Chinese Readability Analysis and its Applications on the Internet

LAU Tak Pang

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Computer Science and Engineering

Abstract of thesis entitled:

Chinese Readability Analysis and its Applications on the Internet

Submitted by LAU Tak Pang
for the degree of Master of Philosophy
at The Chinese University of Hong Kong in October 2006

Readability assessment is a method to estimate the difficulty of a piece of writing, and it is widely used in the educational field to assist instructors in preparing appropriate materials for their students. Unlike English, which has a long history of readability research, Chinese, one of the most important languages nowadays, has not yet received much attention in similar research. In the first part of this thesis, we conduct an advanced Chinese Readability analysis. First, we analyze the potential factors affecting Chinese Readability in a systematic way, in which the factors are grouped at various levels. Second, given an input passage, various features of it based on these factors are extracted using advanced Chinese text processing techniques. We then perform regression analysis using advanced machine learning technique. We employ Support Vector Regression (SVR) as the modeling technique due to its superior performance in solving other regression problems. Experimental result shows that our proposed approach has a satisfactory performance, and has a relatively better performance than the existing approaches using Linear Regression (LR).

Web development can apply readability assessment to develop applications in a more user-oriented way, such as personalized content delivery service. This motivates us to conduct a study on Web Readability. In the second part of this thesis, we propose a bilingual (English and Chinese) readability assessment scheme for Web page and Web site based on textual features. As pages in English or Chinese cover over 70% of the population, and nearly 50% of the Internet users speak in at least one of these two languages, our scheme thus has a high coverage on the Internet community. We conduct a series of experiments with real Web data to evaluate our scheme, and to discover special characteristics of pages and sites having different readability scores. Experimental results show that, apart from just indicating the readability level, the estimated score acts as a good heuristic to figure out pages with low content-values. Furthermore, we can obtain an overall content distribution in a Web site by studying the variation of its readability.

摘要

可讀性評測是用來評估一篇文章深淺度的方法，它廣泛被應用在教育領域，例如協助老師就學生的程度選擇出適當的教育材料。英語在可讀性評測的研究上有著悠久的歷史，但作爲現今其中一種相當重要的語言，漢語在相關的研究則未被受到重視。基於以上原因，本論文將就漢語文章的可讀性作出深入的研究。首先，我們有系統地分析與可讀性相關的潛在因素，並把這些因素分成不同的文字層次。當我們要對一篇文章進行可讀性評測時，我們利用了先進的漢字處理技術，從該文章中擷取出不同文字層次的特徵。有了這些特徵後，我們運用回歸分析法來推斷該文章的可讀性。由於支持向量回歸技術(Support Vector Regression)在解決不同的回歸問題上有著優越的表現，故在這項研究中，我們運用了該技術爲回歸分析法的模型核心。實驗結果顯示，我們提出的方法有著令人滿意的效能，亦較以往的方法(主要以線性回歸技術(Linear Regression)進行分析)爲佳。

可讀性評測是可以應用在互聯網中，發展出一些使用者導向的應用程式，例如個人化內容傳送服務。這誘因促使我們對互聯網內容的可讀性作出探討。在這篇論文的第二部份，我們提出了一個嶄新的方案--互聯網可讀性的評測。該評測爲一雙語方案，對以英語或漢語爲主的網頁及網站作出可讀性評估。據統計資料顯示，以英語及漢語爲

主的網頁達百分之七十，而且約一半的互聯網使用者懂得這兩種語言，可見我們所提出的評測在互聯網社群中有著很高的覆蓋率。我們利用真實的互聯網數據進行了一系列的實驗來評估我們的方案，及觀察當網頁和網站有著不同可讀性時的特徵。實驗結果顯示，我們提出的方案，除了可反映互聯網內容的可讀性外，還可以作爲用來分別出低內容價值的網頁的指標。再者，我們可以透過分析一個網站的可讀性轉變來獲知其內容的分佈。

Acknowledgement

First, I would like to thank my supervisor, Prof. Irwin King, for his guidance, patience, support, and encouragement during my Final Year Project and M.Phil study. His brilliant idea, invaluable advice, and insightful criticism led me to the right research direction and helped me to solve different problems in the study. This dissertation cannot be completed without his effort. In addition to academic, I am also inspired from his working attitude and sharing of life experience, which will surely benefit me a lot for the rest of my life.

Second, I would like to thank Prof. Jimmy Lee and Prof. John Lui for their comments and suggestions on my work. I also learned a lot of problem-solving and logical-thinking skills from Prof. Lee through working in CUPIDE project.

Third, I want to express my gratitude to my friends and colleagues in the department, especially Hackker Wong, Xiang Peng, and CUPIDE team members, for their support and help in these two years. I am very grateful to work with them.

Finally, I would like to thank my family for their love, care and support.

To my dearest family

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we give an overview on the two main themes of the thesis (1) Chinese readability analysis and (2) Web readability analysis. We discuss the motivations and our major contributions to the topics. After that, we describe the thesis chapter organization.

1.1 Motivation and Major Contributions

1.1.1 Chinese Readability Analysis

Readability assessment is important in document analysis, and is widely used in educational and instructional technologies. Instructors can make use of readability to develop educational materials which are suitable for students, and to select appropriate materials for students to read independently [25]. Readability assessment can also help in judging the quality of a piece of writing, and assisting writers to locate the possible grammatical and stylistic problems [46].

In this thesis, we focus on Chinese readability analysis for the following reasons.

1. Significance of the Chinese Language. According to the CIA world fact book [13], the Chinese language has the largest number of speakers; about 15% of the world's population speak some form of Chinese as their native language. Furthermore, it is estimated that more than half of the world's published books are in Chinese [19], and the volume is increasing rapidly, especially in the Internet community [23]. As a result, the Chinese language will become more and more important in the future, and hence there is a strong incentive for analyzing the readability of Chinese passages.

2. Lack of Chinese Readability Research. Unlike English, which has a long history of readability research [15, 20, 47, 24, 22, 18, 5], there are only few research works on Chinese readability analysis [81, 32, 31]. In our work, we try to improve previous results by including a broader range of possible readability factors and employing advanced computational techniques.

3. Representativeness of Chinese as a Non-alphabetic Language. Chinese is one of the important non-alphabetic, since it consists of distinct Chinese character called *Hanzi*. Our research on Chinese readability can benefit research into other non-alphabetic languages such as Japanese and Korean.

4. Advanced Chinese Language Processing Techniques. In the past, the performance of Chinese Language processing (e.g. Chinese word segmentation) has not proved very successful due to its complexity and ambiguity. But with the recent advance in natural language processing using statistical and corpus-based methods, processing performance is now adequate for us to carry out higher level analyses, such as the readability analysis studied in this thesis.

Based on the above motivating factors, we present our proposed Chinese readability analysis. The main contributions of our work are as follows:

1. Perform Systematic Readability Factor Analysis. In the analysis, we consider a wider range of possible factors affecting readability than previous works, and then group them in a systematic way. We classify each of the possible factors in one of the following levels: *1. Sub-character Level*, *2. Character Level*, *3. Word Level*, *4. Phrase Level*, and *5. Sentence Level.* Among these categories, *sub-character level* is the category ignored by previous works. As a result, our research is more "complete" than previous works.

2. Apply Advanced Chinese Text Processing in Feature Extraction. In order to accurately extract features based on the discussed factors, advanced Chinese text processing techniques are employed to perform the task. For example, in order to analyze *word level* features, the input passage must first be segmented into words. This process is a non-trivial task, and we apply our proposed *LMR-RC Tagging* [40] segmentation technique, which has shown good performance in the Second Chinese Segmentation Bakeoff [75].

3. Apply Sophisticated Machine Learning Technique in Regression Analysis. We apply Support Vector Regression (SVR) as the modeling technique in our work. The reasons for choosing this technique are that, SVR has shown superior performance in solving various regression problems [12]. Furthermore, it is very powerful in revealing non-linear relations between dependent (readability score) and independent factors (our proposed feature sets). In the experiment, we compare our approach with other modeling techniques used by previous works.

1.1.2 Web Readability Analysis

The World Wide Web contains vast amount of valuable information, but there is not enough guidance for users to find information that is appropriate to their reading ability levels. When a user raises a query, existing search engines mainly return semantically-related materials, but whether the user has sufficient ability to understand the materials is often overlooked. For example, a grade 4 student is doing a project on subject "Internet", and is now looking for related information using search engine. If the search engine returns a set of journal papers describing state-of-the-art Internet technologies, although they are highly related to the subject, it is very likely that the student does not have the ability to understand them. To assist the student better, we propose and investigate the use of readability assessment on Web information.

Although readability has not yet received enough attention in Web technology development, its usefulness cannot be underestimated. The potential applications that can make use of it include:

1. Personalized Content Delivery. Different people have different reading abilities. It is important to provide contents with levels appropriate to users. Otherwise, users may lose interest in visiting the Web pages again. For example, you plan to compose a Web page describing the development of World Wide Web. By creating different versions with various readability levels, people with different backgrounds can enjoy the outcome at various reading levels.

2. Web Page Recommendation. As the amount of Web pages in the Internet is enormous and undergoing a fast growth rate, it is difficult to choose suitable materials for a particular group of people, e.g. students at different grade levels. By referencing the readability level, Web page recommendation can be done in an effective way.

3. Link-based Ranking over Readability. Existing link-based ranking algorithms mainly consider the hyperlink structures between Web pages [53, 36]. But whether the materials are appropriate to users is not known. We believe that by incorporating readability scores into link-based analysis, apart from getting desired materials, users can get the information suitable to their levels at the same time.

Based on the above arguments, we propose a bilingual Web readability assessment scheme [41] in this thesis. The three main contributions of our work are:

1. Bilingual Web Readability Assessment. The proposed readability assessment is capable of handling English and Chinese pages, which have a high coverage in the Internet community. According to a survey on the Internet language usage [23], English and Chinese pages cover over 70% of the total Web pages, and around half of the Internet users speak either in English or Chinese, indicating the significance of the bilingual assessment.

2. Web Page Readability Investigation. We propose the readability assessment of Web pages based on the textual characteristics and support vector regression technique. The difficulties of measuring page readability based on textual characteristics are that, (1) textual contents in Web pages are often shorter, (2) they do not have a complete structure as compared to ordinary document. Furthermore, visual decorations and embedded programming scripts may affect the assessment. These are the reasons why textual readability for Web page has not yet received attention from researchers in the Web community. We believe that Web development can benefit from textual readability, and thus it is necessary to have the first step on the related investigation.

3. Web Site Readability Investigation. Our work is the first study on Web site readability. A good site difficulty indicator can help Web designers to adjust sites' levels in order to suit users with different backgrounds, and users can make use of it to get an idea of whether the site is suitable to them. However, there is a lack of such overall indicator in current Web technology. We extend the Web page readability to site readability, aiming at measuring the overall difficulty of a site.

1.2 Thesis Chapter Organization

This thesis is organized as follows. Chapter 2 introduces the related work in readability analysis, support vector machine, and Chinese word segmentation. Chapter 3 discusses the details of our Chinese readability analysis. We describe a systematic way of analyzing potential readability factors, research methodology, implementation details, and experiments. We then illustrate how readability can be applied on the Internet by proposing Web page and site readability in Chapter 4. Finally, we give the conclusion in Chapter 5, followed by Appendix and Bibliography.

Chapter 2

Related Work

We give a literature review on the three main focuses in this thesis: readability assessment, the support vector machine, and Chinese word segmentation.

2.1 Readability Assessment

Readability assessment is a method of estimating the level of difficulty of a piece of writing. Klare [35] describes the term “readability” in the following three ways: (1) To indicate the legibility of a document. (2) To indicate the ease of reading due to the interest-value or pleasantness of writings, and (3) To indicate the ease of understanding due to the writing style. A readability score can be derived, which represents either (1) a relative difficulty of comprehension, usually on a scale of 0 to 100, or (2) an expected school grade level which a group of people should have in order to understand the writing.

Readability assessment methods can be divided into *computational assessment* and *non-computational assessment*. Computational assessment, or automatic assessment involves the use of statistical techniques such as regression and correlation [54]. The idea is to first extract some easily measurable attributes of the passage, such as sentence length and word length, and then the hard-to-measure readability level is predicted based on them. A typical example is *Formula-based* assessment, which uses an explicit formula to describe the combination of different attributes required to calculate the readability score.

Non-computational assessment estimates readability level based on actual human perception of the passage. As a result, it often requires human intervention during assessment. A typical example is the *Cloze Test*. In Taylor's cloze test [64], every fifth word in a passage is deleted, and subjects with different educational levels are asked to fill in the missing words. If the correctness of a specific group (e.g. Grade-4 students) exceeds 50%, then the passage is assigned with the reading level of that group.

Both categories of assessment have their advantages and disadvantages. [56] discusses the limitations associated with computational assessment of readability. The paper points out that readability formulas only measure "what can be count", and ignore the features which are hard to measure. Furthermore, improving readability scores does not necessarily improve comprehension.

For non-computational methods, because of having human intervention during assessment, they are generally more accurate than computational ones [59]. Furthermore, some other useful measurement, such as *usability* [56], can be carried out. However, non-computational methods are time-consuming and inconvenient. Our study thus focuses on computational methods.

2.1.1 Assessment for Text Document

In this section, we discuss the related work in *computational* readability assessments of text documents. In particular, we discuss factors affecting readability, and the establishment of assessment techniques for the English and Chinese languages.

The English Language

For research in English readability, Spencer [26] suggests the following language-dependent elements are the potential factors affecting readability: syntactic and semantic functions, sentence length, abstract words and concrete words, the number of syllable and short words, and familiar words. Researchers then propose various English readability assessments based on these factors. We discuss some of the assessments from the viewpoints of *complexity modeling* and *assessment formation.*

Complexity Modeling

Most of the existing English assessments model readability based on *Sentence complexity* and *Word/ Vocabulary complexity.* Then the variation among different methods is the way to estimate these two complexities.

The sentence is the basic unit of meaningful communication, and thus sentence complexity is significant in measuring readability. A common way to estimate sentence complexity is by measuring average sentence length, or average number of words per sentence. Dale-Chall [15], Flesch [20], Gunning [24], and Smog [47] adopt this estimation. Another way used by Fry [22] is to measure the number of sentences per 100 words.

There is a larger differentiation between methods of estimating word complexity; the methods can be divided into *wordlist-based* and *syllable-based.* The main idea of *wordlist-based* methods is to first establish a word list consisting basic or "easy" words. Assessments then measure the proportion of words that can be found in the basic word list. Intuitively, the more basic words a passage contains, the easier the passage is. Dale-Chall [15] and Bormuth [5] are two examples of approaches adopting this estimation. Assessments using *syllable-based* estimation assume that word complexity is directly related to the number of syllables in that word: the smaller number of syllables, the easier the word. Examples of this type include Flesch [20] and Farr-Jenkins-Paterson [18]. The Flesch formula measures the average number of syllables per word to indicate the word complexity of a passage. Farr-Jenkins-Paterson simplifies the task of counting syllables by measuring the proportion of monosyllabic words. Table 2.1 summarizes the complexity modeling approaches for English readability assessment.

Table 2.1: Complexity Modeling in English Readability Assessments.

Complexity	**Estimation**	**Example Assessment**
Sentence	Average Sentence Length	Dale-Chall [15], Flesch [20], Gunning [24], Farr-Jenkins-Paterson [18], McLaughlin [47], Bormuth [5]
	Number of Sentences in Sample	Fry [22]
Word	Wordlist-based	Dale-Chall [15], Bormuth [5]
	Syllable-based	Flesch [20], Farr-Jenkins-Paterson [18], Gunning [24], McLaughlin [47], Fry [22]

Assessment Formation

Formula-based and *Non-formula-based* are two formations of existing assessments. *Formula-based* assessments use explicit formulas to calculate the readability scores. The simplest way to establish the formula is by linear regression, in which different factors, such as the average sentence length and number of syllables described in the previous section, interact linearly. Dale-Chall [15] and Flesch [20] are examples of this. The Bormuth formula [5] uses non-linear regression, in which some of the factors are raised to several powers. *Non-formula-based* assessment is less common, and *Fry's Readability Graph* [22] is one such approach. In the graph shown in Figure 2.1, the X-axis is

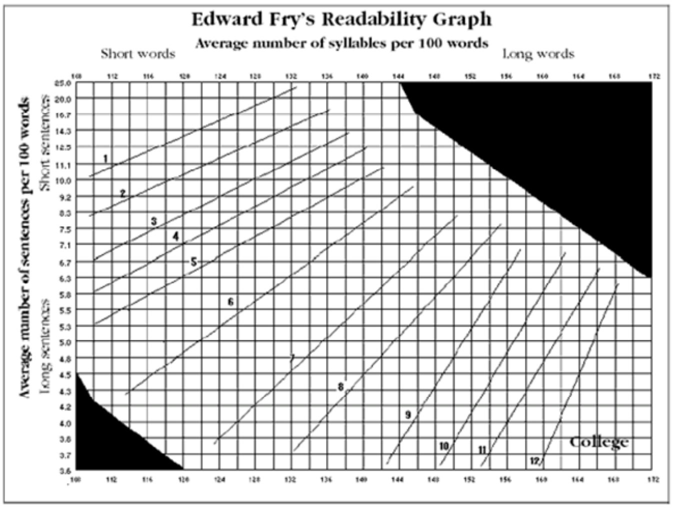


Figure 2.1: Fry's Readability Graph

the average number of syllables per 100 words; the Y-axis is the average number of sentences per 100 words. The graph contains 15 regions with unequal separation, indicating different readability levels. The level of a passage is then found by locating the region in which the corresponding factor values reside.

The Chinese Language

There are only a few studies of Chinese readability research in the literature, and to the best of our knowledge, the earliest work is by *Yang* [81]. Yang investigates various factors affecting Chinese readability, and some of them are unique to the Chinese language. The factors in his final equation after feature selection include (1) average number of strokes of characters, (2) proportion of symmetrical characters, (3) proportion of words in the basic word list, (4) average number of characters per word (average word length), (5) average sentence length, (6) average phrase length, and (7) proportion of full sentences (sentences with both subject and predicates). After collecting the statistics from a set of randomly selected materials (including news, magazine articles, book chapters, etc.) with manually predefined readability levels, regression analysis is then applied to obtain a formula consisting of a linear combination of these factors. The scale of the readability score is from 0 (hard) to 50 (easy).

Jing [32] tries to estimate the readability according to the scale of school grade level in Taiwan. He analyzes the contents of Chinese literature textbooks from first grade to twelfth grade. The factors investigated include (1) total number of characters per article, (2) average sentence length, and (3) proportion of basic characters per article. Similar to Yang, regression analysis is applied to the statistics to obtain the formula. Jing claims that the correlation between the calculated level and the actual grade level was 0.897, indicating that there is high correlation between the actual grade and predicted grade.

Jeng [31] uses *Artificial Neural Network* (ANN) in evaluating the Chinese readability level. After obtaining statistics similar to Yang's work[1], he uses ANN to perform the estimation, instead of applying regression analysis to obtain the readability formula. He compares the ANN model with linear regression model and human judgment. Experimental results show that ANN performs the best in estimating passages extracted from 12 official Chinese language textbooks.

Table 2.2 summarizes the factors considered in the three Chinese readability studies reported to date.

Comparison between English and Chinese Readability Analysis

In this subsection, we compare the differences between English and Chinese readability analysis. We believe that the differences originate in their associated writing systems. English belongs to

[1] Yang and Jeng define the concept of "word" differently. Jeng defines "word" as a single Chinese character (*hanzi*), while Yang defines "word" (Jeng calls this as 'compound word") as a character sequence which is the smallest independently useable part within a sentence. We adopt Yang's definition in this thesis.

Table 2.2: Factors Investigated in Chinese Readability Researches.

Factor	**Yang**	**Jing**	**Jeng**
(Average) Number of Characters		*	*
Average Number of Familiar Characters		*	*
Proportion of Symmetrical Characters	*		
Average Number of Strokes	*		*
Distribution of Stroke Counts per 100 Characters	*		
Average Number of Familiar / Unfamiliar words	*		*
Proportion of Words with Various Number of Characters	*		
Average Phrase Length	*		
Average Sentence Length	*	*	*
Proportion of Full Sentences	*		

the class of alphabetic languages. Each word consists of a finite number of characters. Under this system, the two basic linguistic units are the word and the sentence, and the character level does not contain much information. So we can see that the English assessment approaches discussed above only capture information from the two higher levels.

Chinese, on the other hand, is a logographic language. There is a large number of Chinese characters. Each character has its own meaning, formation and visual characteristics. Based on this variation, we can capture much more information from the character level, and even from the sub-character level as discussed in Section 3.1. As a result, we cannot directly apply the approach used in English analysis to Chinese. We need to consider a broader range of possible factors and apply different text processing techniques in Chinese readability analysis.

2.1.2 Assessment for Web Page

For Web page readability, Hill [28] studies the effects of different combinations of foreground and background colors, font types, and word styles. Experimental results show that there is no one particular combination which can lead to high Web page readability, and they suggest that it is designers' responsibility to consider the effects of different combinations.

Apart from analyzing Web page readability based on visual appearance, Si and Callan [58] investigate the readability of Web page based on text contents. They model the readability estimation as a text categorization problem using machine learning approach. In addition to some surface linguistic factors like the average sentence length and the average word length, they propose the use of language and statistical models to estimate the readability. Although their experimental results show that the approach performs better than formula-based methods, the investigation only limits on the educational science Web pages, which is not applicable on general Web pages.

2.2 Support Vector Machine

Support vector machine (SVM) is a set of supervised machine learning technique based on Statistical Learning Theory, or VC theory, developed by Vapnik [69, 70, 71]. The method is firstly proposed to solve classification problems, and is then extended to regression problems afterwards. As the thesis does not focus on theory of SVM, but rather applications of SVM, we briefly mention characteristics and advantages of SVM, followed by applications employing the technique.

2.2.1 Characteristics and Advantages

In the context of solving classification problem, the motivation of Support Vector Machine is to find a decision plan in order to minimize the *empirical risk*, which describes the extent of how close the estimated result disagrees with the actual class of training data [30]. In other words, the model established in this

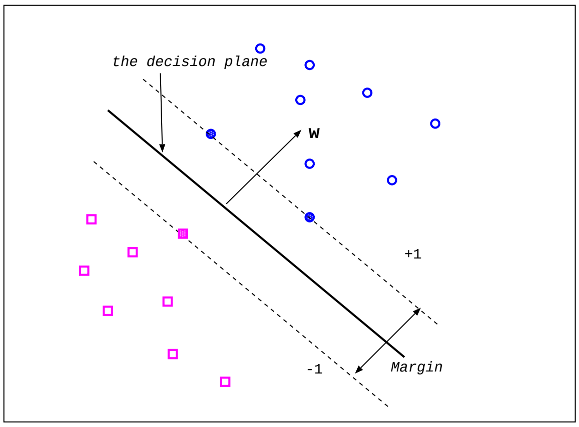


Figure 2.2: Illustration of Support Vector Machine

way is the best one which can fit the training data well.

However, the problem of over-fitting happens if we just aim at finding the best model to fit training data. Such model will fail to predict future unseen data correctly. To minimize the risk of over-fitting, another objective of SVM is to maximize the distance, or *margin*, between classes and the decision plane. Figure 2.2 (extracted from [30]) illustrates the idea of SVM in solving a 2-class classification problem. Points in circle and square represent data belonging to two classes. Although there are many decision planes which can be used to separate the two classes, the one shown in the figure is selected in order to have maximized margin. To conclude, the objective of SVM is to minimize the empirical risk, while at the same time to maximize the margin. The name of Support Vector Machine is resulted from the fact that only several points, called *support vectors*, are contributed in determining the decision plane.

Based on the above characteristics, together with the fact that training SVM leads to *Quadratic Programming*, which in turn belongs to *Convex Programming* problem [9, 10], the advantages of SVM can be summarized as [80] : (1) Theoretical bound on generalization error based on VC theory, (2) Maximum-margin decision hyperplane, (3) Global and unique solution, and (4) Mathematical tractable.

2.2.2 Applications

Because of the aforementioned advantages, SVM has been successfully applied in a wide range of applications and research fields. They include bioinformatics [7, 8], image processing [77, 83], text processing [33, 49], computer security [27], and time series prediction [80, 45].

2.3 Chinese Word Segmentation

Word segmentation, or word tokenization, is an important process in text analysis [65]. In information retrieval, we need to perform indexing by extracting keywords from documents [82]. Other applications involving natural language processing, such as machine translation [85, 76] and text-to-speech synthesis [37, 84] also require word segmentation as a preprocessing step. Word segmentation is a trivial task in languages such as English, but it receives a lot of difficulties in Chinese. In this section, we briefly discuss the difficulties in Chinese word segmentation, followed by some common approaches to solve the problem.

2.3.1 Difficulty in Chinese Word Segmentation

Lack of Word Boundary

Unlike English, in which words are separated by word boundaries like space and punctuation, there are no explicit boundaries in Chinese text. Chinese text is made up of ideographic characters, and a word can comprise one, two or more characters, without explicit indications of “start of a word” and “end of a word". Applying different word boundaries in a Chinese sentence would thus result in completely different meanings. The situation is improved after punctuation is used in modern Chinese documents, but the problem still makes segmentation a difficult task.

Multiple Meanings of Characters

A character carrying multiple meanings is common in natural languages, but the problem is more acute in Chinese text. For example, " 生" can mean produce, grow, live; " 物" can mean matter and content; " 學" can mean school, study, and knowledge. But if we view them as a word, then the meaning of " 生物學" will be clearer, meaning "biology".

2.3.2 Approaches for Chinese Word Segmentation

Different approaches are proposed to solve the aforementioned problems in Chinese word segmentation, and [14] classifies approaches into four categories: (1) Dictionary-based methods, (2) Statistical methods, (3) Syntax-based methods, and (4) Conceptual methods:

1. Dictionary-based methods. This approach uses a dictionary to identify word boundaries. We can view this approach as a greedy method: given a dictionary of frequently used Chinese words, an input Chinese text string is compared with words in dictionary to find the one that matches the greatest number of characters. Figure 2.3 illustrates the idea by segmenting a sentence " 中華人民共和國" using a dictionary { 中華, 華人, 人民, 共和, 共和國}.

   The advantages of this method are that it is efficient and can be easily implemented. But as it is impossible to obtain a dictionary containing all words, and new words are continually evolving, some other approaches are proposed.

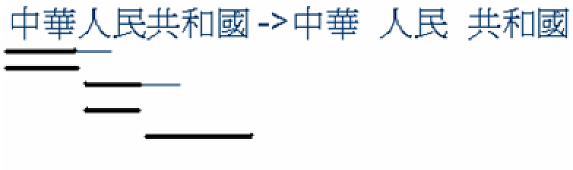


Figure 2.3: Illustration of Dictionary-based Method.

2. Statistical methods. This approach is based on statistical properties and frequencies of characters and character strings in a corpus [14]. Mutual Information (MI) [61, 44] is one of the methods in this category. It is a measure of how strong the two characters are associated, and can be used to measure how likely two characters can be merged as a word. We assume each Chinese character occurs independently to each other. By chance, the probability that the Chinese character A occurs before B is

   $$pr_{chance}(AB) = pr(A) \times pr(B). \tag{2.1}$$

   Let R be an indicator of the chance character A followed by B:

   $$R(A, B) = \frac{pr(AB)}{pr(A) \times pr(B)}. \tag{2.2}$$

   For easier manipulation due to small value of the probability, MI is defined by taking log on R, and the resulting equation is as follow:

   $$MI(A, B) = \log(pr(AB)) - \log(pr(A)) - \log(pr(B)). \tag{2.3}$$

   We can then decide the likelihood of a pair of characters is actually a word. Apart from MI, [43] and [68] apply information entropy in information theory to perform Chinese word segmentation. Our proposed segmentation method, which will be discussed in Section 3.4.1, also belongs to this type.

Statistical methods perform generally better than dictionary-based methods, and it is better in segmenting and extracting unknown words. But disadvantages are that, this system is more complex, and we need to obtain large and high quality corpora in order to obtain good performance.

3. Syntax-based methods. In addition to solely considering character statistics in a large collection of texts to perform word segmentation, syntax-based methods take the syntactic information, such as character class (e.g. noun, verb), into consideration to improve the segmentation. [50] and [11] are examples in this category.

4. Conceptual methods. This approach involves the use of semantic processing to obtain information related to each word in a sentence. The extracted information is stored in knowledge representation scheme [14]. As a result, this approach involves a higher level of language modeling, and domain-specific knowledge is required. [42] is an example belonging to this category.

Chapter 3

Chinese Readability Analysis

We discuss our approach in analyzing readability of Chinese passages in this chapter. First of all, we analyze potential readability factors systematically. Then we illustrate research methodology and procedure. After that, we briefly discuss the Support Vector Regression, the modeling technique used in our analysis, followed by implementation details involving Chinese word segmentation and feature selection using the genetic algorithm. The chapter ends with experimental results and summary.

3.1 Chinese Readability Factor Analysis

In this section, we perform systemic factor analysis by categorizing factors into various language levels. Then we summarize the features which will be used in establishing Chinese readability assessment based on the discussed factors.

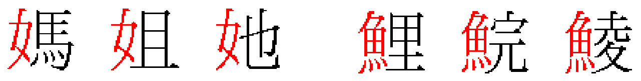
3.1.1 Systematic Analysis

Based on the characteristics of Chinese characters, we categorize the potential factors into different language levels. They are *Sub-character*, *Character*, *Word*, *Phrase*, and *Sentence*. We describe each of them in detail in the following subsection.

Sub-character Level ($\mathcal{R}$)

Introduction

Every Chinese character has a basic identifiable component called a *radical*. The radical is important in the organization and use of Chinese dictionaries. It serves as a basic "category" for each character, and thus the task of looking up a character in a dictionary requires the reader first to identify the corresponding radical.



(a) Characters with radical " 女"

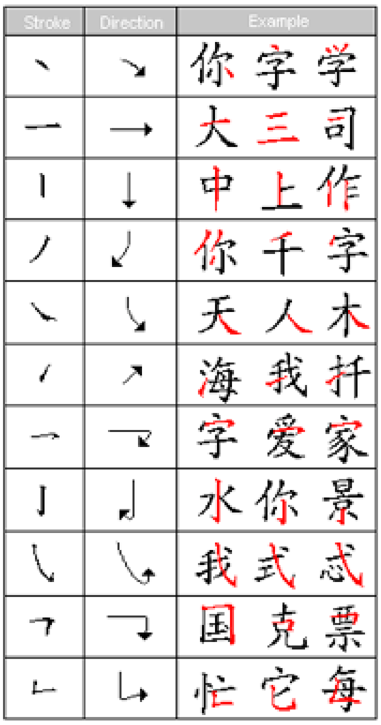
(b) Characters with radical " 魚"

Figure 3.1: Chinese Radical Examples.

Apart from their uses in dictionary lookup, the radical can often help in grouping a list of characters having similar *root* meanings. Figure 3.1 illustrates two groups of characters having the same radicals. The Chinese characters 媽 (in English: mother), 姐 (elder sister), and 她 (she) have the same radical 女 (female), meaning that they refer to entities which have the sense of female. Another example: 鯉 (carp), 鯇 (grass crap), and 鯪 (dace) have the radical 魚 (fish), meaning that they are "fish-related" characters.

Number of Strokes ($\mathcal{R}_{strk}$)

Chinese radicals are composed of basic "symbols" called *strokes*. Figure 3.2 illustrates the basic symbols [17]. In our analysis, we take the number of strokes in a radical into consideration. We expect that if a radical contains more strokes, it will be more complex, thus increasing its difficulty.



Stroke	Direction	Example
㇔	↘	你 字 学
㇐	→	大 三 司
㇑	↓	中 上 作
㇒	↙	你 千 字
㇏	↘	天 人 木
㇀	↗	海 我 扦
㇖	⬎	字 爱 家
㇚	⤶	水 你 景
㇂	⤷	我 式 忒
㇆	⤵	国 克 票
㇗	↳	忙 它 每

Figure 3.2: Basic Strokes in Chinese Characters (extracted and modified from [17]).

Radical Familiarity ($\mathcal{R}_{fam}$)

Radicals can be classified based on ease of interpretation. *Familiar radicals* have obvious meanings. Examples of this type are the aforementioned 女 and 魚. Some other examples include: 肉(meat, flesh) is related to animal body parts; 木(wood) is related to plant; 鳥(bird) is related to bird. If a reader encounters unknown characters, having a familiar radical, one can guess their meanings more easily.

Some radicals do not have obvious meanings. For example, some have simply been created for the sake of classification. We then classify them as *Unfamiliar radicals*, which ordinary readers cannot interpret easily. Examples of this type include: 尢, 匚, 厂, 黹 and 黽.

To conclude, our hypothesis concerning the effect of the radical on readability is as follows: , the more characters there are having familiar radicals, the easier the passage is. We therefore measure the proportion of characters having familiar radicals in our analysis.

Table 3.1 summarizes the possible factors of radicals which affect readability, with examples.

Table 3.1: Summary of Radical Factors.

Example of Radical	**No. of Strokes**	**Familiar or Unfamiliar**
一	1	F
殳	4	U
火	4	F
瓜	5	F
髟	10	U
龍	16	F

Character Level ($\mathcal{C}$)

Introduction

The Chinese character is a logogram used in Chinese-based writing systems. It appears in different languages such as Japanese and Korean. Unlike characters in alphabetical systems like English, each Chinese character has its own meaning, and corresponds to a single syllable. There are two standard character sets used in different Chinese communities: Traditional Chinese and Simplified Chinese; we focus on Traditional Chinese in this thesis.

Number of Strokes ($\mathcal{C}_{strk}$)

The Chinese character is composed of basic “symbols” called *strokes*, shown in Figure 3.2. Intuitively, a character with a higher stroke count will be more complex, and this will increase the difficulty of reading the material. Table 3.2 shows examples of Chinese characters and their stroke counts.

While previous research works have mainly considered total stroke counts of a character, we propose another approach to stroke analysis - *stroke count excluding the radical* ($\mathcal{C}_{strk\_exRad}$). When the radical part of a character can be easily recognized (like the aforementioned familiar radicals), readers may simply ignore that part and just examine the remaining strokes. As a result, it may be possible to discover a relationship between this factor and readability. Table 3.2 also illustrates this quantity.

Table 3.2: Example of Stroke Count.

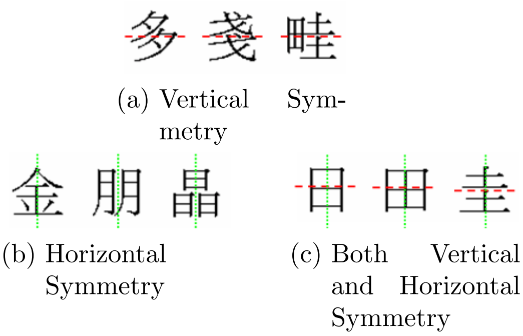
Character	**No. of Strokes**	**No. of Strokes excluding the Radical**
一	1	0
人	2	0
中	4	3
妾	8	5
龍	16	0
豔	28	21

Geometry Complexity ($\mathcal{C}_{symm}$ and $\mathcal{C}_{struct}$)

In this part we analyze the geometry, or shape, of Chinese characters. Yang [81] has conducted an experiment to test students' ability to recognize Chinese characters. He finds that symmetrical characters are relatively easier to recognize. Inspired by this finding, we expand this factor by investigating two geometric characteristics, namely *Symmetry* and *Structure*.

As an extension to *symmetry* ($\mathcal{C}_{symm}$) investigated by Yang, we classify characters into four categories: *Asymmetry*, *Vertical* symmetry, *Horizontal* symmetry, and *Both*. Figure 3.3 illustrates the three cases of symmetry.

Structure ($\mathcal{C}_{struct}$) describes the pattern of a character formed



(a) Vertical Symmetry

(b) Horizontal Symmetry

(c) Both Vertical and Horizontal Symmetry

Figure 3.3: Chinese Symmetry Examples.

from some *sub-parts*, which may or may not be real characters. We classify the structure into seven categories based on arrangement and the number of sub-parts, illustrated in Table 3.3.

Table 3.3: Structure Category.

Cat.	**No. Of Sub-parts**	**Arrangement**	**Examples**
A	1	Single	日, 月, 金
B	2	Vertical	昌, 星, 李
C	2	Horizontal	朋, 欽, 維
D	3	Vertical	賣, 鼻
E	3	Horizontal	砌, 街
F	3	Triangle	雙, 翠
G	>4	Other	綴, 㗊

Character Familiarity ($\mathcal{C}_{fam}$ and $\mathcal{C}_{freq}$)

Familiar characters are common, basic characters. They are well-known and used frequently. If a passage contains a large number of familiar characters, then intuitively the passage should be understood easily. There are different ways to build the familiar character list. For example, it can be obtained directly from the educational department of the government, or by conducting a survey. In our approach, we determine whether a character is familiar or not based on their the frequency of usage ($\mathcal{C}_{freq}$). The procedure for building the list is discussed in Section 3.4.2.

Character Grade ($\mathcal{C}_{grade}$)

Recently, a study [38] has been conducted with the aim of preparing recommendations of Chinese characters for Chinese language learning in primary schools. It lists the recommended grade level of 3000 Chinese characters. For example, a simple character such as “ 一”(one) has a recommended level of grade 1 (primary 1), while the difficult character “ 盪” (swing) has a recommended level of grade 5 (primary 5). So by making use of the list, we obtain an approximate grade level of a passage by calculating the average recommended grade levels of characters. As the list is designed only for primary level, we assume any character not in the list to be level “7”, which is one level higher than primary 6.

Common Characters in All Grade Levels ($\mathcal{C}_{common}$)

Apart from getting useful factors from external sources, such as the aforementioned character grade, we can obtain another list of characters which appear in textbooks of all grade levels that we are going to analyze (i.e. from primary 1 to secondary 4/5). Intuitively, if a character exists in all grade levels, it should be very common and less difficult. This also acts as another “familiar character list”, as discussed previously.

Word Level ($\mathcal{W}$)

Introduction

The concept of “word” in the Chinese language is less clear than in other language systems, and some researchers have even proposed that Chinese “doesn’t have words” [52]. In this thesis, we follow the definition in Yang's work [81] that the Chinese word is the smallest independently useable part within a sentence. We investigate the influences of word formation and familiarity on readability.

Word Length and Pattern ($\mathcal{W}_{len\_char}$ and $\mathcal{W}_{pattern}$)

Word length ($\mathcal{W}_{len\_char}$) is the number of characters in a word. According to [81], the average word length in Chinese text is about 1.5 characters. Intuitively, the longer a word is, the more difficult it is to comprehend.

Apart from word length, the pattern of a word ($\mathcal{W}_{pattern}$) may also affect the difficulty of comprehension. To facilitate our investigation, we classify the words into twelve categories ([A-L]), as shown in Table 3.4, based on word length and pattern. For example, both category B and category C represent the words with two characters, but for category B, the two characters are the same, while for category C, the two characters are different. " 媽媽" and " 母親" (both have the meaning of "mother") are examples of the two categories respectively.

Word Familiarity ($\mathcal{W}_{fam}$)

As for radicals and characters, a familiar word list can be built for readability analysis. The proportion of familiar words in a document is measured to predict the readability. The ways of building a familiar word list include: whether characters inside the word can be found in the familiar character list, the number of characters involved, frequency of the word's usage, and finally from a survey. We discuss our way of building the list in Section 3.4.2.

Table 3.4: Word Pattern.

Cat.	**Word Length**	**Unique Char**	**Pattern**
A	1	1	A
B	2	1	AA
C	2	2	AB
D	3	1	AAA
E	3	2	AAB, ABA, BAA
F	3	3	ABC
G	4	1	AAAA
H	4	2	AAAB, AABA, ABAA, BAAA
I	4	2	AABB, ABAB
J	4	3	AABC, ABAC, ABCA, BAAC, BACA, BCAA
K	4	4	ABCD
L	≥ 5	–	Other

Common Words in All Grade Levels ($\mathcal{W}_{common}$)

By analogy with the concept of the common character list $\mathcal{C}_{common}$ discussed in the previous section, a list of common words which appear in all grade level textbooks can also be built. For example, 小心 (careful) and 清潔 (clean) are two common words. If there is a larger proportion of common words, then the passage should be easier to comprehend.

Phrase Level ($\mathcal{P}$)

Introduction

In our analysis, a *phrase* is an incomplete sentence, that is, a sequence of words delimited by a "comma"(，). It serves as an intermediate level between word and sentence.

Effect of Idioms ($\mathcal{P}_{idiom}$)

Chinese idioms are commonly used to illustrate a relatively complex concept using only a few characters. The typical length of a Chinese idiom is four. Whether a passage contains idiom may affect its readability. As each idiom often has a story behind it, its semantic meaning may be *deeper* and more complex than the *surface* meaning. So its effect on readability will operate in two completely different manners. On the one hand, if a passage contains well-known idioms, it will be easier to comprehend as fewer characters are needed to present a complex concept. But on the other hand, if a passage contains rarely-used idioms, readers cannot fully understand its meaning unless they know the story behind the idioms. As a result, the passage's difficulty will increase. Taking an English idiom as an analogue, the idiom "Burn one's bridges" (or "burn one's boats") has an origin related to Julius Caesar. In 49 B.C. the Roman emperor Caesar commanded the burning of all bridges and boats after the army passed the Lupigen River, showing his determination to win the war. So the idiom has the meaning of "to cut oneself off from all means or hope of retreat". If a reader does not know the story behind the idiom, he/she cannot fully understand the meaning.

Phrase Length ($\mathcal{P}_{length}$)

We consider phrase length to be the the total number of strokes, characters and words. Intuitively, the longer the phrase, the more complex the phrase is, and thus it is more difficult to comprehend.

Sentence Level ($\mathcal{S}$)

Introduction

A sentence is a group of words (or a single word) that expresses a complete thought or idea. It usually contains a subject (either explicit or implicit) and a predicate containing a finite verb.

Sentence Length ($\mathcal{S}_{length}$)

We consider sentence length to be the number of strokes, characters, words, and phrases in our analysis. As for phrase length, the longer the sentence, the more complex the sentence is, and thus it is more difficult. We collect average sentence length in our investigation.

Sentence Structure

The complexity of sentence structure has a direct effect on readability. The more complex a sentence, the harder it is to comprehend. Previous research has made little investigation of sentence structure, instead using sentence length to infer sentence structure complexity. In Yang's work [81], he studies the effect of whether a sentence is a *full sentence* ($\mathcal{S}_{fullsent}$) (sentence having both subject and predicate), but we think that a deeper investigation of sentence structure is necessary. We consider the number of word classes involved in a sentence ($\mathcal{S}_{tag}$). A sophisticated Part-of-Speech (POS) tagger [48] is used to assist the analysis. If a sentence contains many word classes, then the sentence will be more complex. As a result, this quantity acts as an indicator of sentence complexity.

3.1.2 Feature Extraction

After describing the possible factors affecting Chinese readability in previous sections, we list the features, or attributes, which are used in the assessment. Features are various statistics obtained from a passage, such as the average sentence length, the average number of familiar words, etc. Table 3.5 shows the complete list of features.

Table 3.5: Summary of Extracted Features.

Index	Factor	Feature Names
Sub-character Level		
1-2.	$\mathcal{R}_{stroke}$	Average and Standard deviation of number of radical strokes per Chinese character
3-4.	$\mathcal{R}_{fam}$	Proportion of familiar and unfamiliar radicals
Character Level		
5-6.	$\mathcal{C}_{strk}$	Average and Standard deviation of number of strokes per Chinese character
7-8.	$\mathcal{C}_{strk\_exRad}$	Average and Standard deviation of number of strokes excluding the radical per Chinese character
9-10.	$\mathcal{C}_{fam}$	Proportion of familiar and unfamiliar characters
11-15.	$\mathcal{C}_{symm}$	Proportion of Symmetrical, Non-symmetrical, Vertical, Horizontal and Both Symmetrical characters
16-22.	$\mathcal{C}_{struct}$	Proportion of characters belonging to *Structure Category*[A-G]
23-24.	$\mathcal{C}_{grade}$	Average and Standard deviation of character grade
25-26.	$\mathcal{C}_{common}$	Proportion of common and uncommon characters
27-28.	$\mathcal{C}_{freq}$	Average and Standard deviation of character frequency of occurrence
Word Level		
29-30.	$\mathcal{W}_{fam}$	Proportion of familiar and unfamiliar words
31-32.	$\mathcal{C}_{stk}$, $\mathcal{W}_{length}$	Average and Standard deviation of number of strokes per word
33-34.	$\mathcal{W}_{length}$	Average and Standard deviation of number of characters per word
35-46.	$\mathcal{W}_{pattern}$	Proportion of words belonging to *Word Pattern Category*[A-L]
47-48.	$\mathcal{W}_{common}$	Proportion of common and uncommon words
Phrase Level		
49.	$\mathcal{P}_{idiom}$	Proportion of phrases containing idioms
50-55.	$\mathcal{P}_{length}$	Average and Standard deviation of number of strokes, characters, words per phrase
Sentence Level		
56-61.	$\mathcal{S}_{length}$	Average and Standard deviation of number of strokes, characters, words per sentence
62.	$\mathcal{S}_{fullsent}$	Proportion of full sentences
63-64.	$\mathcal{S}_{tag}$	Average and Standard deviation of number of unique POS tags in sentence

3.1.3 Limitation of Our Analysis and Possible Extension

In this subsection, we discuss our limitations and some possible further extensions to the task of readability factor analysis. We discuss the issue at each level as follows.

For the Sub-character level, we analyze only the radical part of the character. However, apart from the radical, there are other sub-parts, or sub-components, constituting a character. As a result, these could also be taken into consideration at this level.

For the Character level, the first limitation is that our analysis focuses only on Traditional Chinese, so one of the extensions is to cover Simplified Chinese texts. Another extension is related to character familiarity. Other measures of character familiarity, such as extending the recommended level of Chinese characters to secondary school grades, can be carried out to make the evaluation better.

For the Word level, as the quality of statistics collected at this level depends strongly on the word segmentation performance, the segmentation technique should be improved. In particular, it is important to segment name entities and newly evolved words correctly, so as to make the readability analysis more robust and less sensitive to changes over time.

For the Phrase level, in our analysis, we only consider whether an idiom exists in the phrase. But actually, as with the radical, character, and word levels, we can also differentiate the familiarity of idioms. For example, a similar grade recommendation list can also be constructed for the Chinese idioms.

Finally, for the Sentence level, first of all, more sophisticated natural language processing technique should be applied to analyze the sentence structure. Second, sentence analysis should be enhanced to the syntactic level, such as grammatical analysis, or even to the semantic level, in order to capture more useful factors.

3.2 Research Methodology

In this section, we discuss the research methodology used for Chinese readability analysis. We describe the readability definition involved, data acquisition, text processing, feature extraction, regression analysis, and the evaluation method.

3.2.1 Definition of Readability

The purpose of our work is to estimate the readability level Y (dependent variable) of a piece of *Traditional* Chinese text X based on its features (independent variables). Note that the term *readability level* has different meanings in the literature, and we adopt the following definition.

Definition 1 *The Readability level Y of a Traditional Chinese passage X represents the grade level at which people belonging to this level will find the passage appropriate to them.*

For example, if $Y = 5$, then X is suitable for people of grade level 5. To simplify the wording, we say that the readability level is the grade level of the passage.

We adopt the Hong Kong education system for the grade level. This is scaled between [1 to 13], where [1 to 6] represents primary 1 to primary 6, and [7 to 13] represents secondary 1 to secondary 7. However, as secondary 4 and 5 relate to the Hong Kong Certificate of Education Examination (HKCEE) syllabus, and the syllabus of secondary 6 and 7 is different from that of previous years (it is called “Chinese Language and Culture”), we merge secondary 4 and 5 to one level, and omit secondary 6 and 7. Thus the resulting scale will be [1 to 10].

Before continuing the discussion, it is necessary to point out that, the readability discussed is just one of the tools to evaluate the quality of a passage based on the textual features. General qualities of a passage, such as understandability, comprehensibility and usability (as discussed in [56]) should consider more factors, such as grammatical correctness, document layout and organization. As a result, one should not solely depend on the readability measure when judging the quality of a passage.

3.2.2 Data Acquisition and Sampling

The training data used in establishing our Chinese readability assessment is based on passages in Chinese language textbooks of primary schools and secondary schools [86, 87]. In Hong Kong, each grade level has two semesters, and one textbook is used for each semester. In each textbook, there are about 10 - 30 chapters. The higher the level, the smaller the number of chapters. Each chapter contains an *article*, in the form of poems, drama scripts, argumentative articles, etc. In our study, the contents of the selected articles are transferred to electronic format by making use of a scanner and OCR software. After correcting recognition errors manually, the articles are used as training data.

After preparing the articles, we need to sample some appropriate articles used for training and testing purposes. In our analysis, we only select general articles in the data set to avoid articles with special formats, such as poems, dialogues and scripts, which may upset the analysis. As a result, we omit chapters in primary 1 textbooks because they are only simple sentences. The selected articles are passed to the next steps for further analysis.

To summarize, Table 3.6 shows details of the data set extracted from different textbooks.

Table 3.6: Summary of Data Set.

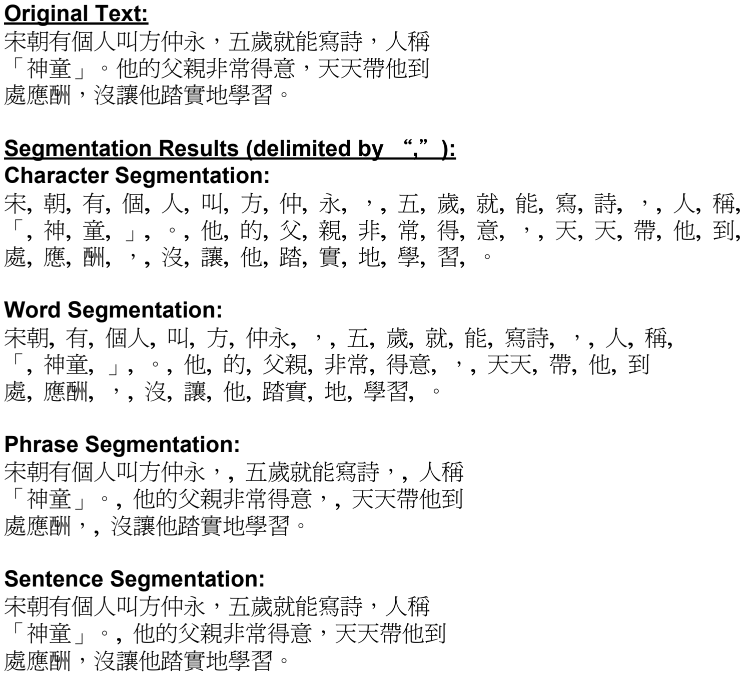
Actual Level	Grade Level	No. of Textbooks	No. of Selected Articles
Pri 2	2	2	25
Pri 3	3	2	22
Pri 4	4	2	20
Pri 5	5	2	20
Pri 6	6	2	19
Sec 1	7	2	21
Sec 2	8	2	19
Sec 3	9	2	19
Sec 4/5	10	4	11
			Total: 176

3.2.3 Text Processing and Feature Extraction

After compiling the data set, each passage then undergoes text processing, which is a pre-processing stage ahead of feature extraction. The main text processing is *segmentation*. As described in Section 3.1.1, factors are categorized into *Radical*, *Character*, *Word*, *Phrase*, and *Sentence*. As a result, in order to extract features from each category, it is required to segment the text according to these categories.

Segmentation by radical is accomplished by table lookup using segmented characters, which is a trivial task. For word segmentation, we employ a technique called *LMR-RC Tagging*, to be discussed in Section 3.4.1. For phrase segmentation, we use the punctuation mark *Comma* (，) as the delimiter, which is a practice adopted in [81]. Finally for sentence segmentation, we apply common delimiters such as *Chinese full stop* (。), *Question mark* (？), *Exclamation mark* (！), end of *Chinese Quotation* (「」), and *Semi-colon* (；). Figure 3.4 shows the segmentation results of a sample text.

After segmenting a passage into different levels, we extract the features based on the discussion in Section 3.1. The list of



Original Text:
宋朝有個人叫方仲永，五歲就能寫詩，人稱
「神童」。他的父親非常得意，天天帶他到
處應酬，沒讓他踏實地學習。

Segmentation Results (delimited by “,”):
Character Segmentation:
宋, 朝, 有, 個, 人, 叫, 方, 仲, 永, ，, 五, 歲, 就, 能, 寫, 詩, ，, 人, 稱,
「, 神, 童, 」, 。, 他, 的, 父, 親, 非, 常, 得, 意, ，, 天, 天, 帶, 他, 到,
處, 應, 酬, ，, 沒, 讓, 他, 踏, 實, 地, 學, 習, 。

Word Segmentation:
宋朝, 有, 個人, 叫, 方, 仲永, ，, 五, 歲, 就, 能, 寫詩, ，, 人, 稱,
「, 神童, 」, 。, 他, 的, 父親, 非常, 得意, ，, 天天, 帶, 他, 到
處, 應酬, ，, 沒, 讓, 他, 踏實, 地, 學習, 。

Phrase Segmentation:
宋朝有個人叫方仲永，, 五歲就能寫詩，, 人稱
「神童」。, 他的父親非常得意，, 天天帶他到
處應酬，, 沒讓他踏實地學習。

Sentence Segmentation:
宋朝有個人叫方仲永，五歲就能寫詩，人稱
「神童」。, 他的父親非常得意，天天帶他到
處應酬，沒讓他踏實地學習。

Figure 3.4: Illustration of Text Segmentation.

extracted features is shown in Table 3.5.

3.2.4 Regression Analysis using Support Vector Regression

We apply Support Vector Regression (SVR) [60] as the modeling technique in our analysis. It is because SVR is good at exploring nonlinear relationships between dependent and independent variables, and also because it has superior performance in solving other regression problems. We discuss the basic concepts of SVR in Section 3.3.

3.2.5 Evaluation

To evaluate our proposed scheme, we measure the training accuracy and cross-validation accuracy based on standard metrics in regression analysis. Let Y_i and $\hat{Y_i}$ be the actual and predicted levels of passage X_i respectively, N be the number of testing passages. The following shows the metrics.

- Max. Prediction Error (MPE):

$$MPE = \max_{1 \leq i \leq N} |Y_i - \hat{Y}_i|. \tag{3.1}$$

- Mean and Standard deviation of Absolute Error (MAE and $STDDEV\_AE$):

$$\text{Let } AE_i = |Y_i - \hat{Y}_i|, \tag{3.2}$$

$$MAE = \frac{\sum_{i=1}^{N} AE_i}{N}, \tag{3.3}$$

$$STDDEV\_AE = \sqrt{\frac{\sum_{i=1}^{N}(AE_i - MAE)^2}{N-1}}. \tag{3.4}$$

- Mean and Standard deviation of Squared Error (MSE and $STDDEV\_SE$):

$$\text{Let } SE_i = (Y_i - \hat{Y}_i)^2, \tag{3.5}$$

$$MSE = \frac{\sum_{i=1}^{N} SE_i}{N}, \tag{3.6}$$

$$STDDEV\_SE = \sqrt{\frac{\sum_{i=1}^{N}(SE_i - MSE)^2}{N-1}}. \tag{3.7}$$

- Pearson Correlation Coefficient and Squared Correlation Coefficient (r and r^2):

$$r^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)}{\sum_{i=1}^{N}(Y_i - \bar{Y}_i)}. \tag{3.8}$$

In addition to the standard metrics, we follow Jeng's approach [31] of using a metric called *Hit Rate* (HitRate), which is defined as follows:

Definition 2 *Hit Rate is the proportion of testing passages with prediction errors less than a predefined error ϵ. Mathematically,*

$$\textit{Let } f(Y_i, \hat{Y_i}) = \begin{cases} 1 & \textit{if } |Y_i - \hat{Y_i}| < \epsilon \\ 0 & \textit{otherwise} \end{cases} \tag{3.9}$$

$$HitRate \pm \epsilon = \frac{\sum_{i=1}^{N} f(Y_i, \hat{Y_i})}{N}. \tag{3.10}$$

3.3 Introduction to Support Vector Regression

In this section, we introduce the Support Vector Regression (SVR) technique based on discussion in [60]. We describe the basic concept of SVR and a technique used for non-linear extension.

3.3.1 Basic Concept

Given the training data $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset X \times Y$, where X is the space of input patterns, and Y is the outcome in real number. In the context of Chinese readability assessment, $x_i \subset X$ is a feature introduced in Section 3.1, like the average number of strokes per character and the average number of characters per word; $y_i \subset Y$ is readability score in terms of grade level. Our goal in applying SVR technique is to find a function $f(x)$ (in Equation (3.11)) that has at most ε deviation to the actual outcome y, i.e., $\forall x_i, |f(x_i) - y_i| \leq \varepsilon$. We first describe $f(x)$ in the case of linear regression, and extend to non-linear regression using kernel technique in next subsection.

$$f(x) = \langle w, x \rangle + b \text{ with } w \in X, b \in \mathbb{R}, \tag{3.11}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, and b is a real number. Among all the feasible $f(x)$, we want the function as *flat* as possible so as to reduce the function complexity and retain generality in prediction. It is achieved by finding w as small as possible. This is done by minimizing the norm, i.e. $\|w\|^2 = \langle w, w \rangle$ [60]. Equation (3.12) defines the corresponding optimization problem.

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2, \\ &\text{subject to} \begin{cases} y_i - (\langle w, x \rangle + b) \le \varepsilon \\ (\langle w, x \rangle + b) - y_i \le \varepsilon \end{cases} . \end{aligned} \tag{3.12}$$

But it is difficult to have a f which can approximate all pairs of (x_i, y_i) with precision ε. As a result, in order to make the optimization possible, we allow some data points to break the restriction, and *slack variables* ξ and ξ^* are introduced to represent errors. Equation (3.13) defines the modified optimization problem, in which the errors are minimized at the same time.

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi + \xi^*), \\ &\text{subject to} \begin{cases} y_i - (\langle w, x \rangle + b) \le \varepsilon + \xi \\ (\langle w, x \rangle + b) - y_i \le \varepsilon + \xi^* \\ \xi, \xi^* \ge 0 \end{cases} . \end{aligned} \tag{3.13}$$

The constant C is the tradeoff between flatness of f and the amount of error toleration. Under this setting, only data points outside the ε region will contribute to the cost. We then solve this optimization problem using classic method utilizing *Lagrangian Multiplier* [21]. This technique involves solving the original objective function (called primal function) indirectly using a dual set of variables, called lagrange multipliers. Equation

(3.14) defines the corresponding Lagrangian.

$$\begin{aligned} L :=& \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi + \xi^*) - \sum_{i=1}^{l}(\eta_i\xi + \eta_i^*\xi^*) \\ &- \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + (\langle w, x\rangle + b)) \\ &- \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - (\langle w, x\rangle + b)). \end{aligned} \tag{3.14}$$

Here α_i, α_i^* (denoted by $\alpha_i^{(*)}$), and η_i, η_i^* (denoted by $\eta_i^{(*)}$) are Lagrange multipliers, and they need to satisfy positivity constraints, i.e., they need to be greater than or equal to zero.

After having the Lagrangian L, the problem is converted to an unconstrained optimization problem. We set the partial derivatives of L respect to primal variables (w, b, ξ_i, ξ_i^*) to zero in order to obtain the optimal solution:

$$\partial_b L = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0, \tag{3.15}$$

$$\partial_w L = w - \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)x_i = 0, \tag{3.16}$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0. \tag{3.17}$$

Substituting Equations (3.15), (3.16) and (3.17) into Equation (3.14) results in the following dual optimization problem, which can then solved by various numerical methods:

$$\begin{aligned} &\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases}, \\ &\text{subject to} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) \text{ and } \alpha_i, \alpha_i^* \in [0, C]. \end{aligned} \tag{3.18}$$

We obtain the following observation during the derivation of Equation (3.18).

1. According to condition in Equation (3.17), the dual variables $\eta_i^{(*)}$ can be rewritten as $\eta_i^{(*)} = C - \alpha_i^{(*)}$, and are then eliminated in Equation (3.18);

2. According to condition in Equation (3.16), w can be rewritten as follows

$$\begin{aligned} w &= \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)x_i, \text{ and thus} \\ f(x) &= \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\langle x_i, x\rangle + b. \end{aligned} \tag{3.19}$$

   This expression is then called Support Vector expansion [60], in which $f(x)$ can be described in terms of dot products between training data. We do not need to compute w explicitly when evaluating $f(x)$. This observation is useful for non-linear extension using kernel technique.

3.3.2 Non-Linear Extension using Kernel Technique

So far we only discuss the case of applying SVR in linear regression problem, and we now discuss the use of kernel technique to extend the application in non-linear cases.

The basic idea of non-linear extension is to preprocess the training data by mapping them into another feature space, in which the data in that space will behave linearly. An example is given in [60] about quadratic feature: Consider the map $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ with $\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, where x_1 and x_2 are the components of $x \in \mathbb{R}^2$. Training the linear SVR using preprocessed features would yield a quadratic function.

Equation (3.20) shows the corresponding $f(x)$.

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle \Phi(x_i), \Phi(x) \rangle + b. \tag{3.20}$$

Although the use of mapping in data preprocessing seems reasonable, computational complexity will greatly increase after preprocessing data, and thus making the method infeasible. So, instead of explicitly calculating the resulting mapped features, implicit mapping via kernel k is proposed by [6] based on the following observation on the discussed quadratic feature:

$$\begin{aligned} \langle \Phi(x), \Phi(x') \rangle &= \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2), (x_1^2, \sqrt{2}x_1x_2, x_2^2) \rangle \\ &= \langle x, x' \rangle^2. \end{aligned} \tag{3.21}$$

As discussed in previous subsection that SVR is only interested in dot product of the x_i, it is sufficient to know the $k(x, x') := \langle \Phi(x), \Phi(x') \rangle$ rather than the mapping Φ explicitly. We can observe in Equation (3.21) that, the kernel function $\langle x, x' \rangle^2$ is much simpler than Φ, and thus making the non-linear extension of SVR feasible. Equation (3.22) restates the new $f(x)$.

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b. \tag{3.22}$$

Common kernels used in SVR include *Polynomial kernel*, *Radial basis function kernel* and *Sigmoid kernel*. In our experiment, we compare performances of different kernel functions in prediction of Chinese readability.

3.4 Implementation Details

3.4.1 Chinese Word Segmentation

Chinese word segmentation is a non-trivial task because no explicit delimiters (like spaces in English) are used for word separation. As the task is an important precursor to many natural

Table 3.7: Tags used in LMR Tagging scheme.

Tag	Description
L	Character is at the beginning of the word (or the character is the leftmost character in the word)
M	Character is in the middle of the word
R	Character is at the end of the word (or the character is the rightmost character in the word)
S	Character is a ”single-character” word

Original sentence:	我喜歡吃西瓜，他喜歡吃士多啤梨。										
After segmentation:	我	喜歡	吃	西瓜	，	他	喜歡	吃	士多啤梨	。	
Tagging:	S	L R	S	L R	S	S	L R	S	L MMR	S	

Figure 3.5: Example of LMR Tagging.

language processing systems, it receives a lot of attentions in the literature for the past decade [78, 62]. In our implementation, we propose and apply a statistical approach based on the works of [79], in which the Chinese word segmentation problem is first transformed into a tagging problem, then the Maximum Entropy classifier is applied to solve the problem. We further improve the scheme by introducing correctional treatments after the first round tagging. Two different training methods are proposed to suit our scheme.

Chinese Word Segmentation as Tagging

One of the difficulties in Chinese word segmentation is that, Chinese characters can appear in different positions within a word [79], and *LMR Tagging* was proposed to solve the problem. The basic idea of LMR Tagging is to assign to each character, based on its contextual information, a tag which represents its relative position within the word. Note that the original tag set used by [79] is simplified and improved by [50]. We shall then adopt and illustrate the simplified case here.

The tags and their meanings are summarized in Table 3.7. Tags *L*, *M*, and *R* correspond to the character at the beginning, in the middle, and at the end of the word respectively. Tag *S* means the character is a “single-character” word. Figure 3.5 illustrates a Chinese sentence segmented by spaces, and the corresponding tagging results.

After transforming the Chinese segmentation problem to the tagging problem, various solutions can be applied. The *Maximum Entropy model* (MaxEnt) [4] [55] was proposed in the original work to solve the LMR Tagging problem. In order to make MaxEnt succeed in LMR Tagging, *feature templates* used in capturing useful contextual information must be carefully designed. Furthermore, it is unavoidable that invalid tag sequences will occur if we just assign the tag with the highest probability. In the next subsection, we describe the feature templates and the measures used to correct the tagging.

Two-Phase LMR-RC Tagging

In this section, we introduce our *Two-Phase LMR-RC Tagging* used to perform Chinese word segmentation. The first phase, *R*-phase, is called *Regular Tagging*, in which similar procedures as in the original LMR Tagging are performed. The difference in this phase as compared to the original one is that, we use extra feature templates to capture characteristics of Chinese word segmentation. The second phase, *C*-phase, is called *Correctional Tagging*, in which the sentences are re-tagged by incorporating the regular tagging results. The models used in both phases are trained using MaxEnt model.

Regular Tagging Phase In this phase, each character is tagged similar to the way in the original approach. In our scheme, given the contextual information (x) of the current character, the tag

(y^*) with highest probability will be assigned:

$$y^* = \arg\max_{y \in \{L,M,R,S\}} p(y|x). \tag{3.23}$$

Table 3.8: Feature templates used in R-phase. Example used is “32 個蘋果”.

Index	**Feature Type**	**Example** – Features extracted of character “ 個”
1	Characters within a window of ±2	C_{-2}=“3”, C_{-1}=“2”, C_0=“ 個”, C_1=“ 蘋”, C_2=“ 果”
2	Two consecutive characters within a window of ±2	$C_{-2}C_{-1}$=“32”, $C_{-1}C_0$=“2 個”, C_0C_1=“ 個蘋”, C_1C_2=“ 蘋果”
3	Previous and next characters	$C_{-1}C_1$=“ 2蘋”
4	Current character is punctuation	–
5	ASCII characters within a window of ±2	A_{-2}, A_{-1} (as “3” and “2” are ASCII)
6	Current and character in window ±1 belong to different types	D_{-1} (as “2” is digit, but “ 個” is letter)

The features describing characteristics of Chinese segmentation problem are instantiations of the feature templates listed in Table 3.8. Note that the feature templates only describe the forms of features, but not the actual features. So the number of features used is much larger than the number of templates.

Additional feature templates as compared to [79] and [50] are templates 5 and 6. Template 5 is used to handle documents with ASCII characters. For template 6, as it is quite common that word boundary occurs in between two characters of different types, this template is used to capture such characteristics.

Table 3.9: Additional feature templates used in C-phase. Example used is "32 個蘋果" with tagging results after R-phase as "SSLMR".

Index	**Feature Type**	**Example** – Features extracted of character " 個"
7	Tags of characters within a window of ±2	T_{-2}="S", T_{-1}="S", T_0="L", T_1="M", T_2="R"
8	Two consecutive tags within a window of ±2	$T_{-2}T_{-1}$="SS", $T_{-1}T_0$="SL", T_0T_1="LM", T_1T_2="MR"
9	Previous and next tags	$T_{-1}T_1$="SM"

Correctional Tagging Phase In this phase, the sequence of characters is re-tagged by using the additional information of tagging results after the R-phase. The tagging procedure is similar to the previous phase, except extra features (listed in Table 3.9) are used to assist the tagging.

Training Method Two training methods are proposed to construct models used in the R- and C-phase: (1) *Separated Mode* and (2) *Integrated Mode.* Separated Mode means the models used in the two phases are separated. The tagging model for R-phase is called R-model, and the model for C-phase is called C-model. Integrated Mode means only one model, I-model is used in both phases.

The training methods are illustrated now. First of all, training data are divided into three parts, (1) Regular Training, (2) Correctional Training, and (3) Evaluation. Our method first trains using observations extracted from Part 1 (observation is simply the pair $(context, tag)$ of each character). The created model is used to process Part 2. After that, observations extracted from Part 2 (which include previous tagging results) are used to create the final model. The performance is then evaluated by processing Part 3.

Let O be the set of observations, with subscripts R or C indicating the sources of them. Let $TrainModel : O \rightarrow P$, where P is the set of models, be the "model generating" function. The two proposed training methods can be illustrated as follow:

1. Separated Mode

$$\begin{aligned} R\text{-}model &= TrainModel(O_R), \\ C\text{-}model &= TrainModel(O_C). \end{aligned}$$

2. Integrated Mode

$$I\text{-}model = TrainModel(O_R \cup O_C).$$

The advantage of the Separated Mode is that, it is easy to aggregate different sets of training data. It also provides a mean to handle large training data under limited resources, as we can divide the training data into several parts, and then use the similar idea to train each part. The drawback of this mode is that, it may lose the features' characteristics captured from Part 1 of training data, and the Integrated Mode is proposed to address the problem, in which all the features' characteristics in both Part 1 and Part 2 are used to train the model.

We need to obtain basic character and word lists in order to measure the proportion of familiar and unfamiliar characters and words. We try to build the two lists by using the entire Traditional Chinese character and word lists obtained from [66, 67].

3.4.2 Building Basic Chinese Character / Word Lists

The character list contains both the stroke number and frequency of occurrence of each character, while the word list contains the frequencies of common words. Table 3.10 shows their information content.

Table 3.10: Details of Character and Word List.

List	**No. of Entries**	**Encoding**	**Information**
Character	13,060	Big5	Stroke, Frequency
Word	138,614	Big5	Frequency

Algorithm 1 Building Basic Character and Word Lists.

```
Input: Complete Character List $C$ and Word List $W$
Output: Basic Character List $C_B$ and Word List $W_B$
Initialize set $C_B$ and $W_B$ = empty set
Sort-Descending-Order-by-Frequency($C$)
$C_B$ = First-1500-Entries($C$)
for all $w$ in $W$ do
    if for each character $c'$ in $w$, $c' \in C_B$ then
        Append($W_B$, $w$)
    end if
end for
Sort-Descending-Order-by-Frequency($W_B$)
$W_B$ = First-5600-Entries($W_B$)
return $C_B$, $W_B$
```

We determine whether a character is basic based on its frequency of occurrence. Assuming characters with high frequency of occurrence is the interpretation of “basic” or “familiar”, our final basic character list is thus composed of the 1,500 characters with the highest frequencies.

We build the basic word list based on the basic character list created in the aforementioned way and the frequency of occurrence. We assume basic words are those consisting of purely basic characters and having high frequencies. For example, a basic character list contains characters “ 老” (old) and “ 師” (teacher), then the word “ 老師” (teacher) is included in the basic word list if it also has high frequency of occurrence. 5,600 words are then selected in this way to build the word list. Algorithm 1 shows the algorithm of the lists building. In the algorithm, we first sort the character list C according to frequency in

Algorithm 2 Full Sentence Detection.

```
Input: Sentence $S$
Output:Yes if $S$ is full sentence, No otherwise
$Tags$ = POS ($S$)
$flag = 0$
$result$ = NO
for $i$ : 1 to Num(Tags) do
    if $Tags[i] == N$ AND $flag == 0$ then
        $flag = 1$
    end if
    if $Tags[i] == V$ AND $flag == 1$ then
        $result$ = YES
    end if
end for
return $result$
```

descending order, then we obtain the basic character list C_B by getting first 1500 entries. After that, we examine each word to determine whether each character in it is in basic character list. If it is the case, then the word is put in the potential basic word list. By getting the first 5600 entries after sorting in descending order of frequency, the basic word list W_B is obtained.

3.4.3 Full Sentence Detection

Yang defines a *full sentence* as a sentence with both subject and predicate; we use the Chinese Part-of-Speech (POS) Tagging approach to decide whether a sentence is a full sentence. A proper approach to the detection should involve sentence structural parsing, but to simplify the implementation, we make use of the fact that it is not usual for a Chinese full sentence for the predicate (verb) to come before the subject (noun), and so we consider that a sentence is a full sentence only when noun-related tags come before verb-related tags. The POS tagger is implemented by the Fujitsu research group, and details of the tagger can be found in [48]. Algorithm 2 shows the algorithm used for judging a full sentence. In the algorithm, we first get the part-of-speech tags sequence by the function $POS()$. Then we determine whether the noun-related tags come before the verb-related tags. If it is the case, a result YES will be returned.

3.4.4 Feature Selection Using Genetic Algorithm

In Section 3.1, we have systematically listed out various potential factors related to Chinese readability based on the characteristics of the Chinese language. However, not all factors discussed are useful in the estimation, as some of them may be useless, redundant, or even noisy. In this section, we describe a method to select significant features which are useful in estimating Chinese Readability. In addition to improving the estimation accuracy, performing feature selection can reduce the computational complexity as the feature dimension is reduced, and at the same time, we can discover the factors which have stronger relation to readability.

In this work, we apply the Genetic Algorithm (GA) as the feature selection method. The GA is a general search algorithm that imitates the evolution processes in nature [72]. It is an effective technique in solving various optimization problems [2, 74, 34]. Feature selection using GA has received wide attention in the literature [51], and various research works have demonstrated the advantages of this approach. As our work does not focus on the field of the evolutionary computing and the genetic algorithm, we implement the approach in a basic, conventional way, with some necessary modifications to suit our problem. We describe the approach in the remaining section.

Problem Definition of Feature Selection

Feature selection (FS) is to search for a subset of d features from the entire D features which gives the best regression performance. Let F be the entire feature set, and $|F| = D$ be the number of features in the set. Feature selection process is to search for a feature subset $F' \in F$, with $|F'| = d \leq D$, such that an optimization criterion (e.g *mean square error* or *squared correlation coefficient*), denoted by J, can be achieved. Mathematically,

$$FS(F, d) = \arg\max_{F' \in F} J(F') \text{ s.t. } |F'| = d. \tag{3.24}$$

GA Implementation Details

A generic GA approach to solve a problem involves several basic modules. They are *Chromosome Encoding*, *Population Initialization*, *Crossover*, *Mutation*, *Fitness Evaluation*, *Selection*, and *Termination.* We describe each of them in the context of feature selection. The implementation is mainly following [51].

Chromosome Encoding

A chromosome c in the feature selection problem is a string of D binary digits (gene g), where D is the total number of features in the feature space. Each binary digit in the string represents a feature. If the digit is equal to 1, it means the corresponding feature is selected for regression analysis. On the other hand, the feature is discarded if the corresponding digit is 0. The number of selected features is determined by the input parameter d.

For example, according to Table 3.5, $D = 64$ in our Chinese readability analysis. If $d = 3$ and $\{5, 7, 14\}$-*th* digits in the chromosome are set to 1, that means *Proportion of unfamiliar radicals*, *Average number of strokes per Chinese character*, and *Standard deviation of unfamiliar characters* are the features selected for the regression analysis.

Population Initialization

The first population (a set of chromosomes) is generated randomly. Let P be the population, and $|P| = N_p$ be the population size. The initialization process is shown in Algorithm 3. In the algorithm, for each chromosome c in the population P, we randomly assign 0 and 1 to each gene g in c.

Algorithm 3 Population Initialization.

```
Input: Population size $N_p$
Output: Initial population $P$
Initialize $P$
for all $c$ in $P$ do
    for all $g$ in $c$ do
        $g = random(\{0, 1\})$
    end for
end for
return $P$
```

Crossover and Mutation

Crossover and mutation are GA operators used to mimic natural genetic evolution, such that better and better chromosomes will evolve when the GA proceeds [72].

For crossover, we adopt the m-point crossover proposed in [51]. Under this operator, m cutting points are randomly chosen, then each segment is copied out alternately from two parents to form two offsprings. For example, "01|11|00|01" and "11|10|10|00" are the two chromosomes which are undergoing 3-point crossover, with "|" indicating the cutting points. The two offsprings are "01|10|00|00" and "11|11|10|01", and are then added back to the population. We use a parameter *crossover probability* r_c to control the probability of whether two chromosomes should undergo crossover. Algorithm 4 shows the details. In the algorithm, we first generate a random number between 0 and 1. If it is smaller than r_c, two offsprings $c1'$ and $c2'$ are generated according to the copy mechanism discussed (function *alternate-copy*()).

Algorithm 4 $m-$point Crossover.

```
Input: Parents $c1$ and $c2$, No. of crossover point $m$ and Crossover prob. $r_c$
Output: Offsprings $c1'$ and $c2'$
if $random([0,1]) < r_c$ then
    $(c1', c2') = alternate\text{-}copy(c1, c2, m)$
end if
return $(c1', c2')$
```

Mutation is used to increase the chromosome diversity. For each chromosome in the population, each of its gene has a probability called *mutation rate* r_m of changing its value. Mutated chromosomes are added to the population afterwards. Algorithm 5 shows the details. In the algorithm, we generate a number between 0 and 1 for each gene in the chromosome. If it is smaller than r_m, the corresponding gene will be negated.

Algorithm 5 Mutation.

```
Input: Chromosome $c$ and Mutation rate $r_m$
Output: Mutated Chromosome $c'$
c' = c
for all $g$ in $c'$ do
    if $random([0,1]) < r_m$ then
        $g = \bar{g}$
    end if
end for
return $c'$
```

Chromosome Correction

Chromosome correction is to make the chromosome to satisfy the preset value of d. After crossover and mutation, the number of selected features in the offsprings may break the subset size requirement. To correct this problem, 0-1 and 1-0 conversions will be made randomly until the requirement is satisfied.

Algorithm 6 Roulette Selection.

```
**Input:** Population $P$, Number of selection $N_{sel}$
**Output:** New population $P'$
$evaluate(P)$
$P = sort\text{-}non\text{-}increasing\text{-}by\text{-}fitness(P)$
calculate $pr_i$ for each $c$ in $P$
**for** $j = 0$ to $N_{sel}$ **do**
    $ran = random([0, pr_n])$
    choose $c_i$ such that $pr_{i-1} < ran < pr_i$
    add $c_i$ to $P'$
**end for**
**return** $P'$
```

Fitness Evaluation, Selection, and Termination

The fitness of a chromosome is the regression performance obtained using the selected features represented by that chromosome. In our experiment, we apply three different metrics to perform the feature selection process as discussed in Section 3.2.5. These are as follows: J_1: mean squared error (MSE), J_2: squared correlation coefficient (r^2), and J_3: HitRate±0.5. The qualities of the three feature subsets are then compared.

Chromosome selection for the next generation is performed such that better chromosomes can have a higher chance of being selected, and it is achieved by applying *Roulette Selection* scheme. Under this scheme, selection probability of a chromosome is proportional to its fitness non-linearly. After sorting the chromosomes in descending order of fitness, the i-th chromosome will be assigned a number $Pr(i)$ generated using nonlinear function:

$$Pr(i) = q(1-q)^{i-1}, q \in [0,1]. \tag{3.25}$$

The accumulative number pr_i will be calculated for each chromosome:

$$pr_i = \sum_{j=1}^{i} Pr(j), pr_0 = 0. \tag{3.26}$$

Then we generate a random number *ran* within $[0, pr_{N_P}]$. The chromosome selected will be the one with $pr_{i-1} < ran < pr_i$. Algorithm 6 describes the algorithm. In the algorithm, we first evaluate each chromosome c in the population based on the selected fitness function. After sorting the chromosomes in descending order of fitness values, we calculate pr_i for each c. The selection of c_i then depends on the corresponding pr_{i-1} and pr_i values and the random number *ran*. After N_{sel} iterations, the new population P' is returned.

The advantages of the roulette-selection are that, on the one hand, fitter chromosomes can have a higher chance to survive for the next generation; on the other hand, less fit chromosomes can still have a chance to survive, and thus the population diversity can be increased.

The GA stops when the termination condition is satisfied. To keep our implementation simple, the condition is a preset maximum number T of generation.

3.5 Experiments

In this section, we present experimental settings, procedures, and results of the proposed Chinese readability analysis. In particular, we conduct experiments by using SVR, feature selection using GA, and Chinese word segmentation using LMR-RC Tagging scheme at different configurations, aiming to discover the best prediction models. We then compare our approach with Linear Regression (LR), which is the modeling technique commonly used by previous works. All experiments are conducted in a machine with the configuration shown in Table 3.11. For SVR, we adopt and modify the JAVA interface in the *libsvm library* [12] as the core. We use this library because it is well documented and provides several useful tools for finding the best SVR settings. For LR, we apply the Matlab routine to run the

Table 3.11: Testing Environment.

CPU	Intel Pentium 4 3.2 GHz
RAM	4.0 GB
Operating System	RedHat Linux Fedora Core 4
Harddisk Size	300GB
Programming Language	Java SDK 1.5.06 and Matlab 7.1.0 sp3

experiments.

We implement the GA feature selection routine in the JAVA programming language [63]. We search for the optimal feature subset by repeating the experiment with different numbers of selected features d and fitness functions J.

3.5.1 Experiment 1: Evaluation on Chinese Word Segmentation using the LMR-RC Tagging Scheme

Objective

In this experiment, we verify and evaluate the proposed LMR-RC Tagging scheme in Chinese word segmentation. In particular, we compare the performance of the scheme under different configurations: (1) Regular tagging only, (2) Regular and Correctional tagging trained with separated mode, and (3) Regular and Correctional tagging trained with integrated mode.

Methodology

We conduct closed track experiments on the Hong Kong City University (CityU) corpus in The Second International Chinese Word Segmentation Bakeoff [75] to evaluate the proposed methods. The training data are split into three portions. Part 1: 60% of the data is trained for R-phase; Part 2: 30% for C-phase training; and Part 3: the remaining 10% for evaluation. The evaluation part is further divided into six parts to simulate actual size of test document used in the bakeoff. The MaxEnt classifier is implemented using the Java opennlp maximum entropy package from [3], and training is done with feature cutoff of 2 and 160 iterations. Feature cutoff means a feature will only be used if it happens in training data for more than 2 times.

We carry out four sub-experiments for each evaluation data. For Experiment A, data are processed with R-phase only. For Experiment B, data are processed with both R- and C-phase, using Separated Mode as training method. For Experiment C, data are processed similar to Experiment B, except Integrated Mode is used. Finally for Experiment D, data are processed similar to Experiment 1, with both Part 1 and Part 2 data are used for R-model training. The purpose of Experiment D is to observe whether the proposed scheme can perform better than just the single Regular Tagging under the same amount of training data.

Evaluation Metric

The performance of word segmentation is measured in *Recall* and *Precision*. Let W and W_{seg} be the set of words in testing corpus (ground true) and words in segmentation result respectively. Recall is than defined as the number of correctly segmented words ($|W \bigcap W_{seg}|$) divided by the number of words in testing corpus ($|W|$):

$$Recall = \frac{|W \bigcap W_{seg}|}{|W|}. \tag{3.27}$$

Precision is defined as the number of correctly segmented words ($|W \bigcap W_{seg}|$) divided by the number of segmented words ($|W_{seg}|$):

$$Precision = \frac{|W \bigcap W_{seg}|}{|W_{seg}|}. \tag{3.28}$$

In other words, we can view recall as a quantitative measure, while precision as a qualitative measure of the segmentation result.

F-measure [57] is introduced to obtain a single-valued performance metric, and it is defined as the harmonic means of recall and precision:

$$F\text{-}measure = \frac{2 \times Recall \times Precision}{(Recall + Precision)} \tag{3.29}$$

Result and Discussion

Table 3.12 summarizes the experimental result measured in F-measure. The bold entries represent the best results.

Table 3.12: Experimental Result of CityU Corpus Measured in F-measure. Best entries are bold-faced.

Data Set	**Exp A**	**Exp B**	**Exp C**	**Exp D**
1	0.918	0.943	**0.949**	0.947
2	0.913	0.939	**0.943**	**0.943**
3	0.912	0.935	**0.939**	0.937
4	0.914	0.940	**0.943**	0.942
5	0.921	0.942	**0.945**	**0.945**
6	0.914	0.941	**0.945**	0.942

From the results, we obtain the following observations.

1. Both Integrated and Separated Training modes in Two-Phase Tagging (Exp B and Exp C) outperform single Regular Tagging (Exp A). It is reasonable as more data are used in training.
2. Integrated Mode (Exp C) still performs better than Exp D, in which same amount of training data are used. This reflects that extra tagging information after R-phase helps in the scheme.
3. Separated Mode (Exp B) performs worse than both Exp C and Exp D. The reason is that the C-model cannot capture enough features' characteristics used for basic tagging. We believe that by adjusting the proportion of Part 1 and Part 2 of training data, performance can be increased.

4. Under limited computational resources, in which constructing single-model using all available data (as in Exp C and Exp D) is not possible, Separated Mode shows its advantage in constructing and aggregating multi-models by dividing the training data into different portions.

We have participated in the closed track of the Second International Chinese Word Segmentation Bakeoff organized by SIGHAN workshop of Association for Computational Linguistics [1]. The meaning of closed track is that, we can only use the training data provided to perform segmentation on the testing data, other materials such as dictionary are not allowed in this track. We submit multiple results for CityU, MSR and PKU corpora by applying different aforementioned tagging and training methods.

The official BakeOff2005 results are summarized in Table 3.13, sorted by F-measure value. Different methods are indicated by keys: (1) F - Regular Tagging only, all training data are used; (2) P1 - Regular Tagging only, 90% of training data are used; (3) P2 - Regular Tagging only, 70% of training data are used; (4) S - Regular and Correctional Tagging, Separated Mode; (5) I - Regular and Correctional Tagging, Integrated Mode. In the table, R_{OOV} and R_{IV} are recall rate of out-of-vocabulary (OOV) and recall rate of in-vocabulary (IV), measuring performances of segmenting words which can and cannot be found in the training data respectively.

In addition to results of our submissions, we also list the topline and baseline performances, and results of the best, the median, and the worst participants. Baseline scores are generated via maximal matching using only words from the training data, while topline scores are generated via maximal matching using only words from the testing data. Detailed discussions of the bakeoff can be found in [16].

Bakeoff results show that our approach performs better than

Table 3.13: Official BakeOff2005 results, sorted by F-measure.

Corpus	R	P	F	R_{OOV}	R_{IV}	**Key**
CityU	0.988	0.991	0.989	0.997	0.988	Topline
	0.941	0.946	0.943	0.698	0.961	Best
	0.937	0.922	0.929	0.698	0.956	I
	0.915	0.940	0.928	0.598	0.94	Median
	0.938	0.915	0.927	0.658	0.961	F
	0.936	0.913	0.925	0.656	0.959	P1
	0.925	0.896	0.910	0.639	0.948	P2
	0.882	0.790	0.833	0.000	0.952	Baseline
	0.814	0.711	0.759	0.227	0.86	Worst
MSR	0.991	0.992	0.991	0.998	0.990	Topline
	0.962	0.966	0.964	0.717	0.968	Best
	0.965	0.935	0.950	0.189	0.986	Median
	0.946	0.933	0.939	0.587	0.956	F
	0.941	0.932	0.937	0.624	0.950	S
	0.955	0.912	0.933	0.000	0.981	Baseline
	0.898	0.896	0.897	0.327	0.914	Worst
PKU	0.985	0.988	0.987	0.994	0.985	Topline
	0.953	0.946	0.950	0.636	0.972	Best
	0.922	0.934	0.928	0.728	0.934	Median
	0.918	0.915	0.917	0.621	0.936	I
	0.926	0.908	0.917	0.535	0.950	F
	0.917	0.903	0.910	0.600	0.937	P2
	0.904	0.836	0.869	0.059	0.956	Baseline
	0.843	0.737	0.786	0.153	0.885	Worst

the baseline scores, and perform the best in CityU corpus. Performance difference between our approaches to the best participants is about 2%, showing that our approach is comparable to other advanced techniques. Based on the results, we are confident that our approach is suitable in higher level text analyses, including Chinese readability studied in this work.

3.5.2 Experiment 2: Initial SVR Parameters Searching with Different Kernel Functions

Objective

In this experiment, we try to find out the best kernel function available in SVR for our problem. Four common kernel functions are available in the libsvm library. They are *linear*, *polynomial*, *radial basis function*, and *sigmoid*. As the kernel function plays an important role in SVR for nonlinear regression, it is necessary to select the best one for further experiments. Another purpose of this experiment is to obtain an initial parameters setting which can give good results, as SVR is quite sensitive to the parameter settings [29].

Methodology

We perform this experiment using a tool called *gridregression.py*, provided in the libsvm library. This program conducts a grid search on the parameters used in SVR: C, *gamma*, and *epsilon*. This program acts as a good tool for initial coarse evaluation during the search for optimal model settings.

We repeat this experiment using the four aforementioned kernel functions. The best parameter settings and the corresponding performance, measured in *MSE* of 5-fold cross-validation, are recorded and compared. Note that for regression analysis, we scale the target grade level (dependent value) from range the [2,10] to [0,1], so the measured MSE is in the scaled range. As

Table 3.14: Experiment 2 Settings.

Parameters	**Values**
SVR Method	epsilon-SVR
Kernel	Linear, Polynomial, RBF, Sigmoid
Cross-validation	5-fold CV
Performance Metric	MSE

Table 3.15: Experiment 2 Result - Comparison of Different Kernel Function for SVR. Best entries are bold-faced.

	Parameters			**Performance**
Kernel	Cost (c)	Gamma (g)	Epsilon (p)	MSE
Linear	0.5	0.0625	0.0625	0.03213
Polynomial	2.00	0.0625	0.0625	0.0298
RBF	1.00	0.2500	0.0625	**0.0238**
Sigmoid	16	0.0078	0.125	0.0308

we have not yet performed the feature selection process, all 64 features listed in Table 3.5 are used in this experiment. Table 3.14 summarizes the experimental settings.

Result and Discussion

Table 3.15 shows the results of Experiment 2, with the best entry shown in bold. We find that SVR with the kernel function RBF achieves the best performance among the four tested kernel functions; RBF is about 15% better than the second-best in terms of MSE. This shows that the RBF kernel is the most suitable for our Chinese readability estimation problem. As a result, we apply the RBF kernel and the corresponding parameter values in the remaining experiments.

Another observation is that this initial experiment provides the baseline performance of our scheme before further optimizations. By converting the MSE values to our original grade level range [2, 10], the deviation is about 1.6 to 2.0 depending on the kernel function used. So roughly speaking, our scheme can estimate passage readability with an error range of ± 1.3 grade level.

3.5.3 Experiment 3: Feature Selection Using Genetic Algorithm

Objective

In this experiment, we search for the best feature subset for Chinese readability analysis under different optimization criteria by using different fitness functions. We also aim to learn more about the importance of different features in readability regression analysis. Future workers can then have better control over the readability prediction process by making use of this information.

Methodology

We repeat the experiment at different numbers of target subset size d to find out the optimal subset size. We also apply three fitness functions, as described in Section 3.4.4: J_1: MSE, J_2: Squared Correlation Coefficient, and J_3: HitRate ± 0.5. The fitness score of each chromosome is the average of ten times 10-fold cross-validation (CV), so as to minimize any effect caused by the random nature of genetic algorithm and CV.

For SVR, we apply the RBF kernel and the corresponding optimized parameter values obtained from the previous experiments. Tables 3.16 and 3.17 summarize the settings of SVR and GA feature selection routine.

Result and Discussion

Figures 3.6, 3.7 and 3.8 show the experimental result of feature selection using epsilon-SVR at different numbers of selected

Table 3.16: Experiment 3 Settings - SVR Parameter Values.

Parameters	**epsilon-SVR**
Kernel Function	RBF
Cost (c)	1.0
Gamma in kernel function (g)	0.25
Epsilon (p)	0.0625

Table 3.17: Experiment 3 Settings - GA Feature Selection Parameter Values.

Parameters	**Values**
No. of Selected Features	[5,50] out of 64
Population Size	30
Crossover Rate	0.8
Mutation Rate	0.2
No. of Cross-over Points	3
No. of Good Chromosomes	15
No. of Generations	30
Fitness Functions	J_1:MSE, J_2:Sq. Correlation Coefficient, J_3:HitRate ±0.5

features. The X-axis is the the target number of feature subsets, while the Y-axis is the three tested fitness functions: MSE, Squared correlation coefficient (SCC), and HitRate±0.5 (H0.5). We make the following observations based on the results:

Observation 1: Performance Rising. Fitness functions J_1 and J_2, both show a high performance gain when the number of features d is increased from 5 to 15. For J_1, the MSE drops from about 1.53 at $d = 5$ to 1.19 at $d = 15$, which is a percentage decrease of 22%. For J_2, the SCC increases from about 0.76 at $d = 5$ to 0.81 at $d = 15$, which is a percentage change of 6.6%. But J_3 reaches its maximum at $d = 10$, then drops gradually from $d = 15$ onwards. The H0.5 increases from 0.43 at $d = 5$ to 0.45 at $d = 15$, which is a percentage change of 4.7%.

Observation 2: Performance Retention. All three

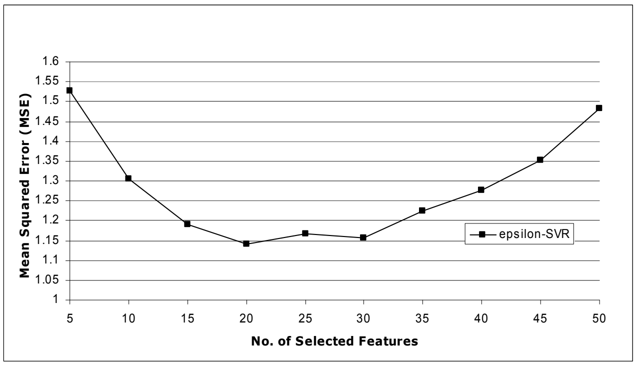


Figure 3.6: Experimental Result of Feature Selection using Fitness Function J_1.

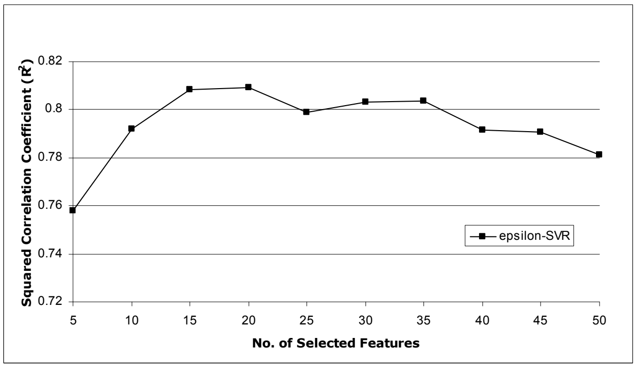


Figure 3.7: Experimental Result of Feature Selection using Fitness Function J_2.

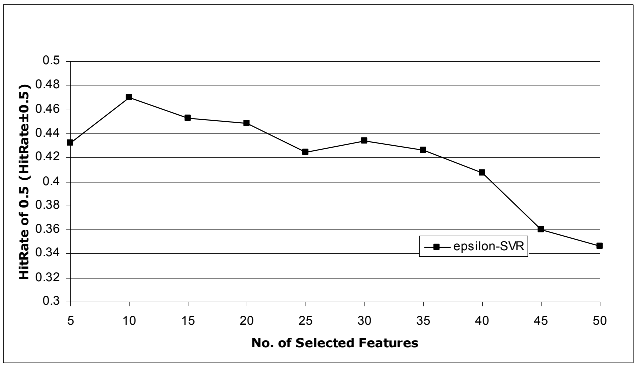


Figure 3.8: Experimental Result of Feature Selection using Fitness Function J_3.

fitness functions retain a relatively stable performance between $d = 15$ and $d = 30$, and then the performance drops afterwards. For J_1, the performance remains within the range of around 1.15 to 1.2; for J_2, the performance remains at around 0.8 to 0.81; for J_3, the performance remains at 0.4 to 0.45. The fluctuations inside these ranges can be explained by the random nature of GA and CV.

Observation 3: Performance Dropping. After $d = 35$, the performances of the three functions start to drop gradually. We can see from Figure 3.6 that the curve is actually approaching the baseline performance obtained from the last experiment. The curve in J_2 shows similar behavior to J_1. There is a slight difference between J_3 and others, in that the performance at $d = 50$ drops to the minimum, whereas the other two functions do not.

Discussion. Based on the three observations, we can conclude that if $d = 5$ to 15, the regression performance is not good enough as the number of features is not sufficient. But if the number of features is too large, e.g. after $d = 45$, those features will generate noise and upset the regression performance. So the optimal value of d lies in the range $d = 15$ to $d = 30$, in which range the performance remains at a relatively steady level. As the number of features affects the complexity of text analysis and regression complexity, we select $d = 15$ for the remaining experiments. This value provides an acceptable balance between performance and number of features. The feature indices in the three corresponding feature sets are as follows:

1. Feature Set 1: Using Fitness Function J_1 (MSE)
   $\{4\ 7\ 10\ 15\ 17\ 24\ 27\ 33\ 39\ 41\ 43\ 44\ 46\ 47\ 59\}$

2. Feature Set 2: Using Fitness Function J_2 (Squared Correlation Coefficient)
   $\{1\ 3\ 4\ 6\ 8\ 18\ 20\ 24\ 27\ 31\ 39\ 47\ 59\ 62\ 63\}$

3. Feature Set 3: Using Fitness Function J_3 (HitRate±0.5)
   {4 5 17 18 24 26 27 32 35 44 46 47 48 51 60}

3.5.4 Experiment 4: Training and Cross-validation Performance using the Selected Feature Subset

Objective

After selecting the best feature subsets in the previous experiment, we evaluate the training and cross-validation performance in this experiment. We compare the performances of the three subsets obtained for the different metrics as described in Section 3.2.5.

Methodology

The three selected feature subsets are input into the SVR learning algorithm, and the training and cross-validation performances are recorded and compared. We employ the leave-one-out cross-validation (LOO_CV) method, as it gives the best cross-validation accuracy. Before the actual training, we search again for the best SVR parameters by using the aforementioned *gridregression.py* tool, as the best parameters may be different for different feature sets. To summarize, Table 3.18 shows the settings in Experiment 4.

During the experiment, we discovered that the prediction results of some training samples have large variations between their correct and predicted grade scores. As our task is dealing with natural languages, for which it is difficult to obtain a set of consistent training data, we manually inspected those samples with large variations, and decided whether we should prune out those samples from training data. The details of the pruning process will be discussed in the next subsection. We tested the filtered sample data again, and compared the results with the original ones.

Table 3.18: Experiment 4 Settings - SVR Parameter Values.

Parameters	**Feature Set 1**	**Feature Set 2**	**Feature Set 3**
Kernel Function	RBF		
Cost (c)	0.5	1.0	1.0
Gamma in kernel function (g)	1.0	1.0	1.0
Epsilon (p)	0.00391	0.01563	0.03125

Result and Discussion

Table 3.19 shows the results of regression analysis before sample data filtering using the feature sets obtained in Experiment 4. Both *training* and *leave-one-out* (LOO-CV) accuracies are reported. In the table, bold-faced values indicate the best training performance among the three feature sets, while both bold-faced and italicized values indicate the best LOO_CV performance.

Observation 1: Comparison of Training and Cross-validation Performances. According to the results, the training performance is better than the cross-validation performance for all evaluation metrics, which is under our expectation. This is because, for training, the test data is actually taking part in the training process, such that when the same piece of data is presented, the predicted grade level should be similar to the correct level. The training performance can achieve a squared correlation coefficient of around 0.9, and a HitRate±1.0 of around 0.8 in all three feature sets, showing that the regression can successfully cover the majority of the training data.

Cross-Validation performance is lower than that of training performance because, in LOO_CV, the test data is isolated from the training data, so the training process may not be able to capture the characteristics of the isolated data. The squared correlation coefficient is around 0.8 for LOO_CV, which is a percentage decrease of 10%, and the HitRate±1.0 is 0.7, which is a percentage drop of 12%. If we look at the HitRate carefully, it is

Table 3.19: Experiment 4 Results - Training and CV Performance (Before Filtering). Best entries in training performance are bold-faced. Best entries in LOO-CV performance are bold-faced and italicized.

Metrics	**Feature Set 1**		**Feature Set 2**		**Feature Set 3**	
	Train	LOO-CV	Train	LOO-CV	Train	LOO-CV
Max Prediction Error	**3.2603**	***3.3679***	3.6759	4.8550	3.4669	4.8928
Mean Absolute Error	0.5417	***0.7985***	**0.4362**	0.8018	0.5131	0.8574
Stdev. Ab. Error	0.7096	***0.7357***	0.6099	0.7395	**0.5974**	0.7856
Mean Sq. Error	0.7942	***1.1758***	**0.5602**	1.1867	0.6181	1.3487
Stdev. Sq. Error	1.7238	***1.9946***	**1.4988**	2.5604	1.6036	2.6412
Correlation	0.9362	***0.9029***	**0.9557**	0.9019	0.9506	0.8878
Sq Correlation	0.8765	***0.8153***	**0.9133**	0.8135	0.9037	0.7883
HitRate:±0.5	0.6136	***0.4659***	**0.7386**	0.4318	**0.7386**	0.4299
HitRate:±0.6	0.6420	0.5170	**0.7727**	***0.5284***	0.7500	0.4943
HitRate:±0.7	0.6989	0.5511	**0.8068**	***0.6136***	0.7841	0.5625
HitRate:±0.8	0.7273	0.5966	**0.8125**	***0.6477***	0.8011	0.6080
HitRate:±0.9	0.7614	0.6477	**0.8239**	***0.6705***	0.8068	0.6477
HitRate:±1.0	0.7841	0.6705	**0.8352**	***0.7216***	0.8182	0.6818
HitRate:±1.5	0.8864	***0.8466***	0.9148	0.8409	**0.9205**	0.8220
HitRate:±2.0	0.9489	0.9261	**0.9716**	***0.9432***	**0.9716**	0.9167
HitRate:±2.5	0.9773	0.9659	**0.9886**	***0.9830***	0.9773	0.9659
HitRate:±3.0	0.9830	***0.9830***	**0.9943**	***0.9830***	0.9886	***0.9830***
HitRate:±3.5	**1.0000**	***1.0000***	0.9943	0.9830	**1.0000**	0.9886
HitRate:±4.0	**1.0000**	***1.0000***	**1.0000**	0.9943	**1.0000**	0.9943

Table 3.20: Summary of Data Set after Filtering.

Grade Level	**No. of Selected Articles (Before Filtering)**	**No. of Selected Articles (After Filtering)**
2	25	25
3	22	19
4	20	18
5	20	18
6	19	19
7	21	19
8	19	18
9	19	17
10	11	10
Total	176	163

found that HitRate$\pm X + 0.5$ of LOO_CV is close to HitRate$\pm X$ of Train, for $X \geq 2.0$ e.g. for Feature Set 3, HitRate±1.5 of LOO_CV (0.822) is close to HitRate±1.0 of Train (0.8182). Based on this observation, we can say that the model will generally adding an error of 0.5 level to an unseen datum, which results in ±1 grade level if we round off the values to the nearest integer.

Observation 2: Comparison between Performances of the Three Feature Sets. After comparing the training and cross-validating performance, we now investigate performances of feature sets obtained by the three fitness functions. We focus on the LOO_CV results in this section. For statistical metrics like mean absolute error, mean squared error, and squared correlation, feature set 1 and feature set 2 have similar performances, and they are better than feature set 3. On the other hand, for the HitRate-related metrics, feature set 2 out-performs the others. This is because set 1 and set 2 are obtained from fitness functions J_1 MSE and J_2 Squared Correlation Coefficient, which try to optimize the overall distribution performance. In contrast, for fitness function J_3 HitRate±0.5, although it tries to find out a subset such that as many training data as possible can be estimated correctly, it appears that the function fails to optimize the global performance, thus causing the poorer result for set 3.

Observation 3: High Training Error and Data Filtering. Although our model gives satisfactory results in estimating readability with an error of ±1 grade level range, we also discover that the prediction results of some training data have a large deviation from their correct levels, causing a Maximum Prediction Error of around 3.5. Direct inspection of articles with large deviations suggests that the following are the possible reasons for the errors.

1. There are some articles on specific topics, such as a discussion of “marine life”, in primary school level texts, in which some difficult characters and terms are introduced, causing a large variation in prediction results.
2. For the secondary level, articles are extracted from famous, popular authors, and it is common that several articles by the same author are being selected for different grade levels, causing ambiguity.
3. According to a professor[1] in the Faculty of Education at CUHK, Hong Kong Chinese language textbooks may show an inconsistency in difficulty, even when the texts are nominally intended for the same level.

To observe the results without the problematic articles, we eliminate from our training data those articles which have differences greater than 2.0 between the correct and predicted grade levels. Table 3.20 shows the details of the filtered sample data.

The filtered data set are input for training and cross-validation again. Table 3.21 shows the results, and Table 3.22 shows the

[1] Prof. HO Man-koon, Associate Professor in the Department of Curriculum and Instruction, CUHK

percentage change resulting from the data filtering.

Table 3.21: Experiment 4 Results - Training and CV Performance (After Filtering). Best entries in training performance are bold-faced. Best entries in LOO-CV performance are bold-faced and italicized.

Metrics	**Feature Set 1**		**Feature Set 2**		**Feature Set 3**	
	Train	LOO-CV	Train	LOO-CV	Train	LOO-CV
Max Prediction Error	**1.8284**	***2.1732***	2.0259	2.5832	1.8481	2.5457
Mean Absolute Error	0.3992	***0.6821***	**0.3575**	0.7267	0.3916	0.7100
Stdev. Ab. Error	0.4903	***0.5520***	0.4596	0.5580	**0.3682**	0.5774
Mean Sq. Error	0.3983	***0.7681***	0.3378	0.8376	**0.2881**	0.8353
Stdev. Sq. Error	0.7233	***1.0428***	0.7880	1.2014	**0.5811**	1.2236
Correlation	0.9700	***0.9386***	0.9741	0.9323	**0.9777**	0.9323
Sq Correlation	0.9408	***0.8810***	0.9489	0.8692	**0.9558**	0.8691
HitRate:±0.5	0.6687	0.4908	0.7791	0.4540	**0.7730**	***0.5031***
HitRate:±0.6	0.7117	***0.5337***	**0.8098**	0.5031	0.7975	0.5276
HitRate:±0.7	0.7607	***0.5951***	**0.8405**	***0.5951***	0.8344	0.5828
HitRate:±0.8	0.8037	0.6319	0.8589	***0.6442***	**0.8712**	***0.6442***
HitRate:±0.9	0.8221	0.6564	0.8834	***0.6994***	**0.8896**	0.6871
HitRate:±1.0	0.8650	0.7362	0.8957	***0.7526***	**0.9080**	0.7423
HitRate:±1.5	0.9509	***0.8998***	0.9509	0.8773	**0.9693**	0.8793
HitRate:±2.0	**1.0000**	***0.9816***	0.9939	0.9571	**1.0000**	0.9693
HitRate:±2.5	**1.0000**	***1.0000***	**1.0000**	0.9939	**1.0000**	0.9939
HitRate:±3.0	**1.0000**	***1.0000***	**1.0000**	***1.0000***	**1.0000**	***1.0000***
HitRate:±3.5	**1.0000**	***1.0000***	**1.0000**	***1.0000***	**1.0000**	***1.0000***
HitRate:±4.0	**1.0000**	***1.0000***	**1.0000**	***1.0000***	**1.0000**	***1.0000***

Observation 4: Discussion of results obtained After data filtering. The overall performances of the three feature sets increase after performing filtering. In particular, the statistical metrics like maximum prediction error, mean absolute error and mean squared error have the greatest improvement. It is under our expectation that as after removing those data with error larger than 2.0, the maximum prediction error should be smaller than 2.0. The squared correlation coefficient in LOO_CV also shows an improvement of about 9%, indicating that our prediction model has a high correlation with the

Table 3.22: Experiment 4 Results - Percentage Change After Filtering).

Metrics	**Feature Set 1**		**Feature Set 2**		**Feature Set 3**	
	Train	LOO-CV	Train	LOO-CV	Train	LOO-CV
Max Prediction Error	-43.92%	-35.47%	-44.89%	-46.79%	-46.70%	-47.97%
Mean Absolute Error	-26.31%	-14.57%	-18.03%	-9.36%	-23.68%	-17.20%
Stdev. Ab. Error	-30.90%	-24.97%	-24.64%	-24.55%	-38.37%	-26.50%
Mean Sq. Error	-49.85%	-34.67%	-39.69%	-29.42%	-53.40%	-38.06%
Stdev. Sq. Error	-58.04%	-47.72%	-47.42%	-53.08%	-63.76%	-53.67%
Correlation	+3.60%	+3.95%	+1.93%	+3.37%	+2.84%	+5.00%
Sq Correlation	+7.34%	+8.06%	+3.90%	+6.85%	+5.77%	+10.26%
HitRate:±0.5	+8.98%	+5.34%	+5.48%	+5.13%	+4.65%	+17.01%
HitRate:±0.6	+10.84%	+3.23%	+4.80%	-4.80%	+6.34%	+6.73%
HitRate:±0.7	+8.85%	+7.98%	+4.17%	-3.02%	+6.41%	+3.61%
HitRate:±0.8	+10.51%	+5.92%	+5.71%	-0.55%	+8.74%	+5.96%
HitRate:±0.9	+7.98%	+1.35%	+7.23%	+4.32%	+10.26%	+6.08%
HitRate:±1.0	+10.32%	+9.81%	+7.24%	+4.29%	+10.97%	+8.88%
HitRate:±1.5	+7.28%	+6.28%	+3.95%	+4.33%	+5.31%	+6.98%
HitRate:±2.0	+5.39%	+5.99%	+2.29%	+1.47%	+2.92%	+5.74%
HitRate:±2.5	+2.33%	+3.53%	+1.15%	+1.11%	+2.33%	+2.89%
HitRate:±3.0	+1.73%	+1.73%	+0.57%	+1.73%	+1.15%	+1.73%
HitRate:±3.5	+0.00%	+0.00%	+0.57%	+1.73%	+0.00%	+1.15%
HitRate:±4.0	+0.00%	+0.00%	+0.00%	+0.57%	+0.00%	+0.57%

remaining data.

We also observe improvements on HitRate-related metrics as expected, although the percentage change is not as large as the statistical metrics. After filtering, our model can predict the readability within ±1 grade level (HitRate±1.0) at around 75% of correctness, and up to around 90% for ±1.5 grade levels (HitRate±1.5).

Discussion. Based on the experimental results, our measure is good at estimating readability ±1 grade level (indicated by a HitRate of 1 and 1.5, around 90 - 93%). Furthermore, the high correlation coefficient demonstrates that our regression analysis is significant. The reason that it is difficult to predict grade level exactly (indicated by low HitRate±0.5) is that articles of two successive levels are quite similar. In fact, we found one article which belonged to primary 2 level in an old edition of a textbook was shifted to primary 3 in a later edition. Some other sources of error may also upset the prediction, such as: (1) errors introduced from text processing, like sentence and word segmentation and POS tagging, (2) ambiguity of the Chinese language, and (3) difficulties of natural language processing.

3.5.5 Experiment 5: Comparison with Linear Regression

Objective

We compare our approach using SVR with Linear Regression (LR), which is a common technique employed in previous readability studies. LR applies the method of least squares [39] in order to minimize the sum of squares of the residuals between observed and predicted values. we omit its details here as this is a well-known and well-implemented technique.

Methodology

The feature set we used is Set 1, as its HitRate±1.5 is the highest among the 3 sets, meaning that it can estimate readability ±1 grade level correctly. We then apply SVR and LR on the reduced feature set. The setting of SVR is the same as that in Experiment 4. For LR, we apply the MATLAB routine to perform the prediction.

Table 3.23: Comparison between SVR and LR. Best entries in training performance are bold-faced. Best entries in LOO-CV performance are bold-faced and italicized.

Metrics	**SVR**		**LR**	
	Train	LOO-CV	Train	LOO-CV
Max Prediction Error	**1.8284**	***2.1732***	4.1007	4.8232
Mean Absolute Error	**0.3992**	***0.6821***	1.0264	1.1442
Stdev. Ab. Error	**0.4903**	***0.5520***	0.7591	0.8762
Mean Sq. Error	**0.3983**	***0.7681***	1.6262	2.0721
Stdev. Sq. Error	**0.7233**	***1.0428***	2.3960	3.3065
Correlation	**0.9700**	***0.9386***	0.8636	0.8249
Sq Correlation	**0.9408**	***0.8810***	0.7457	0.6804
HitRate:±0.5	**0.6687**	***0.4908***	0.2638	0.2331
HitRate:±0.6	**0.7117**	***0.5337***	0.3129	0.2883
HitRate:±0.7	**0.7607**	***0.5951***	0.3988	0.3374
HitRate:±0.8	**0.8037**	***0.6319***	0.4724	0.3988
HitRate:±0.9	**0.8221**	***0.6564***	0.5276	0.4847
HitRate:±1.0	**0.8650**	***0.7362***	0.6012	0.5337
HitRate:±1.5	**0.9509**	***0.8998***	0.7853	0.7423
HitRate:±2.0	**1.0000**	***0.9816***	0.8773	0.8344
HitRate:±2.5	**1.0000**	***1.0000***	0.9387	0.9141
HitRate:±3.0	**1.0000**	***1.0000***	0.9816	0.9693
HitRate:±3.5	**1.0000**	***1.0000***	0.9877	0.9816
HitRate:±4.0	**1.0000**	***1.0000***	0.9939	0.9816

Result and Discussion

Observation 1: Training Performance. From the results shown in Table 3.23, we find that SVR performs better than LR in all evaluation metrics. This is due to the strength inherited from the Support Vector Machine. The results also reveal that non-linear relationships exist between readability level (dependent variable) and features (independent variables). As LR fails to capture such relationships, it has poorer performance.

Observation 2: Cross Validation Performance. The result shows the performance of models in predicting the readability of an unseen passage. In other words, it indicates their generalization powers. From the result, SVR again outperforms LR, as with the previous observation on training performance. The mean absolute error and Hit Rate ±1 suggests that our readability assessment based on SVR is good at predicting an unseen passage within ±1 grade level precision. When the precision is loosened to ±2 grade levels, our approach can achieve a correctness of about 90%.

3.6 Summary and Future Work

In this chapter, we demonstrate our work on Chinese readability analysis. Firstly, we perform Chinese readability factor analysis in a systematic way, in which various features are grouped in different language levels. Then we apply advanced Chinese text processing techniques to increase the accuracy in extracting features of the text. Finally we make use of Support Vector Regression to perform the regression analysis due to its superior performance in solving other problems. Experimental results show that the proposed approach has a satisfactory performance.

For future work, we are now planning to conduct a large-scale experiment in collaboration with the Faculty of Education at CUHK. In particular, we would like to obtain more testing materials, such as Chinese textbooks from mainland China, Taiwan, and other Chinese districts, and student compositions.

Chapter 4

Web Readability Analysis

We study and discuss Web page and site readability in this chapter. We investigate Web page readability as *comprehension difficulty* and *grade level*. Comprehension difficulty is a score, for example, ranged from 0 (hardest) to 100 (easiest), estimating the degree of difficulty in comprehending a passage. Grade level, on the other hand, is a score representing the grade level of a group of people found the passage suitable to them. Comprehension difficulty and grade level are two common explanations of the term "readability".

We apply existing readability formulas (Flesch [20] and Yang [81]) to estimate Web page readability in the analysis of comprehension difficulty. We perform experiments on real Web pages to observe their behaviors.

After proposing the Chinese readability analysis using SVR (discussed in Chapter 3), we apply the assessment in estimating Web page readability in terms of grade level. We perform experiments on the same set of data as in the previous analysis. We then compare results obtained in the two measures.

We extend the idea of Web page readability and propose a Web site readability assessment scheme. We model a Web site as a rooted tree, in which the root is index page of a site. Then the scheme, which consists of three assessments, is based on pages at different *tree levels* relative to the root. Experiments are conducted in order to study the behaviors of Web sites having different readability scores.

4.1 Web Page Readability

4.1.1 Readability as Comprehension Difficulty

We adopt and modify the Flesch reading ease [20] and Yang [81] formula (discussed in Chapter 2) to evaluate comprehension difficulty of English and Chinese Web pages. The definition of the page readability is as follows.

Definition 3 *Page Readability (Comprehension Difficulty) of a Web page $p \in P$, denoted by r_p is defined as:*

$$r_p = \begin{cases} -84.6X_{E_1} - 1.015X_{E_2} + 206.835 & \textit{if } lang(p) = 0, \\ \\ \begin{array}{r} 2 \times \{13.90963 + 1.54461X_{C_1} + \\ 39.01497X_{C_2} - 2.52206X_{C_3} - \\ 0.29809X_{C_4} + 0.36192X_{C_5} + \\ 0.99363X_{C_6} - 1.64671X_{C_7}\} \end{array} & \textit{if } lang(p) = 1, \end{cases} \tag{4.1}$$

where X_{E_i} and X_{C_i} are the factors, $lang : P \rightarrow \{0, 1\}$ is a mapping from page to its language:
- *X_{E_1}: Average number of syllables per word;*
- *X_{E_2}: Average sentence length;*
- *X_{C_1}: Proportion of full sentence;*
- *X_{C_2}: Proportion of words in Chinese basic word list;*
- *X_{C_3}: Average number of stroke of characters;*
- *X_{C_4}: Number of characters with stroke number = 5 (in a sample of 100 characters) ;*
- *X_{C_5}: Number of characters with stroke number = 12 (in a sample of 100 characters) ;*
- *X_{C_6}: Number of characters with stroke number = 22 (in a sample of 100 characters) ;*

- *X_{C_7}: Number of characters with stroke number = 23 (in a sample of 100 characters) ;*
- *$lang(p)$: 0 for English, 1 for Chinese.*

The calculated score will be truncated if it is not in the range of [0,100], i.e. negative scores will be truncated to 0, where as scores larger than 100 will be truncated to 100.

We choose *Flesch reading ease* formula in evaluating English Web page for two reasons. Firstly, Flesch reading ease is well-established. Many other assessments, such as Dale-Chall, Farr et al. and Fry's Readability Graph are highly correlated to Flesch [31, 81], showing its representativeness in the field. Secondly, it is widely used in various applications, including the popular word processor Microsoft Word.

Among the three Chinese readability assessments (Yang, Jing and Jeng), we choose Yang formula, which is normalized to the scale of 0 (hard) to 100 (easy) as in Flesch formula, for Chinese page assessment because of its good balance in ease of computation and quality of factors. Although Jing formula can be easily computed, it considers the number of words in a page (page length). As we think that page length should not affect its readability in terms of text contents (otherwise one can create a page with very short length to get high readability), we do not choose the Jing formula. For Jeng formula, although it gives good performance according to experimental results [31], the use of artificial neural network hinders its application in terms of efficiency. On the other hand, Yang formula achieves good balance between ease of computation and factors involved. As a result, we adopt this for Chinese Web page assessment.

Although the two formulas consider different factors and apply different evaluation methods, it is reasonable to use them together for preliminary investigation as they represent readability estimations for their own communities. Better normalization between the two readability scores should be done to improve the assessment, and this is left for our future work.

4.1.2 Readability as Grade Level

English Web Page

As in comprehension difficulty for English Web pages, we adopt a similar assessment as Flesch reading ease, called *Flesch-Kincaid Grade Level* [20] as the assessment of English part. Flesch-Kincaid Grade Level (FK) is a formula-based assessment to estimate readability of a piece of text based on US grade level scale. The formula of the assessment is as follows:

$$FK(p) = 11.8X_{E_1} + 0.39X_{E_2} - 15.59 \tag{4.2}$$

The parameters X_{E_1} and X_{E_2} are same as that in Definition 3.

Chinese Web Page

We apply the proposed Chinese readability analysis described in Chapter 3 as the basis of Chinese Web page readability estimation. The basic idea of our method is first extracting some text features like average sentence length and the average number of strokes, then the features are input to Support Vector Regression (SVR) Model for regression analysis. We denote our method as $SVR(p)$ for the sake of easier discussion. Table 4.1 is a replication of Table 3.5 to show the features being used in the analysis.

After discussing approaches used for both languages, the definition of this page readability is as follows.

Definition 4 *Page Readability (Grade Level) of a Web page $p \in P$, denoted by r_p is defined as:*

$$r_p = \begin{cases} FK(p) & \mathit{if}\ lang(p) = 0, \\ SVR(p) & \mathit{if}\ lang(p) = 1, \end{cases} \tag{4.3}$$

Table 4.1: Summary of Chinese Readability Features. (Replication of Table 3.5)

Index	**Factor**	**Feature Names**
Sub-character Level		
1-2.	$\mathcal{R}_{stroke}$	Average and Standard deviation of number of radical strokes per Chinese character
3-4.	$\mathcal{R}_{fam}$	Proportion of familiar and unfamiliar radicals
Character Level		
5-6.	$\mathcal{C}_{strk}$	Average and Standard deviation of number of strokes per Chinese character
7-8.	$\mathcal{C}_{strk\_exRad}$	Average and Standard deviation of number of strokes without radical per Chinese character
9-10.	$\mathcal{C}_{fam}$	Proportion of familiar and unfamiliar characters
11-15.	$\mathcal{C}_{symm}$	Proportion of Symmetrical, Non-symmetrical, Vertical, Horizontal and Both Symmetrical characters
16-22.	$\mathcal{C}_{struct}$	Proportion of characters belonging to *Structure Category*[A-G]
23-24.	$\mathcal{C}_{grade}$	Average and Standard deviation of character grade
25-26.	$\mathcal{C}_{common}$	Proportion of common and non common characters
27-28.	$\mathcal{C}_{freq}$	Average and Standard deviation of character frequency of occurrence
Word Level		
29-30.	$\mathcal{W}_{fam}$	Proportion of familiar and unfamiliar words
31-32.	$\mathcal{C}_{stk}$, $\mathcal{W}_{length}$	Average and Standard deviation of number of strokes per word
33-34.	$\mathcal{W}_{length}$	Average and Standard deviation of number of characters per word
35-46.	$\mathcal{W}_{pattern}$	Proportion of words belonging to *Word Pattern Category*[A-L]
47-48.	$\mathcal{W}_{common}$	Proportion of common and non-common words
Phrase Level		
49.	$\mathcal{P}_{idiom}$	Proportion of phrases containing idioms
50-55.	$\mathcal{P}_{length}$	Average and Standard deviation of number of strokes, characters, words of phrase
Sentence Level		
56-61.	$\mathcal{S}_{length}$	Average and Standard deviation of number of strokes, characters, words per sentence
62.	$\mathcal{S}_{fullsent}$	Proportion of full sentences
63-64.	$\mathcal{S}_{tag}$	Average and Standard deviation of number of distinct POS tags in sentence

where $lang : P \rightarrow \{0, 1\}$ is a mapping from page to its language: lang(p): 0 *for English,* 1 *for Chinese. The calculated score is in a scale of [2,10], which follows the range used in $SVR(p)$. Score not in the range is truncated.*

4.2 Web Site Readability

Web site readability is an indicator of overall difficulty level of a site, and it is defined over page readability mentioned in the previous section. We propose three site readability assessments: (1) *Exact-Level*, (2) *In-Level*, and (3) *Out-Level*, aiming at describing the site from different angles of page composition. We first define some preliminary concepts, *Web site*, *root page* and *page level*, or simply level, before continuing the discussion.

Definition 5 *Web site, denoted by $s \in S$, is a group of Web pages with the same domain name in the URLs. Root page of a Web site, denoted by p_0, is a user-specified page where the crawling of the site starts. Page level of a Web page p in a Web site, denoted by lv, is the minimum number of traversal reaching p starting from the root page of the Web site through hyperlinks. We use $level : P \rightarrow \mathbb{N} \cup \{0\}$ be the mapping from page to its level. Root page has page level* 0.

Based on above definitions, we define site readability in terms of pages at different page levels, and this gives rise to the three aforementioned assessments:

Exact-Level Readability is to indicate the average readability of pages at a particular level. By using this metric, Web authors can decide how should the readability change with levels. Take Online Teaching Site as an example. Teachers should probably want to teach some simpler things at the beginning, and then increase the difficulty level gradually. By analyzing the changes

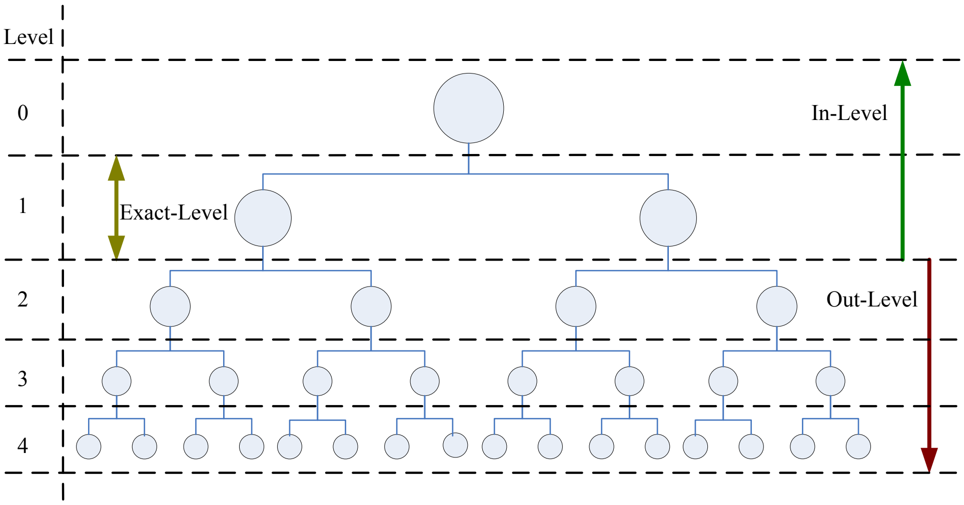


Figure 4.1: Illustration of Exact-Level, In-Level, and Out-Level Site Readability at level=1.

of Exact-Level readability along levels, teachers can then prepare and arrange materials in proper order.

In-Level Readability of a site gives the average readability of Web pages starting from root page up to pages at specified level. This is an overall indicator of a site difficulty. By using this metric, users can get a general idea of whether the site is suitable to themselves first before start browsing it.

In contrast to *In-Level Readability*, *Out-Level Readability* of a site gives the average readability of Web pages starting one level higher than the specified one, up to pages with *maximum available level*, which is determined by the depth of crawling. In other words, it is an difficulty indicator of remaining pages after browsing a site for some times. Users may make use of this metric as one of factors to decide whether he or she should continue browsing the site. For example, a user finds that the currently browsing page is difficult, but by referencing the high Out-Level score, he or she may stay at the site as coming pages are easier.

Figure 4.1 depicts the three assessments, and we define them formally as follows:

Definition 6 *Exact-Level Site Readability of a Web site s at level lv, denoted by $r_{s,lv,e}$:*

$$r_{s,lv,e} = \begin{cases} \frac{\sum_{\forall p_i level(p_i)=lv} r_{p_i}}{n_{lv}} & \textit{if } (n_{lv} \neq 0), \\ 0 & \textit{otherwise.} \end{cases} \tag{4.4}$$

where n_{lv} is the number of pages with levels equal to lv.

Definition 7 *In-Level Site Readability of a Web site s at level lv, denoted by $r_{s,lv,i}$:*

$$r_{s,lv,i} = \begin{cases} \frac{\sum_{\forall p_i level(p_i)<=lv} r_{p_i}}{n_{lv-}} & \textit{if } (n_{lv-} \neq 0), \\ 0 & \textit{otherwise.} \end{cases} \tag{4.5}$$

where n_{lv-} is the number of pages with levels smaller than or equal to lv.

Definition 8 *Out-Level Site Readability of a Web site s at level lv, under the maximum available level m, denoted by $r_{s,lv,o}$:*

$$r_{s,lv,o} = \begin{cases} \frac{\sum_{\forall p_i m \geq level(p_i)>lv} r_{p_i}}{n_{lv+}} & \textit{if } (n_{lv+} \neq 0) \textit{ and } (lv \neq m), \\ 0 & \textit{otherwise.} \end{cases} \tag{4.6}$$

where n_{lv+} is the number of pages with levels greater than lv, and within the maximum available level.

4.3 Experiments

In this section, we describe a series of experiments to evaluate our proposed scheme for Web readability assessment. The main experiments are: (1) Analyzing characteristics of Web pages having different page readability scores (in both comprehension

Table 4.2: Testing Environment

	Machine 1	**Machine 2**
CPU	Intel Pentium 4 3.2 GHz	AMD Athlon A64 3000+
RAM	4.0 GB	
Operating System	RedHat Linux Fedora Core 4	
Harddisk Size	300GB	
Programming Language	Java SDK 1.5.06 and Python 2.4.2	
Task	Text Processing, Readability Evaulation	Web Crawling

difficulty and grade level), and (2) Investigating the variation of Web site readability scores at different page levels. Table 4.2 summarizes the testing environments. We make the following assumptions throughout experiments.

1. As different types of document can be retrieved using HTTP protocol, our Web crawler will only get documents with file extensions of ". htm" / ". html", ".php", ".jsp" and ".asp", which are typical file types for Web pages.
2. For Chinese Web readability, we consider Traditional Chinese only. If Simplified Chinese pages are detected, they will first be converted to Traditional Chinese before the readability assessment. Furthermore, we only consider pages with the proportion of Chinese content exceed 50%. We make this assumption because during experiments, we find that for a page with both English and Chinese contents, even for a small portion of Chinese, such as translation of a particular term, the coding detector will still report it as a Chinese page. This assumption is made to ensure validity of the readability assessments.

Table 4.3: Summary of Web Sites Tested.

Site	**Max Level**	**URL**
CSE	10	`http://www.cse.cuhk.edu.hk`
CUHK	5	`http://www.cuhk.edu.hk`
HKGOV	7	`http://www.gov.hk`
XANGA	5	`http://www.xanga.com`

4.3.1 Experiment 1: Web Page Readability Analysis - Comprehension Difficulty

Objective

The goal of this experiment is to investigate readability level in terms of comprehension difficulty of real Web pages crawling from different sites. We try to discover special characteristics of pages having different readability scores, and observe the effects of those characteristics on page readability.

Methodology

We test the proposed page assessment using pages from the following sites:

1. *Department of Computer Science and Engineering, CUHK* (CSE). This site is used to simulate a small-size organization. It contains general information of the department, course homepages, and personal homepages of staffs and students.
2. *The Chinese University of Hong Kong* (CUHK). This site is used to simulate a large-size organization. It contains general information of the university and pages from different administrative and academic units.
3. *The Hong Kong Government* (HKGOV). As a government organization, we expect that pages in this are relatively formal and regular than other. Hence, we test it to observe

Table 4.4: Statistics of Experiment 1

	CSE	**CUHK**	**HKGOV**	**XANGA**
English				
Num. of Pages	4561	1249	46345	62822
Average Score	38.28	22.70	18.50	14.75
Std. Dev.	23.96	23.51	21.92	19.69
Chinese				
Num. of Pages	71	164	25184	–
Average Score	54.71	54.43	51.88	–
Std. Dev.	5.67	6.86	7.67	–
Both				
Num. of Pages	4632	1413	71529	–
Average Score	38.53	26.39	30.26	–
Std. Dev.	23.87	24.44	24.21	–

whether the page readability can distinguish the formality of pages.

4. *Xanga.com* (XANGA). Xanga.com is a free Web site that provides users a place to publish their articles. As a result, it contains passages written by people with different backgrounds. We only test English readability for this site as it targets for English users.

Table 4.3 summarizes information of the tested sites.

After crawling pages, we measure the number of pages crawled, average readability scores, and score distributions to study their behaviors.

Result and Discussion

Table 4.4 summarizes the statistics of tested sites. Figures 4.2, 4.3, and 4.4 show the comprehension difficulty distributions of pages in English, Chinese, and both languages respectively for each site.

Observation 1. For comprehension difficulty, from Table

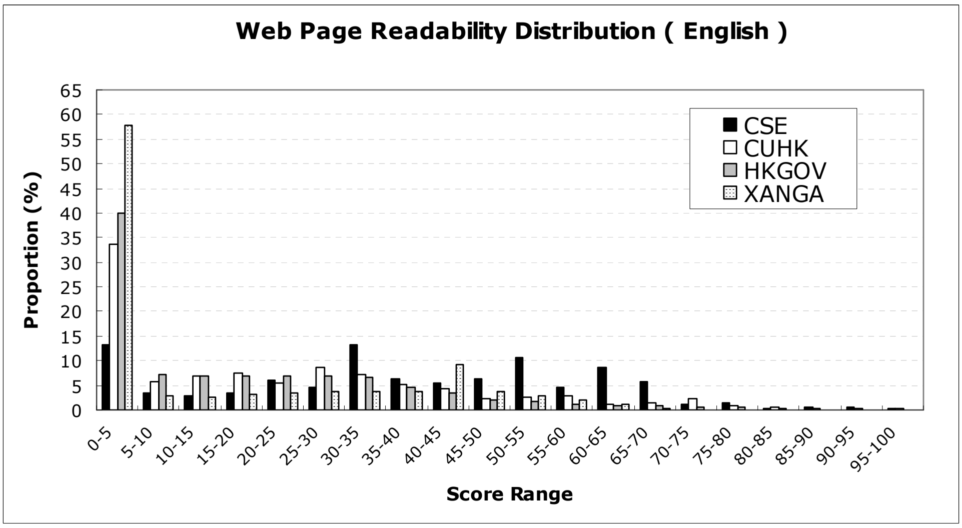


Figure 4.2: English Pages Readability Distribution (Comprehension Difficulty).

4.4, we find that the average readability score of English pages for CSE is about 40, while for CUHK, HKGOV, and XANGA are about 15 to 25, which are lower than we expected. To investigate this, we study the distributions in Figure 4.2, and observe that there are large portions of pages having scores lie in the range of 0 to 5. This explains the phenomenon of low average scores for the three sites. Our next step is to study the characteristics of low-scored pages.

Table 4.5: Manual Examination on 100 Pages in CUHK with Score Ranges: 0-5 and 5-100.

Score Range	**Page Type**		
	Index	Passage	Others
[0,5]	28	17	5
(5,100]	13	32	5

To study the characteristics of low-scored pages, we manually examine 50 randomly selected pages of scores less than or equal to 5, 50 English pages of scores greater than 5 from CUHK. We classify the pages into three types: (1) *Index pages*, (2) *Passage*

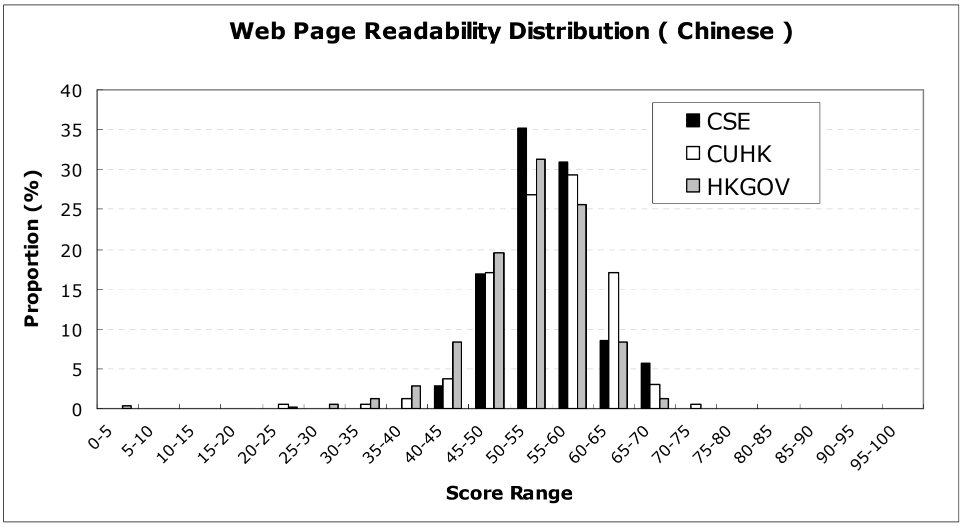


Figure 4.3: Chinese Pages Readability Distribution (Comprehension Difficulty).

pages, and (3) *Others*. *Index pages* are introductory pages which contain hyperlinks linking to internal pages. *Passage pages* contain regular articles. *Others* are pages which mainly contain non-textual contents such as scripts, images, and videos. Table 4.5 summarizes the results.

From Table 4.5, we observe that index pages generally receive low readability score than passage pages. It is because only index terms will remain after removing HTML tags and extracting raw texts from those pages, and a long sentence will form because there are no separators such as full stop to delimit the index terms. Based on Flesch Reading Ease (the English part in Equation (4.1A)), a long sentence will reduce comprehension difficulty. For the remaining discussion, we ignore low-scored pages and discuss the general distributions of the sites.

Observation 2. For English comprehension difficulty distribution (Figure 4.2), we find that CSE has a larger portion of pages in high score ranges, such as 50-55, 60-65 and 65-70, than the other sites. It is because CSE contains personal pages of staffs and students, in which the contents are more compre-

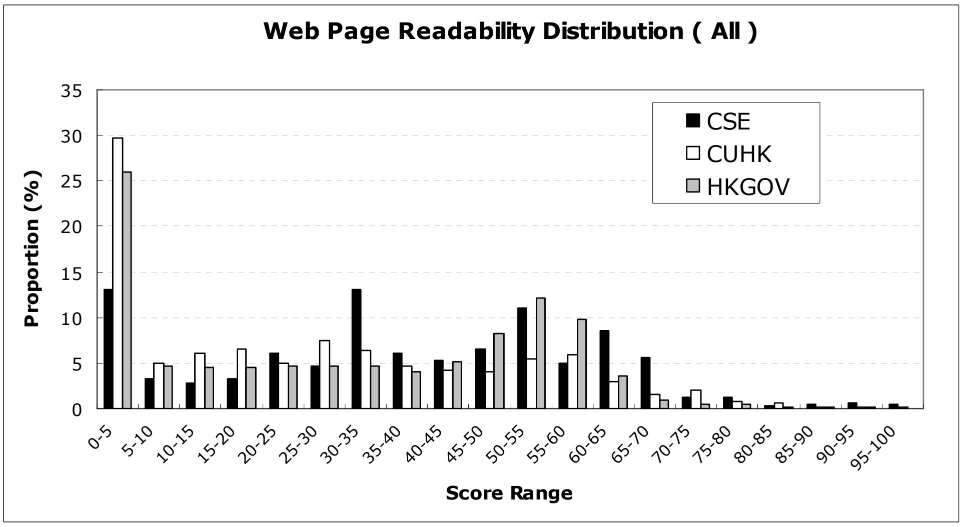


Figure 4.4: Readability Distribution of Pages in Both Languages (Comprehension Difficulty).

hensible than formal articles published in CUHK and HKGOV, where pages from them mainly locate in the range 10 - 40. Xanga has a relatively even distribution from 10 to 60, except a sharp rise in 40-45, meaning the page authors are from different backgrounds.

Observation 3. For Chinese comprehension difficulty distribution (Figure 4.3), pages in all three sites mainly locate in the range 50 to 60, indicating that Chinese passages are generally with similar difficulty. The effect of low-scored “index-page” which happens in English readability does not occur here. It is because, by referring to Yang formula (the Chinese part in Equation (4.1)), sentence length is not considered, and the most significant factor is the proportion of basic Chinese words. This also indicates that page contents are composed in words with similar difficulties, and are suitable to general users.

We conclude this part by studying Figure 4.4, in which both English and Chinese pages are taken into considerations. As English pages dominate the data set, and those pages with score ranged from 0 to 5 take the largest portion, indicating that there are large amount of pages with low content-values. For the three Web sites being investigated, pages in CUHK and HKGOV have readability levels generally lower than that of CSE. The explanation to this phenomenon is the formalness and technicalness of articles in the two sites.

4.3.2 Experiment 2: Web Page Readability Analysis - Grade Level

Objective

Similar to Experiment 1, the goal of this experiment is to investigate readability of real Web pages in terms of grade level. We try to discover special characteristics of pages having different readability scores, and compare the results with previous experiment.

Methodology

We apply the Web page grade level assessment to the same data set used in the previous experiment. We also plot the graph of comprehension difficulty with grade level to observe the relationship between them.

Result and Discussion

Table 4.6 summarizes the statistics of tested sites. Figures 4.5, 4.7, and 4.9 shows the grade level distributions of pages in English, Chinese, and both languages respectively for each site.

Observation 1. Similar to English grade level distribution in Figure 4.5, it is expected that there is a large portion of high-scored pages in grade level difficulty distribution. It is because both Flesch reading ease and Flesch-Kincaid grade level use the same factors in the formulas: the number of syllables per word and average sentence length, with the difference in the sign of

Table 4.6: Statistics of Experiment 2

	CSE	**CUHK**	**HKGOV**	**XANGA**
English				
Num. of Pages	4561	1249	46345	62822
Average Score	9.28	9.41	9.67	9.96
Std. Dev.	1.42	1.67	1.36	0.28
Chinese				
Num. of Pages	71	164	25184	–
Average Score	4.28	2.97	3.18	–
Std. Dev.	1.57	0.95	1.25	–
Both				
Num. of Pages	4632	1413	71529	–
Average Score	9.20	8.66	7.38	–
Std. Dev.	1.54	2.61	3.36	–

their coefficients. So low score in Flesch reading ease (meaning relatively more difficult) results in high score in Flesch-Kincaid grade level (appropriate to people with higher grade level of education).

To study relation of the two readability measures of readability, we plot comprehension difficulty against grade level of the tested sites in Figure 4.6. From the result we find that the two measures are negatively correlated, as they both depend on same variables. Furthermore, pages having comprehension score less than approximately 50 would result in grade level greater than 10, indicating that the range of comprehension difficulty having large grade level is large. This shows that Flesch-Kincaid grade level assessment is not suitable to differentiate pages with low comprehension scores, which are common in the Web.

Observation 3. For Chinese grade level distribution in Figure 4.7, the scores fall in the range of primary grade level (level 2 to 6) for the three sites, which generally agrees to comprehension difficulty that page contents are suitable to general users.

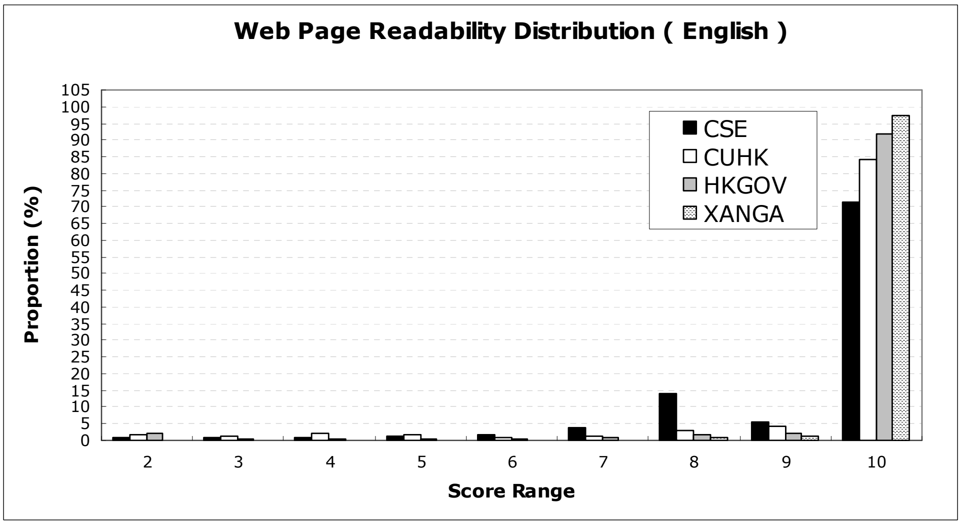
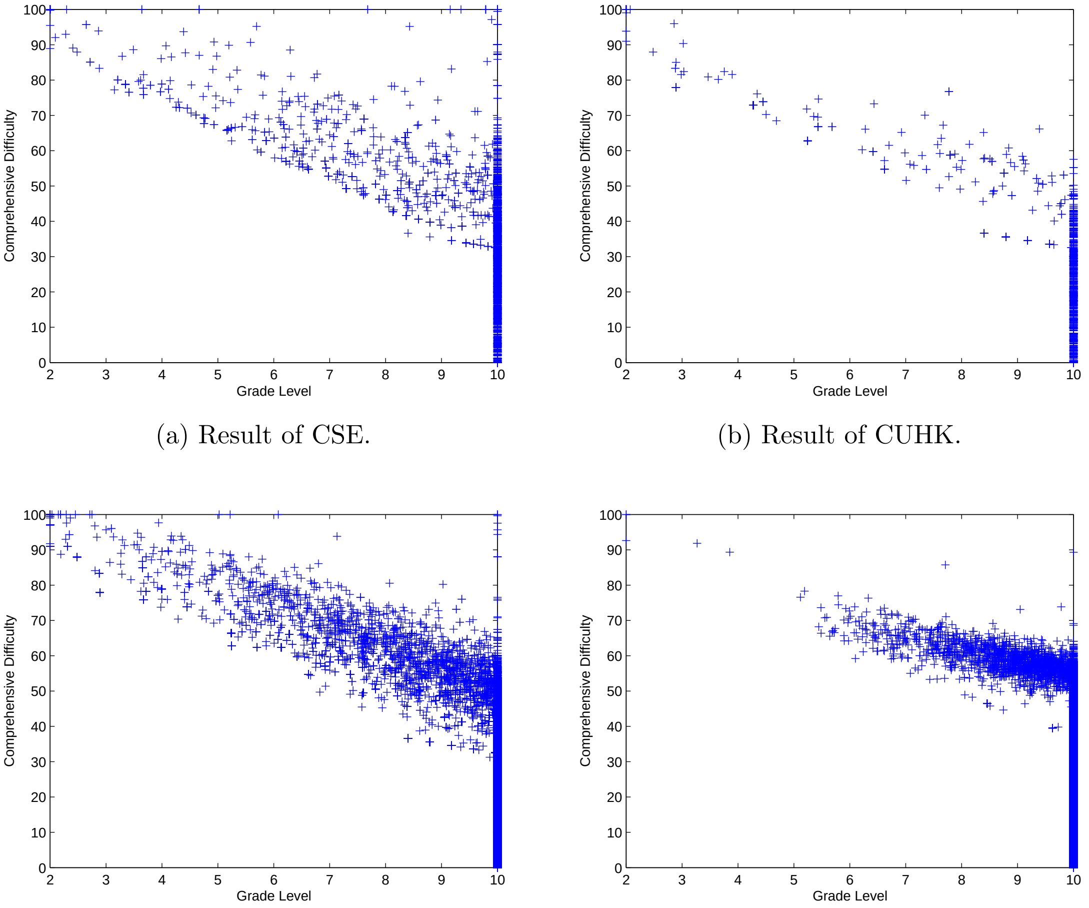


Figure 4.5: English Pages Readability Distribution (Grade Level).

To explain the fact that scores mainly distribute among primary school levels, we need to refer to the nature of training data used in our readability analysis. In the training data of secondary school level, actually the passages are composed by famous authors, in which ordinary people are not able to write. As a result, it is reasonable that pages mainly locate in primary school level.

We plot the graph of comprehension difficulty using Yang formula against grade level using our proposed SVR method in Figure 4.8 to study their relationship. Unlike the case in English that the two measures show a negative correlation, there is no distinguish relationship the case of Chinese. It is because the two approaches are based on different variables: the most significant factor in Yang formula is proportion of basic words, while in our proposed method, we consider factors from different language levels. As a result, it is reasonable that there is no correlation between them, e.g. a passage having more basic words would not necessarily mean low grade level. We need to further consider other factors to determine the result.



(a) Result of CSE.

(b) Result of CUHK.

(c) Result of HKGOV.

(d) Result of XANGA.

Figure 4.6: Comprehension Difficulty against Grade Level (English Web Page).

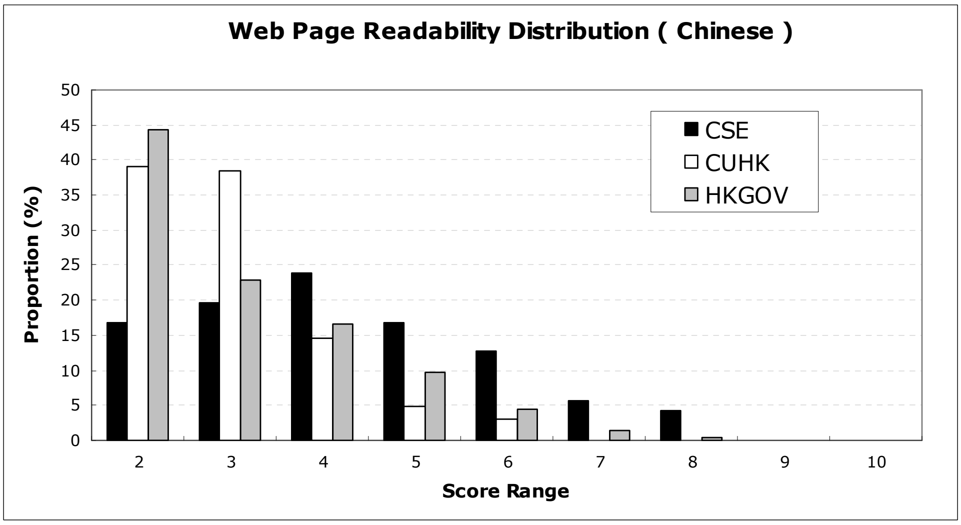
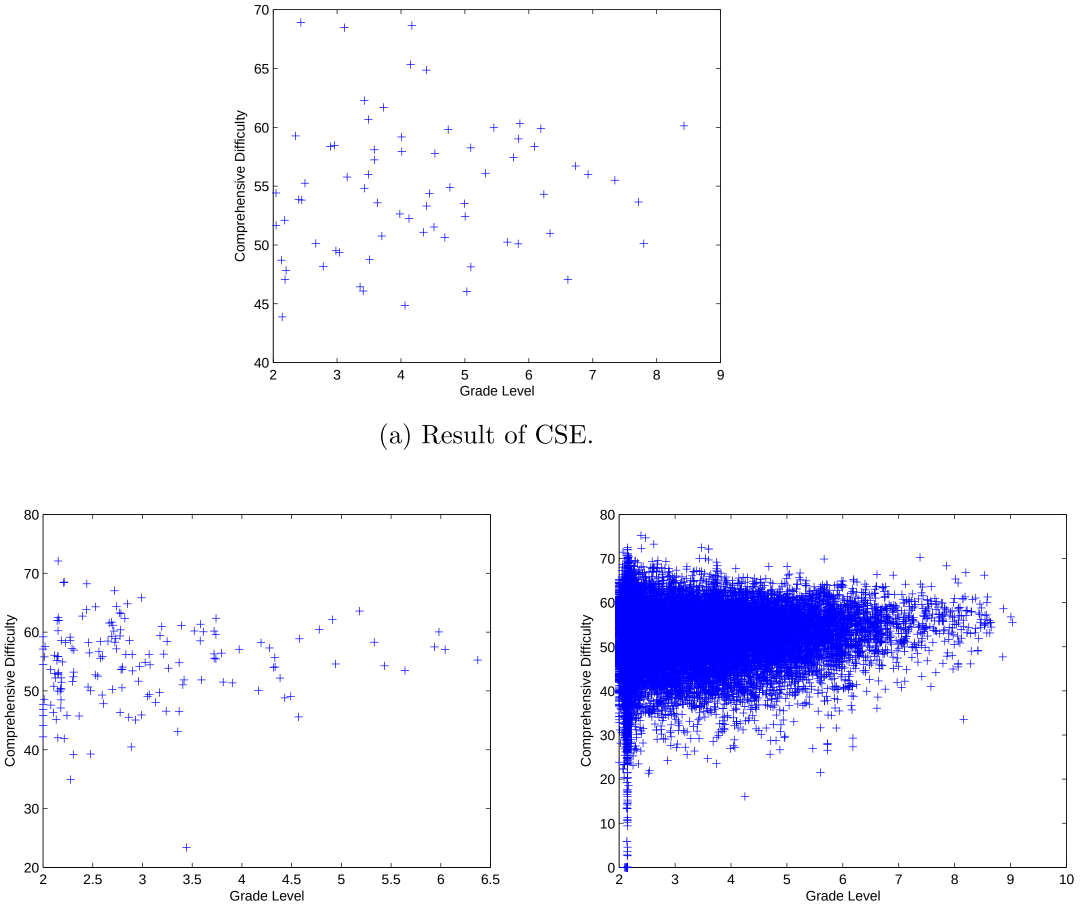


Figure 4.7: Chinese Pages Readability Distribution (Grade Level).

4.3.3 Experiment 3: Web Site Readability Analysis

Objective

The goal of this experiment is to evaluate the proposed Web site readability assessments in comprehension difficulty scale with the real Web data. The reasons of applying comprehension difficulty are that, firstly Flesch-Kincaid assessment is not suitable to estimate pages with low comprehension score, and pages easily fall in the range of greater than level 10. Secondly, scales used by Flesch-Kincaid and our proposed SVR method do not fit probably. Flesch-Kincaid tends to estimate pages with high grade level, while our method estimates pages at lower level due to the training data being used. On the other hand, two approaches measuring comprehension difficulty have a better scale, so we apply the measurement in this section. We investigate factors affecting site readability, and the variation of readability against pages at different page levels.



(a) Result of CSE.

(b) Result of CUHK.

(c) Result of HKGOV.

Figure 4.8: Comprehension Difficulty against Grade Level (Chinese Web Page).

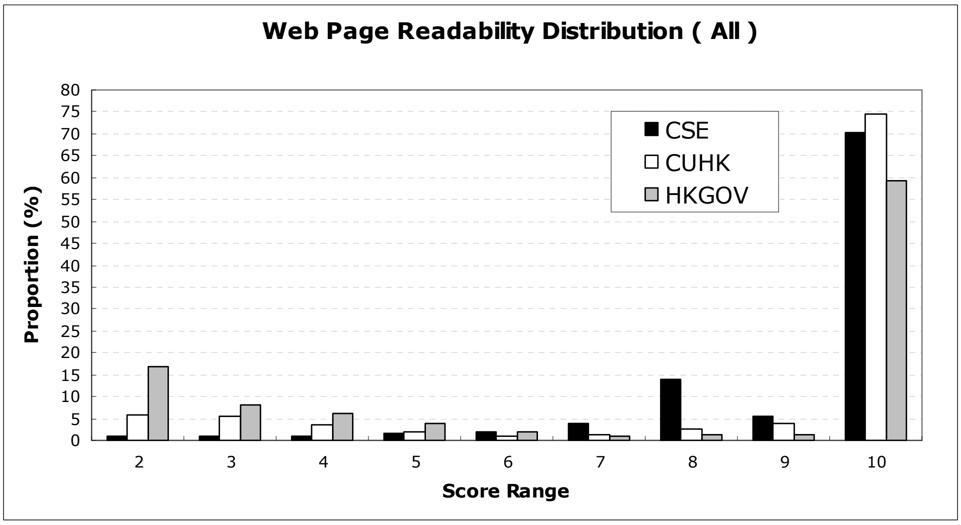


Figure 4.9: Readability Distribution of Pages in Both Languages (Grade Level).

Methodology

The data set used in this experiment is same as that in Experiment 2, in which the four Web sites, CSE, CUHK, HKGOV, and XANGA are being investigated (Table 4.3). We then study the three proposed assessments: (1) Exact-Level, (2) In-Level, and (3) Out-Level site readability against page level. Apart from Xanga data set, in which only English pages are available for investigation, we use both English and Chinese pages in CSE, CUHK, and HKGOV to estimate site readabilities.

Result and Discussion

Figures 4.10, 4.11, 4.12, and 4.13 show results of the four sites. We discuss the case of CSE in detail as all the tested sites show a similar behaviors, in which there is a fluctuation of readability score against levels.

Observation 1. CSE shows a readability behavior of dramatic change. We find that for Exact-Level score, there is a rise in level 4 - 5, and drop again in level 6. After studying the pages

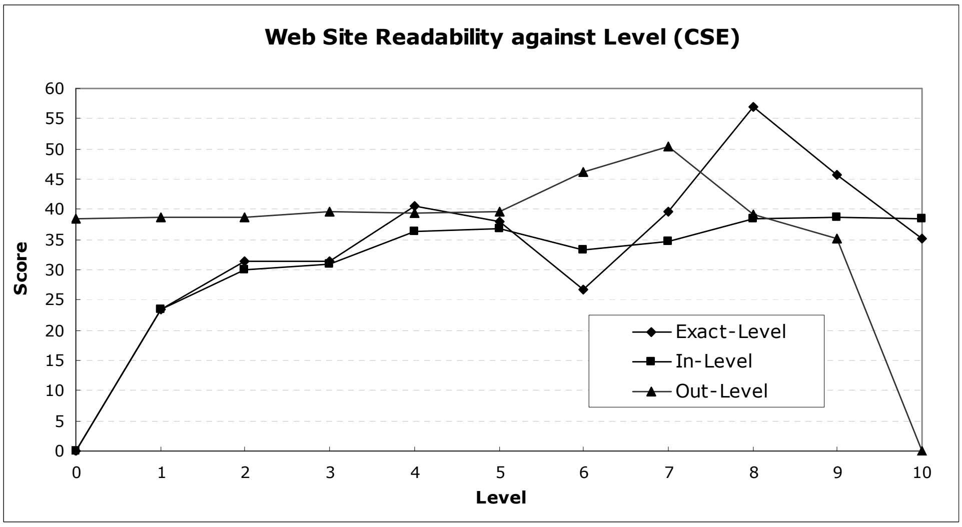


Figure 4.10: Site Readability of CSE.

in these levels, we find that level 4 - 5 are the levels where the personal homepages reside. So based on the argument in Experiment 2 discussion, in which personal homepages have score generally higher than official articles and index pages, these two levels receive a higher score than level 0 - 3. For level 6, we find that as this level follows the personal homepages, authors would like to put more non-textual information such as images, videos, etc. in this level. As a result, there is a drop of readability score.

But we find one drawback in current readability during the investigation. Although the score reach its maximum at level 8, after randomly examining pages with score greater than 75 in this level, we find that the high score is contributed by documents like programming codes, which are commonly used as tutorials in CSE courses. As programming codes contain a lot of short words such as “if”, “then”, “for” (the keywords of programming languages), such short words will have small number of syllables, and thus favor the calculation of Flesch Formula. To overcome the problem, we need to design some rules to eliminate these types of content when applying the readability formula.

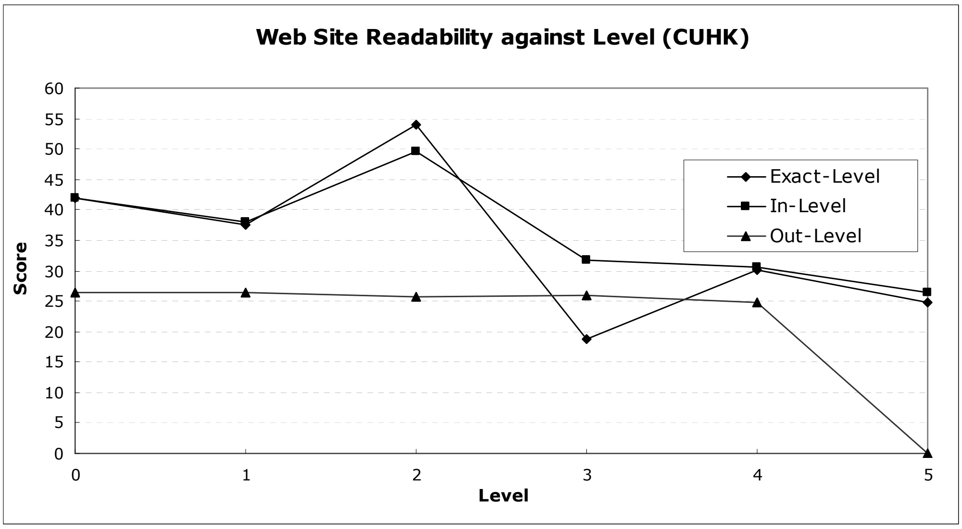


Figure 4.11: Site Readability of CUHK

Observation 2. In-Level score generally follows the trend of Exact-Level, but with a smoother variation. As the In-Level score indicates readability of a site starting from root page up to a specified level, it has a "smoothing" effect on Exact-Level.

Observation 3. Out-Level score shows the readability of a site for pages after the specified level. We discover that Out-Level readability generally has a smaller degree of variation than Exact-Level and In-Level site readabilities, and it approaches the Exact-Level score at higher level. For example in CUHK, the Out-Level score stays around at 25, which is close to the Exact-Level score at 30. It is because in a large Web site, number of pages at higher levels is much greater than lower levels, so their scores would dominate the calculation which involves averaging. This simulation result reflects that Out-Level readability may only reflect difficulties of page levels having larger amount of pages. This suggests that we need to apply other measures, such as random sampling, when considering the number of pages in the calculation.

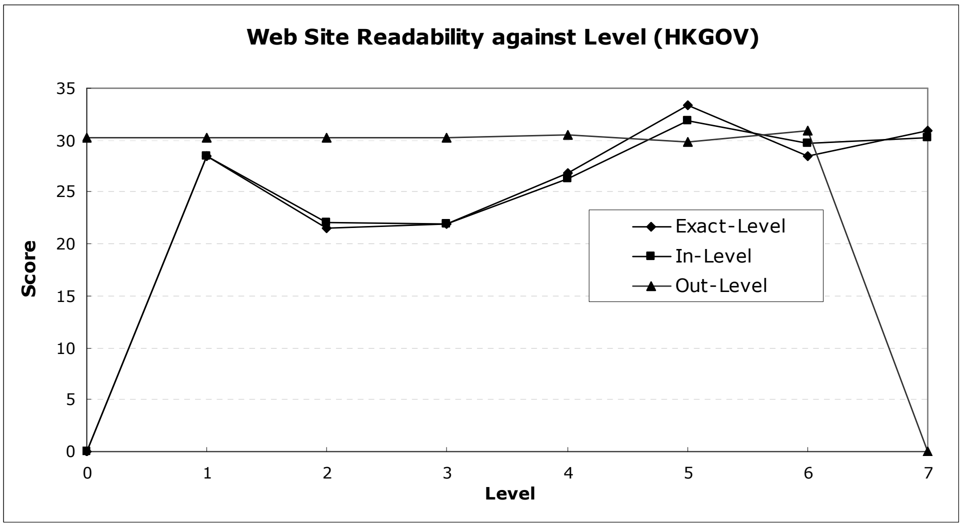


Figure 4.12: Site Readability of HKGOV.

4.4 Summary and Future Work

In this chapter, we propose a bilingual readability assessment scheme for Web page and site in English and Chinese languages. For page readability assessment, unlike other researches which mainly focus on visual appearance, our scheme utilizes textual features to assess readability scores for English and Chinese pages respectively. We believe that our work is the first study on applying Chinese readability assessment in Web application. Furthermore, we propose three Web site readability assessments, Exact-Level, In-Level, and Out-Level scores, based on readability of pages at different page levels. Our preliminary experimental results show that the assessments, apart from measuring difficulty in comprehending the pages, can also serve as a heuristic in figuring out low content-valued pages such as index pages. Furthermore, by studying the variation of site readability against page level, we can also get an overall picture of content distribution within a site. Our readability assessments can help designers to create Web pages and sites which are more structural and user-friendly.

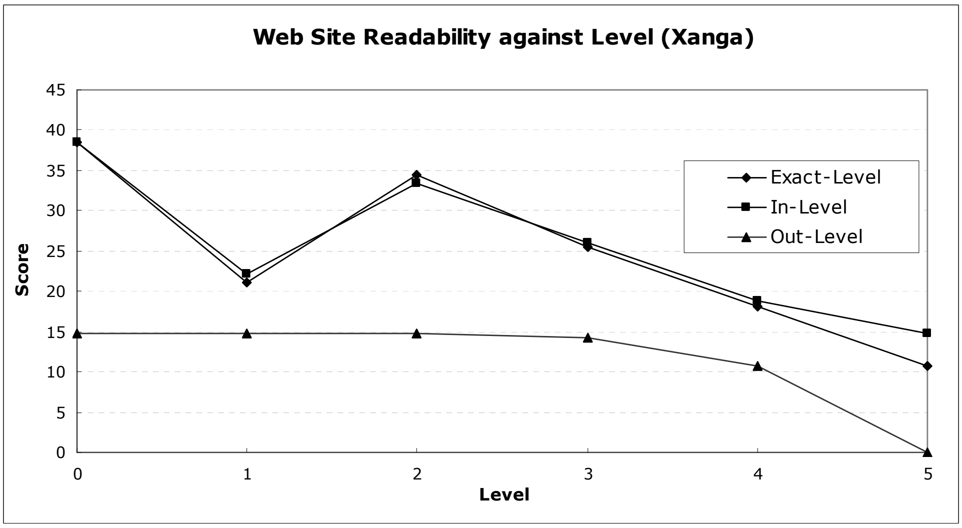


Figure 4.13: Site Readability of XANGA.

Our future work is mainly in three directions: (1) establishing better readability formulas for Web, (2) performing experiments with larger scale, and (3) applying the proposed scheme in other Web-related fields. For the first direction, although the assessment used are proven to be effective in readability evaluation, we still need to establish better formulas to tackle Web specific problems. For example, we need to extend our scheme to handle multilingual pages other than just English and Chinese languages, to evaluate pages with mixed languages, etc. For second direction, we need to carry out larger scale experiment on Web sites of different categories, such as news, sports, company, articles in Wikipedia [73] etc., so as to discover more valuable characteristics of Web Readability. Finally for the third direction, we are now investigating the application of Web readability in Web Pages Recommendation System. When a user submits a query, our system will be able to return the pages which are not only relevant, but also appropriate to the user's ability level.

Chapter 5

Conclusion

In this thesis, we give a literature review on readability analysis, support vector machine, and Chinese word segmentation. Then we conduct analyses on Chinese readability assessment and propose an application of readability estimation on the Internet. Our work is summarized as follows.

For the literature review, we study related works of the readability analysis, the support vector machine, and Chinese word segmentation. Research works on English readability assessment has a long history in the literature, and the assessment methods can be classified based on the viewpoints of complexity modeling and assessment formation. Research works on Chinese readability can be dated back to Yang's work in 1971, but it does not have enough attention afterwards. We then introduce two works of Web readability by Hill, and Si and Shan. For support vector machine, we discuss its characteristics, advantages, and applications. For Chinese word segmentation, we briefly mention some difficulties in performing the task, and approaches to solve the problems.

For Chinese readability analysis, based on the motivation that Chinese language is becoming more important nowadays, but is lacking of research works for readability analysis, we improve the existing approaches by (1) analyzing potential factors affecting Chinese readability in a systematic way, and (2) applying advanced text processing and machine learning techniques.

In the analysis, we obtain the training data from the Chinese language textbooks in the primary and secondary school levels. We then apply our proposed LMR-RC Tagging approach in performing Chinese word segmentation to extract features. Regression analysis, using the Support Vector Regression (SVR), and feature selection process, using genetic algorithm (GAFS), are applied to perform readability assessment modeling.

We then evaluate the proposed work by measuring the performances of applying LMR-RC Tagging approach in Chinese word segmentation, and SVR and GAFS in the readability estimation. Furthermore, we compare the proposed work with the existing approaches, which mainly use the Linear Regression as the modeling technique. According to the experimental results, our method can successfully estimate the readability at ±1 grade level, and it is better than the approaches using LR.

For Web page analysis, based on the motivation that readability can potentially benefit the development of user-oriented Internet applications such as personalized content delivery service, we analyze the application of readability assessment on the Internet. We propose a novel bilingual Web page and site readability assessment scheme, which employs both existing and our proposed readability assessment.

For Web page readability, after removing some useless components in a Web page (such as HTML tags and scripts) and detecting the language being used, the readability of the remaining textual contents are evaluated. We try to measure readability in terms of comprehension difficulty and grade level using the Flesch reading ease, the Flesch-Kincaid grade level, the Yang's Chinese readability assessment, and our proposed approach.

For Web site readability, we first model a Web site as a tree, with the index page as the root. Then three assessments, Exact-Level, In-Level, and Out-Level are proposed based on the readability of pages at different levels (depth of the tree). Exact-Level score indicates the average readability of pages at a particular level. In-Level score gives the average readability of Web pages starting from root page up to pages at the specified level. Out-Level score indicates the readability of pages at the remaining levels. We then briefly discuss their potential usages.

We perform experiments on evaluating our proposed scheme using real Web data. At the same time, we try to discover some special characteristics of pages and sites having different readability scores. Experimental results indicate that, in addition to indicating readability level, the estimated scores act as a good heuristic to figure out pages with low content-values. Furthermore, we can obtain an overall content distribution in a Web site by studying the variation of its readability.

Appendix A

List of Symbols and Notations

Table A.1 shows symbols and notations appeared in the thesis.

Table A.1: Lists of Symbols and Notations

Symbol	**Meaning**
	Chinese Readability Analysis
c	Chromosome in GA / Parameter "cost" in SVR
C, C_B	Entire and basic character list
$\mathcal{C}$, $\mathcal{C}_{name}$	Character level factors
d	Number of features in F'
D	Number of features in F
F	Entire feature set
F'	Selected feature subset
g	Gene in GA / Parameter "gamma" in SVR
$HitRate \pm \epsilon$	Hit Rate
J_i	Fitness functions used in Genetic Algorithm of feature selection
L, M, R, S	Tags used in LMR-RC Tagging scheme
m	Number of points in m-point crossover
MAE	Mean Absolute Error
MPE	Maximum Prediction Error
MSE	Mean Squared Error
N	Number of testing passages

Symbol	**Meaning**
N_g	Number of good chromosomes
N_{gen}	Number of generations
N_p	Population size
N_{sel}	Number of selections
p	Parameter "epsilon" in SVR
P	Population in GA
$\mathcal{P}$, $\mathcal{P}_{name}$	Phrase level factors
r, r^2	Pearson Correlation Coefficient and Squared Correlation Coefficient
$\mathcal{R}$, $\mathcal{R}_{name}$	Sub-character level factors
r_c	Crossover rate
r_m	Mutation rate
$\mathcal{S}$, $\mathcal{S}_{name}$	Sentence level factors
$STDDEV\_AE$	Standard deviation of Absolute Error
$STDDEV\_SE$	Standard deviation of Squared Error
W, W_B	Entire and basic word list
$\mathcal{W}$, $\mathcal{W}_{name}$	Word level factors
X	A piece of Chinese text / Independent variables
Y	Readability level (Grade level) / Dependent variable
Y_i	Actual grade level of passage X_i
$\hat{Y}_i$	Predicted grade level of passage X_i
Web Readability Analysis	
lv	Page level
n_{lv}	Number of Web pages with levels equal to lv
n_{lv+}	Number of Web pages with levels greater than lv, and within the maximum available level
n_{lv-}	Number of Web pages with levels smaller than or equal to lv
p	A Web page
p_0	Root page of a Web site
P	A set of Web pages
r_p	Web page readability score
$r_{s,lv,e}$	Exact-Level Site Readability of Web site s at level lv
$r_{s,lv,i}$	In-Level Site Readability of Web site s at level lv

Symbol	**Meaning**
$r_{s,lv,o}$	Out-Level Site Readability of Web site s at level lv
s	A Web site
S	A set of Web sites
X_{C_i}	Factors in Chinese readability assessment
X_{E_i}	Factors in English readability assessment

Appendix B

List of Publications

Here is a list of publications during my study:

1. **Tak Pang Lau** and Irwin King. Bilingual Web Page and Site Readability Assessment. In Proceedings of the 15th International World Wide Web Conference, Pages 993-994, 2006.

2. **Tak Pang Lau** and Irwin King. Two Phase LMR-RC Tagging for Chinese Word Segmentation. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Pages 183-186, 2005.

3. Wan Yeung Wong, **Tak Pang Lau**, and Irwin King. Information Retrieval in P2P Networks Using Genetic Algorithm. In Proceedings of the 14th International World Wide Web Conference, Pages 922-923, 2005.

4. Irwin King and **Tak Pang Lau**. Advanced Chinese Readability Analysis using Support Vector Regression. In preparation to submit to ACM Transactions on Asian Language Information Processing (TALIP), 2006.

5. Irwin King and **Tak Pang Lau**. Bilingual Web Readability Assessment. In preparation to submit to ACM Transactions on Information Systems (TOIS), 2006.

6. Irwin King, Wan Yeung Wong, and **Tak Pang Lau**. A Genetic Algorithm for Query Routing in Hybrid Peer-to-Peer Networks. Submitted to IEEE Transactions on Evolutionary Computation for review, 2005.
7. Dexter Chi Wai Siu and **Tak Pang Lau**. Distributed Ranking Over Peer-to-Peer Networks. In Proceedings of the 13th International World Wide Web Conference, Pages 356-357, 2004.
8. Wan Yeung Wong, **Tak Pang Lau**, Irwin King, Michael R. Lyu. A Tutorial on RDF with Jena. Book chapter to appear in Advances in Electronic Business Vol. 2, Idea Group, 2007.

Here is a list of projects and publications involved:

1. Chi Chung Mak, Andy Chi Chung Chan, Irwin King, and Jimmy Ho-Man Lee. The Chinese University Plagiarism IDentification Engine (CUPIDE) System. Third award in the 9th Challenge Cup, Fudan University, Shanghai, 2005.
2. Chi Chung Mak, Andy Chi Chung Chan, Irwin King, and Jimmy Ho-Man Lee. The Chinese University Plagiarism IDentification Engine (CUPIDE) System. Champion in the Vice-Chancellor's Cup of Student Innovation, The Chinese University of Hong Kong, 2005.
3. Chi Chung Mak, Andy Chi Chung Chan, Irwin King, and Jimmy Ho-Man Lee. The Chinese University Plagiarism IDentification Engine (CUPIDE) System. Champion in the IEEE CI Final Year Project Competition, 2005.
4. Haixuan Yang, Irwin King, and Michael R. Lyu. Predictive Random Graph Ranking on the Web. To appear in Proceedings of the 2006 IEEE World Congress on Computational Intelligence, 2006.

5. Haixuan Yang, Irwin King, and Michael R. Lyu. Predictive Ranking: A Novel Page Ranking Approach by Estimating the Web Structure. In Proceedings of the 14th International World Wide Web Conference, Pages 944-945, 2005.

Bibliography

[1] ACL. Web site of SIGHAN. *http://www.sighan.org/*.

[2] C. W. Ahn and R. S. Ramakrishna. A genetic algorithm for shortest path routing problem and the sizing of populations. *Transactions on Evolutionary Computation*, 6:6:566–579, 2002.

[3] J. Baldridge, T. Morton, and G. Bierner. The opennlp maxent package in Java. *http://maxent.sourceforge.net*, 2004.

[4] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[5] J. R. Bormuth. Readability: A new approach. *Reading Research Quarterly*, 1:79–132, 1966.

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *D. Haussler, editor, Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–152, 1992.

[7] P. S. Bradley, O. L. Mangasarian, and W. N. Street. *Advances in Neural Information Processing Systems -9- (NIPS*96)*, chapter Clustering via concave minimization, pages 368–374. MIT Press, Cambridge, MA, 1997.

[8] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, M. Ares, and J. D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. In *Proceedings of the National Academy of Sciences*, volume 97, pages 262–267, 1997.

[9] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[10] C. Burges and D. Crisp. *Advances in Neural Information Processing Systems*, volume 12, chapter Uniquess of the SVM Solution, pages 223–229. MIT Press, Cambridge, MA, 2000.

[11] C. Chang and C. Chen. A study of integrating Chinese word segmentation and part-of-speech tagging. *Communications of COLIPS*, 3(1):69–77, 1993.

[12] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *Software available at* ***http://www.csie.ntu.edu.tw/~cjlin/libsvm***, 2001.

[13] CIA. The world fact book. ***http://www.odci.gov/cia/publications/factbook/print/xx.html***, 2006.

[14] Y. Dai, T. E. Loh, and C. S. G. Khoo. A new statistical formula for Chinese text segmentation incorporating contextual information. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pages 82–89, 1999.

[15] E. Dale and J. S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27:11–20, 1948.

[16] T. Emerson. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN*

Workshop on Chinese Language Processing, pages 123–133, 2005.

[17] EuroAsiaSoftware. Learn Chinese. `http://www.euroasiasoftware.com/english/chinese/learn/grundstreckeng.html#Grundstreck:`, 2004.

[18] J. N. Farr, J. J. Jenkins, and D. G. Paterson. Simplification of Flesch reading ease formula. *Journal of Applied Psychology*, 35:333–337, 1951.

[19] S. C. Fen and L. David. Chinese language. *Microsoft Encarta 2006 [DVD], Redmond, WA: Microsoft Corporation*, 2005.

[20] R. F. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233, 1948.

[21] R. Fletcher. *Practical Methods of Optimization.* John Wiley and Sons, New York, 1989.

[22] E. B. Fry. Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*, 21(3):242–252, 1977.

[23] Global-Reach. Global internet statistics: Sources and references. `http://global-reach.biz/globstats/refs.php3`, 2004.

[24] R. Gunning. *The technique of clear writing.* New York: McGraw-Hill, 1952/1968.

[25] T. S. Hansell. Readability, syntactic transformations, and generative semantics. *Journal of Reading*, 19(7):557–562, 1976.

[26] A. J. Harris and M. D. Jacobson. A framework for readability research: Moving beyond Herbert Spencer. *Journal of Reading*, 22(5):390–398, 1979.

[27] K. A. Heller, K. M. Svore, A. D. Keromytis, and S. J. Stolfo. One class support vector machines for detecting anomalous window registry accesses. In *3rd IEEE Conference Data Mining Workshop on Data Mining for Computer Security*, 2003.

[28] A. L. Hill and L. F. V. Scharff. Readability of websites with various foreground/background color combinations, font types and word styles. *Engineering Psychology and Cognitive Ergonomics*, 4:123–130, 1999.

[29] C. W. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification. *Available at http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf*, 2003.

[30] K. Huang. *Learning From Data Locally and Globally.* PhD thesis, Department of Computer Science and Engineering, The Chinese University of Hong Kong, 2004.

[31] C. C. Jeng. Chinese readability using artificial neural networks. *Dissertation for Doctor of Education, Northern Illinois University*, 2001.

[32] Jing 荊, 溪昱 中文國文教材的適讀性研究: 適讀年級的推估. 教育研究資訊, 3(3):113–127, 1995.

[33] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In C. Nsedellec and C. Rouveirol, editors, *10th European Conference on Machine Learning*, pages 137–142, 1998.

[34] I. King, W. Y. Wong, and T. P. Lau. A genetic algorithm for query routing in hybrid peer-to-peer networks. *Submitted to IEEE Transactions on Evolutionary Computation for review*, 2005.

[35] G. R. Klare. Table for rapid determination of Dale-Chall readibility scores. *Educational Research Bulletin*, 31:43–47, 1952.

[36] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999.

[37] C.-C. Kuo. A Chinese text-to-speech system with text preprocessing and confidence measure for practical usage. In *Proceedings of 1997 IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications*, volume 2, pages 583–586, 1997.

[38] Language Centre HKBU. A study of the Chinese characters recommended for the subject of Chinese language in primary schools. `http://alphads10-2.hkbu.edu.hk/~lcprichi/`, 2003.

[39] J. M. Lattin, J. D. Carroll, and P. E. Green. *Analyzing Multivariate Data*. Duxbury Press, 1 edition, 2002.

[40] T. P. Lau and I. King. Two-phase LMR-RC tagging for Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 183–186, 2005.

[41] T. P. Lau and I. King. Bilingual Web page and site readability assessment. In *Proceedings of the 15th International World Wide Web Conference*, pages 993–994, 2006.

[42] I. Liu. Descriptive-unit analysis of sentences: Toward a model natural language processing. *Computer Processing of Chinese and Oriental Languages*, 4(4):314–355, 1990.

[43] K. Lua. From character to word - An application of information theory. *Computer Processing of Chinese and Oriental Languages*, 4(4):304–312, 1990.

[44] K. Lua. Experiments on the use of bigram mutual information in Chinese natural language processing. In *Presented in the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL)*, 1995.

[45] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618, 2003.

[46] D. R. McCallum and J. L. Peterson. Computer-based readability indexes. In *Proceedings of the ACM'82 Conference*, pages 44–48, 1982.

[47] H. G. McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

[48] Y. Meng, H. Yu, and F. Nishino. A lexicon-constrained character model for Chinese morphological analysis. In *IJCNLP*, pages 542–552, 2005.

[49] T. Nakagawa and Y. Matsumoto. Detecting errors in corpora using support vector machines. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, 2002.

[50] H. T. Ng and J. K. Low. Chinese part-of-speech tagging. One-at-a-time or all-at-once? Word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 277–284, 2004.

[51] I.-S. Oh, J.-S. Lee, and B. R. Moon. Hybrid genetic algorithms for feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1424–1437, 2004.

[52] J. L. Packard. *The Morphology of Chinese: A Linguistic and Cognitive Approach.* Cambridge University Press, 2000.

[53] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Libraries*, 1998.

[54] R. D. Powers, W. A. Summer, and B. E. Kearl. A recalculation of four adult readability formulas. *Journal of Educational Psychology*, 49:99–105, 1958.

[55] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, pages 133–142, 1996.

[56] J. Redish. Readability formulas have even more limitations than klare discusses. *ACM Journal of Computer Documentation (JCD)*, 24(3):132–137, 2000.

[57] C. J. V. Rijsbergen. *Information Retrieval.* London, Butterworths, 2 edition, 1979.

[58] L. Si and J. Callan. Information retrieval and text mining: A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576, 2001.

[59] H. D. Simons. Reading comprehension: The need for a new perspective. *Reading Research Quarterly*, 6(3):338–363, 1971.

[60] A. Smola and B. Sch. A tutorial on support vector regression. *NeuroCOLT2 Technical Report NC2-TR-1998-030*, 1998.

[61] R. Sproat and C. Shih. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351, 1990.

[62] R. Sproat, C. Shih, W. Gale, and N. Chang. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3):377–404, 1996.

[63] Sun Microsystems. JAVA programming languague. *http://java.sun.com/*, 2006.

[64] W. Taylor. Cloze procedures: A new tool for measuring readability. *Journalism Quarterly*, 53:415–433, 1953.

[65] W. J. Teahan, R. McNab, Y. Wen, and I. H. Witten. A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3):375–393, 2000.

[66] C. H. Tsai. Frequency and stroke counts of Chinese characters. *http://technology.chtsai.org/charfreq/*, 2005.

[67] C. H. Tsai. A review of Chinese word lists accessible on the internet. *http://technology.chtsai.org/wordlist/*, 2005.

[68] C. Tung and H. Lee. Identification of unknown words from a corpus. *Computer Processing of Chinese and Oriental Languages*, 8:131–145, 1994.

[69] V. N. Vapnik. *Estimation of Dependencies Based on Empirical Data.* Nauka, Moscow, 1979.

[70] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[71] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[72] L. D. Whitley and M. D. Vose. *Foundations of Genetic Algorithms 3.* Morgan Kaufmann, 1995.

[73] Wikipedia. Main page — wikipedia, the free encyclopedia, 2006.

[74] W. Y. Wong, T. P. Lau, and I. King. Information retrieval in P2P networks using genetic algorithm. In *Proceedings of the 14th International World Wide Web Conference*, pages 922–923, 2005.

[75] S. Workshop. Second international Chinese word segmentation bakeoff. *http://sighan.cs.uchicago.edu/bakeoff2005/*, 2005.

[76] D. Wu and H. Wong. Machine translation with a stochastic grammatical channel. In *Proceedings of the 17th international conference on Computational linguistics*, volume 2, pages 1408–1415, 1998.

[77] H. Wu, H. Lu, and S. Ma. A practical SVM-based algorithm for ordinal regression in image retrieval. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 612–621, 2003.

[78] Z. Wu and G. Tseng. Chinese text segmentation for text retrieval: achievements and problems. *Journal of the American Society for Information Science*, 44(9):532–542, 1993.

[79] N. Xue and L. Shen. Chinese word segmentation as LMR tagging. In *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pages 176–179, 2003.

[80] H. Yang. Margin variations in support vector regression for the stock market prediction. Master's thesis, Department of Computer Science and Engineering, The Chinese University of Hong Kong, 2003.

[81] S. J. Yang. A readability for Chinese language. *Ph.D. Thesis for Mass Communication, University of Wisconsin*, 1971.

[82] R. B. Yates and B. R. Neto. *Modern Information Retrieval.* ACM Press, 1 edition, 1999.

[83] G. Zhu, D. liang, Y. Liu, Q. Huang, and W. Gao. Improving particle filter with support vector regression for efficient visual tracking. In *International Conference on Image Processing(ICIP2005)*, pages 422–425, 2005.

[84] Y. Zhu, L. Zhao, Y. Xu, and Y. Niimi. A Chinese text to speech system based on TD-PSOLA. In *Proceedings of 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, volume 1, pages 204–207, 2002.

[85] C. Zong, B. Xu, and T. Huang. Interactive Chinese-to-English speech translation based on dialogue management. In *Proceedings of the ACL-02 workshop on Speech-to-speech translation*, volume 7, pages 61–68, 2002.

[86] 今日中國語文編委會. 今日中國語文, volume 1-12. 教育出版社有限公司, 1991.

[87] 宋琦, 譚帝森, 漢聞, 陳佳榮. 新編中國語文, volume 1-10. 齡記出版社有限公司, 1991.