Design and Implementation of OCR Text Correction by Crowdsourcing on Historic Newspaper Archive

Student Name: Megha Gupta
IIIT-D-MTech-CS-DE-13-MT11024
Indraprastha Institute of Information Technology, New Delhi

Advisor: Dr. Haimonti Dutta*
Co-Advisor: Manoj Pooleery*
*Affiliated to The Center for Computational Learning Systems,
Columbia University, New York

Submitted in partial fulfillment of the requirements
for the degree of MTech in Computer Science
with Specialization in Data Engineering

July 21, 2013

1 Abstract

Newspapers are the first draft of history, they have always been a rich source of information for historians, lay researchers, scholars, etc. With the advent of digitized newspapers, its accessibility to the general public has increased. The usability of these old historic newspapers has tremendously increased owing to OCR devices. However the text generated from the OCR devices is often garbled, affecting the efficiency of the retrieved results.
Our goal is to create a web based project that allows a patron to read and edit the articles. The task for editing the electronically translated newspapers is crowdsourced. This crowd based project would help in correcting the OCR errors thus improving the efficiency of retrieved results.
This report describes the OCR Text Correction System, its functions and workflow in section 2 and 3. Section 4 and 5 elaborates on its design model and System specifications respectively. Implementation details with real screenshots and the challanges faced during development are mentioned in Section 6.

Contents

2 Introduction

2.1 Problem Description

This is a web based crowdsourcing project where patrons can edit the electronically translated or OCR-ed text of old historic newspapers present in the holdings of California Digital Newspaper Collection (CDNC) [1]. The OCR-ed text is often garbled due to inaccurate OCR devices. The accuracy of a device depends on the following factors: variations in the quality of input paper, size and style of font, column layout, etc. In our case, the quality of input paper is very poor as the newspapers date back to 1846. This adversely affects the retrieval effectiveness, hence the techniques for cleaning the OCR [2] need to be improvised. Often such techniques involve laborious and time consuming manual processing of data. By correcting the garbled text, we will be able to improve the retrieved results thus making our users satisfied.

2.2 Scope & Objectives

The objective of this project is to build a software that is usable, intuitive, simple and functions well consistently. It utilizes the "Wisdom of the Crowd"to gather information about the historical newspaper articles and magazines, and utilize Machine Learning for further analysis of data.
Our main focus here is on developing a web interface equipped with a text correction tool for enabling the users to correct the OCR errors as they come across them. By crowdsourcing this project, we accomplish more with fewer resources and build a user community in process.
The software is responsible for a host of functions. Its basic functions are searching and displaying images of issues, search by issue name and date range, user management functions, text correction functionality.

2.3 Software Functions

The project functions can be categorized into the following:

- User management functions
  These encompass all the activities to manage the users for the application. These include:

1. Registration module: Ability to enter and validate user demographic information
2. User Security module: Ability to authenticate and authorize the user
3. User tracking module: Ability to track user activity, prepare statistics (how long users spent on each module, how many articles they corrected, participation in discussions etc)

- Article related functions
  1. Articles search: this provides the capability to retrieve an article based on a set of criteria like article topic, date range etc.
  2. OCR correction: this interface provide users with the capability to correct the OCR displayed along side a given article image. A basic level of authentication (based on a CAPTCHA model) will be required before the user can correct the text.
  3. Tag specification: this interface will allow the user to specify tags corresponding to a given article. This functionality will also require a basic level of authentication.

- Application administration functions

3 Software Project Plan

This project was divided into three phases:
Phase 1 included laying out UI components, database connectivity, searching and displaying user requested issues with their images.
Phase 2 included basic functionality text correction tool, user management functions.
Phase 3 included advance searching and functionality of tool.

3.1 Workflow of the System

- In Phase 1, the first task was to design and build the layout of the interface. The user interface components in Vaadin can roughly be divided in two groups: components that the user can interact with and layout