Category

Based on this, we define a library retrieval information query as an 2-tupule given as:

query=[A, B]......................................................(1)

where A represents the usual search word, sentence or phrase. B is a list of user preferences which lies between 1 to n. B is dependent on the database capacity and classification metadata present in a given library.

The efficiency, speed and quality of information retrieval for a given digital library therefore depends so much on B. For this reason, RSE model will focus on B aims to optimize B using the principal component analysis, predicting the prioritized B using the KNN algorithm and sorting the retrieved search results based on the prioritized B. The details are presented the following sections.

#### 2.1.1 Selecting the appropriate user preferences for a search query

The number of user preferences required for the RSE model for a given library is computed

using the principal component analysis (PCA) approach for reduction of factors. The PCA is a familiar data analysis statistics that is used for dimensionality reduction. Simply put, the PCA reduces the number of variables for a research to a manageable few without losing much information. Hence, for an n number of B user preferences, the PCA will definitely reduce the preferences to n-i remaining preferences, where i =1,2,3...,n-1.
The steps involved in carrying out a principal component analysis as listed by the authors in [26] is described as follows:

a) Standardizing the range of continuous initial variables required for the PCA
b) Compute the covariance matrix to identify correlations between the variables
c) Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
d) Create a feature vector to decide which principal components are significant
e) Recast the data along the principal components axes

Hence, for a given library, selecting all the user preferences in the database may result to a clumsy, slow and inefficient search. On the other hand, a random selection of user need may also result to in efficient search and subsequent prediction. The PCA is therefore a major tool for the RSE model.

### 2.1.2 Predicting the prioritized user need for information retrieval

Determining and selecting the appropriate user preferences to be used as search criteria is one step while another is being able to make use of the selected user preferences to determine the priority need upon which the search is to be subsequently arranged. A suitable tool for such prediction is the K-nearest neighbour algorithm. KNN algorithm stores available observations and uses it to classifier a new one depending on how they are related to existing ones. The KNN algorithm is a popular algorithm for prediction in [27], [28], [29]. In this work, the KNN algorithm was also employed to predict the right class for a new search query based on the user need parameters entered by the library user. The algorithm thus predicts the most significant user need upon which the final output will be presented.

### 2.1.3 Presenting the search result of RSE model

The method of displaying the results of a search query is equally important to the library user. Before initiating a search, the library user already has some expectation based on the preferences for reading/requesting for such book. The user thus may not be interested in the book that has the highest similarity to the search query entered but he/she will be interested on the books that relate to the preferences in mind. Unlike other information retrieval search engines which displays search results based on the similarity index ratio observed between related items, the RSE model rather displays results sorted in the order of the most significant user need. The KNN algorithm predicts the most significant user need. It is upon this that the information displayed on the screen for the library user is based. After all, the user is not interested in the most similar result for his/her search, but on the need at hand for which the search was initiated.

## 2.2 The RSE model

Based on the concepts described, the RSE model is made up of 5 distinct modules which work together to produce the RSE improved result. The modules are;

a) Parameter input:- This module allows users to enter the search query as well as the user preferences. The user preferences are real numbers which is automatically scaled using equation 1.
b) Database:- The database stores the books and related parameters. The search query searches for books with given search criteria and locates them in the database.
c) Decision engine:- This module is the intelligent section of the system which takes n input parameters (user's preferences examples include; Rt, Pg, Py, Pt, Pr) and produces a single output classification based on KNN algorithm. The output computed depends on the weight of interest entered for the preferences.
d) Sort engine:- After the search results are extracted from the database, the sort engine sorts the extracted results based on the output of the decision engine. The difference between the PR mode of ranking and the RSE is established here. While the PR mode ranks output based on the computed ranks, the RSE is interested in ranking based on the most important user preference entered.
e) Display search outcome:- This module is responsible for displaying the search results

and associated links in a search page. The search page is made easier for the user and is sorted in order of priority based on the classification output.

A conceptual model is presented as shown in Figure 1.

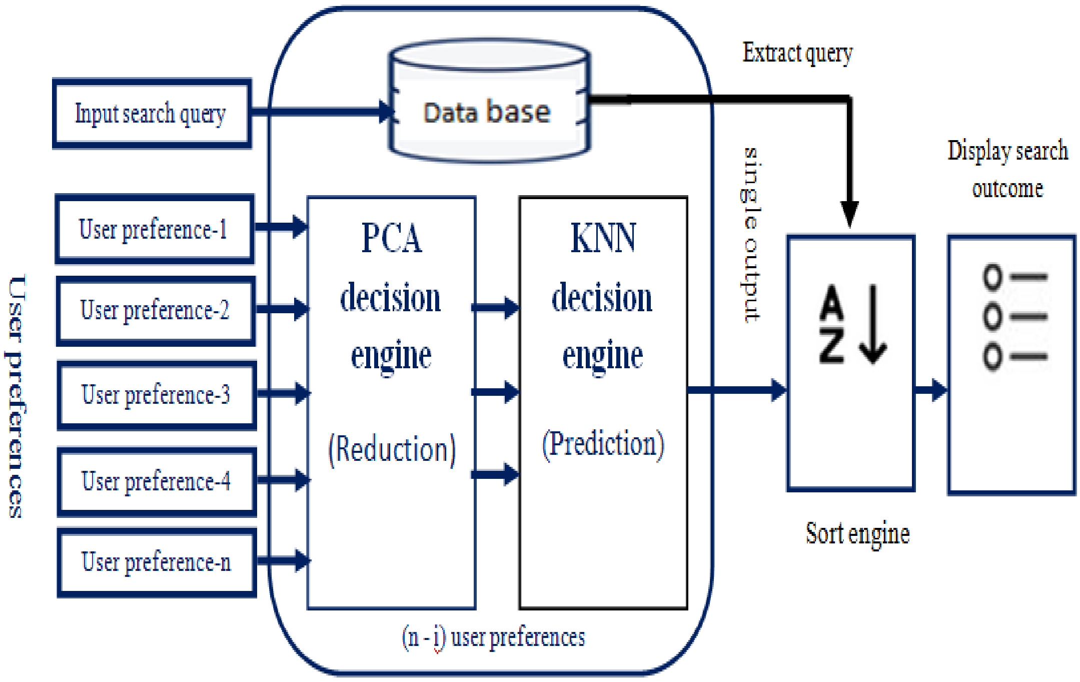

**Figure 1: Conceptual RSE model**

Figure 1 shows the conceptual model of the improved search engine for recommending books based on user preferences. The system allows users to enter a search query and the associated user preferences. These preferences reflects the user's interest, capability and need. Hence, it allows novice readers to limit their search to books with fewer pages, it allows those looking for foundational historical books to obtain search results based on older years of publication, it allows those who prefer books that have higher or lower ratings to make their choice and it also gives room for those who want affordable books which have a high percentage of popularity to make their selection. The search query is the usual search criteria common with traditional search engines. When it is entered, the search query is routed to the database to retrieve the required query. On the other hand, the user preferences values entered are sent to the decision engine the PCA and KNN respectively. The PCA reduces the number of preferences entered to only the significant ones. After the initial reduction, the KNN then analyzes the various input and comes up with a priority single output classification for the search. The sort engine thus orders the output retrieved from the database based on the decision engine's classification. The result is then presented to the user with emphasis on the preference that matters most and not on the Page rank as in the traditional approach.

### 2.3 Dataset and variables

In order to empirically justify the model, the popular goodreads dataset was used. The dataset is originally a huge set of 52,478 records and 25 fields. However, some of the fields were not relevant to this study and some were also not having complete records. A data cleaning exercise was carried out on the data for the purpose of extracting only the necessary fields with complete records. At the end, a total of 5 fields each having a total of 35,963 records were extracted and used for the study. The fields and interpretation are presented in Table 1 as shown. The interpretation of the variables were based on [30].

**Table1: Selected variables and meaning**

| S/no | Field attribute | Meaning |
|---|---|---|
| 1 | Rating (Rt) | Global goodreads rating |
| 2 | Pages (Pg) | Number of pages |
| 3 | PublishYear (Py) | Publication date |
| 4 | Popularity (Pt) | Popularity of book among redaers |
| 5 | Price (Pc) | Book's price |
| 6 | Rating rank **(derived)** | Rank of individual rating scores |
| 7 | Pages rank **(derived)** | Rank of individual number of pages |
| 8 | PublishDate ranking **(derived)** | Rank of individual date of publication |
| 9 | LikedPercentage rank **(derived)** | Rank of individual percentage of likeness |
| 10 | Price rank **(derived)** | Rank of individual book price |
| 11 | Outcome **(derived)** | Classification for each record |

The derived variables represent the additional variables needed for the study.
Hence, application of equation 1 to 3 will produce the data layout as shown Table 2.

**Table 2: Layout of the Extracted Data**

| S/no | Rating | Pages | PublishYear | **LikedPercentage** | Price | Outcome |
|---|---|---|---|---|---|---|
| 1 | $Rt_i$ | $Pg_i$ | $Py_i$ | $Pt_i$ | $Pc_i$ | $Co_i$ |
| 2 | - | - | - | - | - | - |
| n | $Rt_n$ | $Pg_n$ | $Py_n$ | $Pt_n$ | $Pc_n$ | $Co_n$ |

### 2.4 Model specification

In this section, we define some variables and subsequent equations associated with the work. Firstly, the variables are of different scale and there will be need to convert all the variables to a common scale thereby normalizing the data appropriately. In this work, all the variables were scaled between 1-100 percent. Again, some variables were derived for the purpose of ascertaining the classification outcome for each record. Being that the dataset did not originally come with a classification which is prerequisite for KNN algorithm, the classification field was therefore computed also.

a) The scaling factor ($S_f$):- This represents the scaling interval used on the data.

$S\_f = (b - a)\ (S - \min(S))/(\max(S) - \min(S)\ ) + a$ ……………………………….(2)

Where, a = lower range limit, b = upper range limit, S = variable to be scaled.

b) The ranking factor ($R_f$):- This represents the ranking criteria used for determining the classification outcome for a record. This field is needed in order to employ the KNN algorithm. Hence,

$R_f = \max(Rt_i, Pg_i, Py_i, Pt_i\ Pc_i)\ , 1 \leq i \leq n$…………………………………………...(3)

Where, Rt, Pg and others have their uual meaning as denoted in Table 1.

c) The classification outcome ($C_o$):- The classification outcome denotes the computed outcome for each record and it is computed in consideration of the rank weight of all the variables. The smallest rank is designated as the outcome. Hence,

$C_o = \min(R_{f(i)}), 1 \leq i \leq n$…………………………….............(4)

d) The Euclidean distance for finding the K-th minimum distance is given as;

$d(x, y) = \sqrt{\sum_{i=1}^{m}(x_1 - y_1)^2}$…………………………………………………………..…(5)

where $x_1$ and $y_1$ represent the distance between observations.

## III. DATA ANALYSIS

### 3.1 Data Preparation

Some of the fields in the goodreads dataset were reported as being incomplete [30] necessitating data cleaning. Some missing values, special characters and incomplete records were removed. The initial extracted data comprises of different scales. The rating (Rt) variable was on a scale of 1-5, the book pages (Pg) ranged from 1 to 3848. The publication year (Py) ranged from 1900-2021, the popularity (Pt) rating was on a scale of 27-100 percent while the price ranged from 0.84-898.64 dollars. In addition, the extracted data does not have a classification field which is a necessary for using KNN. In order to scale the variables to be of the same scale, equation 1 was used ($S_f$). After obtaining the scaled data, the rank ($R_f$) for each variable were also obtained for every record in the dataset. Subsequently, the maximum rank for each record is determined and the variable having the first position was also recorded using equation (2). Finally, the classification ($C_o$) for each record were now obtained using equation 3. The data was then modeled using KNN algorithm.

### 3.2 Analysis

The 35,963 extracted records were used to test the efficacy of the RSE model. Firstly, a Principal component analysis (PCA) was carried out on the data to ascertain the factors that would be most appropriate for the model with minimum loss of information. Out of the initial five (5) variables, only three (3) were extracted from the PCA using Varimax rotation with Kaiser normalization. The extracted variables include; Year of publication (Py), Popularity of book (Pt) and Price of book (Pr). Based on this, further analysis focused on the three variables instead of the initial five (5). The appropriate value of k for the model was determined. Several methods has been suggested for finding the appropriate value of k for a KNN algorithm. The authors in [31] proposed a novel method for finding k using cross validation approach. The authors in [32] also proposed the method of cross validation to determine the appropriate k. In [33], the authors proposed the use of expectation Maximation (EM) algorithm while others approach the problem of identifying the appropriate k from the perspective of the dataset in question [34]. However, in this work, a visual approach was used. The approach involves plotting different values of k and the corresponding error term as suggested in [35]. Having identified the appropriate k for the model, the data was divided into 75:25 ratio with 75% serving as train data while 25% served as the test data. The KNN algorithm was thus employed to model the book data. The accuracy of the model was computed and the confusion matrix also computed. The analysis was done using Python programming.

### 3.3 RSE Implementation algorithm

The implementation algorithm defines the programming steps required for developing the RSE system. The algorithm is as follows:

```
1.  Input Section
1.1 Input s_query // Search query [data type-string]
1.2 Input Py // prefered publication year [data type-interger (1900-2021)]
1.3 Input Pt // prefered popularity [data type integer (27-100)]
1.4 Input Pr // prefered book price [data type real number (0.84- 898.64)]
2.  Retrieval Section
2.1 for (i=1,i++,i<=number of extractions){
2.2     $\mathrm{database(S\_query)} \xrightarrow{\mathrm{yields}} \mathrm{R\_query(i)}$ // Retreive search query from database [datatype-string]
2.3 }
3.  Decision section using KNN
3.1 Determine the parameter K which is the number of nearest neighbours
3.2 Calculate the distance between the query-instance ( Py, Pt, Pr) and all the training samples
3.3 Sort the distance and determine nearest neighbours based on the k-th minimum distance
    // $\mathrm{d(x,y)} = \sqrt{\sum_{i=1}^{m}(x_1 - y_1)^2}$
3.4 Gather the categories of the neatest neighbours
3.5 Use simple majority of the category of nearest neighbours, a single prediction output $C_o$ is
        produced
4.  Sort section
4.1 $\mathrm{R\_query} \xrightarrow{\mathrm{sort\ on\ Co}} \mathrm{SR\_query}$ // Produce sorted retrieved query results
5.
    $\mathrm{SR\_query} \xrightarrow{\mathrm{Co\ ascending}}$ Display search page and links
    // Display sorted retrived queries based on $C_0$
        output
```

### 3.3 RSE Implementation architecture

The RSE can be implemented on the Internet framework of any digital library based on Figure 2 as shown.

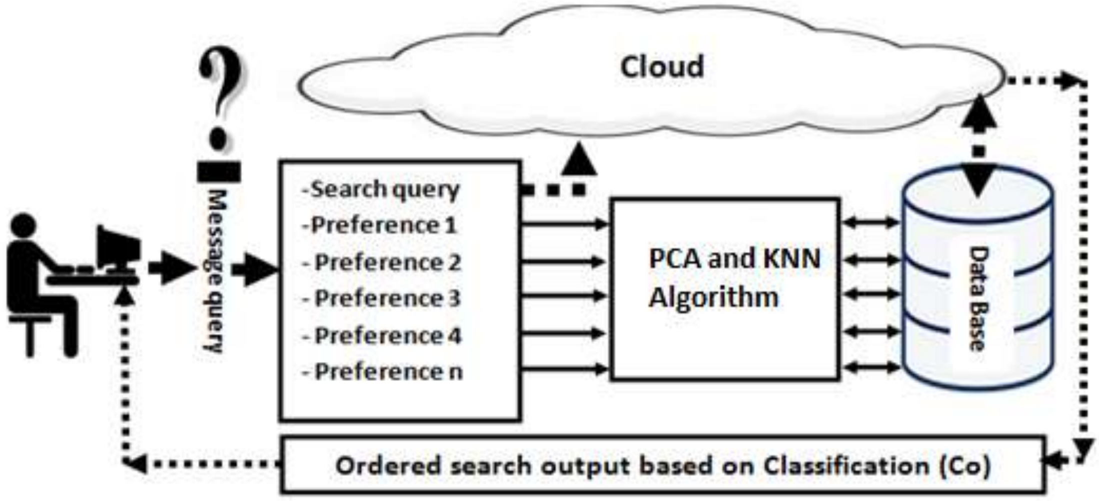

**Figure 2: Implementation architecture of RSE**

Figure 2 shows the implementation architecture of the RSE system. The user enters a search query alongside the preferences. The search query goes to the Internet based database for retrieval while the preferences form a query instance for the KNN algorithm. The query instance is an n-object input which culminates to a single output $C_o$ after the KNN algorithm. The retrieved query is then sorted based on the output of the KNN algorithm and presented in a web page to the user.

## IV. RESULTS

The results of the PCA test carried out on the data is presented in Table 3 as shown.

**Table 3: Component Score matrix of the PCA**

| Preferences | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Rating | .518 | .003 | .028 |
| Pages | -.003 | .571 | .318 |
| **PubYear** | -.013 | -.016 | **.911** |
| **Popularity** | **.533** | -.083 | -.037 |
| **Price** | -.083 | **.794** | -.226 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

The highlighted preferences are the ones extracted from the PCA and they represent the variables having the highest component score per column. For column 1, we have component score of 0.533 corresponding to Popularity, for column 2, we have component score of 0.794 corresponding to Price and for column 3 we have 0.911 corresponding to Publication year. The implication of this is that if any other variable which is not part of the extracted ones is used for the model, it will definitely reduce the accuracy of the model. Again, the extracted variables represent the variables that could be used without much information loss on the whole data. The scree plot and how they were selected is also shown in Figure 3.

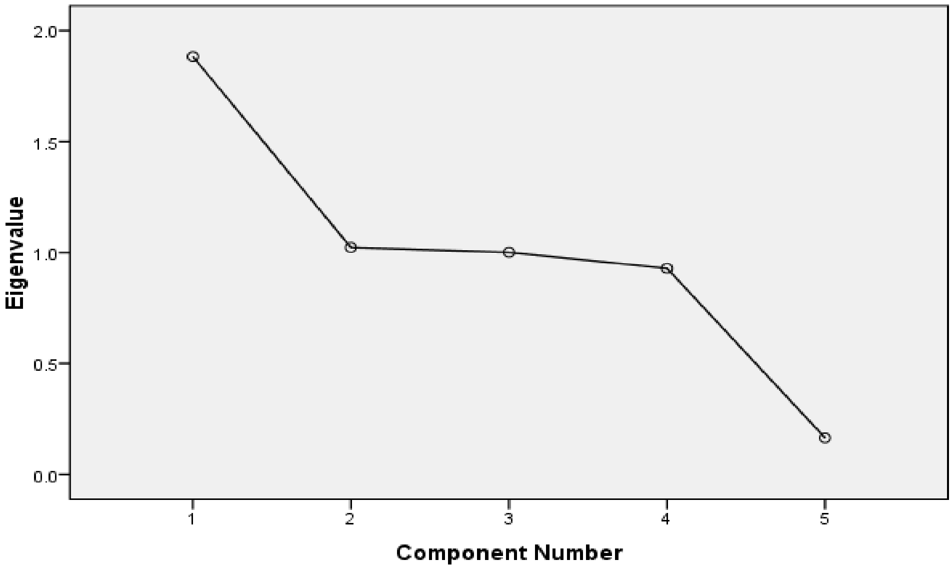

**Figure 3: Scree Plot of component selection**

From Figure 3, the PCA simply extracted all the variables having eigenvalue greater than or equal to 1. Based on the figure, there are only 3 variables that crossed the range, hence they were extracted.

On the same note, the 3-dimension scatter plot of the 3 variables were plotted to picture the trend and relationship as shown in Figure 4.

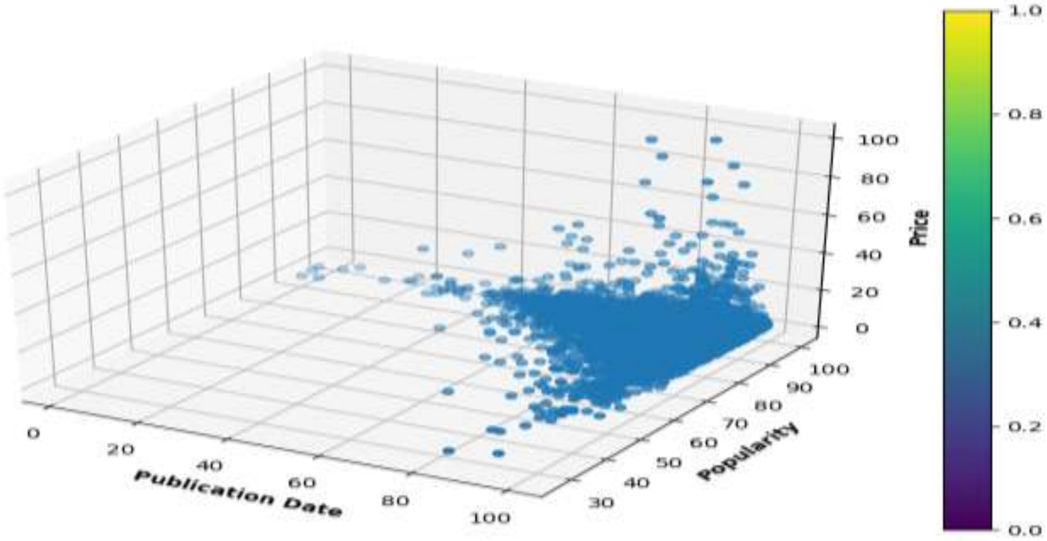

**Figure 4: Scatter plot of the data**

From Figure 4, the plot obviously showed some positive trend indicating some level of correlation.

Based on this, we proceeded to model the data using the KNN algorithm. Firstly, the appropriate k was determined by plotting the different k against the error term as shown in Figure 5.

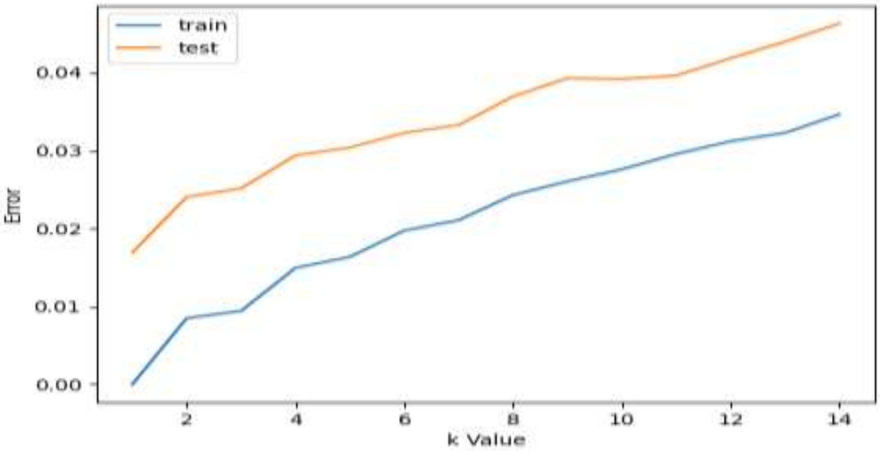

**Figure 5: Determining appropriate k**

From Figure 5, test error term begins to reduce from k = 2 to 3. At k = 3, it begins to rise again. The same applies to the train error term. Since our objective is to find a balance for k between the test and train data, we select k = 3 where such balance occurred. Another good k will be 5, however, we chose k = 3 where the balance first occurred.

Hence using k = 3, the model was computed and the results are shown in Table 4.

**Table 4: Summary of the KNN decision engine**

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **Popularity** | 0.96 | 0.99 | 0.98 | 2794 |
| **Price** | 0.98 | 0.95 | 0.97 | 3157 |
| **Publishdate** | 0.98 | 0.99 | 0.98 | 3048 |
| | | | | |
| **Accuracy** | | | 0.97 | 8991 |
| **Macro avg** | 0.97 | 0.98 | 0.97 | 8991 |
| **Weighted avg** | 0.98 | 0.97 | 0.97 | 8991 |

From Table 4, it could be seen that the model recorded a prediction accuracy of 0.97 with k = 3. It predicted the Popularity category with a precision of 0.96 while Price and Publishdate both recorded a precision of 0.98 based on the support values which were correctly predicted to be true.

Further test was carried out to test the performance metric of the model in predicting the required user preference, a confusion matrix was plotted for that purpose as shown in Figure 6

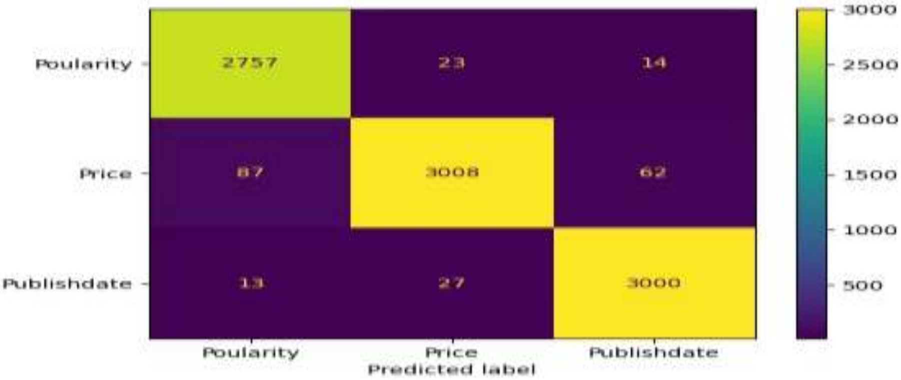

**Figure 6: Testing the efficacy of the decision engine**

A confusion matrix can be used to measure the performance metric of a model when it is defined in terms of true positive (tp) which indicates the number of positive examples classified accurately, true negatives (tn) which shows the number of negative examples classified accurately, false positive (fp) which represents the number of actual negative examples erroneously classified as positive and false negative (fn) which is the number of actual positive examples erroneously classified as negative [36]. Based on this, it could be ascertained from Figure 6 that 2,757 observations out of a total of 2794 were predicted correctly for user preference Popularity. Out of a total of 3159, 3008 was correctly predicted for user preference Price and out of 3040, a total of 3000 was predicted correctly. On the overall, the model recorded a 0.97 accuracy in prediction.

## V. IMPLICATION OF FINDINGS

Based on the results, it is obvious that for any digital library, the most significant user preferences can be determined and subsequently used for the RSE modeling for that library. Specifically, the empirical data used in this work (“goodreads_bbe_dataset") show that the preferences that most users of digital libraries consider when they search for books include the year of publication, the popularity of the book and the price of such book. Understandably, a digital library user will have at the back of his mind that if he must buy a book off the shelves of a library, the price will be a determining factor. Again, most researchers consider the year of publication when they search for books. Current publications will naturally represent current innovations. However, a researcher may once in a while be interested in historical or foundational books that defines a concept. In such cases, the user’s preference will be on books published earlier. It can also be ascertained that surprisingly, users of digital library seem to be more interested in how popular a book is among the readers than how high the book is rated. It was observed that book rating was not extracted from the PCA. It was also observed that people do not usually consider the number of pages in books before searching for them.

In general, the RSE model focuses more on the user need of the reader than just finding a list of related books based on search query. Being that the model recorded a very high accuracy, it gives credence to the work as a means of improving a book search based on user preferences. Hence, any digital library can adopt the model and build it into their search engine in order to provide user satisfaction.

## VI. CONCLUSION

In this work, we have clearly revealed the limitation of a traditional search engine using PageRank in searching for books in a digital library. An improved search engine using RSE model was thus proffered. The model was further tested using empirical library data. Hence, the major user preferences for digital library users were identified to be Year of publication, Popularity of books and the cost Price of the books. However, the user preference for individual digital library may differ and can easily be determined and subsequently used to implement the RSE model for such library. The accuracy of the model on the empirical data used also gives credence to it. Hence, an algorithm and implementation architecture were presented in order to facilitate its implementation for any given digital library. The significance of the work is that prior to this work, most search engines in digital libraries does not consider iser preferences when they search their database for books. Results of search queries are simply based on keywords only. However, with the introduction of the RSE model, digital libriries should begin to employ a much intelligent approach to ascertain the user preferences of books rather than limiting them to books that they may really not have interest in.

## REFERENCES:

[1] Seadle, M. & Greifeneder, E. Defining a digital library. Library Hi Tech. 25. (2007), 169-173. http://dx.doi.org/10.1108/07378830710754938.

[2] National Science Foundation. Digital Libraries Initiative: Available Research, US Federal Government (1999). Available at: http://dli2.nsf.gov/dlione/

[3] Irsa Arma Perdana , Lantip Diat Prasojo. Digital Library Practice in University: Advantages,Challenges, and Its Position. Advances in Social Science, Education and Humanities Research, volume 401 International Conference on Educational Research and Innovation (ICERI 2019),. Available at: https://www.atlantis-press.com/article/125934015.pdf

[4] Eisenberg, M. Information Literacy: Essential Skills for the Information Age. DESIDOC Journal of Library & Information Technology, 28 (2008).39-47. http://dx.doi.org/10.14429/djlit.28.2.166.

[5] Gladney, Henry & Ahmed, Zahid & Ashany, Ron & Belkin, Nicholas & Fox, Edward & Zemankova, Maria. Digital library: Gross

structure and requirements: Report from a workshop. IBM Research Report, RJ 9840. (1994), Available at: https://www.researchgate.net/publication/228602799_Digital_Library_Gross_Structure_and_Requirements_Report_from_aWorkshop

[6] Dode, Albi & Hasani, Silvester. PageRank Algorithm. (2017), 2278-661. http://dx.doi.org/10.9790/0661-1901030107.

[7] Borgs, C., Brautbar, M., Chayes, J.T., & Teng, S.. A Sublinear Time Algorithm for PageRank Computations. WAW (2012). http://dx.doi.org/10.1007/978-3-642-30541-2_4

[8] Kim, S.J., & Lee, S.H. An Improved Computation of the PageRank Algorithm. ECIR. (2002), http://dx.doi.org/10.1007/3-540-45886-7_5

[9] Jeh, G., & Widom, J.. Scaling personalized web search. WWW '03. (2003) DOI:10.1145/775152.775191

[10] Burke, Robin & Felfernig, Alexander & Göker, Mehmet. Recommender Systems: An Overview. Ai Magazine. 32. (2011) 13-18. http://dx.doi.org/10.1609/aimag.v32i3.2361. ISBN 9780128183663, https://doi.org/10.1016/B978-0-12-818366-3.00005-8.

[11] Melville, P., & Sindhwani, V. Recommender Systems. Encyclopedia of Machine Learning.(2010), DOI:10.1007/978-0-387-30164-8_705

[12] Konstan, J.A. Introduction to recommender systems. SIGMOD Conference. (2008). http://dx.doi.org/10.1145/1376616.1376776.

[13] Kurmashov, N., Latuta, K.N & Nussipbekov, A.. Online book recommendation system.1-4.(2015). http://dx.doi.org/10.1109/ICECCO.2015.7416895.

[14] Ali, Z., Khusro, S. & Ullah, I.. A Hybrid Book Recommender System Based on Table of Contents (ToC) and Association Rule Mining. (2016) http://dx.doi.org/10.1145/2908446.2908481.

[15] Rana, C., & Jain, S.K.. Building a Book Recommender system using time based content filtering. WSEAS TRANSACTIONS on COMPUTERS, (2012). Available at: https://wseas.com/journals/computers/2012/54-571.pdf

[16] Sushama R, Darshana B & Pooja, M.. Book Recommendation System. International Journal for Innovative Research in Science & Technology|( IJIRST) 1 (11), (2015) 314-316.Availableat: https://www.academia.edu/15972400/BOOK_RECOMMENDATION_SYSTEM

[17] Omisore M. O. & Samuel O. W. Personalized Recommender System for Digital Libraries.International Journal of Web-Based Learning and Teaching Technologies 9(1) 92014),18-32. https://doi.org/10.4018/ijwltt.2014010102

[18] Bhagyashree R., Nikhitha J., Mahalakshmi T., Ujwala,R.. Book Recommendation System with relevant Text Audiobook Generation. International Journal of Creative Research Thoughts (IJCR), Volume 9 (7) (2021), 398-404. Available at: https://ijcrt.org/papers/IJCRT2107170.pdf

[19] Gupta Vishal and Pandey Shriram , Recommender Systems for Digital Libraries: A review of concepts and concerns. Library Philosophy and Practice (e-journal). 2417 (2019). https://digitalcommons.unl.edu/libphilprac/241

[20] Imandoust, S.B. & Bolandraftar, M. Application of K-nearest neighbor (KNN)approach for predicting economic events theoretical background. Int J Eng Res Appl. 3,(5),(2013),.605-610.Availableat www.ijera.com

[21] Prihatmono, M. W. ., Arni, S., Iin, J. N. ., & Moeis, D. Application of the KNN Algorithm for Predicting Data Card Sales at PT. XL Axiata Makassar. Conference Series, 4(1), (2022), 59–64. https://doi.org/10.34306/conferenceseries.v4i1.692

[22] Novitasari,H.B., Hadianto, N., Sfenrianto, Rahmawati, A., Prasetyo, R., Miharja, J., Gata, W. K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo), J. Phys.: Conf. Ser. 1367 012008, (2019). http://dx.doi.org/10.1088/1742-6596/1367/1/012008

[23] Onwuchekwa, Edeama O & Jegede, Olumakinde Richard (2011), Information Retrieval Methods in Libraries and Information Centers, African Research Review, 5(6), (2011) 108-120. http://dx.doi.org/10.4314/afrrev.v5i6.10

[24] Ekenna, Margaret-Mary & Iyabo, Mabawonku. Information Retrieval Skills and Use of Library Electronic

Resources by University Undergraduates in Nigeria, Information and Knowledge Management, ISSN 2224-5758 (Paper) ISSN 2224-896X (Online) 3(9) (2013),.6-15.Retrievedfrom: https://core.ac.uk/download/pdf/234671455.pdf

[25] Svenonius, E. The intellectual foundation of Information Organization, Cambridge, MA, MIT Press (2000).

[26] Zakaria Jaadi. A Step-by-Step Explanation of Principal Component Analysis (PCA), (2021). Retrieved from: https://builtin.com/data-science/step-step-explanation-principal- component-analysis

[27] ]Theerthagiri, P., Jacob, I. J., Ruby, A.U. & Yendapalli, V. Prediction of COVID-19 Possibilities using KNN Classification Algorithm. (2020), http://dx.doi.org/10.21203/rs.3.rs-70985/v2.

[28] Amra, Ihsan & Maghari, Ashraf. Students performance prediction using KNN and Naïve Bayesian. (2017), 909-913. http://dx.doi.org/10.1109/ICITECH.2017.8079967

[29] Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi and Mohammed K. Ali Shatnawi International Journal of Business, Humanities and Technology3(3),(2013)32-44. https://www.ijbhtnet.com/journals/Vol_3_No_3_March_2013/4.pdf

[30] Lorena C.L & Sergio C.P. Best Books Ever Dataset (1.0.0) [Data set]. Zenodo.(2020), https://doi.org/10.5281/zenodo.4265096

[31] Masoud Maleki, Negin Manshouri, Temel Kayıkçıoğlu. A Novel Simple Method to Select Optimal k in k-Nearest Neighbor Classifier. International Journal of Computer Science and Information Security, 15(2), (2017),464-469.

[32] Zhongguo, Y., Hongqi, L., Liping, Z., Qiang, L., & Ali, S. A case based method to predict optimal k value for k-NN algorithm. Journal of Intelligent & Fuzzy Systems. 33. (2017), 1-10. http://dx.doi.org/10.3233/JIFS-161062.

[33] Zulkarnain L., Poltak., S. and Herman M., Optimization of K Value at the K-NN algorithm in clustering using the expectation maximization algorithm, IOP Conf. Ser.:Mater. Sci. Eng. 725 012133, 2020.

[34] Iman P., What Affects K Value Selection In K-Nearest Neighbor, International Journal of Scientific & Technology Research 8(7), (2019), 86-92.

[35] Islam, Mohammed J & Wu, Q. M. Jonathan & Ahmadi, Majid & Sid-Ahmed, Maher. Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers. JCIT. 5. (2010), 133-137. http://dx.doi.org/10.1109/ICCIT.2007.148.

[36] Ajay Kulkarni, Deri Chong, Feras A. Batarseh. 5 - Foundations of data imbalance and solutions for a data democracy, Data Democracy, Academic Press,(2020), 83-106,