A Case Study in the Design of a Source Extraction System
Pranav Raikundalia, Abdullah Mitkar, Shrikant Jagtap, Prof. Uday Joshi
BE Student, Associate Professor
Department of Computer Engineering
K. J. Somaiya College of Engineering, Mumbai, Maharashtra, India

Abstract:
The systems prevailing in the world today consists of character recognition systems and tools that aid in finding text on the internet. But to our knowledge, there is no system that uses a combination of both of these systems and provides services to the user which is required at various places. Our system aims at performing character recognition and finding the source of the text. The system takes image document as an input, processes the image and extracts all the text it can from the image. It then searches the internet for the source of the text and determines whether the text is a part of a technical paper, book, news article or anything of that sort.

Keywords: Character Recognition, finding text, documents, image.

I. INTRODUCTION

Many-a-times, a person finds a piece of page torn from a book or a news article and finds it interesting and wishes to read more about it. But since the page is part of something it does not contain information about the source of the page and also possible tells very little about the remaining contents of the original source that this text is a part of. Hence, the person wishes to find the source of the text and also the remaining contents in the source. Finding the source of such a piece that is found is very much impossible. Consider, you go to an old library and you find a piece of page torn but it contains something interesting and you wish to read more about it. But since it a library it is not possible to go searching for the book that this page fits in.

Also old newspapers are used for a lot of other purposes and if one happens to find something interesting in a piece of torn newspaper, it is very difficult to find which paper it is and on what date it was published. One option is that user can search the text on the internet and if there is a softcopy of the text then the person can find the source and read more. But the text can be very large and it is difficult to type and search the whole text on the internet. So times like this, when a person has to resort to tedious measures to find something calls for need of a system that does the person’s job of finding the source for him.

Our system reduces all the efforts involved in this entire process. It takes an image of the text and does the rest of the job for the person. It tells the person the source of the text and provides an option to the source’s soft copy available on the internet. This paper provides a methodology for implementation of such a system that combines the abilities of two currently prevailing system into one system and aims at reducing the errors and glitches in those systems.

II. RELATED WORK

A. Current work in OCR
Due to emergence of internet and virtualization, all the historic data is being converted into digital form to store virtually. There is no question of retyping the whole text again for conversion since the data is large and huge in number. So there are systems being developed that perform recognition of such documents and extract the text from them and store the text. Also there is work going on in making these systems more efficient in recognizing more and diversely written languages around the world.

OCR systems refer to the systems that convert hand-written or machine-written text into digital form suitable for storing. The systems prevailing generally use the traditional paradigm for creation which includes grey-scale conversion, binarisation, character extraction and character recognition. There are a lot of available OCR applications. They are used in invoice imaging, legal industry, banking, healthcare, CAPTCHA, automated number recognition and handwriting recognition.

B. Current work in text finding
There is no such thing as text finders on the internet to our best of knowledge. But there are so many search engines that can be used for finding text on the internet and finding the source of the text.

It is as simple as typing or copying the text from one source and pasting it in the search bar and searching on the internet. The search engine then displays the text, it various source and sometimes a gist of the entire text. Hence searching the text on the internet can be helpful in finding the source provided that the text is available in soft copy. Hence, this system attempts in reducing this gap of typing the text or copy/pasting the text by simply taking a picture of the entire text.

III. PROPOSED SYSTEM

Normally, people tend to use the Internet to gain a wide range of information but this software or web application will give the user exactly what it is looking for. As explained in the abstract, the user may find some extract from a particular paperback or maybe some newspaper, and he/she is not able to guess where it is from.
The system that we developed will help them in the same. The output of this system will return the exact location or link of the extract if available on the internet. This reduces the effort of users where they need to type in the words, instead they upload a picture. Also it reduces the effort the user has to take to check and scroll down all the links given by Google, Yahoo or any search engine for that matter.

The software is a solution to two major problems:-

A. Converting Image to Text

This technique which is commonly known as Optical Character Recognition is used to identify the text in the image and store it in a document. This is the first step to our system’s implementation. The system has a two-step confirmation process of the text that is converted.

First the image is uploaded and the text in the image gets stored in a document. This document is displayed to the user to confirm that it is the appropriate and accurate text from the picture. Secondly, the user can also add his own keywords or correct some text that is misinterpreted by the computer. Optical Character Recognition using MATLAB is around 90% accurate, so it might give wrong text interpretation to some extent.

This might happen rarely if the text is clearly visible in the picture. So the user has been given the ability to correct some text if it does not match the text in the picture.

B. Searching for links

This function of the system is invoked automatically as and when the user confirms the text. This confirmation invokes the JAVA API and programs for text searching and extraction. Web mining algorithms are used for searching the text and extracting it from the web pages.

[Diagram of use case for user input, image to text]

IV. SYSTEM ARCHITECTURE

OCR algorithms were studied and we came to a conclusion that OCR till date is best implemented using MATLAB. The efficiency and accuracy of MATLAB is more than any other individual algorithm used to develop an OCR application. So, OCR was developed using MATLAB libraries in Python. Correlation algorithm for Relevance ranking was modified by us to identify the link that has the most relevant data. It can also be said that in this algorithm we extracted the data from the links and matched the string with our string to get one link that is the most visited, and appropriate link. This algorithm eased gives a single link as an output and all other links are being neglected.

Input: Relevant links in an array list and string to be searched

Output: Best link

Step 1: Extraction of data from each link

Step 2: Store data in array list

Step 3: Create relevance Counter ‘RC’ and initialize it to 0

Step 4: Compare the Strings with the original string to be searched and increase counter accordingly

Note: comparing should resume from where the last searched word had been found. This increases the efficiency because the order of searched words remains the same and also reduces the time for comparing.

Step 5: Save the RC to an array where the RC of a link is stored in the same offset as the link.

Step 6: Find the offset of the highest RC and return the offset.
V. COMBINING WORKING COMPONENTS

Final step before testing of the software. Combine all the individual processes developed to fulfill the scope requirements. Java is compatible with PHP and so is Python, therefore, PHP is used as an intermediate language to run JAVA and Python programs using Batch files. MATLAB is run through Python and is also compatible on the web when installed on the server machine.

VI. FUTURE WORK

A lot of data today is available on the internet in digital form. Hence, if we have a soft copy then finding source is easy. However, there may come times when the source is not available on the internet. Hence, in such times what the system can do is provide at least the topic of what the extracted text is about. This can be done by applying text-mining algorithms on the extracted text and performing topic search on it using something as simple as word count or as complex as any NLP text-mining algorithm.

VII. DISCUSSIONS AND LESSONS LEARNED

Many new alternatives to web mining approach have been studied. Also a lot of libraries of JAVA and python coding languages have been discovered and learned by us during the course of the project. Python, a new programming language and one of the most used in the world of programming was learned from scratch by us. An overall view of planning and carrying out tasks as a team was experienced and hence, the project was successfully completed.

VIII. REFERENCES

[1]. (A SURVEY ON WEB MINNING ALGORITHMS 1, Sandhya(M.tech CSE), 2,Mala chaturvedi(M.tech CSE) 1,Jayoti Vidyapeeth Women’s University ,Jaipur 2,Manav Bharti University ,solan.

[2]. Relevance Ranking and Evaluation of Search Results through Web Content Mining G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi. Using Probabilistic Latent Semantic Analysis for Web Page Grouping Guandong Xu1,3, Yanchun Zhang1, Xiaofang Zhou2 1School of Computer Science and Mathematics Victoria University, PO Box 14428, VIC 8001, Australia {xu,yzhang}@csm.vu.edu.au 2 School of Information Technology & Electrical Engineering University of Queensland, Brisbane QLD 4072, Australia zxf@itee.uq.edu.au 3School of Computer Science & Engineering Wenzhou University, Wenzhou 325003, China xgd@wznz.zj.cn