ARTICLE HELPER

Unification of Algorithms to extract special formatted text from portable document format

BY

IMTIAJ AHAMMAD RAHAT
ID: 142-15-3499

LUBNA HOSSAIN
ID: 142-15-3447

ABUL HASNATH LIMON
ID: 142-15-3532

This Report Presented in Partial Fulfilment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Mr. Seraj Al Mahmud Mostafa
Senior Lecturer
Department of CSE
Daffodil International University

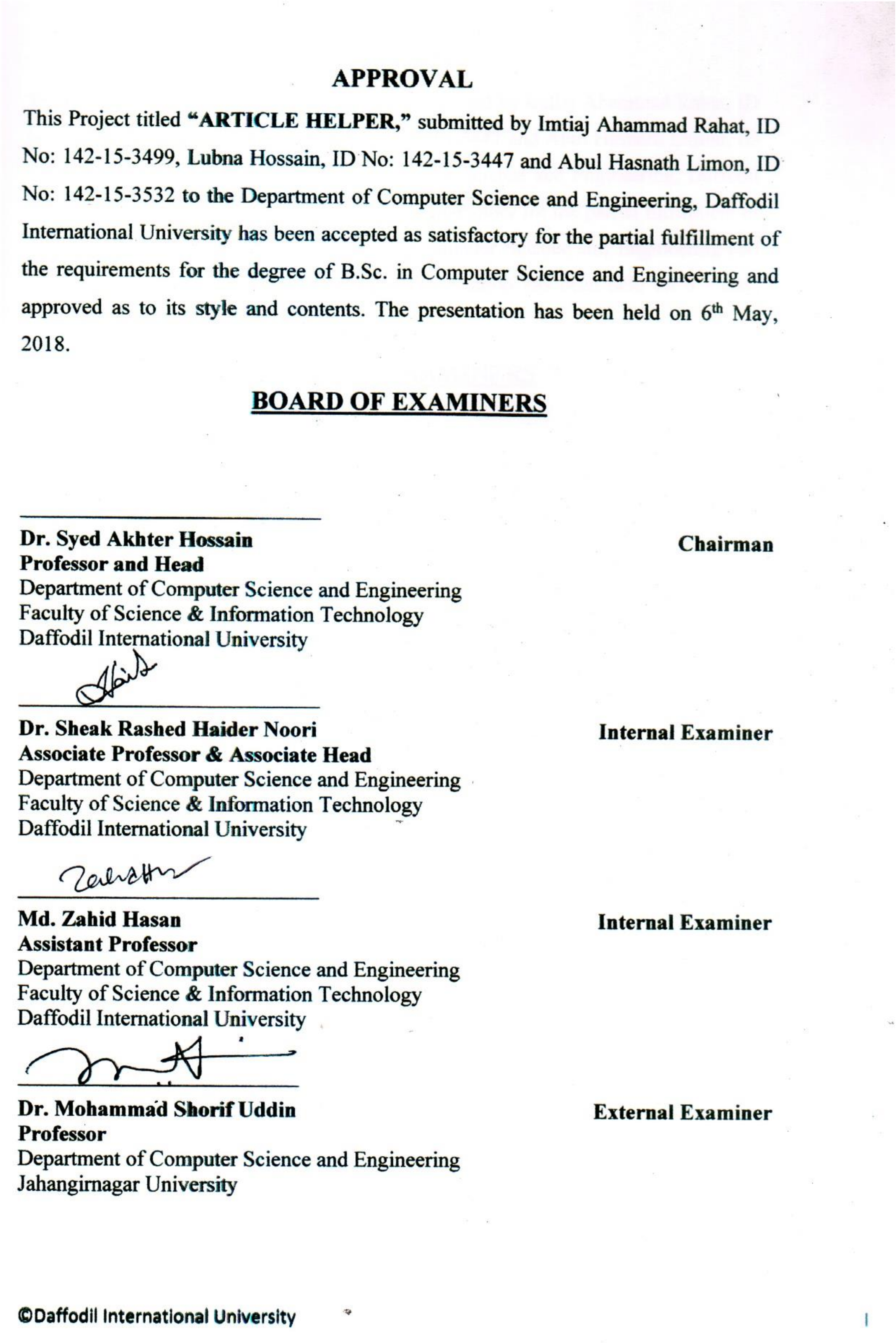
Co-Supervised By

Rubaiya Hafiz
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

6th May 2018

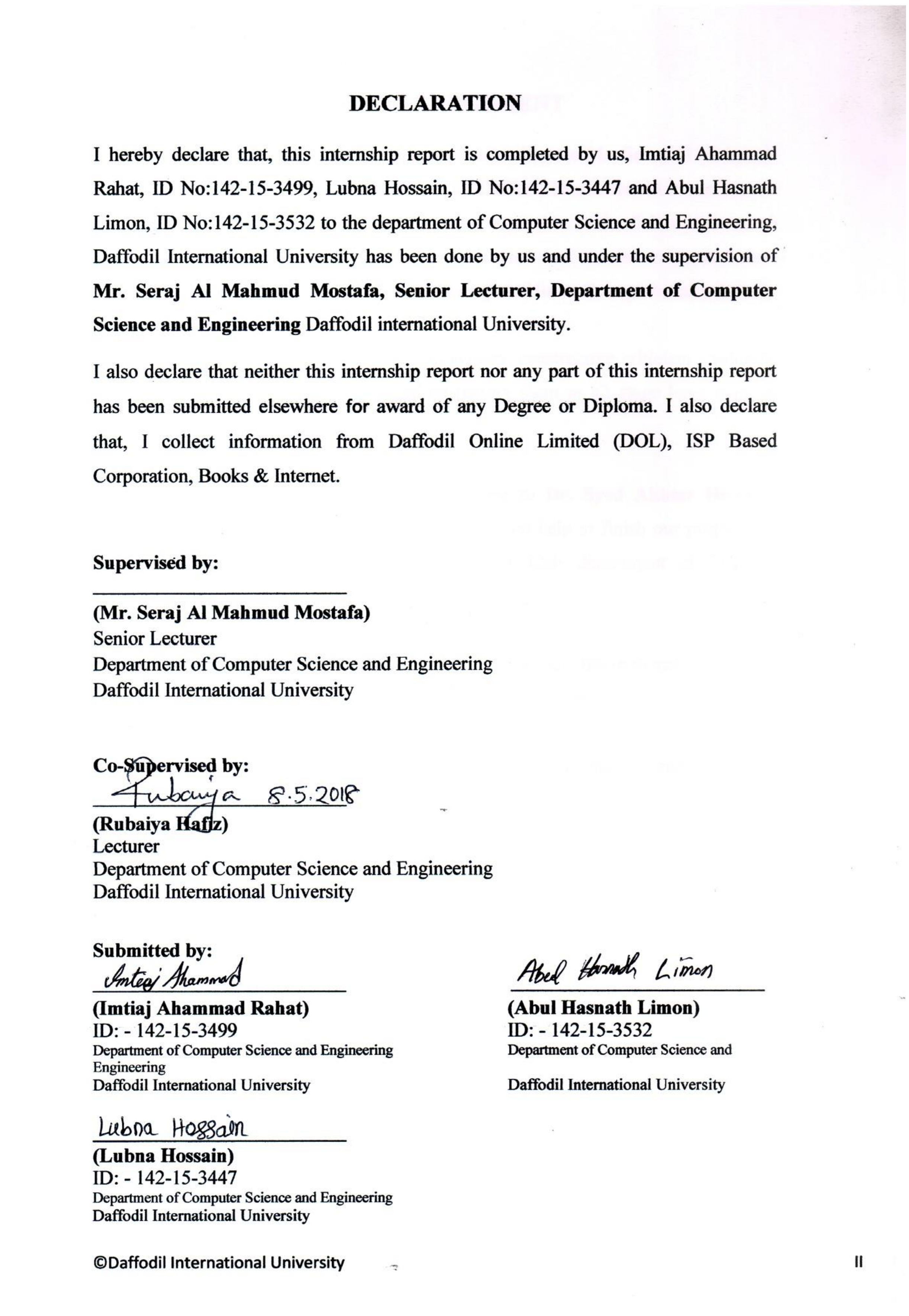


APPROVAL

This Project titled **"ARTICLE HELPER,"** submitted by Imtiaj Ahammad Rahat, ID No: 142-15-3499, Lubna Hossain, ID No: 142-15-3447 and Abul Hasnath Limon, ID No: 142-15-3532 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 6th May, 2018.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Professor and Head** Department of Computer Science and Engineering Faculty of Science & Information Technology Daffodil International University	**Chairman**
Dr. Sheak Rashed Haider Noori **Associate Professor & Associate Head** Department of Computer Science and Engineering Faculty of Science & Information Technology Daffodil International University	**Internal Examiner**
Md. Zahid Hasan **Assistant Professor** Department of Computer Science and Engineering Faculty of Science & Information Technology Daffodil International University	**Internal Examiner**
Dr. Mohammad Shorif Uddin **Professor** Department of Computer Science and Engineering Jahangirnagar University	**External Examiner**



DECLARATION

I hereby declare that, this internship report is completed by us, Imtiaj Ahammad Rahat, ID No:142-15-3499, Lubna Hossain, ID No:142-15-3447 and Abul Hasnath Limon, ID No:142-15-3532 to the department of Computer Science and Engineering, Daffodil International University has been done by us and under the supervision of **Mr. Seraj Al Mahmud Mostafa, Senior Lecturer, Department of Computer Science and Engineering** Daffodil international University.

I also declare that neither this internship report nor any part of this internship report has been submitted elsewhere for award of any Degree or Diploma. I also declare that, I collect information from Daffodil Online Limited (DOL), ISP Based Corporation, Books & Internet.

Supervised by:

____

(Mr. Seraj Al Mahmud Mostafa)
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Rubaiya 8.5.2018
____

(Rubaiya Hafiz)
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Imtiaj Ahammad
____

(Imtiaj Ahammad Rahat)
ID: - 142-15-3499
Department of Computer Science and Engineering
Engineering
Daffodil International University

Abul Hasnath Limon
____

(Abul Hasnath Limon)
ID: - 142-15-3532
Department of Computer Science and

Daffodil International University

Lubna Hossain
____

(Lubna Hossain)
ID: - 142-15-3447
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

We really grateful and wish our profound our indebtedness to **Mr. Seraj Al Mahmud Mostafa**, **Senior Lecturer,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Research papers and journals*" and "*Research organizations and communities and their conferences*" has helped us to carry out this project. His endless patience, scholarly guidance, continual

encouragement, constant and energetic supervision, constructive criticism , valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Syed Akhter Hossain**, Professor and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In today's era, we share our every innovations and implementations with the whole world. It takes us to greater accomplishment to let others know about the advancement that we achieved and also to enhance more. In such conferences, thousands of papers are submitted and also being reviewed. The aim of this project is to make a solution how we can review the submitted papers efficiently. The project proposes a system to mine the papers and bring out the necessary properties to review the papers. It fetches the paper title, keywords and other contents separately such as abstract, introduction, references etc. out of the file. Any paper can also be searched or categorized by their titles or keywords. Both the initial stage on selection of papers and the observation of papers in later stages can be done through this procedure.

TABLE OF CONTETNS

CHAPTER

LIST OF FIGURES

LIST OF TABLES

TABLES **PAGE**

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Ours is a century of innovations and most probably the fastest period we are getting new innovations one after another. One of the most considerable field is technology and the use of technology in all sorts of fields. With the technology no innovations or creativity has any limitation boundaries. We can share our discoveries to all over the world and the people of all over the world can get the information.

At present there are specific communities or organizations which give us a chance to show our work to specific people at specific fields. For every new idea they give a platform to show the work and get reviews from people all over the world. For this purpose we need to publish a research paper or journal. For self-perpetuating and faster feedback we can publish our paper in conferences arranged by those organizations.

Thousands of papers are submitted on those conferences in every year and the organizations need to review those papers. For numerous number of submissions it becomes a tough work to review the papers and get them selected. There must be a solution to categorize these papers, sorts them and create more efficient environment to review them. Our project purposes a solution.

1.2 MOTIVATION

Through the under graduation process, we have gone through some courses that required us to study on new technologies and innovations. It has been also a fact that we have to make projects on new ideas and to review the idea we have been gone through research papers and journals. Only the research papers and journals were the best way to take a deep understanding on any new technology and their new implementations or advancements. Going through this process we came to a situation that we had lots of pdf files in our storage and with the file name we can’t understand the paper type or contents. Also working with thesis paper we got to know about the conferences on research papers and their process of review. So we have decided to make an algorithm to dig out the resources of the papers and make a suitable and efficient environment for reviewing them.

1.3 RATIONALE OF THE STUDY

Research papers or journals play an important role in the technology and innovation fields. And there are numerous number of published papers already and the number of upcoming papers are in growth. So it is a real time complication to get the resources from those papers and mine them, especially for the conferences on research papers. An automated system to get resources from papers and utilize them according to required criteria would be a solution to many real time problems. It may help the reviewer of those papers to make the review in an efficient procedure. Our algorithm requires a criteria of specific structured papers in pdf format and mine the resources to utilize them.

1.4 EXPECTED OUTPUT

The project includes an algorithm that requires bringing out the resources of any paper in pdf format but with some given particular instructions. The paper must be written following those instructions. Such as for every conference on research papers, they give the paper owners a criteria to follow with. For a research paper there are some common things such as the title of the paper, there must be keywords, abstract portion, headings of every portion. If we set some rules on them, we can dig out those resources. In our project we have built a web application which only read the files followed by specific criteria of a paper and digs out its title, keywords, the abstract portion and all the portions with big headings and portions with sub headings inwards the big heading portion. With the keywords it can categorize them. We also can gather all the abstract portions of papers to review them. Every paragraph with separated headings are also can be brought out. The time consumption is one of our main target while we are working with numerous files.

CHAPTER 2

BACKGROUND

2.1 INTRODUCTION

There are different organizations and communities which work on research fields and they provide opportunities to publish papers on their servers and the published papers are shared to all members of that particular community. Each of them follows a criteria for the papers submitted to them. The paper writer needs to follow those instruction in writing the paper. For our study we have taken different papers in pdf file formats and read the file. For reading the file we have used a library named “iTextSharp” which is a. NET library. The library helps us to read the contents in a file in pdf format with the original characteristics such as font, font size, font style etc. Our algorithm then makes an analysis on textual resources and gets the required properties such as title, keywords, abstract and separates all portions with main headings and sub headings.

2.2 RELATED WORKS

There are lots of organizations which offer platforms for publishing research papers and to make them open to review on those. They follows some strict rules for submitting the papers. Most of them use a management system for the paper submissions but they do not mine the properties of a paper by their system. At least none has a system to eliminate papers which do not follow their criteria. In every year these organizations and some others arrange conferences in which thousands of papers are submitted but most of the reviews of those papers are either manual or with static options.

Also many industrial companies acquire a management system for their applicants for jobs. They normally figure out some static information such as type of job positions or experiences from the resumes submitted by the applicants and then sort them or eliminate them automatically. But also they don’t mine the resources as well.

2.3 RESEARCH SUMMARY

Our study has been to make a system that can actually read from files and mine the components from the resources. In our study we use research papers in pdf format and read all the contents. Our algorithm then mines the contents and get some information that we can use to automate a manual reviewing system. For every different format of papers we need to change our criteria for mining the information.

2.4 SCOPE OF THE PROBLEM

Our algorithm provides a solution to a real time complexity. There are many organizations or companies require a documental resources from some sources. Most of the time the number of documents exceeded to limit that it becomes tough to manage them manually or from a static automated system.

For educational institutions, they have to study thousands of applications during their admission periods. If they can sort them with an automated system, that would be a very time consuming and cost consuming solutions.

As well as for the organizations which work with research and innovation. They have to carry out lots of documental resources such as papers or journals, reviews them and stores them.

Also for big industrial companies which require thousands of employee. They need to go through thousands times for documental reviews if we only consider the resumes of the job applicants.

2.5 CHALLENGES

We have established an algorithm to mine the pdf documents on research fields initially. To read the pdf files we have no option but to use a third party library in the application. Our main target was to mine the textual data of the papers, not to make an algorithm to read pdf files. So we have to take the help of that library.

Also for our initial study we have only worked with pdf format files for its popular usage commercially. But with the same library we have options to work with text or word documents which may include in our future advancement.

We have kept the option to change the criteria of the algorithm from which the algorithm mines the information. But in aspect of choosing the papers, it have come to realize that there are lots of different papers with different criteria. So we have to choose few formats to operate our study in spite of every individual formats.

In research papers or journals both contains some figures, tables flow charts etc. We haven't work with these images within the documents since our main target was to mining the textual data. One big challenge for us have been the mathematical terms to mine into information. We have kept mathematical terms aside for our study also.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 INTRODUCTION

Our proposed system is implemented in a web application. But the algorithm we have been using can’t be implemented in all web technologies. Also we are using a third party library which has some limitations. For this reason we need to follow some requirements

3.2 SYSTEM SPECIFICATION

Since it is web application, it has the capability to run on all types of operating systems. Most popular of them are Windows, Mac OS and Linux. Our system as well as the application can be run on all of them with any of web browsers.

3.3 TECHNOLOGY SPECIFICATION

We have used the Microsoft technologies for implementing the system into a web application. The technologies we have used are-

- Microsoft ASP.NET MVC 5.
- Microsoft ASP.NET Razor.

3.4 DEVELOPMENT TOOLS

- Microsoft Visual Studio.
- Microsoft SQL server Management Studio.

3.5 WEB PROGRAMMING

- Languages: C#, HTML, CSS.
- Scripting Language: Java Script, JQuery and Ajax.

- Frame Work: Bootstrap.

3.6 THIRD PARTY LIBRARY

We have used a third party library named “iTextSharp” which is a .NET pdf library [1]. It normally allows to create, adapt, inspect and maintain documents in the Portable Document Format (PDF).

3.7 DATABASE MANAGEMENT SYSTEM

To store and fetch the data we have used a database management system in which we have designed our storage system. It is a tool of Microsoft named “Microsoft SQL Server Management Studio.

CHAPTER 4

APPLICATION DESIGN AND IMPLEMANTATION

4.1 INTRODUCTION

We have designed our application in two phases. At first we have designed our algorithm and then we have designed the application with the algorithm. Our application is a full integration of the whole process including the third party library.

4.2 ARCHITECTURAL DESIGN OF THE INTEGRAL SYSTEM

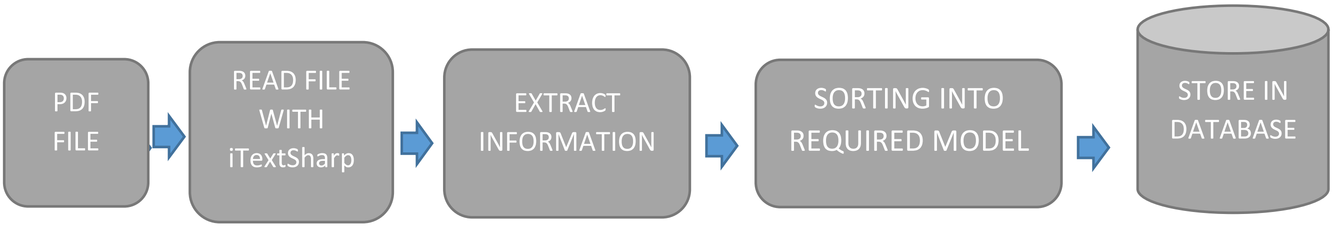


Figure 4.2.1: Basic Architectural Design (part 1)

The basic and main workflow of the whole integrated web application starts from getting a pdf file. If user uploads a file or more the system will take this file. If the files are not in pdf format, system disallows to process. Then system needs to read those files. Reading a pdf file has been done with a third party library “iTextSharp”. Afterwards the system algorithm works with the data given from the library and process the data into information. The information is formed into required model according to needs. And then system stores the processed model into database.

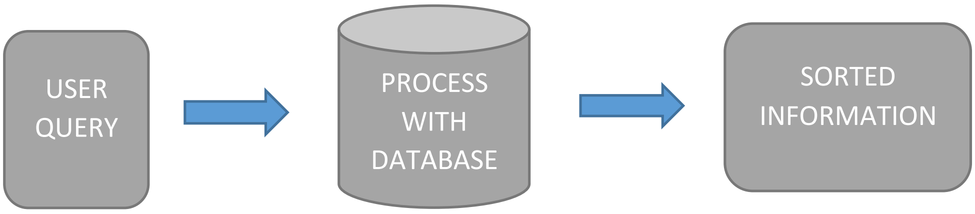


Figure 4.2.2: Basic Architectural Design (part 2)

Once the data is stored in the database, all the main processing has been finished. The data then are ready for the categorization or reviewing both. The data is transferred into information and then we can utilize that information and sort them into any model we need.

4.3 ALGORITHM ARCHITECTURE

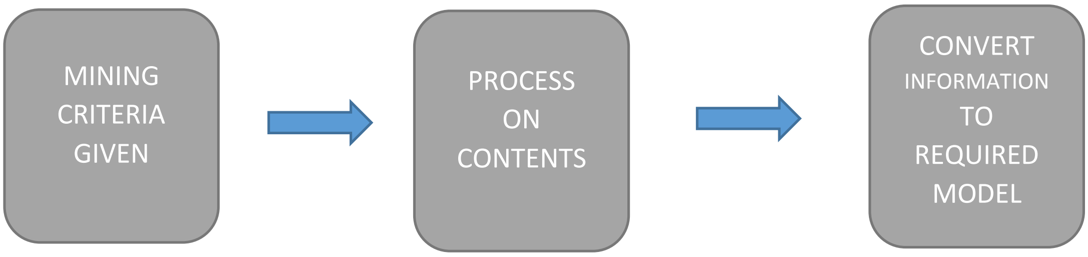


Figure 4.3.1: Algorithm Architecture

The algorithm works on data with its actual properties. From the library we get the required data and then the algorithm turn comes. But before the algorithm starts working, we need to give some criteria or conditional data on how the algorithm works with the data and converts them into information. Afterwards getting the criteria to work with the algorithm process the data. And another important fact is that the output may be needed differently for different purpose. Here the algorithm process to dig the title, keywords, abstract and all other main portions and their sub portions.

4.4 USE CASE DIAGRAM

There are actually two steps for the system for use case diagram. They are-

1. The admin panel sets the criteria for mining the data.
2. The user uploads the file.

4.4.1 USE CASE DIAGRAM FOR ADMIN

This use case diagram shows the initial step before asking users to submit their files to the website. The admin must set a criteria for mining the file contents.

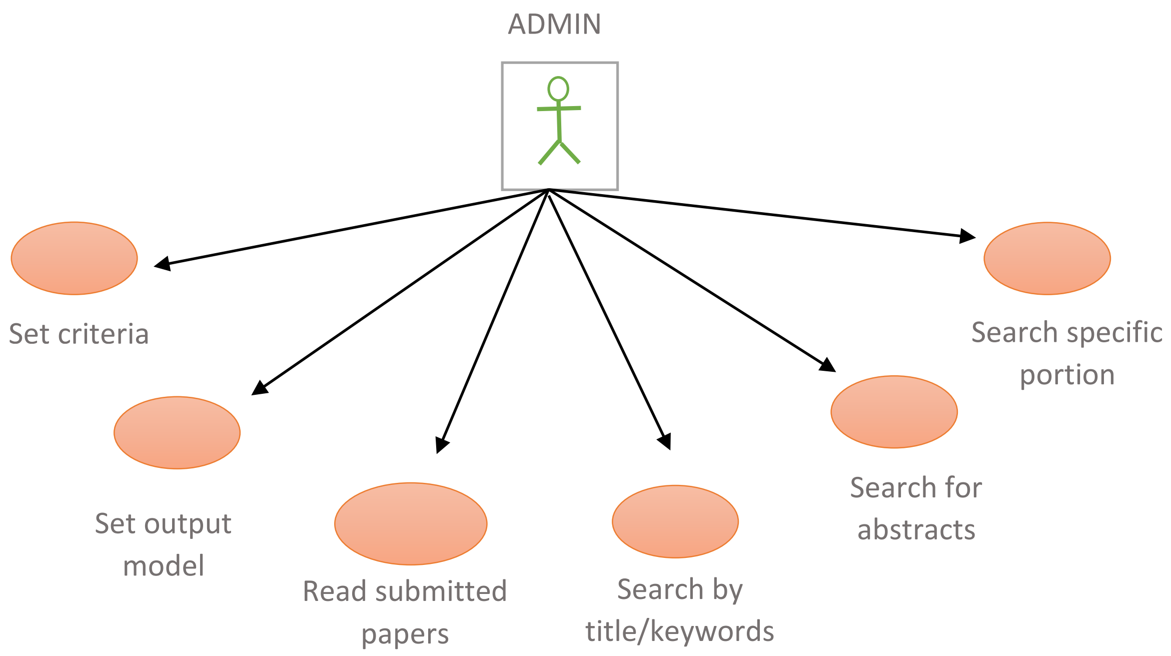
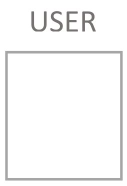


Figure 4.4.1: Use Case Diagram for Admin

4.4.2 USE CASE DIAGRAM FOR USER

This use case diagram shows the user portion. User need to follow the instructions for writing paper and submit the paper.



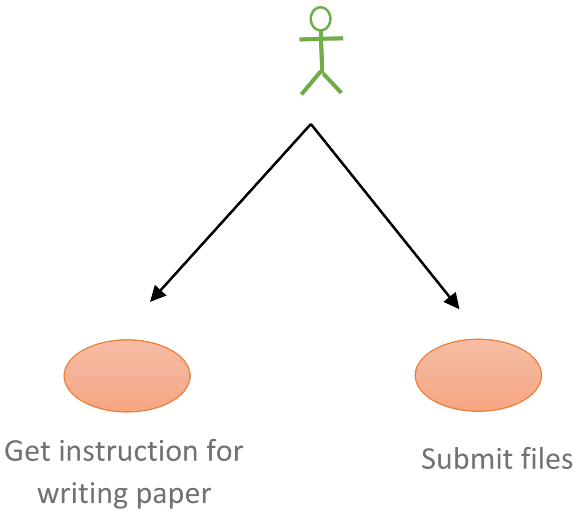
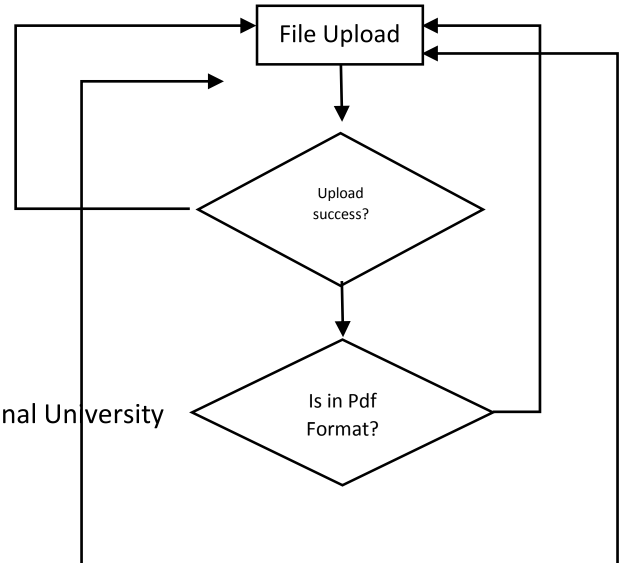


Figure 4.4.2: Use Case Diagram for User

4.5 DATA FLOW CHART



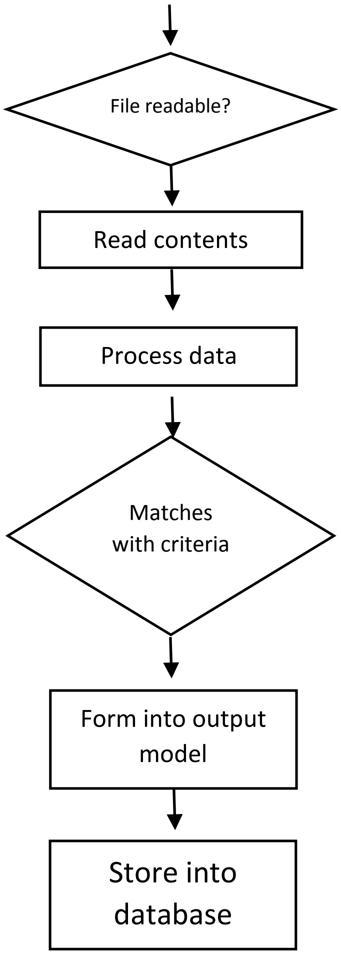


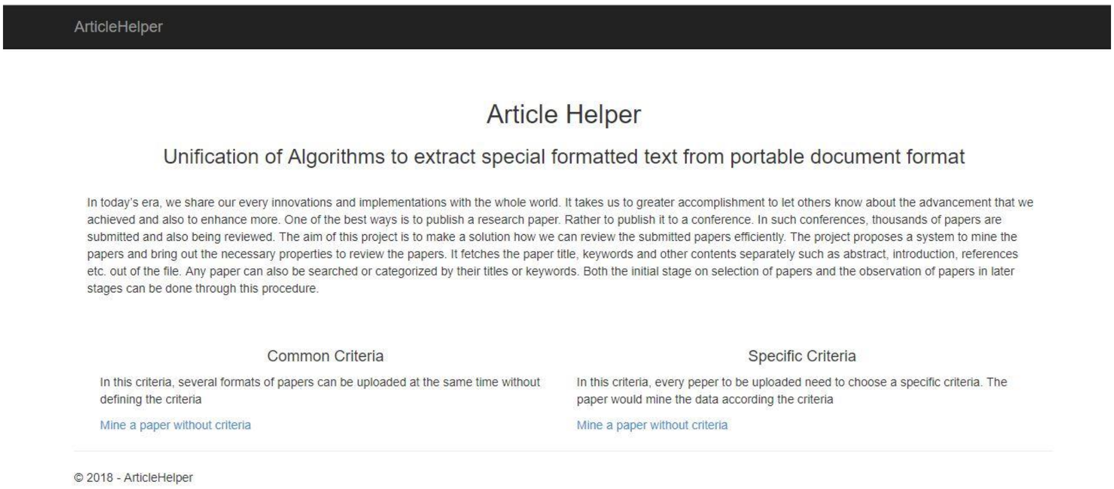
Figure 4.5.1: Data Flow Chart

4.6 PROPOSED GUI OF THE APPLICATION

In our project we put more priority on the algorithm and its solution than the design of the interface. We have designed the interface for testing the system. So most of the pages have a very simple working design.

4.6.1 SETTING THE MINING CRITERIA

This interface shows up with options to select the criteria. There are options to mine the file with common criteria or with specific criteria. Common criteria only absorbs criteria of several papers. And specific criteria mines all the classified text from the paper.



Article Helper

Unification of Algorithms to extract special formatted text from portable document format

In today's era, we share our every innovations and implementations with the whole world. It takes us to greater accomplishment to let others know about the advancement that we achieved and also to enhance more. One of the best ways is to publish a research paper. Rather to publish it to a conference. In such conferences, thousands of papers are submitted and also being reviewed. The aim of this project is to make a solution how we can review the submitted papers efficiently. The project proposes a system to mine the papers and bring out the necessary properties to review the papers. It fetches the paper title, keywords and other contents separately such as abstract, introduction, references etc. out of the file. Any paper can also be searched or categorized by their titles or keywords. Both the initial stage on selection of papers and the observation of papers in later stages can be done through this procedure.

Common Criteria

In this criteria, several formats of papers can be uploaded at the same time without defining the criteria

Specific Criteria

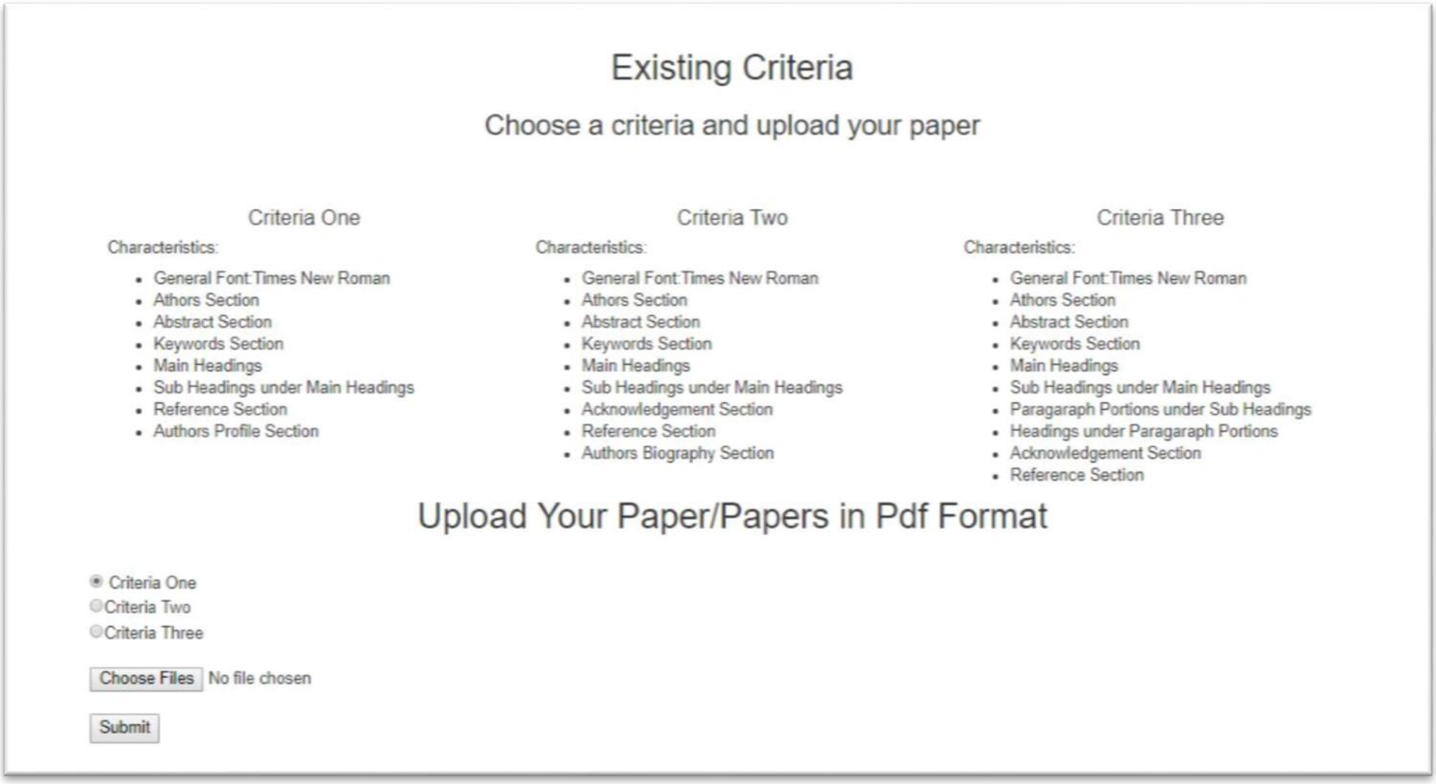
In this criteria, every peper to be uploaded need to choose a specific criteria. The paper would mine the data according the criteria

© 2018 - ArticleHelper

Figure 4.6.1: Setting the Mining Criteria

4.6.2 EXISTING MINING CRITERIA CHARACTERISTICS

There are three different criteria available in the system at present. These three criteria defines three different format of papers. The number of criteria would increase according to the target paper requirement in the future.



Existing Criteria

Choose a criteria and upload your paper

Criteria One

Characteristics:

- General Font:Times New Roman
- Athors Section
- Abstract Section
- Keywords Section
- Main Headings
- Sub Headings under Main Headings
- Reference Section
- Authors Profile Section

Criteria Two

Characteristics:

- General Font:Times New Roman
- Athors Section
- Abstract Section
- Keywords Section
- Main Headings
- Sub Headings under Main Headings
- Acknowledgement Section
- Reference Section
- Authors Biography Section

Criteria Three

Characteristics:

- General Font:Times New Roman
- Athors Section
- Abstract Section
- Keywords Section
- Main Headings
- Sub Headings under Main Headings
- Paragaraph Portions under Sub Headings
- Headings under Paragaraph Portions
- Acknowledgement Section
- Reference Section

Upload Your Paper/Papers in Pdf Format

● Criteria One
○ Criteria Two
○ Criteria Three

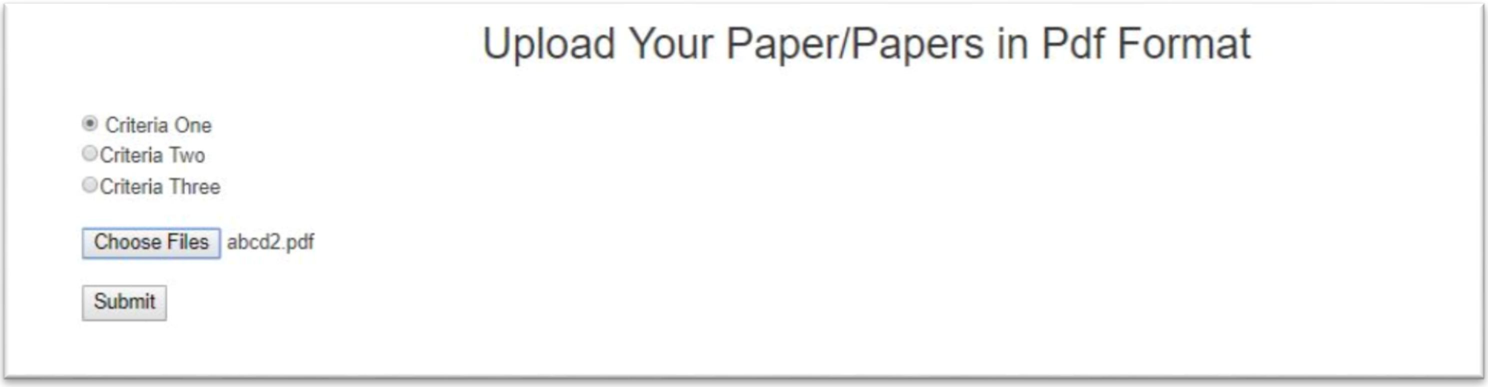
No file chosen

Figure 4.6.2: Existing Criteria Characteristics

4.6.3 UPLOAD PDF FILE WITH SPECIFIC CRITERIA

There are three types of criteria to choose according to the format of the uploading paper. The user need to select an option of criteria and then upload the file and click on the "Submit" button.

Figure 4.6.3: Upload PDF File with Specific Criteria

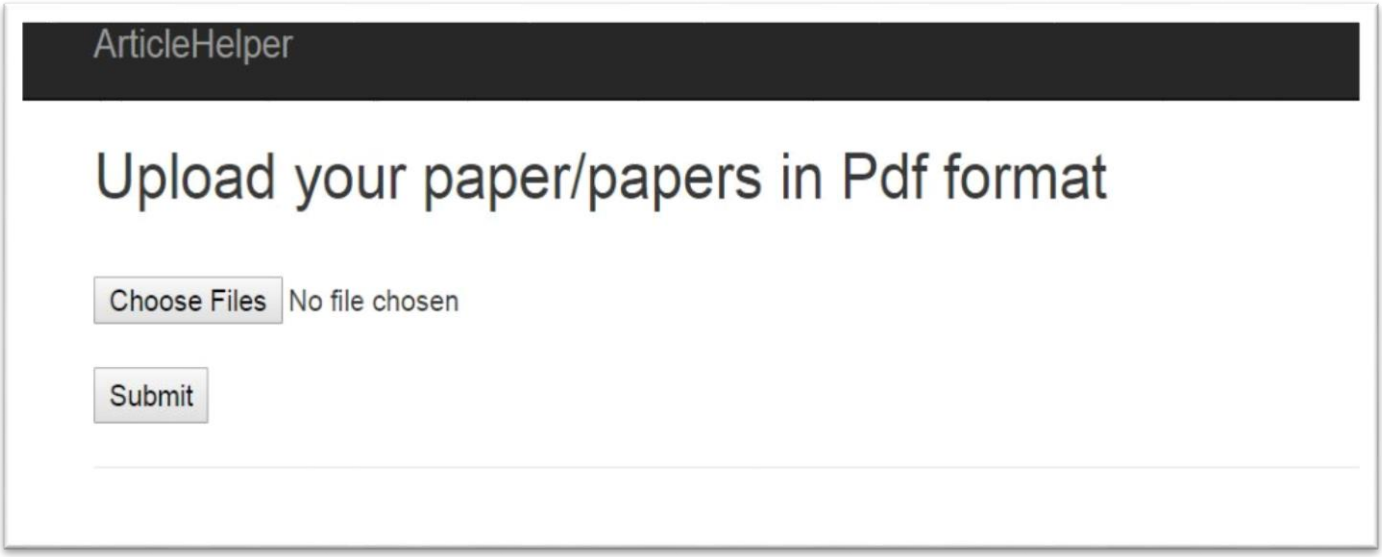


Upload Your Paper/Papers in Pdf Format

● Criteria One
○ Criteria Two
○ Criteria Three

4.6.4 UPLOAD PDF FILE WITHOUT SPECIFIC CRITERIA

Once the paper is uploaded to the database in an efficient format. There is also facility to upload a file without choosing the criteria. Or more specifically with a common criteria. The paper format must be cover in the common criteria to work. In this section, user need to upload the file only.


ArticleHelper

Upload your paper/papers in Pdf format

Figure 4.6.4: Upload PDF File without Specific Criteria

CHAPTER 5

TESTING AND EVALUATION

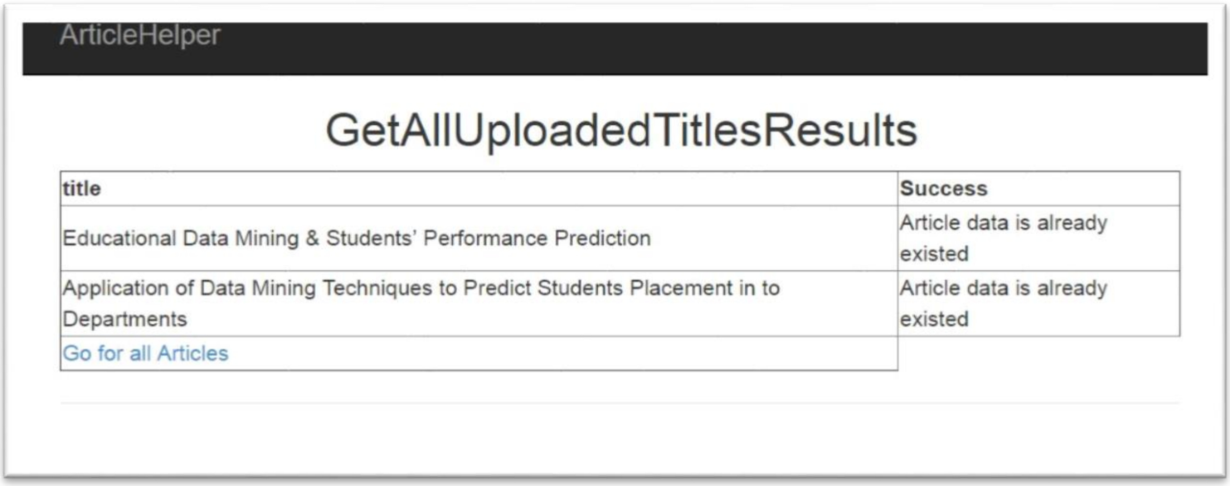
The outcome of the whole integrated project has been represented in this section. Though the whole processing mostly works at the background. In consideration with the application interface and the outcome of the processing some results have been added to this section.

5.1 TESTING THE SOLUTION

Our solution or the system has some gradual steps one by one. Suppose the system at first read the file, then it extract the required information, store to the database and with the help of database we can then review the rest.

5.1.1 FILE UPLOAD ACTION STATUS

The users are to be got the list of papers with their titles and the report of the submission of papers to the storage after uploading the files. It can work for only one file or more simultaneously.



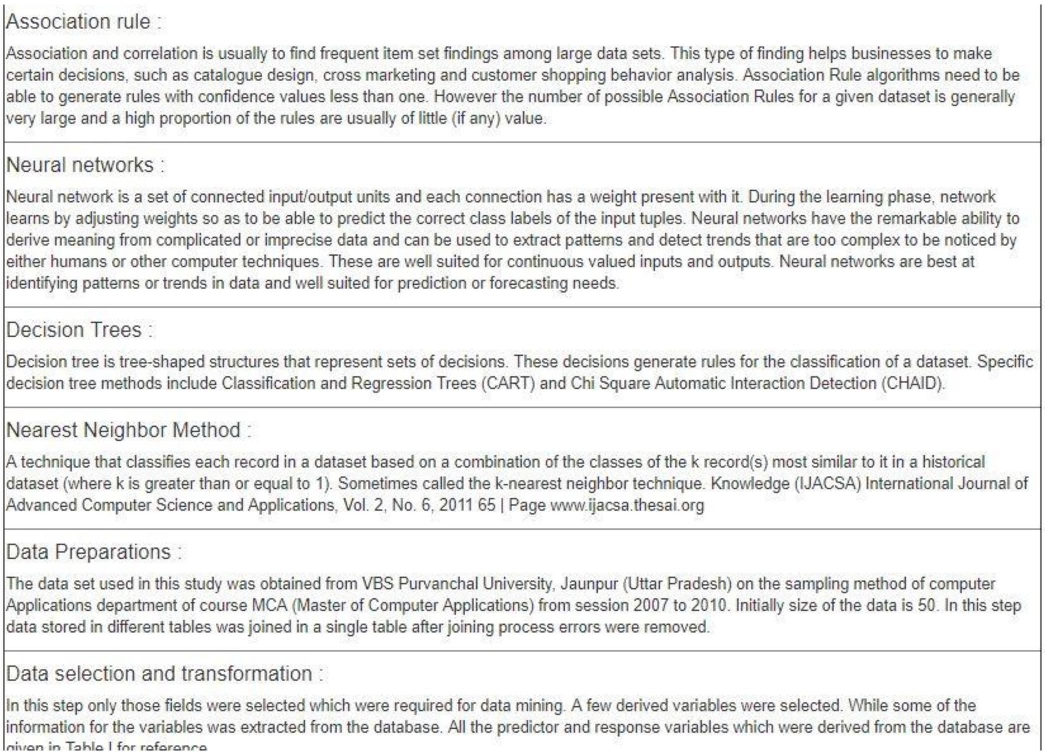
GetAllUploadedTitlesResults

title	Success
Educational Data Mining & Students' Performance Prediction	Article data is already existed
Application of Data Mining Techniques to Predict Students Placement in to Departments	Article data is already existed
Go for all Articles	

Figure 5.1.1: File Upload Action Status

5.1.2 CLASSIFIED CONTENTS SEPARATED FROM PAPER

The users are to be got the list of papers with their titles and the report of the submission of papers to the storage after uploading the files. The paper uploaded with selecting a specific criteria can separate all its classified contents from the paper. All the separated classified contents can be extracted in this section.



Association rule :

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Neural networks :

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Decision Trees :

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Nearest Neighbor Method :

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called the k-nearest neighbor technique. Knowledge (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011 65 | Page www.ijacsa.thesai.org

Data Preparations :

The data set used in this study was obtained from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method of computer Applications department of course MCA (Master of Computer Applications) from session 2007 to 2010. Initially size of the data is 50. In this step data stored in different tables was joined in a single table after joining process errors were removed.

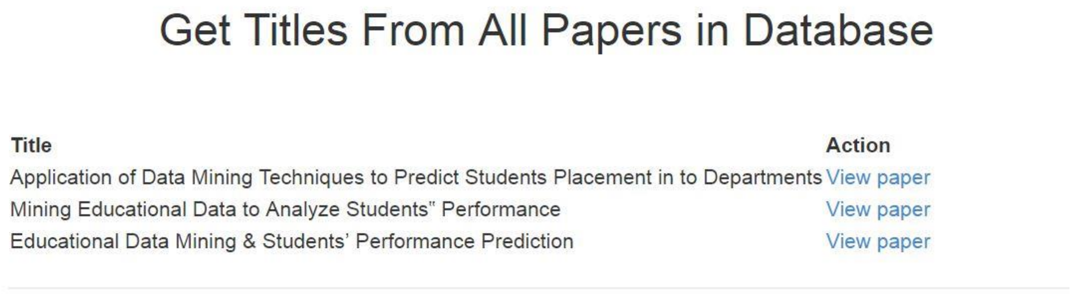
Data selection and transformation :

In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table I for reference

Figure 5.1.2: Classified Contents Separated from Paper

5.1.3 EXTRACTED INFORMATION STORAGE IN DATABASE

The papers that are uploaded and successfully stored in the storage are to be shown in a list with all their titles and users can have options to view the papers in specific output model.



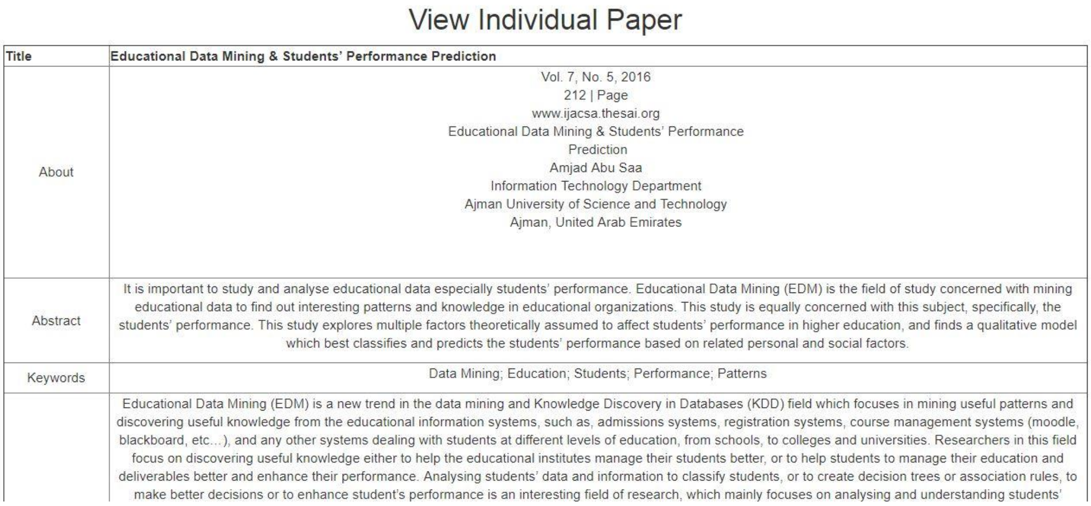
Get Titles From All Papers in Database

Title	Action
Application of Data Mining Techniques to Predict Students Placement in to Departments	View paper
Mining Educational Data to Analyze Students" Performance	View paper
Educational Data Mining & Students’ Performance Prediction	View paper

Figure 5.1.3: Extracted information storage in database

5.1.4 OUTPUT MODEL OF INFORMATION

For every paper in the storage the view option is always open. And the papers are viewed in separate portions that proves the mining successful.



View Individual Paper

Title	Educational Data Mining & Students' Performance Prediction
About	Vol. 7, No. 5, 2016 212 | Page www.ijacsa.thesai.org Educational Data Mining & Students' Performance Prediction Amjad Abu Saa Information Technology Department Ajman University of Science and Technology Ajman, United Arab Emirates
Abstract	It is important to study and analyse educational data especially students' performance. Educational Data Mining (EDM) is the field of study concerned with mining educational data to find out interesting patterns and knowledge in educational organizations. This study is equally concerned with this subject, specifically, the students' performance. This study explores multiple factors theoretically assumed to affect students' performance in higher education, and finds a qualitative model which best classifies and predicts the students' performance based on related personal and social factors.
Keywords	Data Mining; Education; Students; Performance; Patterns
	Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registration systems, course management systems (moodle, blackboard, etc…), and any other systems dealing with students at different levels of education, from schools, to colleges and universities. Researchers in this field focus on discovering useful knowledge either to help the educational institutes manage their students better, or to help students to manage their education and deliverables better and enhance their performance. Analysing students' data and information to classify students, or to create decision trees or association rules, to make better decisions or to enhance student's performance is an interesting field of research, which mainly focuses on analysing and understanding students'

Figure 5.1.4: Output model of information

5.2 EVALUATION

The evaluation process to test the application in each and every step has been shown in the below table.

TABLE 5.2 EVALUATION TABLE

Test Case	Test Input	Expected Outcome	Actual Output	Result
Application test	Application test for browsers below- Google Chrome Mozilla Firefox Microsoft Edge	Success to launch the application	Application has been launched successfully	Pass
File Upload test	A research paper file in pdf format	The application accepts the file	Application has accepted the file	Pass
File Read test	Read the pdf file as input	Get the textual components of the paper	It has read the file and got all the contents	Pass
Paper Mining test	Pdf contents of a paper	Bring out the required information of the paper	It has got required the information	Pass
Storage of Paper information test	A database query to fetch all the papers with title	All the papers in storage should show up with their titles	Papers with their titles have been got from the storage	Pass
Read a paper in required output model from database	A database query to fetch a paper and show in	The paper should be shown in separate	The paper has been shown in separate portions such as title,	Pass

	specific output model	portions with identical terms	abstract, keywords, introduction etc.	

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE

RESEARCH

6.1 SUMMARY OF THE STUDY

In our study we have successfully built an algorithm to mine the textual contents of a paper with specific format. The algorithm works for the files in pdf format. For the time being the algorithm mines on research papers. All the resources got from the paper are stored in database afterwards mining. From the storage the application can then fetch any paper directly. Or it can also fetch any paper with their keywords or title. Also it can fetch every portion of a paper with some identical terms.

6.2 CONCLUSION

We can conclude the project with positivity that a system that can actually mine a document is a solution to numerous problems in the real world. Especially in our era when the amount of data in increasing day by day. Technology has given us the opportunity think and work with not only specific region but to work internationally. When the whole globe is considered, data come to an end with no limitations. And through these process mining on documents would be a greater help to this problem. The algorithm or the system we built can be a solution to the research communities or organization for their conferences. Also by modifying the algorithm a little bit the system can serve an automated review solution system for documents. For example, the system would be a solution to review the job applications received in large companies. Automatically they can sort the applications, categorize them or to eliminate them. This could be a time and cost consuming step in these typical fields.

6.3 IMPLICATION FOR FURTHER STUDY

The whole project was an experiment to solve the problem of facing the great data expansion. Nowadays we are working with the people worldwide. The more our work area would extend, we would need to work with more data. In those data documental data are included.

In our project we had work on only several research paper formats which is a limitation indeed. But the most challenging limitation for us is to mine the images, tables, flow charts and mathematical terms.

We have further plans to work in future. Some of them are-

- To make the algorithm capable for popular formats of research papers.
- To work with images, tables, flow charts within the documents.
- Make a template which will help the job applications sorting.
- To expand the format of files from pdf to all others such as doc files, text files etc.

REFERENCES

[1] iTextSharp .NET Library, available at << https://sourceforge.net/projects/itextsharp/>>, last accessed on April 2, 2018.

[2] Easychair, conference management system, available at << https://easychair.org/conferences.cgi >>, last accessed on April 4, 2018.

[3]. Reading contents from files in C#, available at << https://www.c-sharpcorner.com/blogs/reading-contents-from-pdf-word-text-files-in-c-sharp1>>, last accessed on September 14, 2017.

[4] Read or Create advance PDF report using iTextSharp in C# .NET, available at << https://www.codeproject.com/Articles/686994/Create-Read-Advance-PDF-Report-using-iTextSharp-in >>, last accessed on September 13, 2017.

[5] Text formatting with iTextSharp, available at << https://stackoverflow.com/questions/6882098/how-can-i-get-text-formatting-with-itextsharp >>, last accessed on September 15, 2017.

[6] Change font size of pdf document, available at << https://www.codeproject.com/Questions/742926/How-to-change-font-size-of-pdf-document >>, last accessed on September 20, 2017.

[7] Getting started wi`th iText tutorials, available at << https://developers.itextpdf.com/tutorials >>, last accessed on September 2, 2017.

[8] iTextSharp.text NameSpace, available at << https://afterlogic.com/mailbee-net/docs-itextsharp/ >>, last accessed on October 7, 2017.

[9] Working with existing pdf, available at << https://developers.itextpdf.com/examples/itext-action-second-edition/chapter-6 >>, last accessed on October 14, 2017.

[10] Read and extract searched text from pdf file using iTextSharp in ASP.Net, available at << https://www.aspforums.net/Threads/132819/Read-and-extract-searched-text-from-pdf-file-using-iTextSharp-in-ASPNet/ >>, last accessed on November 23, 2017.

[11] How to search in pdf and extracts results in C#, available at<< https://bytescout.com/products/developer/pdfextractorsdk/how-to-search-in-pdf-and-extract-found-with-pdf-extractor-sdk-in-csharp >>, last accessed on November 23, 2017.

[12] How to find text in pdf and get coordinates in ASP.NET, available at << https://bytescout.com/products/developer/pdfextractorsdk/find-text-and-get-coordinates-pdf >>, last accessed on November 28. 2017.

APPENDIX

Through this project the first challenge was to read the PDF files and extract the texts with their properties. For this we needed to use a third party library. But it was very complex to determine the text properties such as font, font size or font style etc. from the library directly. We needed to modify the whole library to a necessary form.

Also whenever we had started to extract text from research papers, we got to discover that most of the papers are not written by abiding the rules. There were differences within the fonts or styles. Even though we had worked on some papers from same publishers, but found differences between those papers in writing rules and styles. We got to select some papers separately for our purpose.

Our project is primarily a research project. But as we were working with algorithms, we tried to show the output of the project in development of a simple website. We had only implemented the targeted functionality of the project.

In consideration with the mining of the texts, we did not work on the figures, tables and mathematical terms. It was not necessary for our initial step but those data made some complexity through mining the classified properties.

In spite of these problems we did get our primary outcome from our result. We mostly had solved the problems by research more and through programming.