

The University of Hong Kong

Faculty of Engineering

Department of Computer Science

Final Year Project Final Report

Web based tool for Chinese character evolution

Supervisor: Dr. Vincent Lau

Name: Matthew Chan Tsz Ho

University Number: 3035269368

Date of Submission: 27/4/2019

## **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 a set of data to test my web-based tool. I would also like to thank HKU CS department for lending the CS and FYP 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

1. Directory of the main folders of the web-based tool...p.14
2. Desktop layout version 1...p.20
3. Desktop layout version 2...p.20
4. Mobile version...p.21
5. The registration form...p.22
6. Login form...p.22
7. Logout button...p.23
8. Basic Search...p.23
9. Search result...p.24
10. Advanced search...p.25
11. Advanced search example 1...p.26
12. Search result of example 1...p.26
13. Advanced search example 2...p.27
14. Search result of example 2...p.27
15. Advanced search example 3...p.28
16. Search result of example 3...p.28
17. Application of advanced search function (example 4)...p.29
18. Search result of example 4...p.29
19. Direct Search...p.30
20. Autocomplete example...p.30
21. Main page of Chinese character “二”...p.31
22. Input page of Chinese character “二”...p.33
23. Edit history page of Chinese word “圓滿”...p.34
24. The forum page...p.35
25. An email about watch list...p.36

- 26. Example of a valid text file...p.37
- 27. Member center...p.38
- 28. Auto hyperlink example...p.40
- 29. The evolution of the script of the Chinese character “魚”is displayed in the web-based tool...p.42
- 30. The evolution of the attributes of the Chinese character “騎”is displayed in the web-based tool...p.43

## **List of Tables**

1. The database of the web-based tool ...p.18

**Abbreviations**

AJAX - Asynchronous JavaScript And XML

CSS - Cascading Style Sheets

PHP - Hypertext Preprocessor

NoSQL - not only SQL

XML - Extensible Markup Language

## **Table of Contents**

1. Introduction...p.9
2. Project Background...p.10
3. Methodology...p.13
4. Final Progress...p.19
5. Difficulties...p.44
6. Conclusion and future work...p.47
7. Reference...p.49



## 1. Introduction

### 1.1 The culture of Chinese characters

Chinese characters have about 5000 years of history[1]. Chinese words were pictogram in ancient times. However, they are not merely pictures nowadays. They have straight rules to be followed. For example, the order of strokes of every Chinese character is well defined. On the other hand, the pronunciations of Chinese characters are 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, updating or adding the complicated data of Chinese characters by a single group requires much effort.

### 1.2 The evolution of Chinese characters

Chinese characters are evolving intangible cultural. 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 “車” 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. Even in the modern era, there may be new attributes like pinyin, Canjie code and Unicode. New words may appear, and the pronunciation and meaning of a Chinese character or word may change as time goes on[3].

### 1.3 Web-based tool for Chinese characters evolution

In the past when the Internet had not yet been invented, if people wanted to look for the data of a word or a character, they had to read the dictionary or Cihai. It is inconvenient and the data inside the dictionaries may go stale or be inadequate. To store a large amount of ever-updating data, a web-based tool for Chinese characters is needed. It is a convenient tool for people to look for the attribute of the Chinese characters. Visitors can perform several functions including searching the words, exploring related characters and words. Advanced functions can improve the users' experience. For example, visitors can also register as a member and their preference for using the website can be set. The website should also support multi-media that videos and pictures can be uploaded and displayed on the website. To keep the data updated, a user contributed content system can be implemented on the web-based tool. The web-based tool allows the user to contribute and edit the data of the website. To make the website user-friendly, a responsive layout is needed that the website has different layouts in desktop and mobile phone in order to make sure the website functions well in different devices.

### 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 final 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 create a web-based tool for Chinese words and characters. The concept of user-contributed content is applied 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 allows 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 update 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" allows searching by direct character, radical, strokes and pinyin[4]. 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.[5]. Users can search by the exact character or pronunciation. Waveform Audio File is used to play the pronunciations of the characters. It does not support responsive layout.

### 2.2.3 Wikipedia

Wikipedia is also taken as reference for its user-contributed content system[6]. 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. Moreover, the searching systems of the websites are not flexible enough. Some of the websites do not support “NOT” and “OR” query. Even if some of the websites support “NOT” and “OR” query, the searching criteria of the query are limited.

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 manpower used to modify the data in the database, the project implements the idea of “user contributed content”. The content of each character on the website relies on the contribution of the users. Every user can make modifications to the data. It can

increase the number of sources of the website and the efficiency of updating the website. The input system is designed to be easy-to-learn so more people can join the contribution. Furthermore, the search system of the web-based tool is designed for flexible queries. A complexed query with little limitation can be performed in the web-based tool. It can improve the users' experience.

### **3. Methodology**

#### **3.1 Main Programming Languages**

The main programming languages required to construct the web-based 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 and pug are 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 package library that supports many 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. JQuery will also be used to facilitate the usage of JavaScript. Text data will be stored in MongoDB. To improve the efficiency of loading and searching, multimedia will be stored separately on the server and MongoDB will store the description of the multimedia.

#### **3.2 Directory structure**

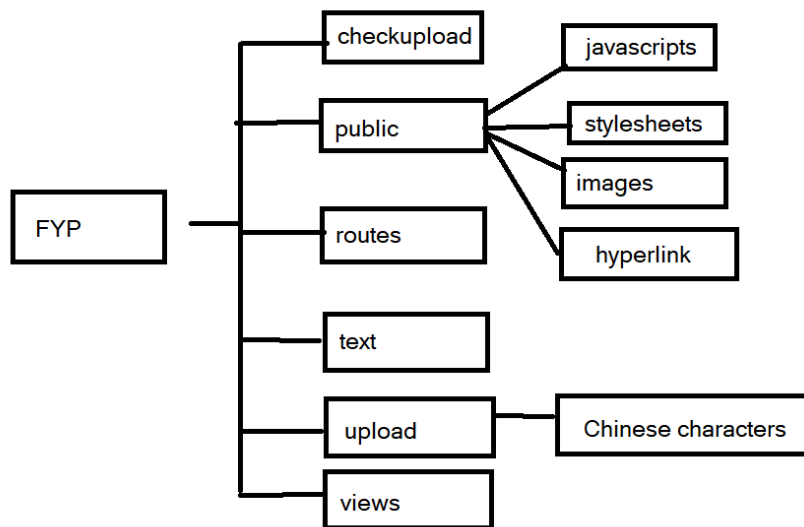


Figure 1. Directory of the main folders of the web-based tool

Figure 1 shows the directory of the web-based tool. Each folder has its own purpose. “Checkupload” is the destination of the .png image that the user uploaded to check similarity with other pictures in the site. The image inside the folder is for temporary use only and it will not display to the public.

In the “public” folder, there are four descendant folders. “Javascript” stores the JavaScript file executed by the browser. “stylesheets” stores the CSS files that decorate the website. “Images” stores the decorative images of the website. “Hyperlink” stores the XML file that will be read when the auto-hyperlink function is performed.

The “routes” stores the index.js and users.js. They are responsible to handle the frontend requests and backend requests.

The “text” folder stores the .txt files uploaded by the users.

The “upload” folder stores the images of the Chinese characters upload by the members. There are descendant folders that are dynamically generated by the system. The images are stored in the corresponding Chinese characters folders. For example, images of “二” are stored in the folder of “二”.

The “views” folder stores the pug files. They are responsible for creating the HTML structure of the front end.

### 3.3 Npm package

To facilitate the functions of the web-based tool, some npm packages are used.

#### 3.3.1 Looks-same

To search for pictures in the directories, a package is used to compare the similarity of the pictures. Looks-same is a npm package that by default it can detect any noticeable differences between .png pictures. If there is no noticeable difference, it will return a result that the two images are the same.

#### 3.3.2 Readline

When a user uploads a text file on the database, the system should be able to read the text file and input the data into the database from the text file. The package “readline” is used to read the text file line by line in a simple way.

#### 3.3.3 Nodemailer

Emails need to be sent when a character in a user’s watchlist is updated. Furthermore, when a user posted a comment in the “contact us” forum, an email will be sent to all the administrators of the website. To create and send an email object, nodemailer is used. An email will be sent after declaring the author, the target, and the content of the email using the nodemailer.

#### 3.3.4 Multer

Users are allowed to upload multimedia files, including .mp4,.mp3,.wav,.jpeg,.jpg,.png. Also, users can upload text files to update Chinese characters and words. Multer is used

to save the uploaded files in the declared position of storage and the name of the uploaded files.

### 3.3.5 Forever

To host the website even after the terminal is closed, the npm package forever is used so that the command "npm start" can be run continuously.

### 3.4 Handle request

There are two JavaScript files called index.js and users.js to handle different requests.

Index.js is used to handle front end request and users.js is used to handle back end request.

For example, when a user browses to an input page of a Chinese character, first the index.js will check the parameter of the URL and then render the input page. When a user inputs some data in a form and submits, the data will be passed to the users.js. Then users.js will update the Chinese character database, history database. It will also check whether any user has registered a watch list on the character and if it does, an email will be sent. After the whole process is completed, users.js will redirect the user to the main page, signaling the update is successful.

### 3.5 Database Structure

Name of the Schema	Purpose of the Schema	Structure	Type
Chinese	Store the Chinese characters and its field	chinesecharacter	String
		field	db.Schema.Types.Mixed
admin	Store the list of	name	String



	administrators		
login	Store the information of the members	name	String
		password	String
		yearofbirth	String
		educationlevel	String
		gender	String
Edit history	Store the time of an edit of a character and the author	Name	String
		chinesecharacter	String
		time	String
watchlist	Store the watch list of the members	chinesecharacter	String
		name	String
Email	Store the email of the members	name	String
		email	String
favoriteword	Store the preference of the word of the members	name	String
		word	String
totallike	Store the number of likes of the words	Word	String
		numoflike	Number
favoritesearchcri	Store the preference of the search criteria of the members	Name	String
		searchcri	String
favoritecontent	Store the preference of the	name	String
		content	String

	search content of the members		
favoritesort	Store the preference of the sorting criteria of the members	name	String
		sort	String
picdescri	Store the description of the multimedia files	filename	String
		wordname	String
		descri	String
forumcomment	Store the comment in the forum from the users	name	String
		hide	String
		comment	String
		forum	String
		time	String
reference	Store the reference material, hyperlink(optional) of the field	word	String
		field	db.Schema.Types.Mixed

Table 1. The database of the web-based tool

The database structure is described in table 1. Most of the types of the fields of the database are strings in order to increase the flexibility of the database's usage.

To create a dynamic database, db.Schema.Types.Mixed is used to create the field of reference and Chinese characters. These two schemas do not have a fixed field. The fields of the schemas can be dynamically increased, deleted and updated.

### 3.6 Frontend construction

To decorate the front end of the web-based tool, free online resources are used. "Font Awesome free" is used to create the icons of the website. For example, the search function of the website in mobile version uses a magnifier from the "Font Awesome free" as a representation. The decorative pictures of the web-based tool have already gained the authors' consent.

To distinct the Chinese words from normal content, free Google font "cwTeXKai" is used to decorate the Chinese words.

## 4. Final Progress

The website is hosted on the FYP server. The link towards the web-based tool is:

<http://fyp18035s1.cs.hku.hk/>

### 4.1 Front end

The front end is specifically designed to make users more comfortable while visiting the website. A responsive layout is implemented to makes the web-based tool fits mobile devices' screen. Decorative pictures will be hidden to avoid causing confusion to the visitors of the website in the mobile version.

Some line breaks are hidden in the desktop version, they will only appear in the mobile version to avoid overflow of different elements and keep the layout clean.

To keep the style of the website consistent, all the pages of the website follow the same sets of layouts.



Figure 2. Desktop layout version 1

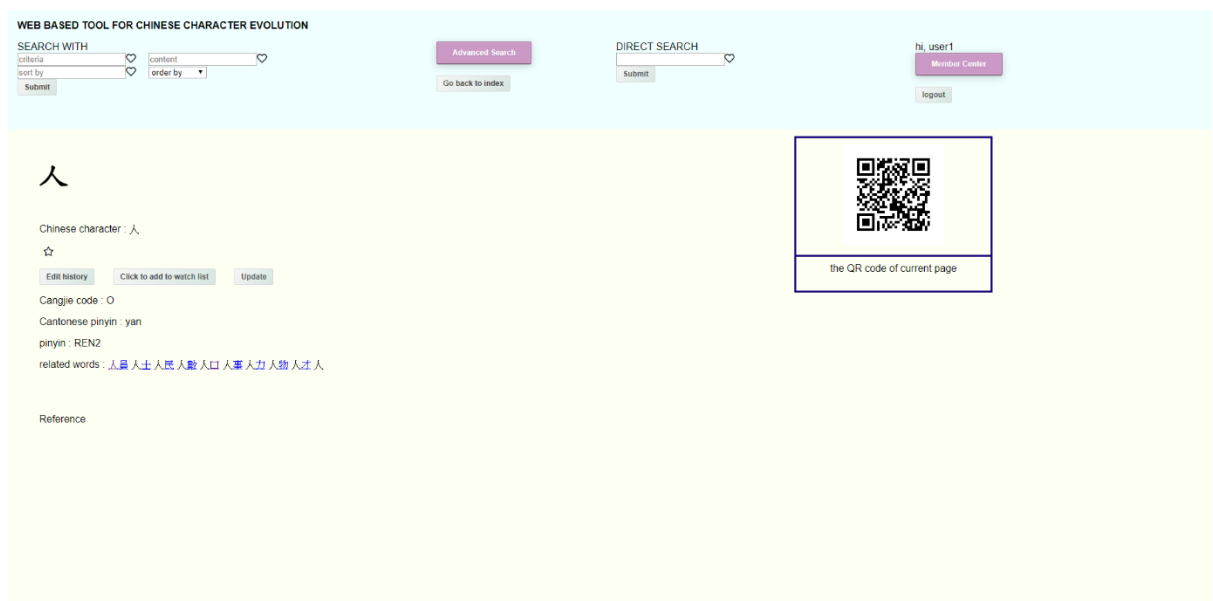


Figure 3. Desktop layout version 2

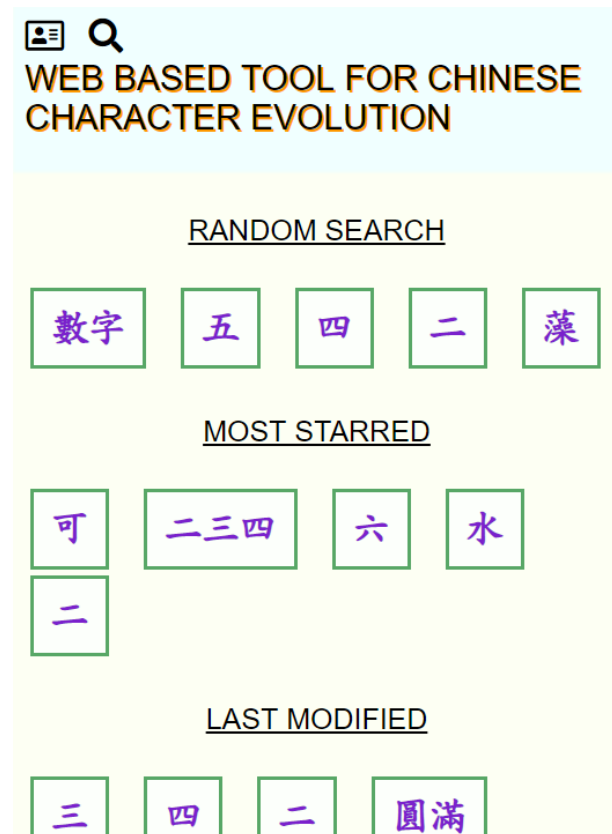


Figure 4. Layout in mobile devices

There are two layouts as shown in figure 2 and 3 in the desktop version. In version 1, there is a large decorative picture on the top of the page. The menu is placed on the left-bottom. The main content is placed on the right bottom. In version two, there is no decorative picture. The menu is placed on the top. The main content is placed on the bottom.

Figure 4 shows the mobile layout of the website. There is no decorative picture. The menu is hidden. On the top of the page, there is a large title of the website. On the top left corner, there are icons. The member card represents the login form. The magnifier represents the search form. The main content is placed in the bottom zone.

## 4.2 Main functions

### 4.2.1 Register

REGISTER

Username  
Desired Username

Password  
Your Password

Please confirm your password  
Confirm Password

Year of birth  
Year

Education level  
Primary education ▼

Email  
Email

Gender  
Male ▼

Sign Up

Back

Figure 5. The registration form

Many of the functions are only available to members only. To become a member, a user can register an account before logging in to any account. Then, a user is required to fill in a form about his username, password, and email as shown in figure 5. The data of the year of birth, education level and gender are also collected for research purpose in the future. Before submitting the form, the system will perform a validation. The user cannot submit a form that has an empty field. The user cannot register an account if his desired user name already exists. After submitting the registration form, the user will be redirected to the index page. Then the user can log in with his created account.

#### 4.2.2 Login and logout

Log in

User name

Password

Submit

[Click here to register an account](#)

Figure 6. Login form

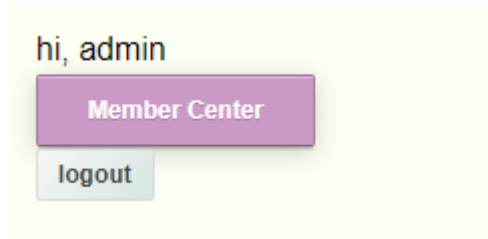


Figure 7. Logout button

The user can log in on every page of the website. A login form will appear as a part of the website on each page as shown in figure 6. When the user submits the form, the system will first check whether all fields in the login form are filled. If the form is valid, it will compare with the database and check whether the username with the submitted password exists. If it exists, the login is successful, otherwise, an error message will return to the user and he will be redirected to the index page. After the user logs into the system, he will be directed to the index of the website. The login form will disappear, instead, a button directing to member center and a log out button will appear as shown in figure 7. Session will be used to store the user's login status so the website can remember the user. After a user has logged in, he can log out whenever he wants by clicking the log out button. After he logs out, the session will be destroyed.

#### 4.2.3 Basic Search

Figure 8. Basic Search

**WEB BASED TOOL FOR CHINESE CHARACTER EVOLUTION**

SEARCH WITH

criteria

content

sort by

order by

Submit

Advanced Search

Go back to index

DIRECT SEARCH

Submit

hi, admin

Member Center

logout

Number of results : 2

Search criteria : pinyin

Search content : er

Sort by : NONE

Order by : default

**SEARCH RESULT**

兒	Number of fields : 3
二	Number of fields : 4

Figure 9. Search result

Like the login form, the basic search form appears on every page of the website. The user can fill in the criteria and content of the search as shown in figure 8. The user can also choose to sort the returned data in the order he wants but it is optional. The system will check whether the criteria and the content are filled in before submitting.

After the user has logged in, a heart icon will appear beside the input area. When the user hovers on the heart, a drop-down list containing the user's preference will pop out. The user can select one of the items inside the drop-down list, and the input field will be automatically filled in with that item.

After the user submits the form, the query will be sent to the system. The system will return all the matching results and direct the user to the search result page. In the search result page, the user can view the query he submitted, the result of the query and the number of fields inside each resulting character as shown in figure 9.

#### 4.2.4 Advanced search

##### 4.2.4.1 Summary of the advanced search



AND:

OR:

NOT:

Sort by:

Figure 10. Advanced search

The user can perform a more flexible search inside the advanced search page. Users can enter the advanced search page by clicking the “advanced search” button on every page of the website.

The advanced search allows users to input “AND”, “OR” and “NOT” field. The “AND” field means the result must satisfy all the requirements stated in the “AND” fields. “OR” fields mean that the result must at least satisfy one of the requirements stated by the user. The “NOT” field means the result must not satisfy any of the requirements. Like the basic search, the user can declare the sorting order optionally. The user can increase the number of searching requirements or leave the searching requirement empty. As an example, the user can search criteria “pinyin” and “e+” in the “AND field” and leave all other searching requirements empty. This query means to search the field “pinyin” that has one or more “e” as shown in figure 10.

To implement the advanced search function, three empty query objects are created in the back-end at the beginning. After the user submits the form, the objects will be filled with the requirements stated by the user. Then the three objects will be grouped into a single

query object and a search is performed in the MongoDB with that query object.

Before submitting the advanced search form, the system will check whether at least one of the fields is filled in. Then like basic search, it will return all the matching results and direct the user to the search result page. In the search result page, the user can view all the searching requirements he submitted, the result of the query and the number of fields inside each resulting character.

#### 4.2.4.2 More examples of the advanced search function

AND:

**add a and**

pinyin

OR:

**add a or**

criteria

criteria

NOT:

**add a not**

criteria

Sort by:

sort by

order by

Figure 11. Advanced search example 1

Number of results : 671	SEARCH RESULT												
AND													
Search criteria (1) : pinyin													
Search content (1) : ^([A-C]+)													
Sort by : NONE													
Order by : default													
	<table><tr><td>昌</td><td>Number of fields : 4</td></tr><tr><td>曝</td><td>Number of fields : 4</td></tr><tr><td>暖</td><td>Number of fields : 4</td></tr><tr><td>昂</td><td>Number of fields : 4</td></tr><tr><td>晟</td><td>Number of fields : 3</td></tr><tr><td>晁</td><td>Number of fields : 3</td></tr></table>	昌	Number of fields : 4	曝	Number of fields : 4	暖	Number of fields : 4	昂	Number of fields : 4	晟	Number of fields : 3	晁	Number of fields : 3
昌	Number of fields : 4												
曝	Number of fields : 4												
暖	Number of fields : 4												
昂	Number of fields : 4												
晟	Number of fields : 3												
晁	Number of fields : 3												

Figure 12. Search result of example 1

Figure 11 is the first example of the advanced function. A user wants to search for Chinese characters that start with one or more “A to C”. the “OR” fields and the “NOT” field are not required in this case, so they are left as blank. After submitting the form, the results will be displayed as shown in figure 12.

AND:

add a and

criteria

content

OR:

add a or

meaning

水+

meaning

草+

NOT:

add a not

criteria

content

Sort by:

sort by

order by ▼

Submit

Figure 13. Advanced search example 2

Number of results : 3

OR

Search criteria (1) : meaning

Search content (1) : 水+

Search criteria (2) : meaning

Search content (2) : 草+

Sort by : NONE

Order by : default

SEARCH RESULT

水	Number of fields : 5
藻	Number of fields : 5
草	Number of fields : 5

Figure 14. Search result of example 2

Figure 13 is the second example of the advanced function. A user wants to search Chinese characters that their meanings include “草” or “水”. “水”, “草” are returned as results as shown in figure 14. “藻” is also returned as one of the results because its meaning includes

both “草” and “水”.

AND:

**add a and**

criteria	content
----------	---------

OR:

**add a or**

meaning	水+
criteria	content

NOT:

**add a not**

meaning	草+
---------	----

Sort by:

sort by
order by ▼
Submit

Figure 15. Advanced search example 3

<p>Number of results : 1</p> <p>OR</p> <p>Search criteria (1) : meaning</p> <p>Search content (1) : 水+</p> <p>Search criteria (2) :</p> <p>Search content (2) :</p> <p>NOT</p> <p>Search criteria (1) : meaning</p> <p>Search content (1) : 草+</p> <p>Sort by : NONE</p> <p>Order by : default</p>	<p><b>SEARCH RESULT</b></p> <table border="1"><tr><td>水</td><td>Number of fields : 5</td></tr></table>	水	Number of fields : 5
水	Number of fields : 5		

Figure 16. Search result of example 3

Figure 15 is the third example of the advanced function. A user wants to search for Chinese characters that their meanings include “水” but not “草”. The result is as shown in figure 16. “草” and “藻” are excluded because the user does not want any result related to “草”.

AND:

**add a and**

pinyin ^B

Cangjie code HU+

OR:

**add a or**

Cantonese pinyin ^b

Cantonese pinyin ^p

NOT:

**add a not**

criteria content

Sort by:

sort by

order by ▼

Submit

Figure 17. Application of advanced search function (example 4)

Number of results : 3

AND

Search criteria (1) : pinyin

Search content (1) : ^B

Search criteria (2) : Cangjie code

Search content (2) : HU+

OR

Search criteria (1) : Cantonese pinyin

Search content (1) : ^b

Search criteria (2) : Cantonese pinyin

Search content (2) : ^p

Sort by : NONE

Order by : default

**SEARCH RESULT**

鼻	Number of fields : 4
癰	Number of fields : 4
邊	Number of fields : 4

Figure 18. Search result of example 4

Suppose a user forgets how to write the word “鼻” on the street. He wants to check how to write the word “鼻” by checking the web-based tool. However, he can only remember the following information, the pinyin of the word “鼻” starts with “B” and the Cangjie code of “鼻” contains “HU”. Finally, he remembers that the Cantonese pinyin of this word starts with either “P” or “B”.

He can input all the above information into the advanced search form. The result is as

shown in figure 18. Users can perform this kind of complicated queries in the web-based tool easily while it is difficult to perform similar queries in other online dictionaries.

#### 4.2.5 Direct search

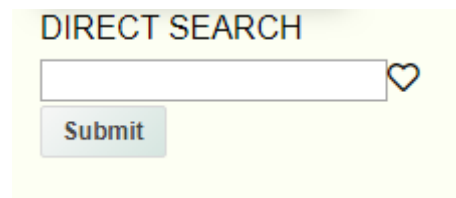


Figure 19. Direct Search

The user can directly enter a Chinese character by submitting the direct search form as shown in figure 19. For example, if the user submits the word “二”, the system will direct the user to the main page of “二”.

The feature “autocomplete” is implemented on this form. When the user types in anything word, a drop-down list containing the result will pop out immediately. Before submitting, the system will check whether the user has typed in anything inside the input field.

After the user logs in to the system, a heart will appear beside the input area. When the user hovers on the heart, a drop-down list containing the user’s favorite words will pop out. The user can select one of the words inside the drop-down list, and the input field will be automatically filled in with that word.

#### 4.2.6 Autocomplete

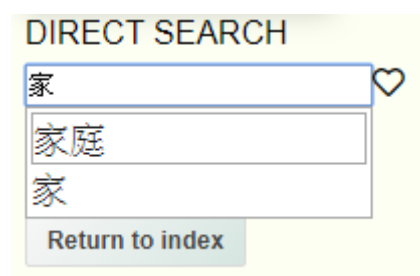


Figure 20. Autocomplete example

Autocomplete can let the user know whether the word he wants to search exists in the database. It can also improve the searching experience because it will show the related result to the user. The main function of autocomplete is implemented by JQuery.ui. JQuery.ui provides a set of widgets, the system only needs to provide the available tags to implement the autocomplete. To get the available tags, the system uses “AJAX” to check any data in the database matches with what the user types. If there are matches, results will be returned to the page. For example, If there are “家” and “家庭” in the database, after the user types “家”, both “家” and “家庭” will be popped out as shown in figure 20.

#### 4.2.7 Main page



Figure 21. Main page of Chinese character “二”

In the main page, a large Chinese word representing the current page will appear as shown in figure 21. For example, in the main page of “二”, a large “二” will appear on the page.

Under the word, the user can go to the edit history or update the current word by clicking the corresponding buttons.

The main content is placed under the buttons. The right-hand side of the page contains the multimedia files. By default, there will be an image of a QR code of the current page in the multimedia zone. The description of the multimedia is under the multimedia file. This page will check whether the word is a valid Chinese word. It must contain at least one or more Chinese characters to be considered as valid. For example, “笑” and “爛 gag” are valid Chinese words as they contain one or more Chinese characters. “Hea” is not considered as valid because it does not contain any Chinese character.

If no user has inputted any data referring to the word, the system will redirect the user to the input page of the word.

After the user logs in, he can choose to add the word to his preference list and watch list. If the user wants to add the word to his preference list, he can click the star. The star is black in color if the word is not inside the user’s preference list. Otherwise, it will be orange in color. Clicking an orange star will remove the word from the user’s preference list. To add the word into the watch list, the user can click on the “Click to add to watch list” button. Then the user will be redirected to manage watch list page. The user can also view the reference materials after logging in. If a hyperlink is declared for that reference material, a blue arrow will be generated and clicking it will redirect the user to that hyperlink.

In the mobile version of the page, the multimedia files will be hidden from the users. Users can choose to open the multimedia files by clicking the button “Click to open multimedia”.

#### 4.2.8 Update and input



WEB BASED TOOL FOR CHINESE CHARACTER EVOLUTION

SEARCH WITH

criteria

content

sort by

order by

Submit

Advanced Search

Go back to index

DIRECT SEARCH

Submit

hi, admin

Member Center

logout

Chinese character: 二

Click here to add a new field

Field name: strokes

Content: 2

Delete

Field name: pinyin

Content: er

Delete

Field name: radical

Content: 二

Delete

Field name: relatedword

Content: 數字

Delete

Submit

Reference

Click here to add a new reference field

Field name: strokes

Reference: yahoo

Hyperlink: www.yahoo.com.hk

Delete

Field name: pinyin

Reference: google

Hyperlink: https://www.google.com.hk/

Delete

Submit

Upload Multi-media Files

Edit and Delete

delete this picture

214\_short1\_happy-happy-fun-joy\_0021\_preview.mp3

change

delete this picture

test.JPG

change

Click here to add a new multi-meida upload field

選擇檔案 未選擇任何檔案

description of the multi-media file

Delete

Submit

Figure 22. Input page of Chinese character “二”

Only members can update the words. If the user has not logged in to the system when he enters the input page, he will be redirected to the index of the page.

Inside the input page as shown in figure 22, the user can modify, add and delete any field of the word. After he finishes modifying, he can click “submit” and the user will be redirected to the main page.

The user can input the reference materials for the word. He can indicate which part of the page needs references in the “Field Name” area and indicate what he is referring. Finally, he can optionally input a hyperlink. On the right-hand side of the input page, the user can add and upload a new multimedia file. He can also modify or delete an existing multimedia file and its description. Only .mp4,.mp3,.wav,.jpeg,.jpg,.png files are accepted as input files. After uploading the files, the original name of the files will be kept. Therefore, the user

cannot upload different files with the same name.

Before submitting, the system will check whether any of the fields, excepting multimedia description and hyperlink field, is empty. If there is an empty field, the user cannot submit the form unless the empty field is filled in.

#### 4.2.9 Edit history



Figure 23. Edit history page of Chinese word “圓滿”

Whenever a user updates a word, by either updating it in the input page or uploading text files, an edit history will be generated in the database. Everyone can view the edit history of the Chinese words as shown in figure 23. A user can view his own edit history in the member center.

An administrator can search for a user and view his edit history.

#### 4.2.10 Forum

The screenshot displays a web interface for a forum. At the top, there is a dropdown menu labeled 'choose forum type' with 'Discussion' selected. Below this is a checkbox labeled 'Only to Administrator'. A large text input area is provided with the placeholder text 'Please input comment here'. A 'Submit' button is located below the input area. Below the submit button is a 'View comment' button. Underneath, there are three comment entries, each in a light blue rounded rectangle. The first entry shows 'Name : user2', 'Time : 2019:04:12:18:54:40', and 'Comment : this is testing'. The second entry shows 'This comment is hidden'. The third entry shows 'Name : admin', 'Time : 2019:04:11:01:42:06', and 'Comment : Hi this is testing'.

Figure 24. The forum page

A forum is created for discussion between users as shown in figure 24. It also allows users to contact administrators. Everyone can view the comment in the forum if the comment is not set as hidden. If the comment is set as hidden, only the author of the comment and the administrators can view it.

Users can only submit a comment after logging in. He can choose the forum type he wants to post, and whether the comment is hidden.

Whenever a user submits a comment in the “Contact us” forum, an email will be sent to all the administrators of the website, reminding them someone has posted a comment on the “contact us” forum.

#### 4.2.11 Watch list

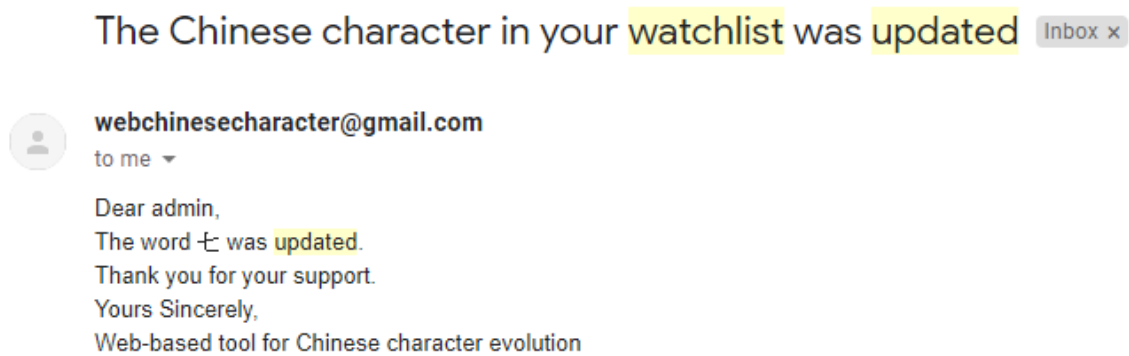


Figure 25. An email about watch list

To add a Chinese word into the watch list, a user can go to the main page of the Chinese word and click the “Click to add to watch list” button after logging in. To remove a word from the watch list, he can go to the member center.

When any user updates a word in the user’s watchlist, by either updating it in the input page or uploading a text file, an email will be generated and sent to the user’s email account, reminding him that one of the words in his watch list was updated as shown in figure 25.

#### 4.2.12 Preference

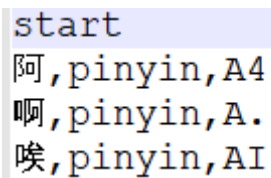
A user can set his preference in the member center. After a preference is set and the user has logged in, the preference will appear in the drop-down list beside the basic search form when the user hovers on the heart.

#### 4.2.13 Upload and read text file

To provide an efficient way for users to update the Chinese words, the system supports uploading .txt files to update data. Only members are allowed to upload .txt files.

The data inside .txt files should strictly follow the rules. The first line should be “start” to indicate that it is the start of the line. The following should be the data. The data should be written in this format: “word, field name, field content, field name, field content”. The

number of lines, field names and field contents are not limited. The following figure is an example.



```
start
阿,pinyin,A4
啊,pinyin,A.
唉,pinyin,AI
```

Figure 26. Example of a valid text file

After uploading the .txt files, the system will first save the file in a folder called “upload”. Then it will start reading the file line by line. After each update, an edit history will be generated. It will also check the watch list database and send emails if necessary. After completing the whole update process, the user will be redirected to the index page.

#### 4.2.14 Search by picture

The system supports the searching of .png files. The user can upload a .png file. This file will be temporarily saved in the “checkupload” folder. Then the system will compare this picture with all other .png files in the upload folder. If there is a match, the system will remember the corresponding Chinese character of the matching .png picture. After finishing the comparison, the system will return all the matching Chinese characters and the user will be redirected to the result page.

#### 4.2.15 Member center

User name : user1  
Gender : male  
Education level : tertiary level  
Year of birth : 1997

Set preference

My watchlist

Change email for notice

Personal history

Change password

Upload textfile

選擇檔案 未選擇任何檔案

Submit

Figure 27. Member center

A member has more power comparing to a user who has not logged in. A member can view the reference materials of the Chinese words, update the Chinese words, set preference, set watch list, view his own update history and submit a comment in the forum. To manage the member account, there is a member center as shown in figure 27. A user can go to the member center by clicking the “member center” from every page of the website after logging in. The member center is as shown in figure 18.

#### 4.2.16 Change email

The user can submit and change his email. Before submitting, the system will check whether the input is a valid email format. If the validation passes, the email will be updated. All emails will be sent to the new email account afterward.

#### 4.2.17 Change password

The user can submit and change his password. The user must input his old password and

confirm his new password to update the password. The system will first validate whether the password is the same as the confirmed password. Then it will check if the old password inputted by the user is the same as the true password of the user. If the validation passes, the password will be updated.

#### 4.2.18 Set admin

To facilitate the management of the web-based tool. There are two types of members, the first type is normal members, the second type is administrators. In the member center, if the user is an administrator, there will be two additional orange buttons. An administrator can check the edit history of the other users. He can also set another account to be an administrator.

When a user submits a comment in the “contact us” forum, all administrator will receive an email. An administrator can also view the hidden comment posted by other users.

#### 4.2.19 Validation

All forms of the website will check validity before submitting. To avoid any error related to null, the system will validate whether all the necessary fields are filled in. Also, it will validate specific types of input. For example, in the register form, the year of birth must be numeric values. The email input must be a valid email format.

#### 4.2.20 Auto-hyperlink

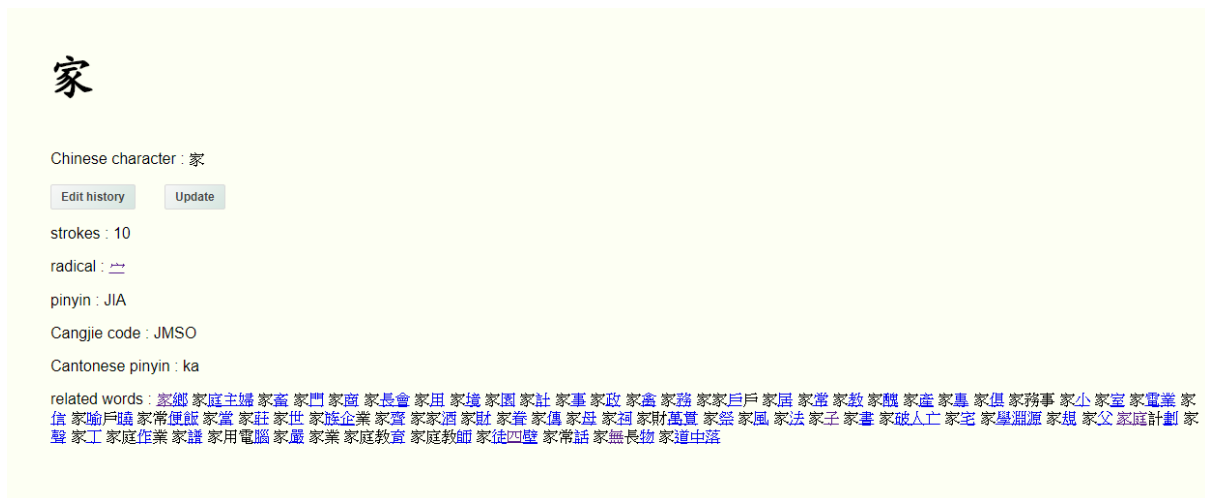


Figure 28. Auto hyperlink example

#### 4.2.20.1 Idea

In the main page of every Chinese character, hyperlinks towards the main page of other Chinese characters will be generated. For example, inside the main page of the word “家”, there is a field “related words” with field content “家庭”. If the word “家庭” exists in the database, the hyperlink towards “家庭” will be generated as shown in figure 28.

To efficiently compare the database and the field content, an idea is applied. The idea is to extract the first character of the Chinese word. The words with the same first character will be grouped. Only the first character will be used to compare with the field content. The following is an illustration of the idea.

First, a list of words is in the database.

[“家”, “家庭”, “家長”, “家庭會議”, “可愛”, “可口”, “可憐”, “五”, “五十”, “五光十色”]

Then the idea is to extract the first word in the database.

[“家”, “家”, “家”, “家”, “可”, “可”, “可”, “五”, “五”, “五”]

Then group the Chinese characters.

[“家”, “可”, “五”]

Then check the field content with the Chinese characters

( “家庭” ).indexOf( “家” )



```
( "家庭" ).indexOf( "可" )
```

```
( "家庭" ).indexOf( "五" )
```

If the result is larger than -1, it means there is a match. Then the idea further compares the content with its children.

Compare ( "家庭" ) with [ "家", "家庭", "家長", "家庭會議" ].

Finally, replace the word with an "a tag" if there is a match.

#### 4.2.20.2 Implementation

To implement the idea, each time a word is inputted into the database, an XML file with the following structure will be generated

```
<hyperlink>
```

```
  <word category=first character>
```

```
    <fullword>fullword</fullword>
```

```
  </word category>
```

```
</hyperlink>
```

The following is an example.

```
<hyperlink>
```

```
  <word category= "家">
```

```
    <fullword>家庭</fullword>
```

```
    <fullword>家長</fullword>
```

```
  </word category>
```

```
</hyperlink>
```

The field content in the main page will read the xml file and compare only with the category attribute first. If there is a match, it will push the "fullword" into an array. Then the system will use replace() to replace any html with <a></a> that matches with the array.

#### 4.3 The relationship between the web-based tool and Chinese character evolution

The attributes of Chinese characters are never fixed. They may increase, decrease, or may even change in the future. This web-based tool is specifically designed to fit the evolution of Chinese characters.



Figure 29. The evolution of the script of the Chinese character “魚”

is displayed in the web-based tool

The script styles of Chinese characters are one of the most significant changes. A static

dictionary cannot store the continually changing scripts. In contrast, multi-media files with description can be uploaded on this web-based tool. Users can clearly understand how the script styles of Chinese characters change over time as shown in figure 29. The number of pictures can be further increased if the script of Chinese characters evolves in the future.

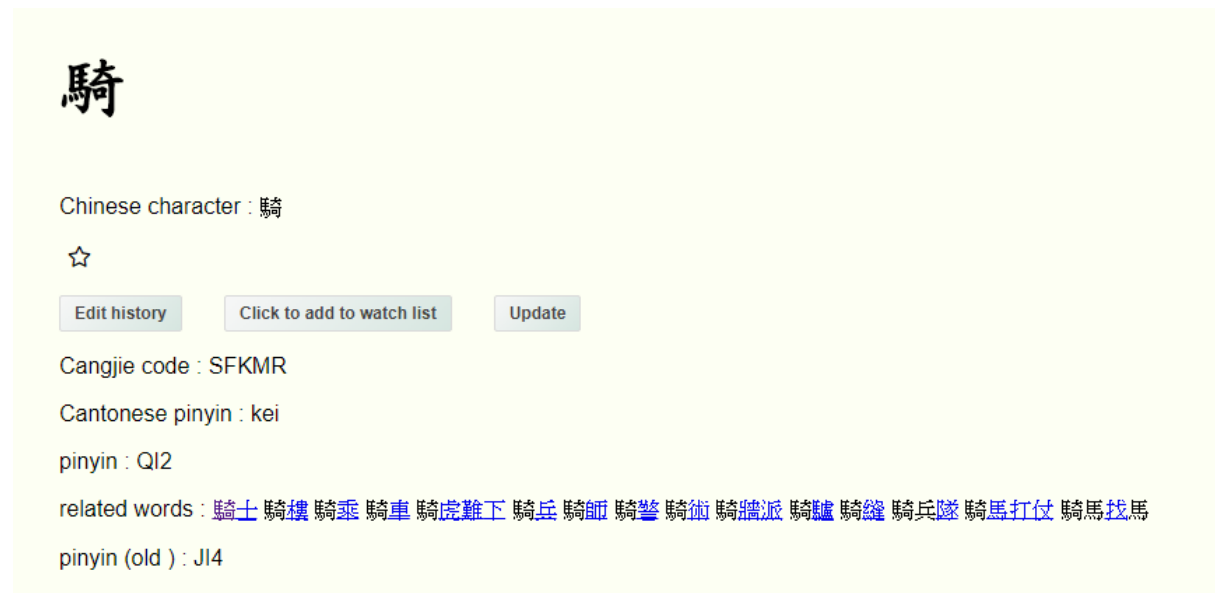


Figure 30. The evolution of the attributes of the Chinese character “騎”

is displayed in the web-based tool

Not only will the script style change, but the pronunciations of the Chinese characters also vary as time progresses. For example, the old pronunciation of the Chinese character “騎” is “JI4”. However, it is now updated to “QI2”. These changes can be displayed through the web-based tool as shown in figure 30.

It is expected that more characters and words may arise in the future. Little maintenance for this web-based tool by humans is needed when there are more data in the database. For example, the hyperlinks, which direct the user to other pages of the web-based tool, are auto-generated by the system. When there is a new word inputted in the database, the hyperlink directing to that word will be automatically generated if that word appears in other pages of the web-based tool. Moreover, the field names in the database of the Chinese characters are flexible. A user can search by a new field name if that field name is

inputted in the database.

## 5. Difficulties

### 5.1 URI-encoding

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.

### 5.2 Last modified

On the index page, there is a part showing five recently modified Chinese character. If the system directly retrieves the edit history data sorted by time, there is a risk that the Chinese characters may be duplicate. For example, if “二” was modified recently twice, in the last modified area, it will appear twice. To get the distinct Chinese character value, `distinct()` can be used in a MongoDB query. However, `distinct()` cannot be used with `sort()` at the same time.

To solve this problem, `aggregate` is used to retrieve and group the data. `aggregate([{$group:{"_id":"$chinesecharacter", "time":{"$max":"$time"}}},{ $sort:{"time":-1}},{$limit:5}])` is used to group the Chinese character, and the time field of the grouped items will use the most recent time. Then the system can display the most recently modified distinct Chinese characters.

### 5.3 Hosting the website

To host the website continuously, the npm package “forever” is used. However, due to a bug in the forever package[7], it is unable to stop the process running the forever command. To solve this problem, to terminate the forever command, “sudo lsof -i:80” is used to check which process is occupying port 80. Then terminate that process with “kill” command.

### 5.4 .txt file in windows notepad

If a .txt is opened and saved in windows notepad, an unicode “&#65279” will be added to the file’s first line. In this case, reading that .txt file will cause an error in the database. To avoid causing problems, the first line of the .txt file is skipped. It is recommended to put “start” in the first line of the .txt file. However, since the first line is skipped, what the first line of the .txt file is will not affect the updating.

### 5.5 Flexible Data

Flexible data is the main feature of the project. The user should be able to add new fields and update or delete existing fields. The following uses Chinese character “二” as an example.

Existing data:

Chinese character: %E4%BA%8C

Content:

Strokes:2, Radical: 一

Since the radical of the existing data is incorrect, a user would like to modify the data to new content.

New content:

Radical: 二, Pronunciation: er

To modify the existing content to the new content, an ordinary system needs to delete the

data of strokes:2, add a data of Pronunciation: er and update the radical:二. It will be complicated especially for the deleting command because each delete command needs to check whether the field is missing. To facilitate the modification, db.Schema.Types.Mixed is used. All the fields inside a Chinese character are treated as an object. Modifying the fields of a Chinese character will simply modify the object inside the Chinese character, which is more convenient.

## 5.6 Search by pictures

To search .png pictures in the database, the system needs to loop through all the directories that save pictures. If there is a match, the matching result will be pushed into a result array. However due to synchronization problem, if the system simply uses for-loop to loop through the directories, the result array will be empty if it is referred to. The following is an illustration.

```
Var result = []  
for (i in all directories)  
  for (all files in i)  
  {  
    If (looks-same)  
      result.push(i)  
  }  
Console.log(result)
```

From the above example, Console.log(result) will always be [] due to synchronization problem. To avoid synchronization problem, promise and recursion are used. For example, all the directories containing .png files will be pushed into an array. Promise is used to make sure that the process of pushing directories is completed before matching.

After that the system will have an array containing different directories, then we use

recursion to compare the uploaded image and the files inside the directories. After the recursion is finished, the result will not be affected by synchronization problem.

## **6. Conclusion and future work**

### **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 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, All the main functions of the website are completed. Many special features are added to the website to improve the user's experience. In the future, the project will listen to user's feedback and perform further modifications.

### **6.2 Future work**

#### **6.2.1 Improve Advanced Search function**

Currently, the advanced search function only supports one "OR" query object. It means that it is not possible to construct queries like `$and : [ { $or : [ { strokes : 10 }, { pinyin : ER4 } ] }, { $or : [ { Cangjie code: HO }, { radical : 水 } ] }`. In the future, the flexibility of the "OR" search function can be further improved that it should support multiple "or" query objects.

Furthermore, since all the data of the Chinese characters are stored as strings. Users are currently unable to perform ranged searches for numbers. For example, if a user wants to search "strokes: 5 to 8". He has to create 4 fields in the "OR" query in the Advanced search page, like "strokes:5 or strokes:6 or strokes:7 or strokes:8". It may bring inconvenience to users and negatively affect users' experience. In the future, the advanced search should support ranged search that search content can be ranged like "1 to 10".

#### 6.2.2 Publicize the web-based tool

The data of the web-based tool relies on the contribution of different users. To make the most use of the functions of the website, the website requires more users. Therefore, the web-based tool needs to be publicized to encourage more users to join the contribution of the data.

#### 6.2.3 Research on the usage of users

Further research can be conducted on how the users use different functions of the website. For example, research can be done on what the preferences of most of the users are. Since the database has the data of the user's personal information like education level, there a relation between the usage habit and the user's background may be discovered if further investigation is performed.

#### 6.2.4 Security issues

The security of the website can be improved to avoid leakage of personal information of the users of the website. Encryption may be necessary to protect the transmission of the data.



## 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](https://www.chiculture.net/index.php?file=topic_description&old_id=0601) [Accessed: 2019, Apr 14].

[2] Evolution of Chinese characters. (n.d.). Omniglot.[Online]. Available: <https://www.omniglot.com/chinese/evolution.htm> [Accessed: 2019, Apr 14].

[3] Epochtimes (2019-02-19)[Online ] Available: <http://www.epochtimes.com/gb/19/2/19/n11054891.htm> [Accessed: 2019, Apr 14]

[4] 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/](https://www.edbchinese.hk/lexlist_en/) [Accessed: 2019, Apr 14]

[5] 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, Apr 14]

[6] Wikipedia (n.d.) [Online] Available: <https://www.wikipedia.org/> [Accessed: 2019, Apr 14]

[7] github (n.d.) [Online] Available: <https://github.com/foreverjs/forever/issues/337> [Accessed: 2019, Apr 14]