*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 12 Issue 2 ǁ February 2024 ǁ PP. 111-115*

# Index-Based Assessment Employing Keyword Extraction Technique

Bhimsen Moharana, Bivas Ranjan Parida

*College Of Engineering Bhubaneswar, Biju Pattnaik University of Technology, Odisha, India*

***ABSTRACT:*** *A library management system increases efficiency for both librarians and users. This system primarily automates traditional libraries. We are developing a new integrated library management system with a simple search interface for all users. A library system that helps librarians manage a university's book collection. This project aims to develop a system that will allow users to easily find books on the topic they require using this application. This project has some additional characteristics to the usual library application, as it discovers the books connected to the topic using a keyword extraction method, so if he/she does not know the name of the book, he/she will be able to find all of the books that cover that topic within the context. In this study, we add a new search module to the standard library system. A property is created from the book's index and will be used in the new search module to find books by topic. We created a standard library management system with a revolutionary topic-based search module. We use optical character recognition to extract text from the book's index, which is in the form of an image file, and then preprocess the text file to acquire the main content. The principal content text files will then be analyzed by a keyword extraction algorithm, which will extract keywords and important phrases. The collected keywords and key phrases are linked to the book in the Backbend, allowing the user to search the book for a specific topic.*

***KEYWORDS:*** *efficiency, clear search interface, keyword extraction*

---

Date of Submission: 04-02-2024 Date of acceptance: 16-02-2024

---

## I. INTRODUCTION

Traditional libraries now allow consumers to search for available resources (also known as books) by title, author, and category. But what if someone wants to know which book offers information on a particular topic?

He would have to trawl through the indexes of every book he believes contains information on the subject he is investigating. This is still not a guarantee of outcomes. In this study, we employ characteristics as tags retrieved from the book index. We don't get everything covered in the book. The book's index highlights the themes it covers. However, reading the entire index may be difficult because it contains numbers, text, and images. As a result, the index has been filtered, and only relevant keywords and key phrases are extracted using the text rank algorithm. In recent years, there has been a surge of interest in improving keyword and key phrase algorithms for usage in a variety of applications. The keyword and key phrase extraction method involves extracting text from the book's image file, preprocessing text to create an input text file for the algorithm, and then extracting keywords and key phrases from the text file using the text Rank algorithm.

## II. LITERATURE REVIEW

Optical character recognition is a science that can help to translate various types of documents, images into easily analyzable and searchable data. OCR engine provides the accuracy for searchable data by tesseract behind the leading commercial engines[1].

Information to a computer system from printed documents or image files is to be st ored tore-utilize information. It helps in automatically retrieve and store the information provided by ocr engine[2].

Text preprocessing is a vital step of text classification and text mining generally. It is used to convert the original text data into raw data structure, and they are served to distinguish between various categories are identified [3].

Keyword, Key-phrases gives the summary of the text or any information for searchable data to users. Automatic keywords or key-phrase extraction techniques helps us to over coming this challenging task[4].

A. Optical Character Recognition

OCR is a process of digitizing a document image, printed text into constituent characters, so that it can bemanipulated by machines [1].

Step 1: To scan the physical form of the document/image using a scanner. When all the pages are scanned,OCR software is used to transform the document basically into 2 colors i.e., black, and white.

Step 2: Characters are then identified using one of two algorithms namely Pattern Recognition and Feature

Detection.

Step 3: After identifying the character using the algorithm, they are converted into ASCII code for manipulations.

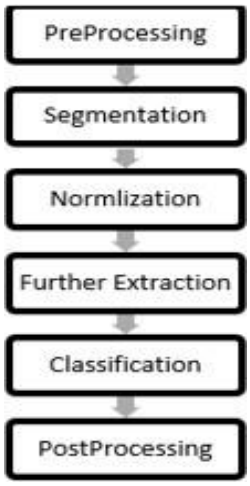



Fig: Phases of Optical Character Recognition.

1. Preprocessing:

The purpose of pre-processing is to put out waste, undesired qualities, or noise in an image withoutlosing any important information.Preprocessing reduces the inconsistent data and noise. It enhances the image and prepares it for the next phases in OCR phases. It gives greater value to the image and gets ready for the next phase in OCR phase[2].

2. Segmentation

Segmentation is the process of segregating text components with in an image background. For appropriate reorganization of the editable text lines from the recognized characters, firstly, segmenting the line of text, then the words are segmented from the segmented line and then from that the characters are segmented[2]. The segmentation process is crucial as it converts the image in the form of m*n matrix[6].

3. Normalization:

The matrix obtained from segmentation is normalized by removing the unnecessary information from the image without losing any important data.

4. Further Extraction:

This is a process of extracting the applicable features from the objects or alphabets to build feature vectors, which are then used by classifiers to find the input units along with objective output unit.

5. Classification:

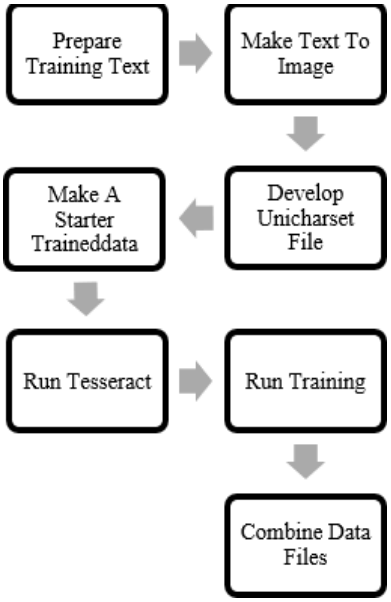
Classification is the process of distributing inputs with respect to detected information to their comparing class to create groups with homogeneous qualities, while segregating different inputs into different classes [2].

6. Postprocessing

In this phase, the incorporation of context and after shaping data in all the phases of Optical Character Recognition framework for increasingthe recognition rate along with an input which is given to early phasesof OCR.

Tesseract OCR Engine:

Tesseract is an open-source OCR engine that was developed between 1984 and 1994 [7]. It is an OCR engine for different operations with the support for Unicode and the ability to recognize more than 100 languages out of the box.




B. Preprocessing Text File

Text preprocessing is an important stage of text classification and text mining. Text preprocessing is a way to show each document as a feature vector which is used to split the text into separate words and used for indexing of the document. Here the keywords are selected through the feature selection.

Text preprocessing stage after taking inputs from text documents, it separates document to features which are called tokenization (words, terms, or attributes) and their weight which are achieved from the frequency of features in the text. After this, it pulls out the non-informative features including full stop, comma, numbers, and special characters. The remaining features are next standardized by reducing them to their root using the stemming process.

C. TextRank Algorithm

TextRank is an unsupervised graph-based technique used to extract the summary of a text. It uses the theory behind thePageRank algorithm [4]. TextRank model can be expressed as a weighted directed graph G= (V, E). The graph consists of a set of points V and a set of edges E and the set of edges E is the su bset of V *V. The weight of the edge of arbitrary two points i and j are Wij. For a given point Vi, In V(i) represents the set of points that point to the point Vi, Out V(i) represents the set of points that point from the point Vi. TR V(i) represents the score of the point Vi obtained by the TextRank model. The formula of TextRank model can be defined as shown below:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

In this formula, d is the same as d in the PageRank algorithm [3]. When using the TextRank algorithm to calculate the score of the points of the graph, it is required to specify the initial value of any given value to the points in the graph and then recursively compute the score of the point until convergence. After each point is convergent, the final score of the point represents the importance of the point in the graph.

## III. SYSTEM DESIGN

We are implementing a web-based library management application and its database. This application performs all basic operations of what a traditional library does. We are implementing a module, while inserting data of the book, to get the index of the book in an image form. The image file will be worked by various modules and will be giving output in the form of keywords and key phrases, which will be further linked to a particular book in the database along with other book details.

SYSTEM FLOW:

1. In the front-end user will give input as animage file of the index.
2. This image file will be processed by tesseractocr and the output will be in the form of a text file.
3. The text file will be further processed by cleaning by extracting symbols and numerical.
4. This processed file will be given as input to the textRank algorithm, which will extract keywords and key phrases.
5. The extracted keywords and key phrases will be linked to the book in the back end.

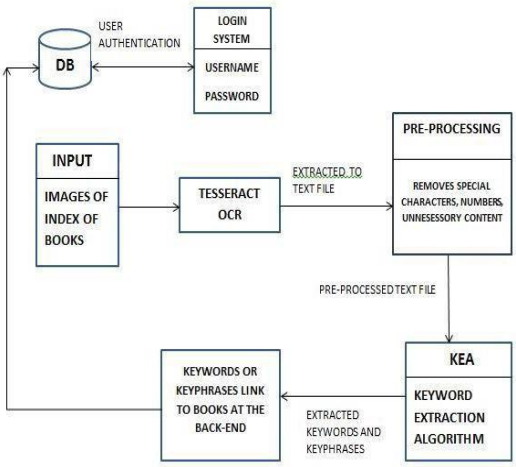



Fig: System Flow Diagram

SUB-FUNCTIONAL MODULES:

1. Enter the image file of an index of the book.
2. Extract keyword and key phrases to link to the book
3. Search the book by the topics (referred to as the extracted keyword and key phrases of the books)

## IV. CONCLUSION

In this study, we investigated OCR (tesseract). It is extremely efficient and accurate when turning a scanned document into a machine-encoded editable text file. We also researched text Rank, a keyword and key phrase extraction algorithm, and plan to use it in our module to extract keywords and key phrases from a book's index. In addition, the module may be improved in the future.

## V. FUTURE WORK

1. Improve the accuracy of extracted keywords and key phrases from the book.
2. Optimize certain keyword extraction algorithm, which can extract keywords and key phrases from the books having small size index.
3. Implementation of intelligent keyword extraction algorithm by providing the algorithm with some knowledge or learning method.

## REFERENCES

[1]. N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," arXiv, no. December2016, 2017.
[2]. K. Hamad and M. Kaya, "A Detailed Analysis of Optical Character Recognition Technology," Int. J. Appl. Math. Electron. Comput., vol. 4, no. Special Issue-1, pp. 244–244, 2016, doi: 10.18100/ijamec.270374.
[3]. A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," Int. J. Comput. Sci. Inf.
[4]. Secur., vol. 16, no. 6, pp. 22–32, 2018.
[5]. M. G. Thushara, T. Mownika, and R. Mangamuru, "A comparative study on different keyword extraction algorithms," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. June, pp. 969–973, 2019, doi: 10.1109/ICCMC.2019.8819630.
[6]. T. Putta, "Semantic Textual Similarity Using Machine Learning Algorithms," no. 8, pp. 10–15, 2017.
[7]. O. D. Trier, a. K. Jain, and T. Taxt, "Feature extraction methods for character r e c o g n i t i o n - - a survey," Pattern Recognit., vol. 29, no. 4, pp. 641–662, 1996.
[8]. S. Rice, F. Jenkins, and T. Nartker, "The fourth annual test of OCR accuracy," 1995 Annu. Rep. ISRI, …, vol.
[9]. 1, no. April, pp. 1–39, 1995, [Online]. Available: http://stephenvrice.com/images/AT-1995.pdf.