The University of Hong Kong

Faculty of Engineering

Department of Computer Science

Final Year Project Interim Report

Web based tool for Chinese character evolution

Supervisor: Dr. Vincent Lau

Name: Matthew Chan Tsz Ho

University Number: 3035269368

Date of Submission: 20/1/2019

Abstract

Chinese characters have a very long history. Each Chinese character has a unique font, and some may have a totally different appearance in the past. There are different Chinese dictionaries on the internet. However, the dictionaries are hard-coded and do not support user-contribution. The quantity of Chinese characters and words is too large that it takes too much effort to make further changes to the database by a single team. Therefore, the idea of user-contributing content is introduced. Users can make changes to existing content or contribute new content to the database. The main purpose of the project is to explore the possibilities of applying the user-contributing-content idea to a web-based tool of Chinese characters. An online dictionary with a dynamic and flexible database will be constructed and maintained by different users. Different special features are provided to the users to improve their user experience.

Acknowledgment

I would like to thank Dr. Vincent Lau for giving me this opportunity to construct a website. He provided me with many suggestions and help. I would also like to thank HKU CS department for lending the CS server to me. Finally, I would like to thank HKU for providing different courses to help me learn many things about computer science.

List of Figures

List of Tables

Abbreviations

NoSQL-not only SQL

PHP- Hypertext Preprocessor

CSS- Cascading Style Sheets

Table of Contents

1. Introduction

1.1 The Evolution of Chinese Characters

Chinese characters have about 5000 years of history[1]. 5000 years ago, people started using pictures to symbol what they saw. Those pictures are called “oracle bone script” nowadays. Oracle bone script is significant in the study of Chinese history and culture. For example, what a cart looked like in the past can be known by studying the word cart in oracle bone script[2]. Every change in Chinese character may represent a significant change in political power in the history of China. For instance, when the Qin Dynasty united the six kingdoms, the ruler formulated the "small seal script". The evolution of Chinese character makes an essential contribution to Chinese history so the data of it shall be well preserved. A platform is needed to store all the data.

Chinese characters are not merely pictures in the modern era. They have straight rules to be written. For example, the order of stokes of every Chinese character is well defined. On the other hand, the pronunciation of Chinese characters is also special that there may be more than one pronunciation for each word depending on the situation or way to use it. Currently, there are several platforms that can perform the task of online dictionaries or databases. However, those platforms do not have the function of user-contributed content. While there are 4000 Chinese characters and 15000 Chinese words commonly used in Hong Kong[2], updating or adding the data Chinese character by a single group requires much effort. If the idea of user-contributed content is implemented on the database, the information in the database can be more up-to-date and the quantity and coverage of data can also be increased.

1.2 Web-based tools for Chinese characters

To store a large amount of data of Chinese characters, this project aims to create a web-

based tool which performs the function of storing the data of Chinese characters with the idea of user-contributed content implemented on it. The basic part of Chinese character, for instance, components, structure, phonetic code, strokes and words will be covered. The data of Chinese Character will be stored in the database and will be displayed to visitors of the website. Visitors can perform several functions including searching the words, exploring related characters and words. Visitors can also register as a member and their preference of using the website can be set. The website also supports multi-media that videos and pictures can be uploaded and displayed on the website. The central management system allows the user to contribute and edit the data of the website. To make the website user-friendly, a responsive layout will be designed that the website has different layouts in desktop and mobile phone in order to make sure the website functions well in different devices.

A suitable database will be designed to suit the use of the website. Not only will the database be able to be supplemented by different users, but it will also be able to expand its fields. To facilitate the efficiency of inputting many Chinese characters, the website allows the reading of text file and the data will be put in the database. The website will construct web pages based on the data in the database. Besides reading the text files, there will also be a page that allows users to contribute their data.

1.3 Outline

The remaining parts of this report are as follows. Section 2 of the report will be about the background information of the project. Section 3 of the report will be about methodology. The technology of implementing the website will be covered. Section 4 of the report will be about the current progress of the project. Difficulties will be followed as section 5. Section 6, which is the last section, will be the conclusion and further work that need to be done.

2. Project Background

2.1 Objective

This project's objective is to apply the concept of user-contributed content on the website. "User-contributed content" means the data of the website is not hard-coded. They are stored in the database and users of the website can contribute data to it. After users contribute data to the database, the backend of the website will dynamically generate a front-end webpage according to the data in the database. An advantage of using user-contributed content is that it can allow public access to the database, the data source and dataset can be extended. It also takes less time to modify the data as anyone can change the database if something in the database is found to be wrong or something is needed to be added.

2.2 Related Websites

There are several existing online platforms that servers the use of a Chinese dictionary.

2.2.1 "Lexical Items with English Explanations for Fundamental Chinese Learning in Hong Kong Schools"

"Lexical Items with English Explanations for Fundamental Chinese Learning in Hong Kong Schools" can search by direct character, radical, strokes and pinyin[3]. There are also several ways of sorting the result. Flash is used to show the stroke sequence of the character. It does not support responsive layout.

2.2.2 "Chinese Character Database: With Word-formations"

"Chinese Character Database: With Word-formations" is a very detailed online dictionary that includes homophone, pronunciation and word example, etc.[4]. Users can search by the exact character or pronunciation. Waveform Audio File is used to play the pronunciation of the character. It does not support responsive layout.

2.2.3 Wikipedia

Wikipedia is also taken as reference for its user-contributed content system[5]. There are links in the paragraphs, and they will link to related pages. Users can log in and edit the existing pages and can also create new pages. Before a user edits anything, he can log in or his IP address will be shown on the page. The previous contribution of a certain user or IP address can be inquired. Responsive layout is implemented.

2.3 Current inadequacy

Currently, the online Chinese dictionaries available on the internet do not support user-contributed content. The websites are static, only the administrator of the dictionaries can modify the data. When there is something needs to be changed or some new data is needed to be inputted to the database, much effort is needed. For example, in the website “Chinese Character Database: With Word-formation”, if an ordinary user wants to add a new “word example” to one of the characters, he may need to fill in the “opinion form” and send to the administrator of the website. Then the administrator may need to contact the technical team to modify the database.

Chinese character may still evolve in the future, more data may need to be inputted into the database. Therefore, a new system should be implemented to improve the online dictionaries.

2.4 Improving the situation

To lighten the man power used to modify the data in the database, the project will implement the idea of “user contributed content”. The content of each character in the website relies on the contribution of the users. Every user can make modification to the

data. It can increase the number of sources of the website and the efficiency of updating the website.

3. Methodology

The main programming languages required to construct the web tool are Node.js, CSS and JavaScript. There will be a database storing the Chinese characters and MongoDB is used. Mongoose and Express are also used to help the implementation of the website. MongoDB is used because MongoDB supports a flexible database. Users can increase or decrease the number of fields of the Chinese characters if needed.

Traditionally, HTML is generated by PHP. However, Node.js is chosen to generate HTML in this project because Node.js is a powerful open source language that runs JavaScript. The advantage of using Node.js over PHP is that node.js has a wide library that supports many built-in functions. It is more convenient to build a website with Node.js that many functions are already implemented in this language. In contrast, using PHP means giving up the flexibility of using node.js. Node.js and Express will generate the required HTML for the basic of the website by using Mongoose to link to the MongoDB. CSS will be responsible for making the style of the website and JavaScript will realize the special effects. Text data will be stored in MongoDB. To improve the efficiency of loading, multimedia will be stored separately on the server and MongoDB will store the thread of the multimedia.

After data is inserted into the database, node.js will use Mongoose to retrieve data from the MongoDB. "POST" method is used to transmit the data instead of "GET" method because the data transferred is expected to be large. "POST" method is more efficient and secure than "GET" method in transferring large data. CSS will be used to place different elements in the correct position on the webpage with color and animation added. JavaScript will be

used to create special effects like sorting the search results and the login system.

Since the field of the data may be dynamically increased. The database structure will use “Types.Mixed” so it can support dynamically increasing fields. The detailed database structure is as follows:

```
var chineseSchema = new db.Schema({
        chinesecharacter: {type: String},
        field : [db.Schema.Types.Mixed]
        },{strict: false});
```

“chinesecharacter ” is used to store the URL encoded format of the Chinese character. The content of the character is stored in “field”. The data should be like this:

```
[ { field: [ [strokes:2, radical:”二”] ],
        _id: 5c31c33df1210638181967f2,
    chinesecharacter: '%E4%BA%8C',
        __v: 0 } ]
```

The above means the URL encoded format of the Chinese character stored is '%E4%BA%8C'. The strokes of the character is 2 and the radical is “二”. It also has an unique id like other data in MongoDB.

4. Current Progress

4.1 Layout

At the current stage, the layouts of the main pages are created.

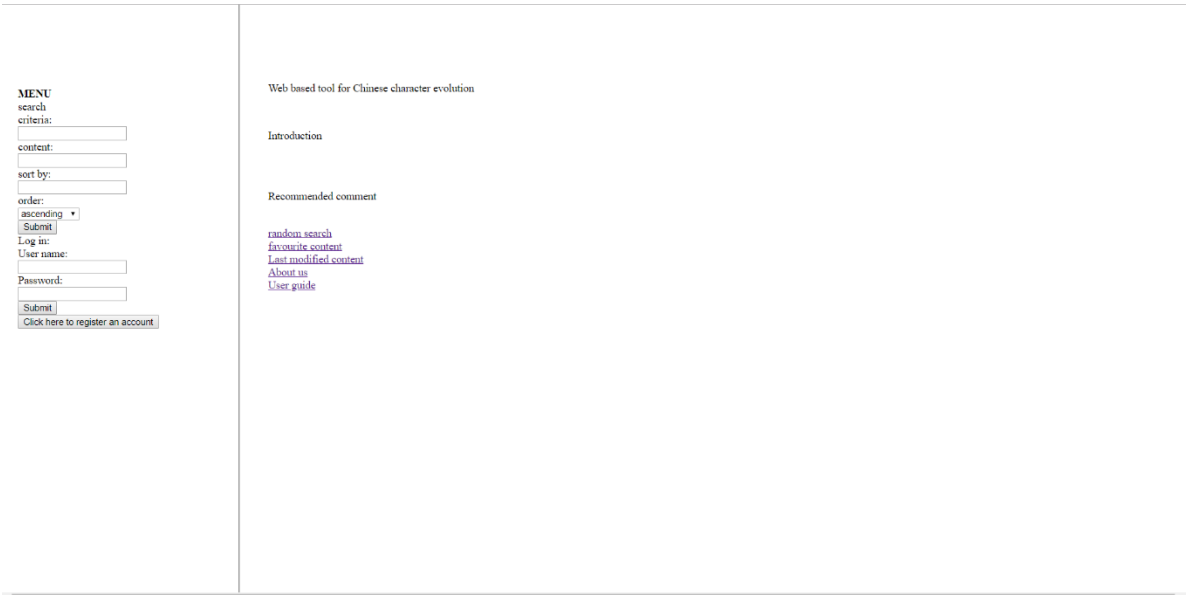

Figure 1. Layout of the index page in desktop computers.

There is a menu on every page of the website. Inside the menu, users can perform different functions. For example, the user can search a specific Chinese character through the menu. Also, there is a log-in form that users can log in to the website through it.

On the right-hand side of the index page. There is a brief introduction to the website. There are also some hyperlinks that will direct the user to different pages of the website. For example, when the user clicks on the “random search link”, he will be redirected to one of the existing Chinese character pages randomly.

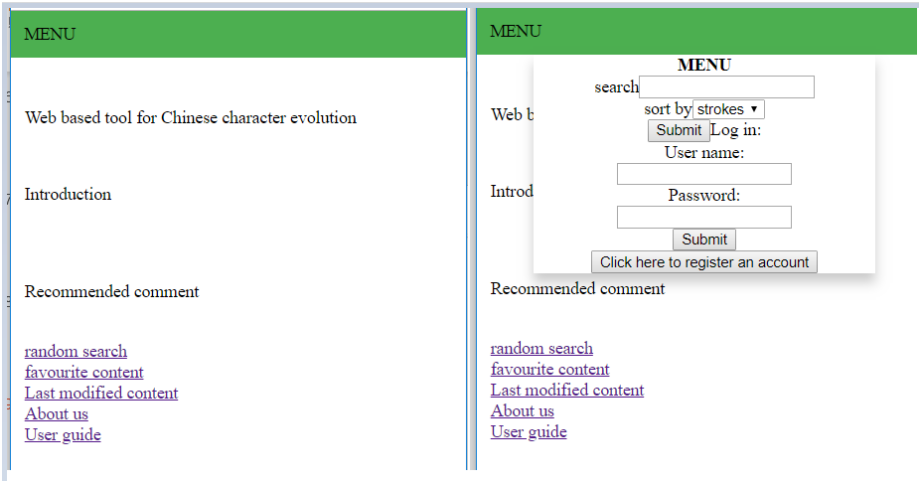

Figure 2. Layout of the index page in mobile devices.

In the mobile devices, the website works the same as in desktop computers except the menu is hidden. When the user clicks on the green menu bar, the menu will pop out and the user can perform the same functions as in desktop computers.

4.2 Input page

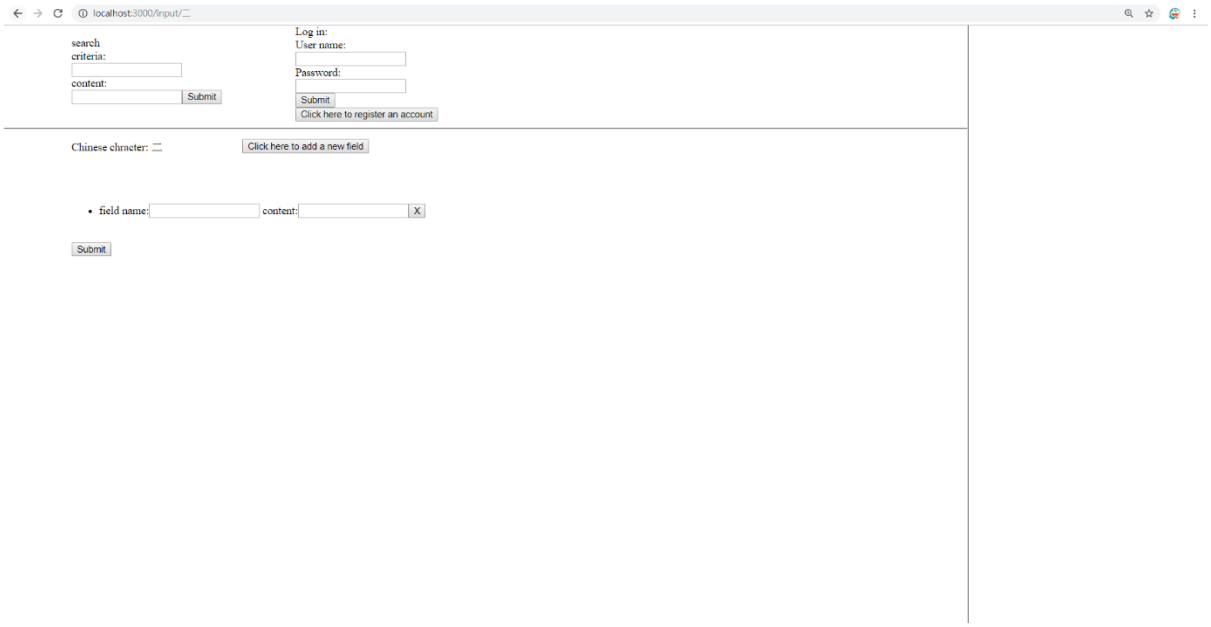

Figure 3. Input page of the character “二”.

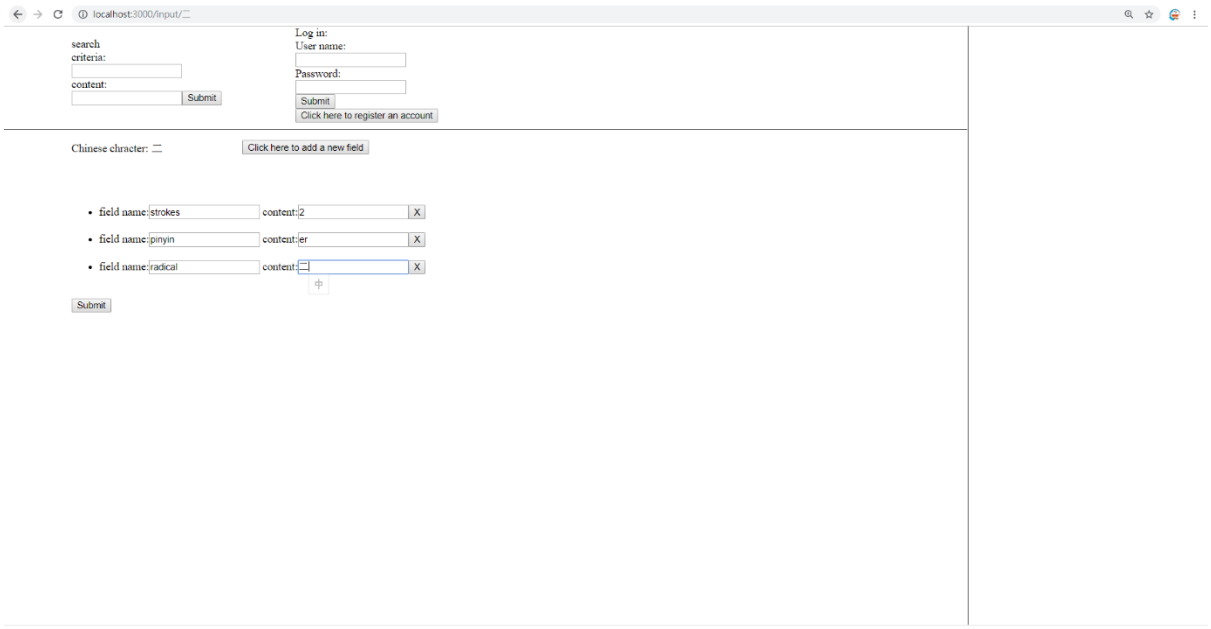

Figure 4. Input page of the character ”二”. User clicked the “add field” button and filled in the form.

Users can input data into the database. The website will check the parameter of the URL of

the input page. For example, if the parameter of the page is “二”, then the page will allow users to input data for the character “二”.

Inside the page, the user can declare the field name and the content of the data. Also, the user can input multiple fields. There is a button “click here to add a new field” on the page. When the user clicks on the button, a new input bar will be created. After the user fills in the form and clicks submit, the data will be transmitted to the backend of the website using “POST” method. The backend of the website will first check if this Chinese character already exists in the database. The character will be updated if the Chinese character already exists in the database. Otherwise, it will create a new Chinese character object and insert to the database.

4.3 Main page

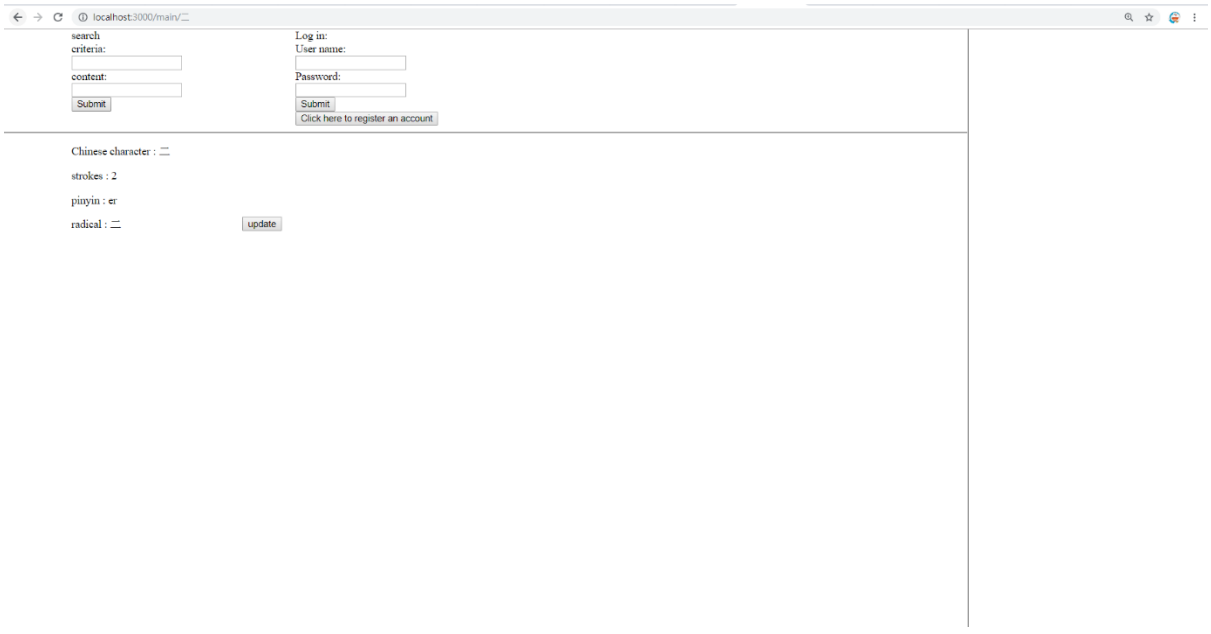

Figure 5. Main page of the character “二”.

In the main page, the webpage retrieves data from the database and output to the users. Like the input page, the webpage will check the parameter of the URL of the main page. For example, if the parameter is “二”, then the website will search the database to see if the

character “二” exists in the database. The page will retrieve the data and output to the user if there is a match. The bottom left corner displays text data, while the right-hand side of the page is reserved for pictures, sound, and videos. If the user wants to edit the page, he can click the “update” button and he will be redirected to the input page of the character.

4.4 Search system

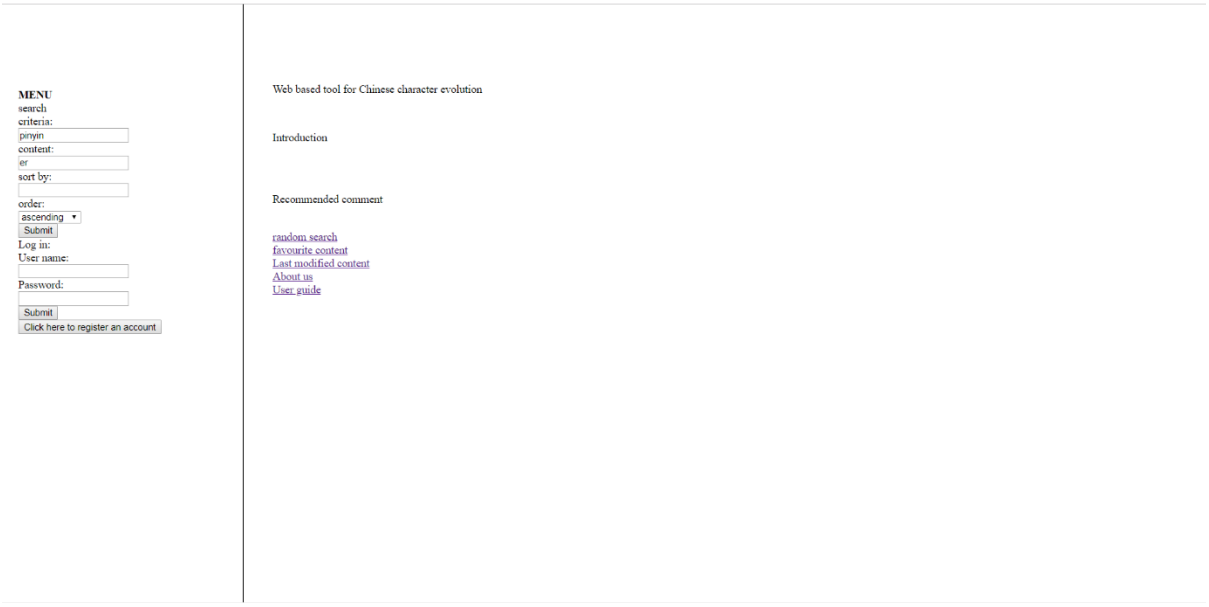

Figure 6. Index page. User is trying to search characters with pinyin “er” in the menu.

Figure 7. The search results that matches pinyin “er”.

The webpage can search and display the results according to what the users want. In the

menu, there is a searching form. After the user fills in the form and clicks submit, the website will search the database and return matching data. Sorting is optional. If the user does not need the result to be sorted, he can leave the field blank. For example, the user wants to search characters with pinyin “er”. He can type “pinyin” in the criteria field and “er” in the content field. The sort-by field is left as blank as the user does not need the result to be sorted. After he clicks submit, the website will return the result as shown as figure 7.

In the search result page, the results will be shown at the bottom right corner. The user can click on the character and he will be redirected to the corresponding page. Moreover, the user can view the number of data fields inside each of the character in the search result page.

4.5 Sorting

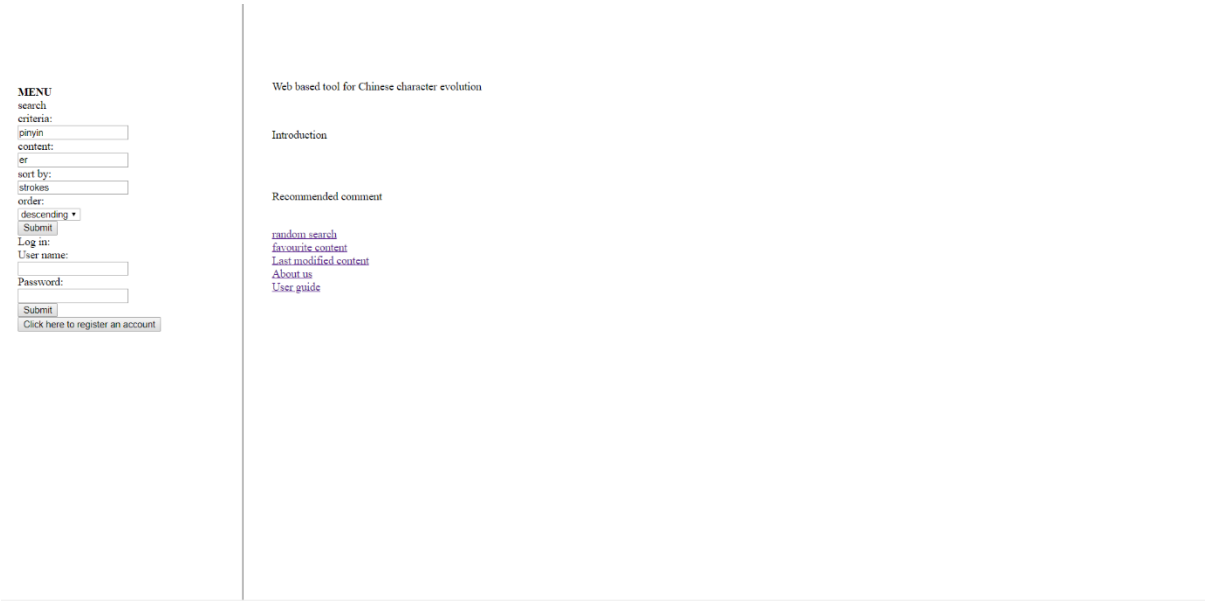

Figure 8. Index page. User is trying to search characters with pinyin “er” in the menu and he wants the results to be sorted by strokes in descending order.

Figure 9. The search results that matches pinyin “er”, sorted by strokes in descending order.

Sorting is allowed during each search. The user can decide how to sort the result and the order of the result. For example, the user wants to search the characters with pinyin “er” while he also wants the result to be sorted by the number of strokes in descending order. He can first fill in the searching form. Then he can type “strokes” in the “sort by” field and choose descending order. After he clicks submit, the result will be sorted by strokes in descending order.

4.6 Login System

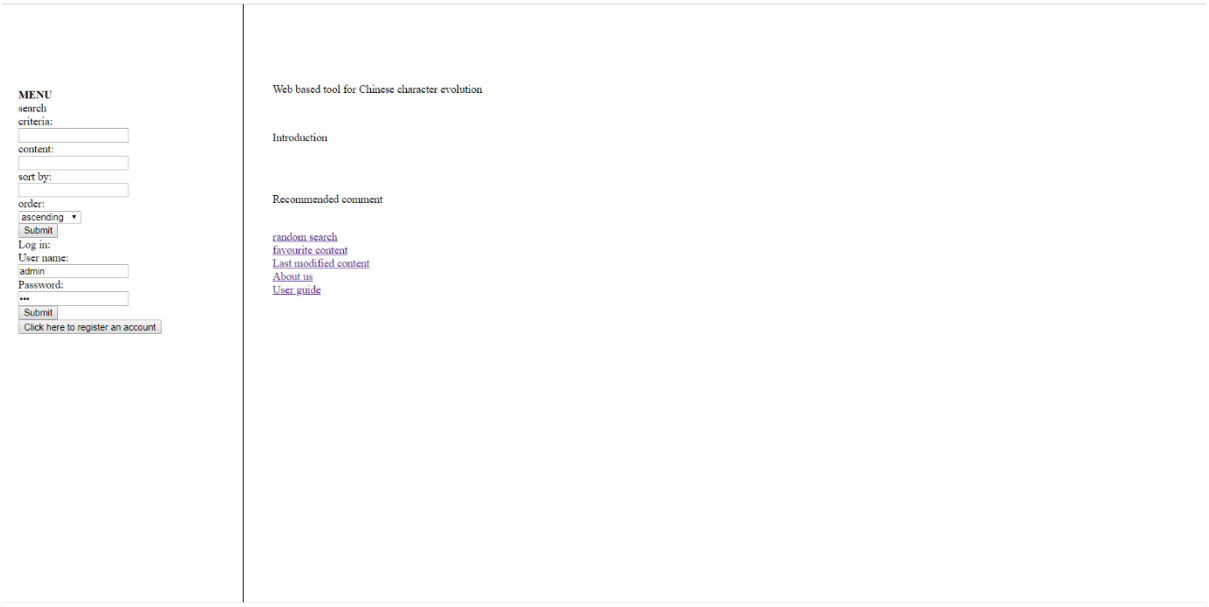

Figure 10. Index page. User is trying to log in.

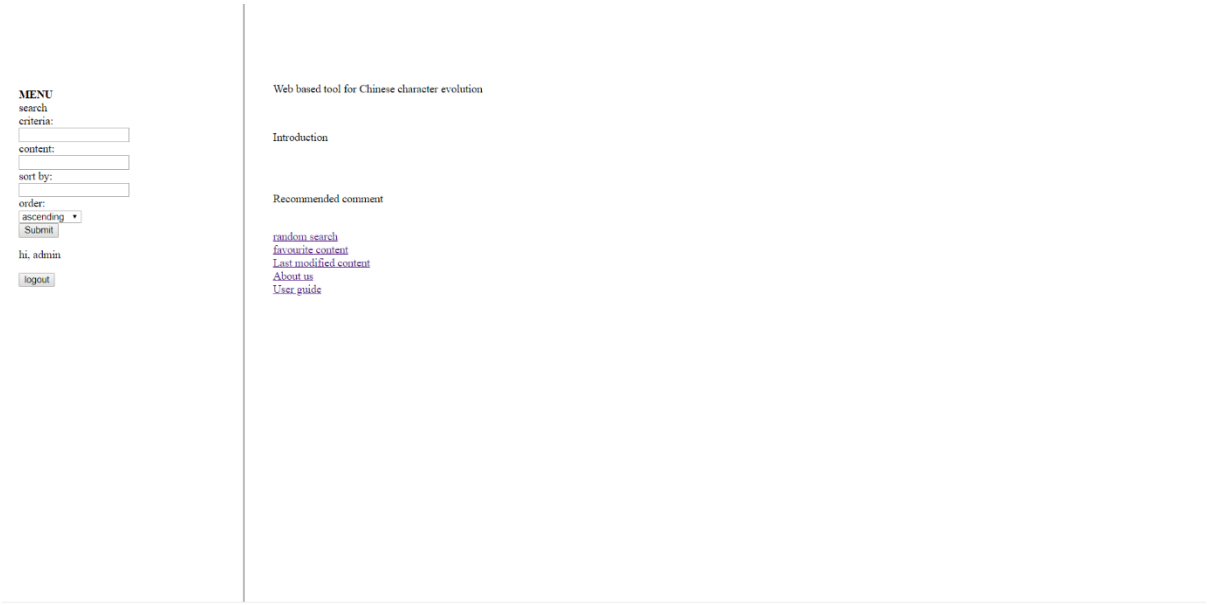

Figure 11. Index page. User has logged in.

The user can register as a member of the website. If the user does not have an account, he can first register as a member in the register page. In the register page, the user can decide the user name and password. The backend of the website will check if the user name already exists. If it is a legal user name, a new user account will be created.

Then the user can log in through the menu of any page of the website. Once he is logged in, the login form in the menu will be changed to “hi, ‘user name’ ”. The user's login status will be saved unless he logs out from the website.

If the user wants to log out, he can click the log out button and he will be logged out. The login form in the menu will appear again.

5. Difficulties

The website uses “GET” method to recognize the Chinese character. For example, http://localhost:3000/input/二 means this is the input page of the character “二”. However, the Chinese character in the URL cannot be directly stored in the database. The Chinese character will be turned into a garbled message in the database.

To solve this problem, the Chinese character in the URL will be encoded before inserting into the database. “EncodeURIComponent” is used to encode the Chinese character to the format that can be stored into the database. When the data is retrieved from the database, “decodeURIComponent” is used to turn the format back into a Chinese character.

6. Conclusion and Future plans

6.1 Conclusion

The project aims to construct a website about Chinese character that allows users to contribute data to the database. Node.js and MongoDB is the main technology that used to construct the website. Currently, there are several existing online dictionaries. The project will take the existing platforms as references and improve their inadequacies.

At the current stage, the layout and the basic function of the main functions of the website

were designed. In the future, the project will focus on the implementation of advanced functions like uploading multimedia data and the design of special features provided to the users.

6.2 Future plans

Special features provided to the users of the website will be the future focus of the project. There may be different levels of users. Some of the users can only view the website and set their preference but he cannot edit the data. Some of the users can both view and edit the data. The highest level of users will be the administrator and they can monitor most of the users of the website.

Moreover, there may still be some minor problems in the basic functions of the website. For example, the login form in the menu does not check whether the input is valid. Further modifications are needed to improve the performance of the basic functions of the website.

Future Tasks	Expected finishing day
Improve the performance of the basic functions	Before the end of January
Preference setting of users	Early February
Edit history of users	Early to Mid February
Implementation of different levels of users. Design of different special features	Before the end of February

Table 1. Time table of future tasks

7. Reference

[1] 郭小武教授. (n.d.). 漢字的起源與變遷.[Online]. Available: https://www.chiculture.net/index.php?file=topic_description&old_id=0601 [Accessed: 2019, Jan 20].

[2] Evolution of Chinese characters. (n.d.). Omniglot.[Online]. Available: https://www.omniglot.com/chinese/evolution.htm [Accessed: 2019, Jan 20].

[3] Hong Kong Education Bureau (n.d.) Lexical Items with English Explanations for Fundamental Chinese Learning in Hong Kong Schools [Online]. Available: https://www.edbchinese.hk/lexlist_en/ [Accessed: 2019, Jan 20]

[4] Prof. Tze-wan KWAN (n.d.) 粵語審音配詞字庫 [Online]. Available: https://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/ [Accessed: 2019, Jan 20]

[5] Wikipedia (n.d.) [Online] Available: https://www.wikipedia.org/ [Accessed: 2019, Jan 20]