Chinese Character Components, Radicals and Phonetics

A dissertation submitted to The University of Manchester for the degree of Master in the Faculty of Engineering and Physical Science

2014

Mingshu Zhao
School of Computer Science
The University of Manchester
Table of Contents

Abstract .......................................................................................................................... 7
Declaration ..................................................................................................................... 8
Intellectual property statement .................................................................................... 9
Acknowledgements ......................................................................................................... 10

1 Introduction ................................................................................................................. 11
  1.1 Motivation .............................................................................................................. 11
  1.2 Aims ....................................................................................................................... 12
  1.3 Objectives ............................................................................................................. 13
  1.4 Structure of dissertation ....................................................................................... 14
  1.5 Summary ............................................................................................................... 15

2 Background ................................................................................................................. 16
  2.1 History of Chinese character ................................................................................. 17
  2.2 Chinese character structures .............................................................................. 20
  2.3 Chinese character components ............................................................................ 23
    2.3.1 Strokes ........................................................................................................... 23
    2.3.2 Radicals ........................................................................................................ 24
    2.3.3 Phonetic component ....................................................................................... 25
  2.4 Pinyin system ....................................................................................................... 26
  2.5 Barriers of westerners learning Chinese ............................................................... 27
  2.6 Effective way of learning Chinese ...................................................................... 28
  2.7 Summary ............................................................................................................. 30

3 Methodology ............................................................................................................... 31
  3.1 Database technology ............................................................................................. 31
    3.1.1 Relational database ....................................................................................... 31
    3.1.2 Entity-relationship model ............................................................................ 32
    3.1.3 Relational database management system ................................................... 34
  3.2 Programming languages ....................................................................................... 34
    3.2.1 Structured query language .......................................................................... 35
    3.2.2 Java ............................................................................................................... 36
  3.3 Tools and methods ................................................................................................. 37
    3.3.1 Database management software .................................................................... 37
    3.3.2 Visual database design tool .......................................................................... 38
    3.3.3 Development environment ............................................................................ 39
3.3.4 Selection of encoding system.............................................39
3.4 Summary .............................................................................40

4 Design and implementation..................................................42
4.1 Design.................................................................................42
  4.1.1 Requirements analysis .....................................................44
  4.1.2 Conceptual structure design .............................................45
  4.1.3 Logical structure design ..................................................47
4.2 Implementation......................................................................55
  4.2.1 Collection of radicals .....................................................55
  4.2.2 Collection of phonetic components .................................56
  4.2.3 Collection of Pinyin.......................................................56
  4.2.4 Collection of elements ...................................................56
  4.2.5 Collection of structures ..................................................57
4.3 Summary.............................................................................61

5 Application development ......................................................63
5.1 Investigation of existed Chinese learning software ...............63
  5.1.1 Hanzi Master ...............................................................63
  5.1.2 Chinese Master ............................................................64
5.2 Conceptual design of the application structure ....................65
5.3 Implementation.....................................................................67
5.4 Summary.............................................................................70

6 Evaluation............................................................................71
6.1 A review of current Chinese database ..................................71
6.2 Contribution.......................................................................72
  6.2.1 Contribution to Chinese teaching ....................................72
  6.2.2 Contribution to the platforms for Chinese enquiries ............73
  6.2.3 Contribution to Chinese research ......................................73
6.3 Performance evaluation........................................................74
  6.3.1 Single table test ............................................................74
  6.3.2 Two-tables joint test ......................................................76
  6.3.3 Three-tables joint test ....................................................77
  6.3.4 Application test ............................................................77
6.4 Summary.............................................................................78

7 Conclusion............................................................................79
7.1 Deliverables .......................................................................79
7.2 Project limitation..................................................................80
7.3 Future work................................................................................................................. 80
  7.3.1 Suggestion of database optimization ................................................................. 80
  7.3.2 Suggestion of learning system development ................................................. 82

Bibliography.................................................................................................................. 83

Word Count: 16203
List of Figures

2.1 Evolution of character ................................................................. 18
2.2: Simplification process of character ............................................. 19
2.3: Chinese character decomposing .................................................. 22
2.4: Eight basic strokes ..................................................................... 24
2.5: Process of Chinese curriculums .................................................... 29
4.1: Database design process ............................................................ 42
4.2: ER model ............................................................................... 46
4.3: One-level structured character ..................................................... 51
4.4: Two-level structured character .................................................... 52
4.5 Three-level structured character ................................................... 52
5.1: Interface of Hanzi Master ............................................................. 64
5.2: Interface of Chinese Master ......................................................... 65
5.3: Process flow ............................................................................. 67
5.4: Convert Chinese character into ASCII code ................................. 68
5.5: Searching a character’s radical ..................................................... 69
6.1: Radical table test ................................................................... 75
6.2: Character table test .................................................................. 75
6.3: Two tables joint test ................................................................ 76
6.4: Three table joint test ............................................................... 77
6.5: Radical search function testing .................................................. 78
List of tables

2.1: Thirteen Chinese character structures ........................................... 21
2.2: Initials and finals in Pinyin system...................................................... 26
4.1: Radical table ...................................................................................... 48
4.2: Phonetic component table ................................................................. 49
4.3: Pinyin table ....................................................................................... 49
4.4: Stucuture table ................................................................................. 53
4.5: Element table ..................................................................................... 54
4.6: Character table ................................................................................... 54
4.7: Eighteen Chinese character structures ............................................. 60
Abstract

Along with China is playing an increasingly important role driving the world’s economy, a growing number of people begin to learn Chinese to improve their competitiveness in business or just out of interest. In recent years, many universities have set Chinese courses in western countries. However, western learners tend to find learning Chinese difficult because the writing system of Chinese and alphabet-based language are significantly different. Furthermore, the available learning resources online are usually defective. Hence it is reasonable to create a Chinese character database to tackle the challenge of westerners learning Chinese.

This dissertation focuses on the research of the history and structure of Chinese characters, investigating the main barriers for western learners then summarizes the efficient way for learning Chinese. After defining the problems, a solution for supporting Chinese learning has been provided in this project, which is the component-oriented Chinese character database.

In summary, this dissertation provides an introduction of the related background knowledge of the technologies that will be used, summarizing the achievement then giving suggestions for future work.
Declaration

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
**Intellectual property statement**

i. The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the dissertation, for example graphs and tables (“Reproductions”), which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see [http://documents.manchester.ac.uk/display.aspx?DocID=487](http://documents.manchester.ac.uk/display.aspx?DocID=487)), in any relevant Dissertation restriction declarations deposited in the University Library, The University Library’s regulations (see [http://www.manchester.ac.uk/library/aboutus/regulations](http://www.manchester.ac.uk/library/aboutus/regulations)) and in The University’s Guidance for the Presentation of Dissertations.
Acknowledgements

I would like to thank Dr. Richard Banach for his invaluable support throughout this project and for his patience in supervising and on-going supporting me. I also would like to thank my partner Yuhang Lei who has supported me greatly on the technology aspect.

Last but not least, I would like to thank my family for providing the funding and giving me endless love.
1 Introduction

Since Chinese is becoming a popular language around the world, a growing number of people start learning Chinese, including many westerners in Europe. However, learning Chinese tends to be a challenge for Europeans because of the significantly different writing systems of Chinese characters and alphabet-based languages used in western countries. Chinese character-related software is likely to support westerners’ learning Chinese.

1.1 Motivation

Chinese has been widely used in multicultural counties in recent years. China towns have been built in most metropolises in western countries. Many advertisements written in Chinese appear in restaurants, on the body of buses and on the show window of luxury shops. Along with Chinese is widely used in Europe, many western residents begin to learn Chinese. Such a phenomenon is probably caused by two main reasons. Firstly, Along with China is playing an increasingly important role driving the world’s economies, plenty of Chinese enterprises invest in western countries. An increasing number of office workers begin to learn Chinese to improve their competitiveness in the cooperation with China. Secondly, the number of Chinese students in western counties tends to rise every year. They have brought Chinese culture and tradition, making westerners become interested in Chinese. According to a recent survey conducted by Chengdu Daily, which is an authoritative newspaper, the number of western learners of Chinese has exceeded one hundred million so far. 58 percent of American believes that in the field of high-paying jobs, the lack of language skills results in their position robbed by foreigners. Many of them believe that mastering Chinese can increase employment opportunities. Mandarin fever runs especially high
In Europe, the annual growth of Chinese learners is 25 percent. In United Kingdom, Most of Britain's private schools have set Chinese classes. In Italy, there are more than forty universities have Chinese courses.\footnote{Mandarin fever: http://www.cdrb.com.cn/html/2013-09/03/content_1915531.htm}

However, it can be a daunting challenge for westerners learning Chinese since the writing system of Chinese is entirely different to any language in Europe. In addition, the existing learning materials online are usually lack of effective guide for beginners. Based on such a situation, a learning system, which can support Europeans Chinese learning could have bright prospect. The building of such a Chinese learning system requires a Chinese character database as basis. Hence it is reasonable to build a Chinese database that also specifies the combination law and pronunciation rules to help learners to get to grips with the components within Chinese characters.

1.2 Aims

Because of the semantic-phonetic writing system of Chinese, learning Chinese calls for knowledge reserve of Chinese characters components. The pronunciation and meaning of a character may both be indicated by the internal structure and components. Generally, a Chinese character consists of two kinds of structural elements: the semantic radical and phonetic component. Radical usually maps to the meaning of a Chinese character, for instance, characters that include radical “火” (huo, fire) are associated with fire in most cases. While phonetic component reflects character’s pronunciation, that is to say characters that include same phonetic may have the same pronunciation to a great extent. Since radicals and phonetics differ in function, the awareness of the two kinds of components can be helpful for beginners’ learning Chinese. Therefor the aim of this project is to build a
Chinese character database with searching function, which specifies the different components within characters to support Chinese learners.

1.3 Objectives

The significance of this project manifests not only in its powerful retrieval function, but also in its excellent support for higher-level software such as an online Chinese learning system. To achieve the aims of the project, specific objectives are essential to determining research direction, collecting relevant materials and making preliminary design of the project. Based on the integration and analysis of the materials collected in early stage, the objectives of the project are summarized as following:

1) A good structure of database will be designed to store data effectively, and greatly support further development.

2) The database will include frequently used Chinese characters.

3) For each character, its detailed information will be listed in the database table.

4) The database will specify different components of characters, such as radicals and phonetic components.

5) The most commonly used characters and character components will be collected through official materials.

6) The database will be evaluated and optimized through typical testing.

7) An application that support search function based on the database will be developed.
1.4 Structure of dissertation

Based on an in-depth study of Chinese characters, this dissertation introduced a solution of building Chinese character database for further Chinese-related system development. The structure of the dissertation is as following:

Chapter 2: Background.

The background chapter introduces Chinese character related knowledge in terms of the history, structure, basic components and Pinyin system. Then the barriers for westerners learning Chinese are summarized according to available surveys. Aiming at those barriers, an effective learning method is proposed and the development direction is determined.

Chapter 3: Methodology

On the basis of obtaining research direction, research methods are introduced in methodology chapter. This chapter is mainly about database technology that will be used in the development. In addition, tools for creating database and programming are selected in this chapter.

Chapter 4: Design and implementation

In the design and implementation chapter, the conceptual and logical models of the database are established. According to the models, data tables are created. After a statistics-based collection, selected data are input into tables.

Chapter 5: Application development

In this section, existed Chinese learning systems are compared and analyzed. By combining the advantages of the current software, the conceptual model of the application is provided. Furthermore, some key code
during the programming is posted.

Chapter 6: Evaluation

After the development, the established database is compared with existed Chinese character databases in terms of their performance. Evaluation chapter refers to the contribution of this project.

Chapter 7: Conclusion

The conclusion chapter mainly summarizes the delivery of this project. It also gives some suggestion about future work.

1.5 Summary

This chapter mainly represents the motivation of doing the project and clarifies the problems that need to be solved. Based on that, the objectives that the project plans to achieve are determined. The whole project will focus on finding out a solution to accomplish the goals. Finally, the structure of the dissertation was introduced for readers to understand the content quickly. Besides, in the next sections, all the mentioned Chinese characters will be annotated with their Pinyin and general meaning so that readers are able to precisely understand the dissertation.
2 Background

Chinese character is a kind of independent language system created by ancient Han ethnic group for the need of communication in the long-term labor and social practices. In the long history of human character development, Chinese is an important and unique writing system. It has significant features in terms of recording methods, combination methods and shaped structures.

Chinese is the oldest independent written language in the world. It is used by the largest population at the same time. In the past years, Chinese has been used by ethnics in East and Southeast Asia (such as Korea, Japan and Vietnam) over a long period of time. For now, in addition to China, Chinese characters are still used in Japan, Korea, Singapore and Malaysia. There is approximately one fifth of the world’s population uses Chinese characters. Summarized by Chinese linguists, Chinese characters possess the following features: (Qiu Zhiwen, 2008)

1) Ideographic writing system

While the alphabets used in western languages are only phonetic, Chinese characters are more ideographic. Chinese characters are composed by symbols, most of which are the abstracts of the objects in real world. Hence Chinese characters can be visually accepted, not necessarily by means of sound.

2) Independent and united figures

Independent means figures will remain the same in combined characters, in other words, characters can be decomposed into some indivisible symbols. For example: character "赢" (ying, win) can be divided into five independent
figures: “亡”, ”日”, ”月”, ”贝”, ”凡”. United means all those figures will be located in a square shape to be a character, no matter how many figures there are in the character.

3) Tridimensional structure

Unlike the horizontally arrange of alphabets in western languages, Chinese characters have plenty of internal structure. For example, character “苹” (ping, apple) has a top-bottom structure while character “呼” (hu, breath) has a left-right structure.

2.1 History of Chinese character

Chinese characters can be considered as the oldest continuously used language system in the world. They are believed to have more than 4500 years history. The origin of Chinese characters is the pictures used by primitive man to record their life. Later, graphical features were used to simplify and replace those pictures, so ideographic symbols appeared. To the late Shang dynasty, those ideographic symbols developed more steady structure. The body of inscriptions on oracle bones from the late Shang dynasty is the earliest evidence discovered that could confirm Chinese origin. These symbols, known as early pictograms, carved on pieces of bone and turtle shell were used to record the procedure of sacrificial ceremony in Shang Dynasty. (Keightley D N, 1996) For now, among the 4500 symbols discovered on the oracle bones, 2000 of them can be read by experts.

After the first emperor of Qin dynasty unified the mainland of China, to consolidate his power, he made uniform standards of writing characters, which is called Seal script. In the later thousands of years, Chinese characters have developed different shapes and writing systems in different

dynasties. Figure 2.1 shows the evolution of character “马” (ma, horse).

<table>
<thead>
<tr>
<th>Inscriptions on Bones</th>
<th>Inscriptions on Ancient Bronze Objects</th>
<th>Big Seal Style</th>
<th>Small Seal Script</th>
<th>Clerical Script</th>
<th>Regular Script Traditional</th>
<th>Regular Script Simplified</th>
</tr>
</thead>
<tbody>
<tr>
<td>骐</td>
<td>马</td>
<td>马</td>
<td>马</td>
<td>马</td>
<td>马</td>
<td>马</td>
</tr>
</tbody>
</table>

Figure 2.1: Evolution of character³

To express more accurately, based on pictograms, a growing number of compound characters were developed. Compound characters usually consist of two or more simple characters to improve the expression of Chinese. For example: in ancient China, there was only one kind of aquatic vehicle, which was represented as “舟” (zhou, boat) in Chinese. So far, “舨” (ban, sampan), “艇” (ting, light boat), “船” (chuan, ship), “舰” (jian, warship) evolved from “舟” to represent different boats. Compound characters have become the most widely used and powerful method of creating Chinese characters. For now, according to the different forming principles, Chinese characters can be divided into four categories. (Norman J, 1988)

1) Pictogram character: Pictograms are usually simple characters. They present external appearance of objects with the lines. For now, pictograms only account for a small percentage of Chinese characters. Over time they have been stylized, simplified and standardized to make them easier to write. Because of that, pictograms are no longer staying the similar shape with the objects they represent. Figure 2.2 shows the evolution of the character “鱼” (yu, fish). At first, it can be regarded as a stick-picture of a fish, but today character “鱼” has been simplified that

³ Chinese etymology, character “马”:
http://www.chineseetymology.org/CharacterEtymology.aspx?submitButton1=Etymology&characterInput=%E9%A9%AC

18
people cannot imagine it as fish any more. (Qiu zhiwen, 2008)

![Figure 2.2: Simplification process of character](image)

2) Self-explanatory characters: Self-explanatory characters were created by abstract methods of making characters. When ancient people could not find a specific object to express their feelings, they used abstract symbols. Take character “凶” (xiong, danger) as example, “凵” stood for a deep hole on the ground, while “×” stood for the feeling of panic. Hence character ”凶” means the situation of danger. (Qiu Zhiwen, 2008)

3) Compound ideograms: which refer to characters that consist of two or more pictographic or Self-explanatory characters to express another meaning. (Sampson G, Zhiqun C, 2013). For example, in Chinese “艹” (cao) means grass and “田” (tian) means cropland, so “苗” (miao) which consist of 艹 and 田 means seeding.

4) Semantic-phonetic compounds: Semantic-phonetic characters were created on the basis of pictograms, self-explanatory character and compound ideograms. So far, semantic-phonetic compounds represent

---

*Chinese etymology-character “魚”:  
http://www.chineseetymology.org/CharacterEtymology.aspx?submitButton1=Etymology&characterInput=%E9%B1%BC*
the largest percentage of Chinese characters. These characters generally consist of two parts: one semantic radical, which indicates the possible meaning of the compound character, and a phonetic component, which map to the pronunciation of the compound character. For instance, in character “桐” (tong, tung), component “木” (mu, wood) is the radical, indicating the character presents a kind of tree. Component “同” indicates that character “桐” and character “同” (tong, same) have similar pronunciation.

A character generally has multiple meanings, as well as a strong ability of forming words. In addition, a lot of characters can be words independently. This leads to a high use efficiency of Chinese characters. About 2,000 commonly used characters can cover more than ninety-eight percent of written expression. The current Chinese writing system is divided into traditional and simplified characters, the former are used by the residents live in Taiwan, Hong Kong, Macau and North America, while the latter are usually used in Mainland China and the Chinese community in Singapore and Southeast Asia.5

2.2 Chinese character structures

Unlike the horizontally arrange of alphabets in western languages, Chinese characters have plenty of internal structure. All components of a character will be located in a variety of ways in a square shape. Such a two-dimensional structure of Chinese character tends to be difficult to be understood than the linear structure of alphabet-based language. Through a statistic analysis of seven thousand commonly used Chinese characters, Professor Fu Yonghe in Beijing University have concluded that there are thirteen kinds of Chinese character structures, which has been

acknowledged by GB18030-2000 standard and ISO/IEC16046 standard. (Lu Jianping, 2010). The thirteen structures include: left-right, left-middle-right, up-down, up-middle-down, full-round, up-three-round, left-three-round, down-three-round, up-left-round, down-left-round, up-right-round, symmetry, and special. The detail information of each structure is listed in table 2.1.

### Table 2.1: Thirteen Chinese character structures

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Name</th>
<th>Frame</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vertical</td>
<td>Upper-lower</td>
<td></td>
<td>“吉” (ji, lucky)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“易” (yi, easy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“惡” (e, evil)</td>
</tr>
<tr>
<td>2</td>
<td>Separation</td>
<td>Upper-middle-lower</td>
<td></td>
<td>“意” (yi, meaning)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“鼻” (bi, nose)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“黄” (huang, yellow)</td>
</tr>
<tr>
<td>3</td>
<td>Horizontal</td>
<td>Left-right</td>
<td></td>
<td>“清” (qing, clear)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“眼” (yan, eye)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“桶” (tong, bucket)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Left-middle-right</td>
<td></td>
<td>“树” (shu, tree)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“难” (nan, difficult) “斑” (ban, spot)</td>
</tr>
<tr>
<td>5</td>
<td>Full enclosure</td>
<td>Complete enclosure</td>
<td></td>
<td>“国”, (guo, country)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“围” (wei, surround)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“囚” (qiu, prisoner)</td>
</tr>
<tr>
<td>6</td>
<td>Enclosure</td>
<td>Upper three-side enclosure</td>
<td></td>
<td>“风” (feng, wind)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“同” (tong, same)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“问” (wen, ask)</td>
</tr>
<tr>
<td>7</td>
<td>Three direction enclosure</td>
<td>Left-hand three-side enclosure</td>
<td></td>
<td>“区” (qu, distract)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“巨” (ju, huge)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“匣” (xia, box)</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Lower three-side enclosure</td>
<td></td>
<td>“画” (hua, draw)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“凶” (xiong, cruel)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“函” (han, letter)</td>
</tr>
<tr>
<td>9</td>
<td>Two direction enclosure</td>
<td>Lower right-hand enclosure</td>
<td></td>
<td>“病”, (bing, ill)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“房” (fang, house)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“庙” (miao, temple)</td>
</tr>
<tr>
<td>No.</td>
<td>Category</td>
<td>Name</td>
<td>Fram</td>
<td>Sample</td>
</tr>
<tr>
<td>-----</td>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>10</td>
<td>Enclosure</td>
<td>Two direction enclosure</td>
<td>Lower left-hand enclosure</td>
<td>“建” (jian, build)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“毯” (tan, blanket)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“连” (lian, link)</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Upper right-hand enclosure</td>
<td></td>
<td>“司” (si, manage)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“句” (ju, sentence)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“式” (shi, formular)</td>
</tr>
<tr>
<td>12</td>
<td>Symmetry</td>
<td>Whole</td>
<td>Whole</td>
<td>“大” (da, big)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“由” (you, through)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“不” (bu, no)</td>
</tr>
<tr>
<td>13</td>
<td>Special</td>
<td>Overlap</td>
<td>Overlap</td>
<td>“爽” (shuang, happy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“噩” (e, bad)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“坐” (zuo, sit)</td>
</tr>
</tbody>
</table>

The structures of Chinese characters may have layers. The character “助” (zhu, help) has one-level structure (left-right) while character “莅” (li, be present) has two-level structure (upper-lower and left-right). Characters can be decomposed on the basis of understanding the structures of Chinese characters. For instance the character “蘑” (mo, mushroom) can be decomposed four times into components in the way showed in Figure 2.3.

![Chinese character decomposing](image)

Figure 2.3: Chinese character decomposing
2.3 Chinese character components

Component is the unit of Chinese orthographic structure, which constitutes a Chinese character. For example: “好” (hao, good) consists of "女"(nv, female) and "子"(zi, son). In this case, "女" and "子" are components. In the past years, when Pinyin system has not been created, people search a character according to its radical in a Chinese dictionary. Hence, each character is definitely associated with one or more radicals. Since semantic-phonetic compounds represent the largest percentage of Chinese characters. Most Chinese characters contain a phonetic component as well. The location of radical and phonetic component may have some internal laws. For example, in terms of the horizontal separation structured characters, the radical parts may appear in the left side to a great extent and in vertical characters, radical usually locates on the top. In addition, the outside parts of enclosure-structured characters are always radicals. (Taft M, Zhu X, Peng D. 1999)

Hence the understanding of the two kinds of components can greatly contribute to Chinese learning. For instance, When a native Chinese speaker see the character "鲂" (fang, gurnard), even though he or she hasn't learned it before, his lexical knowledge can make the reader comprehend that the character might possibly indicate a kind of fish because in Chinese “鱼” (yu, fish) means fish, and that the pronunciation of the character has high possibility to be the same with "方" (fang, square) (though not necessarily the same tone).

2.3.1 Strokes

Stroke is the minimal unit of Chinese character structure. One or more strokes can constitute a component, while one component or more constitute a whole character. For Chinese characters, the least number of a stroke is
one, such as "一" (yi, one) while the character "讕" (zhe, garrulous) that consists of four "龍"(long, dragon) has the most strokes (68 strokes in total). Figure 2.4 shows the basic eight strokes in Chinese character. Basically, all the Chinese characters can be disassembled into the eight strokes.

![Figure 2.4: Eight basic strokes](http://other.allad.com.tw/chinese2/newpage.php)

### 2.3.2 Radicals

Radicals were selected in Han dynasty. At first, radicals were the same components selected in a set of similar-structured characters. Those selected radicals were then used as index of characters. In the whole Chinese character history, various standards of radicals’ selection have appeared. For now, 189 radicals have been included in Xinhua Dictionary, which has the highest approval among different standards. Since radicals are selected for the convenience of searching a set of similar-structured characters, some of them are characters as well while some are not; they are just the compounds of strokes, which cannot be read. For some characters, it

---

is difficult to define their radicals such as those simple characters, then the first stroke will be their radicals.

Along with the Chinese character simplification process, some components changed their appearance when they were used as radicals. For example, character “手” was simplified as “扌” when it was a radical in characters “打” (da, hit), “扔” (reng, drop) and so on. In addition, one radical in different positions tend to have different physical features. For example: character “火” (huo, fire) as a radical, become oblique when locating on the left side (e.g. “烛” zhu, candle). But when appearing at the bottom, it becomes “灬” (huo, fire). Generally, a radical will be associated with one or several meaning. For example: “氵” (shui, water), which is a variant of “水” (shui, water) usually indicate the meaning of water and it can be confirmed in the characters “江” (jiang, river), “湖” (hu, lake), “海” (hai, sea) and “洋” (yang, ocean).

2.3.3 Phonetic component

A phonetic component gives the pronunciation of a semantic-phonetic compound character. For example phonetic component “青” (qing, green) has the similar pronunciation with characters made of it, such as “晴” (qing, sunny), "清” (qing, clear), "精” (jing, smart), "睛” (jing, eye). Some phonetic components only indicate the pronunciation of characters while others imply the meaning as well together with radicals. Take character “婚” (hun, wedding) as an example, “女” (nv, woman) is the radical of it, which implies the character is associated with woman. The phonetic component “昏” (hun, dusk) not only gives the pronunciation, but also points the meaning, because a wedding always happened at dusk in ancient China.

However, unlike radicals, there is no official materials that can define which components are phonetic components because of the huge quantity of
phonetic components. According to a statistic analysis of seven thousand characters in 1993, 5631 characters were semantic-phonetic compounds, 1325 phonetic components were used among them. For now, the most commonly used phonetic components are generally summarized by empirical studies. In fact, phonetic components cannot indicate the pronunciation accurately. For example, character “骨” pronounces “gu”, but its compounds “滑” and “猾” pronounce “hua”. According to a recent survey, among 2500 commonly used characters, 1644 of them are semantic-phonetic compounds, representing about 65 percentages, but only 490 compounds have the same pronunciation with their phonetic components. (Zhang Xichang, 2007)

### 2.4 Pinyin system

Pinyin system is the standard phonetic system of Chinese characters, which is created in 1995. It uses Latin alphabets for Chinese transcript. Pinyin system is now used for elementary Chinese teaching in China and Singapore. The students in the two counties are mandatorily required to understand Pinyin system.

In Pinyin system, each spelling of Chinese characters consists of zero or one initial and one final. The initials, which are usually the first letters of the spellings, are consonants while finals are all possible combine of medials. The initials and all combine of finals are listed in Table 2.2:

<table>
<thead>
<tr>
<th>Initials</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>p</td>
<td>m</td>
<td>f</td>
<td>d</td>
<td>t</td>
<td>n</td>
<td>l</td>
<td>g</td>
<td>k</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>j</td>
<td>q</td>
<td>x</td>
<td>zh</td>
<td>ch</td>
<td>sh</td>
<td>r</td>
<td>z</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>y</td>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>o</td>
<td>e</td>
<td>i</td>
<td>u</td>
<td>ü</td>
<td>ai</td>
<td>ei</td>
<td>ui</td>
<td>ao</td>
<td></td>
</tr>
<tr>
<td>ou</td>
<td>iu</td>
<td>ie</td>
<td>üe</td>
<td>er</td>
<td>an</td>
<td>en</td>
<td>in</td>
<td>un</td>
<td>ün</td>
<td></td>
</tr>
</tbody>
</table>
Unlike other languages, the pronunciation of Chinese has tones, which are described as tonal mark in Pinyin system. For each phonetic symbol, there can be five tones. The tonal marks are always located at the top of the finals. For example, the five tones of spelling ma are: ma, mā, má, mǎ, mà.

Today, Pinyin system has become a tool for foreigners learning Chinese. Since it has some similarities with alphabetic languages, it tends to be relative easy for westerners to grasp law of pronunciation and spelling.

### 2.5 Barriers of westerners learning Chinese

Chinese characters can be well understood by people from Chinese culture circles such as Japanese and Singapore. However, westerners, who have the alphabet-based first language, tend to feel difficult to understand the structure of Chinese characters. This is because in alphabetic writing systems, there is generally a relationship that links the orthography of a phrase typically to its pronunciation, which is known as the grapheme-phoneme correspondence (GPC) rule. On the other hand, Chinese orthography is logographic and composed of radicals, phonetics and other components. (Chen H C, Hsu C C, Chang L Y, et al., 2013) Western learners, whose learning method may depend on phonemes, tend to identify the pronunciation of a character by observing its representation. Hence the deficiency of GPC rules of Chinese characters may become the major barrier for western learners’ comprehension and memorization of those characters.

Furthermore, one character may differ in pronunciation when in different context. Take character “长” for example, when meaning “long”, it pronounces “chang” while meaning “grow”, the pronunciation is “zhang”. Undoubtedly it

<table>
<thead>
<tr>
<th>Finals</th>
<th>Finals</th>
</tr>
</thead>
<tbody>
<tr>
<td>ang</td>
<td>eng</td>
</tr>
<tr>
<td>ue</td>
<td>uai</td>
</tr>
</tbody>
</table>
will increase the difficulty for western learner to grasp the technique for learning Chinese.

2.6 Effective way of learning Chinese

Under the guidance of prototype theory in cognitive psychology, it is prototype rather than appearance of an object that stored in people’s long-term memory. Prototype refers to the internal elements that constitute objects. (Qiu Zhiwen, 2008) For alphabet-based language, letters are the prototypes while for Chinese, The components making up characters are the prototypes, including strokes, radicals and other components. Based on the theory above, component-based Chinese teaching method has been adopted in primary schools in China. Initial learners will be required to memorize some simple characters and radicals. On the basic of understanding essential Chinese character elements, learners can proceed to next step, memorizing some compounds.

Since Chinese characters are more difficult for westerners to understand, it is highly recommended a systematic measure to tackle learning challenge. In terms of learning method, this project mainly takes a cue from a issued paper, which investigated how radical-derived character learning platform benefit international learners of Chinese. The paper suggested that skilled learners who are proficient in the usage of radicals and character formation rules can still improve their recognition ability of Chinese characters. In other words, by knowing one radical, they can further grasp a set of similar characters.

The investigation method in the paper mentioned above included a three-week Chinese learning program, which conducted on two groups: the experimental group and comparison group. The 129 participants contained
one hundred people who took English as their first language, 27 as other alphabetic language native speaker and two who had logographic language as their native language.

Participants in the experimental groups accepted radical knowledge learning every day. Teachers imparted one or more groups of radical-derived characters. On the contrary, the main activities of the comparison group were phrase practice, sentence making and text reading. After three weeks’ learning, a test was conducted in both experimental group and comparison group. The suggested process of curriculum is showed in Figure 2.5.

![Figure 2.5: Process of Chinese curriculums](image)

The scores of two groups were compared by mathematical algorithm. The results suggested that the radical-derived character learning strategy helps
learners understand Chinese orthographic knowledge. (Chen H C, Hsu C C, Chang L Y, et al., 2013)

According to the research above, a radical-phonetic based database of Chinese characters has a broad prospect. First, it can support radical-derived character learning method so that western learners could have sufficient understanding of the internal structure of Chinese characters. On the other hand, the database will have excellent support in higher-level application. For example, it can be used as a basic database of Chinese learning tools aiming at western Chinese learners.

2.7 Summary

The background chapter included an in-depth study of Chinese characters in terms of the history, structures, components, and phonetic system. Based on the introduction of related knowledge of Chinese, a recommended Chinese learning method for westerners was provided. The radical-derived learning method was proved to be effective based on the guidance of prototype theory in cognitive psychology and reasonable experiments. Hence, it can be concluded that a Chinese character database which indicates each character’s components can contribute to study of Chinese. The details of building such a database will be introduced in the following chapters.
3 Methodology

In order to build a useful database that can contribute to future application development, an excellent designed architecture and appropriate tools are essential. In this chapter, the technologies and tools used in the process of development will be elaborated.

3.1 Database technology

Database is a collection of a great quantity of data organized and stored in computers or servers for rapid search and managing. Database technology field has formed a solid theoretical foundation, mature commercial products and a wide range of application area since its birth till now. Along with the increasing management requirements and technology expansion, a variety of data models have been put forward in order to face the challenge of new data form. For now, several data models are widely used according to the different situation and data forms, such as relational model, object-oriented model and semi-structured model. In this project, all the data are organized in a relational database. Hence the following sections introduce mainstream relational database technologies.

3.1.1 Relational database

Data stored in computers or servers has its logical structure, which may be disconnected with its physical storage. In a relational database model, database is a schema that stores information about data and how it is related. Relational database presents information in one or more tables, with each table composed of columns and rows. A record stored in a row while its attributions stored in the corresponding columns.
The existences of relationship can be among columns within one table or among tables. Cross-referencing tables are presented in three ways:

- One-to-one relationship: refers to that one record is related to a different record in another table.
- One-to-many relationship: refers to that one record is related to more than one record in another table.
- Many-to-one relationship: refers to that multiple records in one table relate to one single record in another table.

Tables in relational database should conform to the integrity rules, which are also known as the constraints, to keep the data in the tables accurate and accessible. Integrity rules include two principle rules: entity integrity and referential integrity. Before introducing the two rules, two terms should be clarified first: primary key and foreign key. A primary key is a column of attributions within a table, with each value can uniquely identify one row. A foreign key is a common key that links the primary key column to another table. In other words, foreign key must be the primary key in a relationship in the meantime. In order to ensure the data integrity of tables, the values of primary key and foreign key must follow some constrains. Entity integrity rule defines that the value of primary key must mot repeat and null. Referential integrity rule defines that if a set of attributions is the primary key and foreign key in the same time. Then the value of the foreign key column can be only null or the same value with the corresponding primary key.

### 3.1.2 Entity-relationship model

Entity-relationship (ER) model was first proposed by Peter Chen for describing information and its internal logical relationships. ER model can be
Easily understood by people who know computer little because of its graphically way for presenting entities. Hence ER model is widely used for conceptual design when building a database.

ER model abstracts the real world, generating diagrams, which include three elements: entities, attributes and relationships. An entity is any type of object which is about to be stored in the database. In an ER diagram, the entity type is shown as a box. The useful qualities of each entity can be kept in attributes. One entity type can possess one or more attributes, which can be shown in ovals in an ER diagram. Each entity type can be associated with others for objects have relations in real world. ER diagram describes those relationships in network of the entity types. Relationships between two entities also have types. For instance, the relationship between student entity and teacher entity can be named “teach” and “learn from”. It is essential to name every single relationship so that the diagram can be easily understood by readers.

There can be multiple ways of the entities connecting to others. Therefore the relationship types can be classified by the number of instance of each entity involved.

- **One-to-one type**: in which one instance of entity type A is related to one single instance of entity type B.

- **One-to-many**: in which one instance of entity type A can be related to zero, one or multiple instances of entity type B.

- **Many-to-many**: in which multiple instances of entity type A refer to more than one instances of entity type B.
3.1.3 Relational database management system

Relational database management system (RDBMS) was first introduced by Doctor E. F. Codd in 1970. A RDBMS is one kind of software, which allows data presenting in tables. RDBMSs are used for managing the storage, search, retrieval, security and integrity of data. A RDBMS software contains multiple user interface so that a high-layer applications can focus on the functions provided. When an application or a process need to insert, delete or modify data, it just need make a “call” to the RDBMS. There are varieties of RDBMS commercial products provided by different vendors, some are fee-based while others are open source. All though different products may differ in the manipulating and interface, Each RDBMS software will provide the following four essential functions:

- Data definition function: a RDBMS offers data definition language so that database administrator can easily create database and tables.
- Data manipulation function: a RDBMS offers data manipulation language to support data inserting, deleting and modifying.
- Data query function: a RDBMS offers data query language for supporting rapid search of the database.
- Data control function: a RDBMS offers data control language, which can execute the security test and integrity rules test.

3.2 Programming languages

Two kinds of programming languages will be used in the development steps in this project. Structured Query Language (SQL) is used for building and manipulating the database while java language is used for application development.
3.2.1 Structured query language

SQL is a special-purposed programming language designed for accessing and manipulating the databases stored in RDBMS. In other words, RDBMS is the basis for using SQL. As mentioned above, SQL includes data definition language, data manipulation language, data query language and data control language. Some most essential commands will be needed during building the database, includes:

1) Data definition language:
   - CREATE TABLE: creates a new table to the database.
   - ALTER TABLE: modifies the structure of an existing table.
   - DROP TABLE: deletes a table from the selected database.

2) Data manipulation language:
   - INSERT INTO: adds new data into a table.
   - DELETE: deletes data from a table.
   - UPDATE: modifies data in a table.

3) Data query language:
   - SELECT: retrieves data from one or more than more tables according to the conditional expression.
   - WHERE: lists the constraints when selecting data from tables.

To support the program developing, most RDBMSs provide their ways to combine programming language and the manipulation of SQL. There are two main ways for integrating the programming language and SQL: embedded
SQL and open database connectivity (ODBC).

Embedded SQL is SQL statement written inline with the source code of the host language. The embedded SQL statements will be parsed separately by embedded SQL preprocessor and be replaced by host-language calls. There are varieties of RDBMS products support embedding SQL sentences into advanced programming language, such as C, C++ and so on.

ODBC is a standard of application programming interface (API) developed by Microsoft for accessing database management. The ODBC successfully make database system independent from the operating system. In other words, applications developed using ODBC can be migrated to other platforms with few changes to the source code. Java database connectivity (JDBC) is java-based data access technology designed for java programming. A JDBC-ODBC bridge enables a java-based application get connection to any ODBC-accessible data source.

3.2.2 Java

In the application development step of the project, java language will be used. Java is a class-based, object-oriented programming language. Its open source and portable features make it popular in application development in recent years. An application that is developed in java on one platform can be executed on other platforms without recompiling. It is also known as "write once, run everywhere" feature of java. The portable feature of java is achieved by compiling source code into java bytecode rather than platform-specific machine code. Unlike machine code, java bytecode is intermediate representation that is intended to be interpreted by virtual machine.

Java platform provides an inclusive set of standard class library, including the functions that are common to operating systems. By importing those
standard classes, applications can call the functions defined in those classes to get access to the operating system application-programming interface (API). Among those standard classes, there are functions that support database programming. Java.sql is a package of classes that provides APIs for accessing and processing data stored in a data source. In this project, the following five classes of the java.sql will be imported in the source code. 7

- **Java.sql.Connection**: supporting to get access to a specific database.
- **Java.sql.DriverManager**: managing a set of services provided by JDBC drives.
- **Java.sql.Statement**: an object for executing static SQL statement and returning the results it produces.
- **Java.sql.ResultSet**: a table for representing a set of database results, which are generated by executing SQL statement.
- **Java.sql.SQLException**: An exception that provides information on a database access error or other errors.

### 3.3 Tools and methods

This section is mainly about the selection of tools for database managing and java programming. The features and advantages of selected tools and methods will be introduced.

#### 3.3.1 Database management software

Before creating the database, a RDBMS needs to be selected. For now, the most widely used RDBMSs include DB2, Oracle, MySQL and Microsoft SQL

---

Server. DB2 and Oracle are widely applied for the storage and management of mass data in large enterprises. Considering the small-scale data of this project, MySQL and Microsoft SQL Server Express, which is the free version of SQL Server, are more suitable. However, Microsoft SQL Server Express can only be used on Windows platforms, which greatly impact on its extensibility. For now, MySQL is the second widely used RDBMS. Its completely free and open source makes it very popular among developers. MySQL supports multiple platforms, including Windows, Linux, and Mac OS. It provides API for almost all the mainstream programming language such as C, C++, Java, PHP, Python and so on. Besides, it offers various database connection methods, like TCP/IP, JDBC, and ODBC. MySQL can not only work in the client-server environment as an independent application, it can also be embedded into other software as a database to offer support in multiple languages. Common Chinese characters encoding mode such as GB2312 and BIG5 can be applied in MySQL. Many graphical user interface (GUI) tools for MySQL are available online.

### 3.3.2 Visual database design tool

This project selects MySQL Workbench as the graphic user interface tool for managing the MySQL database. Workbench is a free unified visual tool for developers designing, modeling, creating and managing the MySQL database. It includes comprehensive functions for creating complex ER models, easily administering MySQL environment, configuring servers and executing SQL queries. It enables developers gain better visibility into database. By providing visual console, it improves the development efficiency

---

rather than command-based management interface.\textsuperscript{10}

3.3.3 Development environment

Netbeans is an open-source official integrated development environment for java. It can be installed on all operating systems for efficient coding. Netbeans allows application to be developed by a set of modular source code components called modules. Netbeans supports getting connection with MySQL. After installing JDBC driver, Netbeans enables users graphically configuring the environment variable to connect Netbeans with MySQL. In addition, the built-in SQL editor of Netbeans allows users execute SQL queries directly within the environment.

3.3.4 Selection of encoding system

Since Chinese characters cannot be processed by computer, they must be encoded. An encoding system is usually composed of a code that matches each character to a given repertoire with something else such as a bit pattern and sequence of natural number. The default encoding system of MySQL is Latin1. But Latin1 do not support Chinese characters. Therefore, an encoding system must be selected before creating the database. The next section introduces some commonly used encoding systems that support Chinese characters.

1) UTF-8

UCS Transformation Format—8-bit (UTF-8) is a variable-length encoding system, which can link every character in the unique Unicode character set. UTF-8 system encodes each character using one to four 8-bits bytes.\textsuperscript{11}

\textsuperscript{10} MySQL::Workbench: http://www.mysql.fr/products/workbench/
\textsuperscript{11} UTF-8, a transformation format of ISO 10646 http://tools.ietf.org/html/rfc3629
Specifically, UTF-8 system encodes English character using one byte while encodes Chinese character using four bytes. UTF-8 as a universal encoding system has become the dominant character encoding for the World Wide Web. It is an international cross-platform encoding system.\(^\text{12}\)

2) GBK

Chinese Internal Code Specification (GBK) is an extension of the GB2312 character set for simplified Chinese characters, used in China. GBK system is developed for Chinese characters and is widely used in Mainland China. GBK encodes Chinese characters using one or two bytes. For now CBK system includes 21003 Chinese characters, containing simplified and traditional ones. However, most browser software in western country do not support GBK system.

3) Big5

Big5 is a Chinese character encoding method which is widely used in Taiwan, Hong Kong and Macau for traditional Chinese characters. In fact the characters in Big5 are included in GBK system. Unless there is special requirement, Big5 is not recommended.

Based on the analysis above, GBK system is more space efficient than UTF-8 system in storing Chinese characters. However, it cannot be supported by most website browser in western country, which violates the requirement of the project: to build a database for westerners’ usages. Hence, UTF-8 is finally determined to be the encoding system of the project.

### 3.4 Summary

The methodology chapter interpreted the main technologies that will be used

\(^{12}\) Usage of character encodings for website: http://w3techs.com/technologies/overview/character_encoding/all
in the project so that readers are able to understand the implementation details which will be presented in the next chapter. In addition, some mainstream development tools were compared to select a most proper one to support the database establishment.
4 Design and implementation

The development procedure will be introduced in this chapter, which includes design and implementation steps. First the model of the database will be designed including conceptual and logical structures. In the implementation step, the establishment of the database will be introduced in detail.

4.1 Design

The database design gradually becomes essential throughout the development of database application system for a good structure of database tends to store data more efficiently and meet the data requirements and process requirements of users and applications. In order to establish a superior database model, the design procedure needs to follow some certain steps, which have been showed in the Figure 4.1.

Figure 4.1: Database design process
As described in Figure 4.1, there are four steps in the design procedure:

1) Requirements analysis: the main task in the requirements analysis step is to investigate the business activities of users, determine the scope, quantity and type of data that may be used in the system and clarify the requirements of users when using the system. The requirements may include data requirements which specify the information of users’ description and processing requirements which is associated with the application system.

2) Conceptual structure design: the concrete requirements can only be used after they are summarized into abstract information structure. By means of classification, aggregation and generalization of the information in the real world which users describe, a conceptual data model will be established in step. The model is required to present the structures of each department, data flow among departments and the relationships of the departments. The most effective and widely used way of building a conceptual model is ER model which has been introduced in last chapter. In ER model, objects that users may be interested in will be generalized as entities. The flow of information and relationships among entities will be described as network of multiple entities.

3) Logical structure design: ER model that obtained in the conceptual design is independent of any data model. Therefore, ER model needs to be converted into one specific data model supported by the selected RDBMS. In this project, ER model will be converted into a relational data model. In other words, the entities and their relationships shown in the ER diagram will be presented as tables.
4) Physical design: in this step, the physical computer architecture will be taken into consideration. For example: the access method and access path will be determined in this step.

On the basis of the design above, data will be collected and database will be built. However, in order to perfectly support future application, the database needs to be optimized by running some typical tasks. The structure of database may need to be modified once problem occurs. A good design database usually experiences repeating the design circle for several times. In the following sections, the procedure of database design and database building of this project will be introduced follow the steps in the flow-chart.

4.1.1 Requirements analysis

The main mission in this step is to gain a clear idea of users' actual needs, then generate a paper document for presenting their requirements in a proper way. For this project, users' main activities are analyzed to determine the main functions of the system.

Users' activities analysis: as discussed above, radical-derived character learning method is effective way for beginners improve their recognition of Chinese characters. Hence searching radicals to retrieve characters may be the primary way for beginners using the database. Their activities may include:

- Searching a radical to retrieve a specific character.
- Searching a radical to get associated characters.
- Searching a radical to get its English meaning and Pinyin.
- Retrieving a character to get English meaning and Pinyin.
For skilled Chinese learners, they tend to know a lot about the pronouncing features and structure features of Chinese characters. Hence they may retrieve a character by searching its Pinyin and their main aim of searching a character is likely to get the structure and stroke order of the certain character. Their activities when using the database may contain:

- Searching a Pinyin to retrieve a specific character.
- Searching a phonetic to retrieve a specific character.
- Retrieving a character to get its structure name and stroke order.

In conclusion, to support initial learners and skilled learners of Chinese, the database should contain commonly used characters, radicals, phonetics, Pinyin and their relative information such as English meaning. In addition, the entities need to be connected to each other according their relationships to allow users’ searching across tables.

4.1.2 Conceptual structure design

In this step, a graphical model of information needs to be established on the basis of understanding users’ requirements. Based on the discussion in the requirements analysis step, five entities can be determined, including:

- Character entity: the set of all characters, which has ID, Unicode, GBK code, Stroke number and order and related words as its attributes.
- Radical entity: which includes all the radicals and has ID, Stroke number and order, English explain and related words as its attributes.
- Pinyin entity: that will specify the ID, tone and associated characters of each Pinyin.
- Phonetic entity: in addition to ID and associated characters, each phonetic component will indicate the possible syllables of related characters in the associated syllables attribute.

- Structure entity: in which ID, Chinese name and English name of each kind of structure, samples of each structure will be stored to present Chinese characters’ structure information.

The relationships between entities are shown clearly in the Figure 4.2:

![Figure 4.2: ER model](image)

As described in the ER model, each character is likely to be associated with one or more phonetic components and Pinyin. Every character must correspond to one single radical and one kind of structure. On the other hand, each structure, radical, Pinyin and phonetic may be associated with
multiple characters. Radical and phonetic components will also correspond to one or more Pinyin if they are characters at the same time.

The ER model will directly contribute to further implementation in next steps. Each entity can be converted into tables while the attributes of entities will be the columns of tables. The details of implementation will be introduced in the following sections.

4.1.3 Logical structure design

ER model generated in the conceptual step is an independent data model, which cannot be identified by any RDBMS. Hence in logical structure design step, ER model need to be converted into a specific relational database model. In RDBMS software, relationships are presented as tables. Hence the structure of tables will be established according to the ER model.

The structure of tables could effect on the performance of an application to some extent. Hence the design phase of creating the tables plays an important role in the whole project. As showed in the ER model, the five entities can be directly converted into five tables. In addition, there are two key problems to be solved at the design stage. The first one is how to choose primary key. In the relational database, a primary key is a unique key of each row that can define the characteristics. The essential feature of primary key is that primary keys should be immutable, that is, not changed until the record is destroyed. According to this feature, Unicode is the most appropriate value to be as primary key for the three tables of radical, phonetic and character because Unicode is a universal character set standard which assigns each character a unique value. Since Pinyin is not character and do not have Unicode value, each Pinyin in the table is allocated a unique ID number as its primary key. The second issue is how to make connection among the tables.
The database is supposed to support search function across tables. For example, when users search a phonetic component in the database, he can also get its pronunciation and relevant characters. The solution of the second problem will be detailed in next sections.

1) Radical table

Table 4.1 shows the structure of Radical table. The Unicode value is the primary key of radical table. The Pinyin column is linked to Pinyin table and character table. The value of Pinyin is the Pinyin_ID in the Pinyin table. Hence, Pinyin column can be the foreign key of Radical table. The value of Related_character is the Char_ID in character table. The meaning of each radical will be specified in English_Explain column. Stroke_Number and Stroke_Order column present the number and order of strokes of each radical respectively. If the radical is also a character in character table, the value of If_Char will be 1, otherwise, it will be 0.

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>Varchar</td>
<td>No</td>
<td>PRI</td>
<td>Null</td>
</tr>
<tr>
<td>Radical</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Pinyin</td>
<td>Varchar</td>
<td>Yes</td>
<td>FOR</td>
<td>Null</td>
</tr>
<tr>
<td>English_Explain</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Stroke_Number</td>
<td>Tinyint</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Stroke_order</td>
<td>Carchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>If_Char</td>
<td>Bit</td>
<td>No</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Related_Chracter</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
</tbody>
</table>

2) Phonetic table

Table 4.2 shows the design of Phonetic component table. Similarly, each
phonetic will have Unicode and ID value. The pronunciations of phonetics will be indicated in the Pinyin_ID column. Pinyin_ID is also the foreign key of Phonetic component table. The Associated_Syllables show the possible pronunciation of characters that include the phonetic component. The Associated_Characters include common characters that contain the phonetic.

Table 4.2: Phonetic component table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>Varchar</td>
<td>No</td>
<td>PRI</td>
<td>Null</td>
</tr>
<tr>
<td>Phonetic_ID</td>
<td>Int</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Pinyin_ID</td>
<td>Int</td>
<td>Yes</td>
<td>FOR</td>
<td>Null</td>
</tr>
<tr>
<td>Phonetic_Component</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Associated_Syllables</td>
<td>Text</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Associated_Characters</td>
<td>Text</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
</tbody>
</table>

3) Pinyin table

Table 4.3 shows the details of Pinyin table. The value of tone is from zero to five, corresponding to the five tones of Pinyin system such as ɑ, ā, á, ǎ, à. The Characters column is linked to the character table. The value of it is the Char_ID in character table.

Table 4.3: Pinyin table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinyin_ID</td>
<td>Int</td>
<td>No</td>
<td>PRI</td>
<td>Null</td>
</tr>
<tr>
<td>Write_Code</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Tone</td>
<td>Int</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Characters</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
</tbody>
</table>

To create the above three tables are relatively easy task for radicals,
phonetics and Pinyin are objective existences. But structures are the internal features of Chinese characters. Therefore challenge emerges when designing the structure table, that is, how to decompose the characters then present the structure of each character with the divided elements.

To tackle this problem, this project took a cue from a published thesis, which introduced an effective way of presenting the internal structure of Chinese characters. Proposed by the thesis, Chinese characters can be divided into prototypes, which can be regarded as the basic elements for making up characters. Chinese characters can be presented as the assemblage of the encoding of character prototypes and character structures.

Independent from people’s recognition of Chinese characters, the character prototypes (which will be mentioned as character elements in the following text) are designed for computer to store and decompose Chinese characters. Character elements can be extracted or divided from radicals and characters. There are three principles for character elements extracting:

- Inheriting traditional culture principle: which means that the elements set should include as many pictogram characters and self-explanatory characters as possible because they are the materialized presentation of the ideographic characteristic of Chinese character.

- User-friendly principle: means that the character element set should include most commonly used components such as the radicals and simple characters which are essential for elementary Chinese learning.

- Efficient processing principle: character elements should be generally the smallest inseparable particle of characters in order to improve the coding efficiency and reduce the coding complexity.
Although there is certain degree of difficulty to collect and build a complete set of character elements that can cover all the Chinese characters, a relative practical database of common character elements have been established by researchers, which includes 877 elements in total. Part of the character elements will be selected in this project, being encoded with numbers to build the character element table.

As introduced in the background chapter, there are combination rules when components making up characters. In other words, the internal structures of Chinese characters can be summarized and encoded to build the structure table. In order to tell the encoding of structures from elements, the structures can be encoded with capital English characters.

Chinese characters can be presented as the set of structure encoding and element encoding. The principle of character encoding can be concluded as “from structure encoding to element encoding, from top to bottom, from left to right and from outside to inside”. The principle will be explained in the examples:

- One-level structure characters: Figure 4.3 describes the encoding of character “什”. The encoding of character “什” (shen, what) is “H (47 44)”, in which H stands for the left-right structure, 47 stands for “什 ”

![Figure 4.3 One-level structured character](image-url)
(ren, people) and 44 stands for “丶” (shi, ten).

![Diagram of Two-level Structure Character]

**Figure 4.4: Two-level structured character**

- Two-level structure characters: As showed in Figure 4.4: the encoding of character “同” (tong, same) is “M (57 J (2 17))”, in which M stands for up-three-round structure, 57 stands for “冂”, J stands for up-down structure, 2 strands for “—” (yi, one), 17 stands for “口” (kou, mouth).

![Diagram of Three-level Structure Character]

**Figure 4.5: Three-level structured character**
Three-level structure character: the encoding of character “侗” (tong, naive) is “H (47 (M (57 J (2 17))). The elements within it have been introduced in the last two examples. Figure 4.5 shows the structure of “侗”.

In this project, there will be an element table and structure table to store the element and structure encodings, the set of character encodings will be a column in the character table. By means of connecting the three tables, the internal structures of characters can be clearly described.

4) Structure table

Table 4.4 shows the design of structure table. ID is the order number and primary key of each row. The Name column contains the Chinese name of each kind of structure while the English_name column stores the English names. Code attribute is the structure encoding which will be used in the character encoding. Sample column stores some examples of each structure.

Table 4.4: Structure table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure_ID</td>
<td>Int</td>
<td>No</td>
<td>PRI</td>
<td>Null</td>
</tr>
<tr>
<td>Name</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>English_name</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Code</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Sample</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
</tbody>
</table>

5) Element table

Table 4.5 shows the design of element table. There are only two columns in
the element table. Element_ID column is the primary key and encoding of elements while element column stores character elements.

Table 4.5: Element table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element_ID</td>
<td>Int</td>
<td>No</td>
<td>PRI</td>
<td>Null</td>
</tr>
<tr>
<td>Element</td>
<td>Varchar</td>
<td>No</td>
<td>null</td>
<td></td>
</tr>
</tbody>
</table>

6) Character table

Character table is the most complex table for it contains all information about each Chinese character. It takes Unicode as the unique primary key of rows. Char_ID is the number of character for the convenience of identifying them. The Stroke_number and Stroke_Order describe how to write the character. Radical_ID, Phonetic_ID and Pinyin_ID are foreign keys that have to be the same with the corresponding ones in the Radical, Phonetic and Pinyin tables. Element_Order column stores the combination rules of characters using structure and element encoding in the structure and element tables. Related_Words column stores some common words that include the character. The structure of character table is showed in Table 4.6.

Table 4.6: Character table

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
<th>Null</th>
<th>Key</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicode</td>
<td>Varchar</td>
<td>No</td>
<td>PRI</td>
<td>Null</td>
</tr>
<tr>
<td>Char_ID</td>
<td>Int</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Character</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Stroke_Number</td>
<td>Int</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Stroke_Order</td>
<td>Varchar</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Pinyin_ID</td>
<td>Int</td>
<td>No</td>
<td>FOR</td>
<td>Null</td>
</tr>
<tr>
<td>Radical_ID</td>
<td>Int</td>
<td>No</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Field</td>
<td>Type</td>
<td>Null</td>
<td>Key</td>
<td>Default</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>------</td>
<td>-----</td>
<td>---------</td>
</tr>
<tr>
<td>Phonetic_ID</td>
<td>Int</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Element_Order</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
<tr>
<td>Related_Words</td>
<td>Varchar</td>
<td>Yes</td>
<td></td>
<td>Null</td>
</tr>
</tbody>
</table>

The structures of six tables the Chinese character databases have been introduced above. The remaining work of building the database is to create the six tables in the selected RDBMS and input data in. The details will be introduced in implementation section.

4.2 Implementation

Since the structure of the database has been designed. Tables can be created according to the architecture and data are supposed to be input after that. Hence, the main task this step is to collect data that need to be stored in the database. In this section, data collecting methods will be introduced respectively according to tables.

4.2.1 Collection of radicals

The collection of radicals is relatively easy task because all the radicals are summarized by the Ministry of Education of the People's Republic of China in the Table of Indexing Chinese Character Component. However to enter the radicals into database is not the only task. As mentioned in background chapter, radicals in different positions tend to have different physical features.

For example: “火” (huo, fire) as a radical, become oblique when locating on the left side (e.g. “烛” zhu, candle) but when appearing at the bottom, it becomes ई. Hence, the variants of each radical need to be collected.
4.2.2 Collection of phonetic components

Unlike radicals, there is no official materials define how many phonetic components there are in total. By searching online, an issued paper that focuses on the research of phonetic components has been found. The paper introduced a statistic analysis of the pronunciation of 2500 frequently used characters and their components (if there are). 490 phonetic components had been selected because they had the same pronunciation with the characters composed by them. However, some of the phonetic components found in the paper can only compose one character, then the value of storing them into database can be considered relatively small while the high frequent phonetic components can help learners to detect the pronunciation rules. In the Specification of Common Modern Chinese Character Component and Component Names table, which is issued by Ministry of Education of the People's Republic of China, it has been summarized that the number of times that each component composes characters. Combining the two materials mentioned above, frequently used phonetic components are selected.

4.2.3 Collection of Pinyin

Unlike the other three kinds of data, Pinyin has its internal order. Letters on each position are arranged according to the English alphabetical order, like from “bai” to “ban” and from “bang” to “bei”. For each Pinyin, its tone is from zero to five. Hence the only tasks of collecting Pinyin is to arrange all the Pinyin according to the order mentioned above and assign them an ID number.

4.2.4 Collection of elements

Each Chinese character consists of one or more character elements. Unlike
the letters, which can be regarded as the elements of alphabetic languages, Chinese character elements give the characters meaning. They reflect the ideographic feature of Chinese characters. Hence the selected character elements should include basic symbols that created for recording objects in real world.

Since radicals are components selected to be index of characters, the set of radicals can be included in the element table. In addition, simple characters can be recorded in the element table directly because all the compound characters are composed of radicals and simple characters. Some basic figures which appear in characters frequently will be recorded in the table as well. In summary, character elements contain two kinds of constituents, individual characters and basic symbols. The individual characters include 234 simple characters and 200 radicals (GB18030, 2000). Although the quantity of character symbols tends to be huge, the basic symbols for making up common characters can be selected from the list of standard table of character components issued by China's ministry of education. (GB13000.1, 2009).

4.2.5 Collection of structures

In the background chapter, a standard structure classification method has been described in detail, which includes thirteen structures in total. However, the classification was created based on people’s cognition of Chinese characters. Problems exist in terms of computer-aided information processing:

- It lacks concrete definition and mathematical explaining for each structure. It will bring problems on ranging characters.
- It is difficult to understand the special structure since it may refer to
characters with various appearances.

- It can cover only seven thousand characters. More structures may be summarized if researching more characters.

In order to tackle those problems, a classification scheme which includes eighteen structures is widely used for information processing on computers and each structure has a clear definition:

Individual structure:

- Simple character: opposite to compound characters, simple characters can only be divided into strokes.

Vertical structure:

- Upper-lower: contains two or more components. Characters can be divided into upper and lower parts.
- Upper-middle-lower: contains three or more components. Characters can be divided into upper, middle and lower parts.
- Multi-row: contains four or more than four components. Characters can be divided into multiple rows.

Horizontal structure:

- Left-right: contains two or more components. Characters can be divided into left and right parts.
- Left-middle-right: contains three or more components. Characters can be divided into left, middle and right parts.
- Multi-column: contains four or more components. Characters can be
divided into multiple columns.

Overlapping structure:

- Pyramid overlapping: contains three identical components, which are combined as pyramid structure.

- Double overlapping: contains four identical components, which are combined as symmetrical structure from top to bottom and from left to right.

Enclosure structure:

- Complete enclosure: contains two or more components and can be divided into inside and outside two parts. The outside part is an enclosed component.

- Upper three-side enclosure: contains two or more components and can be divided into inside and outside two parts. The inside part is surrounded by a semi-enclosed component in upper, left and right direction.

- Left-hand three-side enclosure: contains two or more components and can be divided into inside and outside two parts. The inside part is surrounded by a semi-enclosed component in upper, left and lower direction.

- Lower three-side enclosure: contains two or more components and can be divided into inside and outside two parts. The inside part is surrounded by an semi-enclosed component in left, right and down direction.
• Lower right-hand enclosure: contains two or more components and can be divided into inside and outside two parts. One part is surrounded by a component in upper and left direction.

• Upper right-hand enclosure: contains two or more components and can be divided into inside and outside two parts. One part is surrounded by a component in upper and right direction.

• Lower left-hand enclosure: contains two or more components and can be divided into inside and outside two parts. One part is surrounded by a component in upper and right direction.

Embedded structure:

• Frame-based embedded structure: contains 2n+1 components (n=1, 2). Components will be embedded in a fixed frame.

• Mutual embedded structure: contains two or more components. Characters can be divided into two parts and the two parts are mutual embedded.

The frame of each character is showed in Table 4.7. the structured code are letters allocated to structures for character encoding. In the database, one or more samples will be provided for each character.

Table 4.7 Eighteen Chinese character structures

<table>
<thead>
<tr>
<th>Structure code</th>
<th>Frame</th>
<th>Name</th>
<th>Structure code</th>
<th>Frame</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td><img src="image1" alt="Frame" /></td>
<td>Simple</td>
<td>P</td>
<td><img src="image2" alt="Frame" /></td>
<td>Complete enclosure</td>
</tr>
<tr>
<td>H</td>
<td><img src="image3" alt="Frame" /></td>
<td>Upper-lower</td>
<td>Q</td>
<td><img src="image4" alt="Frame" /></td>
<td>Upper three-side</td>
</tr>
</tbody>
</table>

60
6) Collection of characters

List of Frequently Used Characters in Modern Chinese, which is created by the state language commission, is a list of 3500 most frequently used simplified Chinese characters. The list has two parts, one is for frequently used characters, including 2500 characters, and the other is for secondary frequently used characters, including 1000 characters. According to the statistics, 2500 characters can cover 97.97 percent in everyday language. (Zhang Xichang, 2007) Hence, it is reasonable that enter the frequently used characters part of the list into the database.

4.3 Summary

The process of database design and implementation are introduced in detail.
Based on requirements analysis, the first section describes standard procedure of database design, from conceptual structure design to logical structure design. The six tables have been created according to the structures introduced above. In addition, data which are supposed to be stored in database have been collected and input in corresponding tables. The remaining work is to evaluate the performance of the database and for optimization solution deriving.
5 Application development

Based on the built database, this project aims at developing an application that contains searching function of the data stored in the database. Firstly, the application will evaluate the performance of the database by executing some typical searching. Secondly, the application can be regarded as the preliminary version of the further Chinese learning system. In summary, the application is supposed to enable users view the data stored in the database. In addition, the application needs to provide searching function that allows users retrieve data according related information.

5.1 Investigation of existed Chinese learning software

In order to generate excellent product design, this project investigated some available Chinese learning software. Based on analyzing their advantages and disadvantages, a comprehensive conceptual design will be established. In this section, two widely used learning systems with high performance will be compared in terms of their usability.

5.1.1 Hanzi Master

Hanzi Master is Chinese character learning program, which support Mac, Windows and Ipad platforms. The interface of the Hanzi Master is showed in Figure 5.1. Characters can be retrieved according to their stroke number on the top of the window. In the central part of the interface, the detailed information of the selected character will be displayed, including its pronunciation, meaning, evolution process and so on. On the left and right side of the central window, the words composed by the character will be listed with their English explain. In the lower three windows, the characters with the same radical, same phonetic component and same pronunciation
will be displayed respectively.

Figure 5.1: Interface of Hanzi Master

In summary, Hanzi Master provides initial learners with the great support of accessing characters’ components, structures and English meaning. However, the menu layouts and display manners tend to be complicated and confusing for the Chinese beginners.

5.1.2 Chinese Master

Chinese Master is a comprehensive Chinese learning system, providing plenty of functions, including Pinyin courses, character structure courses, daily dialog courses and so on. The strongest feature of Chinese Master is that it includes massive audio resources. For each character, word and sentence, users can get their pronunciation, which enables learners grasp

---

13 Hanzi Master: Features: http://www.learnchinese.cn/interface.htm
characters combined with pronunciation. Figure 5.2 shows the main functions of Chinese Master.

![Chinese Master Interface](http://www.learnchinese.cn/interface.htm)

**Figure 5.2: Interface of Chinese Master**

However, the searching characters function of the software only returns characters’ English explain and Pinyin. It fails to enable users understand characters’ components and structures. Hence it is reasonable to combine the advantages of the two programs in the structure design step.

### 5.2 Conceptual design of the application structure

The conceptual structure will define the included functions of the application.

---

According to the requirements analysis in the database design step, the main activity of targeted user group is searching characters based on their radical, Pinyin or other conditions then get the detailed related information of the characters. Hence the outline of the application is that users can input the searching terms such as radicals and Pinyin then matched results will be displayed on the screen. The detailed information of the conceptual will be introduced in the next paragraph.

At first, users can choose what he or she wants to search, radical, phonetic or character. The procedure of searching radical and phonetic is similar. First, users need to input the stroke number a, then system will select radicals or phonetics in the radical/phonetic table where the Stroke_Number value equals a. All the results will be displayed on the screen. From the results users can select the one he or she wants. The information of the radical or phonetic will be displayed. When searching a character, users firstly need to choose searching method, by radical, phonetic or Pinyin. If users choose radical or phonetic, then similarly, he or she needs to input the stroke number to find the radical or phonetic of the character. From the information displayed on the screen, users can get all the related characters related to the radical or phonetic. Then users need to choose the one he or she wants. Figure 5.3 shows the process flow of the application.
5.3 Implementation

In this section, the detailed implementation methods will be introduced, including the essential source code. Since Chinese characters cannot be processed by machine language, the first challenge will be how to convert
the characters into codes that can be identified by computers. Figure 5.4 shows the function of converting Chinese characters into ASCII code (American standard code for information interchange).

```java
private static String escapeNonAscii(String str) {
    StringBuilder retStr = new StringBuilder();
    for (int i = 0; i < str.length(); i++) {
        int cp = Character.codePointAt(str, i);
        int charCount = Character.charCount(cp);
        if (charCount > 1) {
            i += charCount - 1; //2
            if (i >= str.length()) {
                throw new IllegalArgumentException("truncated unexpectedly");
            }
        }
        if (cp < 128) {
            retStr.appendCodePoint(cp);
        } else {
            retStr.append(String.format("\u%04x", cp));
        }
    }
    return retStr.toString().toUpperCase();
}
```

Figure 5.4: Convert Chinese character into ASCII code

ASCII is an encoding schema that encodes 128 specific characters includes figure 0-9, letter a-z, A-Z and so on. By using ASCII, those characters can be converted into corresponding binary string. Unicode is based on ASCII as well. The above class accepts input Chinese character string then return converted ASCII code.

After solving the Chinese character-converting problem, which means the characters can be processed by computer, the data stored in the database can be retrieved and viewed. The whole searching function can be divided into classes by modules, such as searching radical class and searching Chinese character class. Different classes will differ from each other mainly in the query statement. The methods of writing those classes will remain similar. Hence, this project will provide searching character's radical as an
example. Figure 5.5 shows the source code of searching a character’s radical and phonetic component.

```java
public static void SearchData(myConnection con) throws SQLException {
    Statement stmt = null;
    String character = "";

    String RadicalQuery = "Select * FROM chinese.RADICALS 
    + "join CharTable ON RADICALS.Unicode = CharTable.Radical_ID";

    String PhoneticQuery = "SELECT * FROM chinese.PhoneticTable 
    + "join CharTable ON PhoneticTable.Unicode = CharTable.Phonic_ID";

    // let the user set Character
    character = System.in.read();
    BufferedReader br = new BufferedReader(new InputStreamReader(System.in));
    System.out.print("Enter character: ");
    try {
        character = br.readLine();
    } catch (IOException ex) {
        Logger.getLogger(SearchData.class.getName()).log(Level.SEVERE, null, ex);
    }

    String where = " where CharTable.Unicode = " + escapeNonAscii(character) + "";
    System.out.println(escapeNonAscii(character));
    try {
        stmt = con.getMyConnection().createStatement();
        ResultSet rs1 = stmt.executeQuery(RadicalQuery + 
        " where CharTable.Unicode = " + escapeNonAscii(character) + "";
        String radical = rs1.getString("Radical");
        System.out.println(rs1.getArray("Radical");

        while (rs1.next()) {
            String radical = rs1.getString("Radical");
            System.out.println("Radical component is " + radical);
        }
    }
```

Figure 5.5: Searching a character’s radical

The SQL query statements of radical table and phonetic table are assigned to two string type variables, called RadicalQuery and PhoneticQuery. When this class is called by main function, it will remind users to enter a character. Then the character will be converted into ACSII code by calling converting function introduced above so that machine language can process the character by matching the ASCII code and Unicode of characters stored in database. By calling the connection function-getMyConnection(), java virtual machine will be connected to the MySQL database, then the RDBMS will execute the two queries and return the results. The System.out.println
function will display the radical and phonetic component on the screen.

In order to achieve a comprehensive searching application, the function can be divided into modules. Each module consists of one or more classes. Each class executes a group of SQL query statements. By combining the modules, the application can execute complicated searching task.

Due to the time limitation, the application has achieved the coding part but without a user interface. For now, the application can be operated and tested in the Netbeans environment. The testing details will be provided in evaluation chapter.

5.4 Summary

The application development chapter investigated two useful Chinese leaning platforms. By means of comparing the advantages and disadvantages of the two platforms, the conceptual structure of the application has been established. For supporting future implement, the essential source code of processing database has been provided.
6 Evaluation

The evaluation chapter focuses on clarifying if the deliverables of the project tackle the defined problems. In other words, the character component-oriented database will be compared to the other existed Chinese database to figure out if it make process on contributing to westerners’ Chinese learning. The performance of the database and application will also be tested by executing some typical queries.

6.1 A review of current Chinese database

The Chinese characters informatization has made process in recent 30 years. A variety of Chinese character databases have been developed by international organization for Chinese information system. However, the current existed databases use character as the smallest unit. In other word, all characters in the database will be allocated a unique ID so that system process can draw a character using its ID. Such a model of database can basically meet the requirements of character processing in Chinese system. However, disadvantages exist in the long term:

It does not consistent with people’s cognition of Chinese characters. Based on the discussion in background chapter, people tend to memorize the components first rather than the whole appearance when meeting some characters they do not know. Hence, the elementary Chinese teaching focuses more on the shapes and structures of characters. Initial learners are always required to recognize characters from strokes, radicals to whole characters. However the character-oriented Chinese databases tend to be disjointed with contemporary Chinese teaching. Components cannot be divided and extracted from characters. Users of databases are only able to know set of characters from the very beginning rather than the basic
components.

It cannot satisfy the application demands of the whole society. The universal Chinese character database standard GB18030-2000 contains 27533 characters. (GB18030, 2000) Although it can meet the general requirements of Chinese system, some rare characters are still used in special situations. GB18030-2005 was carried out later as the extension of GB18030-2000, including 70244 characters. However, there can be up to 100 thousand Chinese characters used around the world in total. GB18030-2005 can only account for about seventy percent of the whole set of Chinese characters.

It tend to be difficult to form a steady Chinese character database standard. From 1978 to 2000, four standards have been carried out. Chinese is a growing set of characters. Existed Chinese databases cannot keep pace with character development. Even though expanded standards have been released, new characters will be created as well. (Lu Jianping, 2010)

6.2 Contribution

Compared with most character-oriented Chinese database, the component-oriented Chinese database built in this project has solved the first problem proposed above. In addition, it is able to support Chinese-related systems as well.

6.2.1 Contribution to Chinese teaching

According to the past survey, classified-combination teaching method is generally more effective than other teaching method. (Huang Yaping, 2008). Classification-combination teaching method focuses on the components and combination rules of Chinese characters. Learners will be instructed characters by groups. Each group of characters will possess the same
component. For the recognizable elements of characters are components rather than strokes, such a teaching method tend to reduce the errors in writing of initial learners. In addition, after grasping the relationship between shapes and meanings of radicals and simple characters, learners are able to summarize the combination rules of compound characters. Based on understanding combination rules, learners can be taught massive characters in short time. The component-oriented database perfectly supports classified-combination teaching method for it specify the components of each character. Moreover, it stores the detailed information of components, such as their English explain and stroke order, which can help beginners’ cognize basic components quickly.

6.2.2 Contribution to the platforms for Chinese enquiries

Since this database include common used characters and frequent components, it can be regarded as a foundation of building a Chinese searching platform. The following searching missions can be conducted on the database, which may be the most concerned by learners.

1) Searching a character according to its radical, phonetic component, Pinyin elements, structure and Unicode.

2) Searching a radical according to its stroke number, Pinyin and Unicode.

3) Searching a Phonetic component according to its Pinyin and Unicode.

The functions above can basically meet the general needs of Chinese learners.

6.2.3 Contribution to Chinese research

The establishment of component-oriented Chinese database can also
contribute to other research in Chinese field. For example, it can be used as a tool for Chinese character identifying experiments in cognitive psychology field.

6.3 Performance evaluation

The performance of the database will be tested in this section. For the database will be used in Chinese character searching platform, query language will be conducted on the database. For this step, MySQL Workbench is used as the tool to operate the testing. It is a free official GUI tool as introduced in methodology chapter. Users are able to graphically administer MySQL databases. SQL is used as the query language. The evaluation will aim at testing if the database can support westerners’ learning Chinese. Hence, some possible queries that may be operated most by learners will be conducted.

6.3.1 Single table test

The queries were conducted within one single table to test if the returned results are correct. Two examples will be provided which were operated on radical table and character table. The first query is to search the radicals according to the stroke number. All the radicals which stroke number equals two are correctly returned with their English explain and related words. The results are showed in Figure 6.1
Figure 6.1: Radical table test

The second query was operated on the character table, which searches characters according to the stroke number. As a result, all the matched characters are returned according to the filter sentence: stroke number equals two. The results showed in Figure 6.2 are sorted by their Unicode.

Figure 6.2: Character table test
In the single table test step, six queries were conducted on the six tables respectively. All the matched results are returned successfully.

6.3.2 Two-tables joint test

In this step, the main mission is to test if two tables can connect correctly. Queries will be conducted on two joint tables. For most learners may retrieve characters according to the radical. The test will be conducted on radical and character table.

![Figure 6.3: Two tables joint test](image)

As showed in Figure 6.3, the query language retrieve characters whose Pinyin are written as “li”. The related words are required to be returned together with the results. 5 results are returned, which are exactly the whole five characters in character table with Pinyin written as “li”.

76
6.3.3 Three-tables joint test

The connectivity among three tables are about to be tested in this step. Queries will be conducted on character table, radical table and Pinyin table. The query contains two-steps connection.

![Query Example](image)

As showed in Figure 6.4: first, character table and radical table were connected to retrieve the characters whose radical is “女”. Then character table was connected with Pinyin table to return the Pinyin of selected characters. As showed in picture, right results have been returned.

6.3.4 Application test

The application test will specify one searching function since the operational principle of functions remains similar. In this section, the radical search function will be provided as an example. Figure 6.5 shows the result of
radical search function testing.

“机” (ji, machine) is the input character that needs to be analyzed its radical. After executing the function, the radical “木” (mu, wood) was successfully returned.

6.4 Summary

The evaluation chapter compared the database developed in this project with the current existed Chinese databases, pointing out the advantages and contribution to future work. Moreover, the performance of the database was tested based on some typical queries.
7 Conclusion

Based on an in-depth study of related knowledge of Chinese character and the learning situation of western Chinese learners, this dissertation defined the problems that are supposed to be solved in this project. After researching the technologies and methods used in the establishment of the database, a component-oriented Chinese character database has been achieved, which provides a solution for the defined problems.

7.1 Deliverables

In this project, a component-oriented Chinese database was established. The database includes six tables: character table, radical table, phonetic component table, Pinyin table, character element table and structure table. Connection has been built among tables. For now, 263 characters’ information has been collected and input in the character table. Two hundred radicals summarized by the China Ministry of education have been stored. The phonetic component table has 182 rows collected from available results of statistic analysis-based investigation. The Pinyin table includes 1302 rows. 65 out of 877 selected character elements are input into element table. The structure table stores eighteen character structures concluded in design and implementation chapter. In addition, a huge character table which includes more than ten thousand rows was created, waiting to be completed the detailed information.

The application developed based on the database includes connecting function, converting Chinese character function, view data function and searching function. Though it only executes basic searching tasks, it can be a useful basis of the future developing of Chinese learning system.
7.2 Project limitation

Since Chinese is a huge set of characters, even the number of most commonly used characters can reach two thousand, the mission of collecting data and input the into the database can be tremendous workload, considering the relative short development time. Moreover, it is a challenge to build connection between tables. As introduced above, there are ten columns in character table. Among them, Radical_ID, Phonetic_ID, Pinyin_ID and ElementOrder are the keys to connect character table to the other five tables. However, it is time-wasting to gather information from the five tables and record them in columns for each character.

Hence, although a huge character table which include more than ten thousand rows has been established, which can cover almost all the everyday language, most of the characters are still waiting to be completed their absent information. The only solution of solving this difficulty is to invest time on this project in the future. Besides, due to the impossibility of completing the database, achieving a functional Chinese learning application also becomes unpracticed. However, the main modules of the searching function have been finished.

7.3 Future work

After investigating plenty of available online materials in Chinese research field, some suggestions can be provided for the future study in this project and or other related research.

7.3.1 Suggestion of database optimization

As discussed in the project limitation, the quantity of character table is still small. Hence it is reasonable to contain more frequent characters in the
database. Moreover, the following optimization suggestions may improve the performance of database:

As discussed in the background chapter, some radicals may change their appearances when locating in different part within characters. However, some variants of radicals cannot be input in the computer. Even there is no online resource which specifies those variants. For example: the top part of the character “角” (jiao, horn) is the radical, which is a variant of character “刀” (dao, knife). But it cannot be typed in through keyboards. Hence, the radical table did not include it. Such a problem is still remained to be solved by searching more methods online.

It is better to use 36-ary code for encoding the character elements rather than decimal digits. Since there are hundreds of character elements have been selected so far, to encode the elements using decimal digits can be regarded as impractical scheme. Furthermore, the quantity of elements may increase in the future. Imagine that if a character contains several elements that have large encoding number, then the code of resenting the combination rules of the character can be fairly complicated. The 36-ary coding system is able to solve the problem. It uses ten decimal digits and twenty-six English letters to present one to thirty-six. A two-digit code can cover $36^2$ (1296) elements, which can satisfy the encoding requirement of current data size.

Since Chinese characters have plenty of structures, and in this project eighteen structures are included in the database, users may fail to understand a specific structure only according the English name and samples. Hence it is reasonable to add a figure column to store the frames of structures. This proposition can be achieved in MySQL. The BLOB type can be used for storing the figure samples of structures so that users can
graphically browse the structure information. BLOB stands for binary large object, which is a type of data designed to store binary files.

### 7.3.2 Suggestion of learning system development

Since the application developed in this project can only execute some basic searching tasks, more modules that can operate complex queries need to be involved in. Hence, more classes need to be coded to contain more combinations of SQL query languages.

In addition, to support users greatly, an graphical interface needs to be designed which can present the searching function clearly so that users can use the database conveniently and easily.
Bibliography


(丘志文. 基于认知机理的汉字智能造字之汉字基元研究[D]. 广州: 华南理工大学, 2008.)


(陳學志, 張瓅勻, 邱郁秀, 等. 中文部件組字與形構資料庫之建立及其在識字教學的應用[J]. 教育心理學報, 2011, 43(閱讀專刊))


[15] Lu Jianpin. Research on intelligent Chinese character making without library based on topology and statistic


[28] UTF-8, a transformation format of ISO 10646 [Online] Available:
[29] Usage of character encodings for website [Online] Available:
<http://w3techs.com/technologies/overview/character_encoding/all>

<http://www.mysql.fr/products/workbench/>

[31] Overview package classes [Online] available:
<http://docs.oracle.com/javame/config/cdc/optpkgs/api/jsr169/java/sql/package-summary.html>


[34] Chinese Master: interface. [Image Online] Available at:
<http://www.learnchinese.cn/interface.htm>

[35] Richard Sears, Chinese etymology-character “马”. [Image Online] Available at:
<http://www.chineseetymology.org/CharacterEtymology.aspx?submitButton1=Etymology&characterInput=%E9%A9%AC>

[36] Richard Sears, Chinese etymology-character “鱼”. [Image Online] Available at:
<http://www.chineseetymology.org/CharacterEtymology.aspx?submitButton1=Etymology&characterInput=%E9%B1%BC>