



**IJIRCCCE**

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 2, February 2021

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

**Impact Factor: 7.488**

 9940 572 462

 6381 907 438

 [ijircce@gmail.com](mailto:ijircce@gmail.com)

 [www.ijircce.com](http://www.ijircce.com)

# Index-Based Evaluation Using Keyword Extraction Algorithm

Adarsh Pawar<sup>1</sup>, Tejaswini Nikam<sup>1</sup>, Sakshi Gaikwad<sup>1</sup>, Chanchal Kedia<sup>2</sup>

B.E. Student, Department of Computer Science, KJEE's Trinity Academy of Engineering, Pune, India<sup>1</sup>

Professor, Department of Computer Science, KJEE's Trinity Academy of Engineering, Pune, India<sup>2</sup>

**ABSTRACT:** A library management system enhances the efficiency of both the librarians and the users of the library system. This system mainly provides automation to the traditional library. We are implementing a modern integrated library management system which is a simple and clear search interface for all the users. The library system for assisting a librarian in managing a books library in a university. This project aims in developing a system that will help the users to easily find the book related to the topic he/she needs through this application. This project has some additional features to the normal library application, as it finds the books related to the topic by using of keyword extraction algorithm, so as if he/she does not know the name of the book still through the context he/she will find all the books containing that topic.

In this paper, we are implementing a new search module in the traditional library system. An attribute is created from the index of the book, which will be used in the new search module as search book by topic. We have created a traditional library management system integrated with a new search by the topic module. First, we extract text from the index of the book, which will be in the form of an image file, using optical character recognition, then preprocess the text file to get the main content. These main content text files will be further processed by a keyword extraction algorithm, which will extract keywords and key phrases. These extracted keywords and key phrases will be linked to the book in the backend then the user can search the book by searching for a particular topic.

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

## I. INTRODUCTION

The traditional libraries today have the options to search the available resources (referred to as books), by their title, author, and are sorted by categories.

but what if somebody wants to get to know which book contains the information about the topic that somebody wants? he would have to go through all the indexes of the book he thinks has the information about the topic he wants. this still does not guarantee the results. In this paper, we are implementing attributes as the tags, which are extracted from the index of the book. as we do not get all the topics covered by the book.

The index of the book contains the topics that the book covers. but to read the whole index could be challenging, as it contains numbers, text, images etc. So, the index is filtered and only the required keywords and key phrases are extracted using the text rank algorithm. In recent years we have seen great interest in the keyword and key phrase algorithms getting better, and improvement in various applications by applying them. The keyword and key phrase extraction process includes extraction of the text from the image file of the book, preprocessing text to generate an input text file for the algorithm, extract keywords and key phrases from the text file using the textRank 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 tesseraact behind the leading commercial engines[1].

Information to a computer system from printed documents or image files is to be stored to 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 overcome this challenging task [4].

Semantic textual similarity describes various regression techniques of supervised model to analyze degree of semantic equivalence between two text fragments are word phrases, sentences, paragraphs, or documents [5].

A. Optical Character Recognition

OCR is a process of digitizing a document image, printed text into constituent characters, so that it can be manipulated 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.

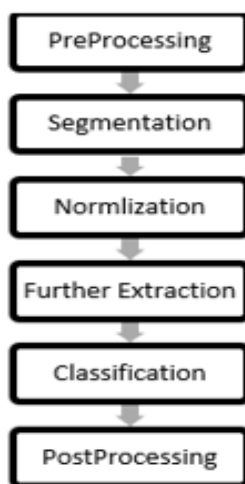


Fig: Phases of Optical Character Recognition.

1. Preprocessing:

The purpose of pre-processing is to put out waste, undesired qualities, or noise in an image without losing 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 within 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  $m \times n$  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].

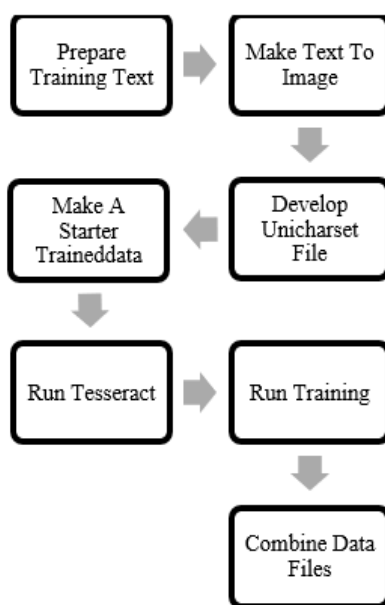


6. Postprocessing

In this phase, the incorporation of context and after shaping data in all the phases of Optical Character Recognition framework for increasing the recognition rate along with an input which is given to early phases of 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 the PageRank 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 subset of  $V * V$ . The weight of the edge of arbitrary two points  $i$  and  $j$  are  $W_{ij}$ . For a given point  $V_i$ ,  $In(V_i)$  represents the set of points that point to the point  $V_i$ ,  $Out(V_i)$  represents the set of points that point from the point  $V_i$ .  $TR(V_i)$  represents the score of the point  $V_i$  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 an image 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.

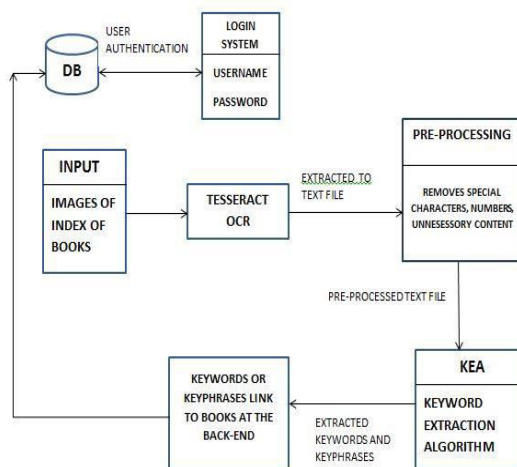


Fig: System Flow Diagram

#### 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 paper, we have studied OCR (tesseract). It is very efficient and accurate while converting a scanned document into a machine-encoded editable text file. We have also studied textRank a keyword and Keyphrase extraction algorithm and will implement it in our module to get keywords and Keyphrase from the index of a book. Also, there is a possibility of improvement in the module in further future.



#### V. FUTURE WORK

1. Improve the accuracy of extracted keywords and keyphrases from the book.
2. Optimize certain keyword extraction algorithm, which can extract keywords and key phrases from the books having smallsizeindex.
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. December 2016, 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. Secur., vol. 16, no. 6, pp. 22–32, 2018.
- [4] 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.
- [5] T. Putta, "Semantic Textual Similarity Using Machine Learning Algorithms," no. 8, pp. 10–15, 2017.
- [6] 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.
- [7] S. Rice, F. Jenkins, and T. Nartker, "The fourth annual test of OCR accuracy," 1995 Annu. Rep. ISRI, ..., vol. 1, no. April, pp. 1–39, 1995, [Online]. Available: <http://stephenrice.com/images/AT-1995.pdf>.



INNO  SPACE  
SJIF Scientific Journal Impact Factor

Impact Factor:  
7.488

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  [ijircce@gmail.com](mailto:ijircce@gmail.com)



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details