Chinese Readability Assessment and Applications

Irwin King

Department of Computer Science and Engineering
The Chinese University of Hong Kong
http://www.cse.cuhk.edu.hk/~king

February 21, 2008
Motivations—The Why

- Significance of Chinese language
  - About 15% of world population use Chinese as native language
  - Half of the worlds published books are in Chinese
  - Volume of Chinese passages in the Internet is growing rapidly

- Lack of Chinese readability researches
  - English has a long history in related studies
  - Few outdated Chinese-related work available

- Representativeness of Chinese as a non-alphabetic language
  - One of the important non-alphabetic language
  - Can benefit similar research for other languages of this type

- Advance in Chinese language processing techniques
  - Chinese language processing is difficult
  - Advance in machine learning and statistical method improves the performance
Motivations–The Why

- **Significance of Chinese language**
  - About 15% of world population use Chinese as native language
  - Half of the world’s published books are in Chinese
  - Volume of Chinese passages in the Internet is growing rapidly

- **Lack of Chinese readability researches**
  - English has a long history in related studies
  - Few outdated Chinese-related work available

- **Representativeness of Chinese as a non-alphabetic language**
  - One of the important non-alphabetic language
  - Can benefit similar research for other languages of this type

- **Advance in Chinese language processing techniques**
  - Chinese language processing is difficult
  - Advance in machine learning and statistical method improves the performance
Motivations—The Why

- **Significance of Chinese language**
  - About 15% of world population use Chinese as native language
  - Half of the world's published books are in Chinese
  - Volume of Chinese passages in the Internet is growing rapidly

- **Lack of Chinese readability researches**
  - English has a long history in related studies
  - Few outdated Chinese-related work available

- **Representativeness of Chinese as a non-alphabetic language**
  - One of the important non-alphabetic language
  - Can benefit similar research for other languages of this type

- **Advance in Chinese language processing techniques**
  - Chinese language processing is difficult
  - Advance in machine learning and statistical method improves the performance
Motivations—The Why

- **Significance of Chinese language**
  - About 15% of world population use Chinese as native language
  - Half of the world's published books are in Chinese
  - Volume of Chinese passages in the Internet is growing rapidly

- **Lack of Chinese readability researches**
  - English has a long history in related studies
  - Few outdated Chinese-related work available

- **Representativeness of Chinese as a non-alphabetic language**
  - One of the important non-alphabetic language
  - Can benefit similar research for other languages of this type

- **Advance in Chinese language processing techniques**
  - Chinese language processing is difficult
  - Advance in machine learning and statistical method improves the performance
The Basic Problem and Processes

The What

- Predicting hard-to-measure dependent variable (readability score) $Y$
- Based on easily measured independent variables (language features) $X$

The How

- Data acquisition—To obtain a set of training data
- Factor analysis—To design a set of potential factors which may affect readability
- Text processing and Feature extraction—To extract statistical information based on the factors
- Regression analysis—To obtain a regression model for readability estimation
- Optimization—Feature selection—To select a subset of significant features
- Evaluation—To evaluate the proposed assessment
The Basic Problem and Processes

- The What
  - Predicting hard-to-measure dependent variable (readability score) $Y$
  - Based on easily measured independent variables (language features) $X$

- The How
  - Data acquisition—To obtain a set of training data
  - Factor analysis—To design a set of potential factors which may affect readability
  - Text processing and Feature extraction—To extract statistical information based on the factors
  - Regression analysis—To obtain a regression model for readability estimation
  - Optimization—Feature selection—To select a subset of significant features
  - Evaluation—To evaluate the proposed assessment
Figure: Overview of Chinese Readability Assessment
Factor Analysis at Different Levels

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。

Sub-character
Character
Word
Phrase
Sentence

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。
Sub-Character-Level Features

- Radical Analysis
  - Basic identifiable component
  - Use in dictionary look-up
  - Can help in grouping similar characters

- Two Major Factors
  1. Number of stroke of radical
  2. Radical familiarity

<table>
<thead>
<tr>
<th>Radical</th>
<th>No. of Strokes</th>
<th>Familiar or Unfamiliar</th>
</tr>
</thead>
<tbody>
<tr>
<td>一</td>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>戈</td>
<td>4</td>
<td>U</td>
</tr>
<tr>
<td>火</td>
<td>4</td>
<td>F</td>
</tr>
<tr>
<td>瓜</td>
<td>5</td>
<td>F</td>
</tr>
<tr>
<td>彩</td>
<td>10</td>
<td>U</td>
</tr>
<tr>
<td>龍</td>
<td>16</td>
<td>F</td>
</tr>
</tbody>
</table>

Figure: Sub-Character-Level Features
Character-Level Features I

- Number of stroke
  - Number of stroke without radical
  - Radical part may be easily recognized
  - Reader may only consider the remaining parts
- Geometry complexity
  - Symmetry
  - Structural

<table>
<thead>
<tr>
<th>Character</th>
<th>No. of Strokes</th>
<th>No. of Strokes without Radical</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>人</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>中</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>星</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>龙</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>數</td>
<td>28</td>
<td>21</td>
</tr>
</tbody>
</table>

Figure: Character-Level Features
Character-Level Features II

- Character familiarity
  - Well known, used frequently
  - Based on the frequency of usage

- Character grade
  - A list of recommended characters from curriculum

- Common characters in all grade level
  - Characters in textbooks for all grades
  - Another way to build familiar character list

<table>
<thead>
<tr>
<th>序号</th>
<th>级别</th>
<th>标准字形</th>
<th>部件</th>
<th>单字</th>
<th>常用度</th>
<th>难度</th>
<th>笔画</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>一</td>
<td>一</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>938</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>乙</td>
<td>部件</td>
<td>乙</td>
<td>1</td>
<td>1020</td>
<td>1303</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>人</td>
<td>部件</td>
<td>人</td>
<td>1</td>
<td>1072</td>
<td>145</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>人</td>
<td>部件</td>
<td>人</td>
<td>1</td>
<td>187</td>
<td>525</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>力</td>
<td>部件</td>
<td>力</td>
<td>1</td>
<td>97</td>
<td>679</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>刀</td>
<td>部件</td>
<td>刀</td>
<td>1</td>
<td>1179</td>
<td>715</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>了</td>
<td>部件</td>
<td>了</td>
<td>1</td>
<td>5</td>
<td>816</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>十</td>
<td>部件</td>
<td>十</td>
<td>1</td>
<td>54</td>
<td>845</td>
<td>2</td>
</tr>
<tr>
<td>1867</td>
<td>希</td>
<td>4</td>
<td>部件</td>
<td>木</td>
<td>1168</td>
<td>2211</td>
<td>8</td>
</tr>
<tr>
<td>1868</td>
<td>兎</td>
<td>4</td>
<td>部件</td>
<td>口</td>
<td>2280</td>
<td>2750</td>
<td>8</td>
</tr>
<tr>
<td>1869</td>
<td>定</td>
<td>4</td>
<td>部件</td>
<td>之</td>
<td>1008</td>
<td>759</td>
<td>9</td>
</tr>
<tr>
<td>1870</td>
<td>甚</td>
<td>4</td>
<td>部件</td>
<td>甘</td>
<td>649</td>
<td>1411</td>
<td>9</td>
</tr>
<tr>
<td>1871</td>
<td>奏</td>
<td>4</td>
<td>部件</td>
<td>夫</td>
<td>961</td>
<td>1678</td>
<td>9</td>
</tr>
<tr>
<td>1872</td>
<td>周</td>
<td>4</td>
<td>部件</td>
<td>十</td>
<td>2263</td>
<td>2560</td>
<td>9</td>
</tr>
<tr>
<td>1873</td>
<td>帝</td>
<td>4</td>
<td>部件</td>
<td>申</td>
<td>921</td>
<td>716</td>
<td>10</td>
</tr>
<tr>
<td>1874</td>
<td>翟</td>
<td>4</td>
<td>部件</td>
<td>田</td>
<td>2211</td>
<td>1346</td>
<td>10</td>
</tr>
<tr>
<td>2861</td>
<td>冀</td>
<td>6</td>
<td>单字</td>
<td>一</td>
<td>2900</td>
<td>2518</td>
<td>10</td>
</tr>
<tr>
<td>2862</td>
<td>耿</td>
<td>6</td>
<td>单字</td>
<td>人</td>
<td>3117</td>
<td>2695</td>
<td>10</td>
</tr>
<tr>
<td>2863</td>
<td>處</td>
<td>6</td>
<td>单字</td>
<td>衣</td>
<td>2202</td>
<td>2720</td>
<td>10</td>
</tr>
<tr>
<td>2864</td>
<td>柯</td>
<td>6</td>
<td>单字</td>
<td>心</td>
<td>3218</td>
<td>2823</td>
<td>10</td>
</tr>
<tr>
<td>2865</td>
<td>朝</td>
<td>6</td>
<td>单字</td>
<td>刀</td>
<td>2832</td>
<td>2839</td>
<td>10</td>
</tr>
<tr>
<td>2866</td>
<td>柳</td>
<td>6</td>
<td>单字</td>
<td>月</td>
<td>3020</td>
<td>2841</td>
<td>10</td>
</tr>
</tbody>
</table>
Word-Level Features

- **Word length**
  - In terms of number of strokes and characters

- **Word pattern**

- **Word familiarity**
  - Build based on the character familiarity list

- **Common word in all grade level**
  - Following the similar concept in character level

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

**Figure:** Word-Level Features
Phrase-Level Features

- **Phrase**
  - Incomplete sentence
  - Delimited by comma ,
  - Intermediate level between word and sentence

- **Two Major Factors**
  1. **Phrase length**
     - In terms of number of strokes, characters, words
  2. **Effect of idiom**
     - For example: 破釜沈舟— to cut off all means of retreat (a similar idiom in English: Burn ones bridges)
Sentence- and Structure-Level Features

- Sentence length
  - In terms of number of strokes, characters, words, and phrases.

- Sentence structure
  - Proportion of full sentence
    - A sentence with both subject and predicates
  - Number of word classes
    - Using Part-of-Speech tagger
    - Complexity of a sentence may increase if it consists of words with many different word classes
Sentence- and Structure-Level Features

- Sentence length
  - In terms of number of strokes, characters, words, and phrases.
- Sentence structure
  - Proportion of full sentence
    - A sentence with both subject and predicates
  - Number of word classes
    - Using Part-of-Speech tagger
    - Complexity of a sentence may increase if it consists of words with many different word classes
## Features Used Summary

<table>
<thead>
<tr>
<th>Index</th>
<th>Factor</th>
<th>Feature Names</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>Sub-character Level</strong></td>
</tr>
<tr>
<td>1-2.</td>
<td>$R_{stroke}$</td>
<td>Average and Standard deviation of number of radical strokes per Chinese character</td>
</tr>
<tr>
<td>3-4.</td>
<td>$R_{form}$</td>
<td>Proportion of familiar and unfamiliar radicals</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Character Level</strong></td>
</tr>
<tr>
<td>5-6.</td>
<td>$C_{stroke}$</td>
<td>Average and Standard deviation of number of strokes per Chinese character</td>
</tr>
<tr>
<td>7-8.</td>
<td>$C_{stroke,rad}$</td>
<td>Average and Standard deviation of number of strokes without radical per Chinese character</td>
</tr>
<tr>
<td>9-10.</td>
<td>$C_{form}$</td>
<td>Proportion of familiar and unfamiliar characters</td>
</tr>
<tr>
<td>11-15.</td>
<td>$C_{sym}$</td>
<td>Proportion of Symmetrical, Non-symmetrical, Vertical, Horizontal and Both Symmetrical characters</td>
</tr>
<tr>
<td>16-22.</td>
<td>$C_{struct}$</td>
<td>Proportion of characters belonging to Structure Category[A-G]</td>
</tr>
<tr>
<td>22-24.</td>
<td>$C_{grade}$</td>
<td>Average and Standard deviation of character grade</td>
</tr>
<tr>
<td>25-26.</td>
<td>$C_{common}$</td>
<td>Proportion of common and non-common characters</td>
</tr>
<tr>
<td>27-28.</td>
<td>$C_{freq}$</td>
<td>Average and Standard deviation of character frequency of occurrence</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Word Level</strong></td>
</tr>
<tr>
<td>29-30.</td>
<td>$W_{form}$</td>
<td>Proportion of familiar and unfamiliar words</td>
</tr>
<tr>
<td>31-32.</td>
<td>$W_{stroke}$</td>
<td>Average and Standard deviation of number of strokes per word</td>
</tr>
<tr>
<td>33-34.</td>
<td>$W_{length}$</td>
<td>Average and Standard deviation of number of characters per word</td>
</tr>
<tr>
<td>35-46.</td>
<td>$W_{pattern}$</td>
<td>Proportion of words belonging to Word Pattern Category[A-L]</td>
</tr>
<tr>
<td>47-48.</td>
<td>$W_{common}$</td>
<td>Proportion of common and non-common words</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Phrase Level</strong></td>
</tr>
<tr>
<td>49.</td>
<td>$P_{idom}$</td>
<td>Proportion of phrases containing idioms</td>
</tr>
<tr>
<td>50-55.</td>
<td>$P_{length}$</td>
<td>Average and Standard deviation of number of strokes, characters, words of phrase</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Sentence Level</strong></td>
</tr>
<tr>
<td>56-61.</td>
<td>$S_{length}$</td>
<td>Average and Standard deviation of number of strokes, characters, words per sentence</td>
</tr>
<tr>
<td>62.</td>
<td>$S_{full}$</td>
<td>Proportion of full sentences</td>
</tr>
<tr>
<td>63-64.</td>
<td>$S_{tag}$</td>
<td>Average and Standard deviation of number of distinct POS tags in sentence</td>
</tr>
</tbody>
</table>

**Figure:** Features Used Summary
Vast amount of valuable information in World Wide Web (WWW)

Getting the appropriate information is very important

Appropriateness of information:
  - Match the query
  - Match the users ability level

Existing search engines often overlook the second factor!
Vast amount of valuable information in World Wide Web (WWW)

Getting the **appropriate information** is very important

Appropriateness of information:
- Match the query
- Match the **users ability** level

**Existing search engines often overlook the second factor!**
Bilingual Web readability assessment
- English and Chinese
- High coverage of Internet community

Web page readability investigation
- Measure text-based readability
- Apply our proposed readability analysis and existing approaches

Web site readability investigation
- Extension of Web page readability
- An overall indicator of a site difficulty
Bilingual Web readability assessment
- English and Chinese
- High coverage of Internet community

Web page readability investigation
- Measure text-based readability
- Apply our proposed readability analysis and existing approaches

Web site readability investigation
- Extension of Web page readability
- An overall indicator of a site difficulty
Bilingual Web readability assessment

- English and Chinese
- High coverage of Internet community

Web page readability investigation

- Measure text-based readability
- Apply our proposed readability analysis and existing approaches

Web site readability investigation

- Extension of Web page readability
- An overall indicator of a site difficulty
Bilingual Web readability assessment
- English and Chinese
- High coverage of Internet community

Web page readability investigation
- Measure text-based readability
- Apply our proposed readability analysis and existing approaches

Web site readability investigation
- Extension of Web page readability
- An overall indicator of a site difficulty
Site readability is an indicator of overall difficulty of a site.

We model a site as a tree, with the index page as the root.

Terms Used:
- **Web Site, $$s$$** — A group of Web pages under the same domain.
- **Root Page, $$p_0$$** — A user specified page where crawling of the site starts, usually it is the index page.
- **Page Level, $$lv$$** of a page $$p$$ — The minimum number of traversals reaching $$p$$ starting from $$p_0$$ through hyperlinks.

**Definition** Page Readability (Comprehensive Difficulty) of a Web page $$p$$ in a Web site, denoted by $$r_p$$ is defined as:

$$
r_p = \begin{cases} 
-84.6X_{E_1} - 1.015X_{E_2} + 206.835 & \text{if } \text{lang}(p) = 0, \\
2 \times \{13.90963 + 1.54461X_{C_3} + 39.01497X_{C_4} - 2.52206X_{C_5} - 0.29890X_{C_6} + 0.36492X_{C_9} + 0.99363X_{C_8} - 1.64671X_{C_1}\} & \text{if } \text{lang}(p) = 1,
\end{cases}
$$

where $$X_{E_1}$$ and $$X_{C_i}$$ are the factors, $$\text{lang} : \mathbb{P} \rightarrow \{0, 1\}$$ is a mapping from page to its language:
- $$X_{E_1}$$: Average number of syllables per word;
- $$X_{E_2}$$: Average sentence length;
- $$X_{C_3}$$: Proportion of full sentence;
- $$X_{C_4}$$: Proportion of words in Chinese basic word list;
- $$X_{C_5}$$: Average number of stroke of characters;
- $$X_{C_9}$$: Number of characters with stroke number = 5 (in a sample of 100 characters);
- $$X_{C_6}$$: Number of characters with stroke number = 12 (in a sample of 100 characters);
- $$X_{C_8}$$: Number of characters with stroke number = 22 (in a sample of 100 characters);
Site readability is an indicator of overall difficulty of a site

We model a site as a tree, with the index page as the root

Terms Used

- **Web Site**, s—A group of Web pages under the same domain
- **Root Page**, \( p_0 \)—A user specified page where crawling of the site starts, usually it is the index page
- **Page Level**, \( lv \) of a page \( p \)—The minimum number of traversals reaching \( p \) starting from \( p_0 \) through hyperlinks

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

\[
r_p = \begin{cases} 
-84.6X_{E_1} - 1.015X_{E_2} + 206.835 & \text{if } \text{lang}(p) = 0, \\
2 \times \left(13.90963 + 1.54461X_{C_1} + 39.01497X_{C_2} - 2.52206X_{C_3} - 0.29809X_{C_4} + 0.36192X_{C_5} + 0.99363X_{C_6} - 1.64671X_{C_7}\right) & \text{if } \text{lang}(p) = 1,
\end{cases}
\]

where \( X_{E_1} \), \( X_{E_2} \), \( X_{C_1} \), \( X_{C_2} \), \( X_{C_3} \), \( X_{C_4} \), \( X_{C_5} \), \( X_{C_6} \), and \( X_{C_7} \) are the factors, \( \text{lang} : P \rightarrow \{0, 1\} \) 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);

**Figure:** Definition of Page Readability
Exact-Level Assessment

- Average readability of pages at a particular level
- By using this, Web authors can decide how should the readability changes with level
- For example, designing Online Teaching Site

Figure: Exact-Level
In-Level Assessment

- Average readability of pages starting from root up to pages at a particular level
- By using this, users can get a general idea of the difficulty of a site
- This is an overall indicator of a site difficulty

Figure: In-Level
Out-Level Assessment

- Average readability of pages starting from one level upper of specified, up to the maximum available level
- An indicator of remaining pages difficulties after browsing a site
- Users can decide whether to stay at the site

**Figure: Out-Level**
Experiments

- To investigate Web page readability level
  - Comprehensive difficulty
  - Grade level
- To evaluate the proposed Web site readability

<table>
<thead>
<tr>
<th>Machine 1</th>
<th>Machine 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td><strong>Intel Pentium 4 3.2 GHz</strong></td>
</tr>
<tr>
<td><strong>RAM</strong></td>
<td>4.0 GB</td>
</tr>
<tr>
<td><strong>Operating System</strong></td>
<td><strong>RedHat Linux Fedora Core 4</strong></td>
</tr>
<tr>
<td><strong>Harddisk Size</strong></td>
<td>300GB</td>
</tr>
<tr>
<td><strong>Programming Language</strong></td>
<td><strong>Java SDK 1.5.06 and Python 2.4.2</strong></td>
</tr>
<tr>
<td><strong>Task</strong></td>
<td><strong>Text Processing, Readability Evaluation</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>Max Level</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSE</td>
<td>10</td>
<td><a href="http://www.cse.cuhk.edu.hk">http://www.cse.cuhk.edu.hk</a></td>
</tr>
<tr>
<td>CUHK</td>
<td>5</td>
<td><a href="http://www.cuhk.edu.hk">http://www.cuhk.edu.hk</a></td>
</tr>
<tr>
<td>HKGOV</td>
<td>7</td>
<td><a href="http://www.gov.hk">http://www.gov.hk</a></td>
</tr>
<tr>
<td>XANGA</td>
<td>5</td>
<td><a href="http://www.xanga.com">http://www.xanga.com</a></td>
</tr>
</tbody>
</table>

**Figure:** Testing Environment and Tested Web Sites
Observation 1

Figure: English and Chinese Web Page Readability Distributions with Range Scores
Observation 2

Figure: English and Chinese Web Page Readability Distributions with Grade Scores
Observation 3

Figure: Web Site Readability of CSE with Levels
Summary & Future Directions

Summary

- Chinese readability analysis and assessment
  - Document forensics
  - Plagiarism detection
  - Teaching and education
- Web page/site readability analysis and assessment
  - Web site analysis
  - Web site management

Future Directions

- Semantic-level, conceptual, high-level analyses
- Writing style analysis, e.g., sms text, poems, technical documents, etc.
- Corpus-based statistical methods
- Temporal readability assessment for tracking
Summary & Future Directions

- **Summary**
  - Chinese readability analysis and assessment
    - Document forensics
    - Plagiarism detection
    - Teaching and education
  - Web page/site readability analysis and assessment
    - Web site analysis
    - Web site management

- **Future Directions**
  - Semantic-level, conceptual, high-level analyses
  - Writing style analysis, e.g., sms text, poems, technical documents, etc.
  - Corpus-based statistical methods
  - Temporal readability assessment for tracking
On-Going Research Projects

Web Ranking/Classification Related
- **MatchSim**: link-based web page similarity measurements (WI’07)
- **Diffusion rank**: ranking web pages based on heat diffusion equations (SIGIR’07)
- Web text classification (WWW’07)

Social Networks and Web/Opinion Mining
- **Recommender system**: accurate recommendation based on sparse matrix (SIGIR’07)
- **Feature-opinion association** for sentiment analysis: ”The quality is low” vs. ”The noise level is low”
- **User credibility analysis**: ”the enemy of my enemy is my friend!” phenomena, etc.
- **Expert finding** in DBLP bibliography database

Machine Learning
- Transductive Support Vector Machine (NIPS’07)
- Global and local learning (ICML’04, JMLR’04)
On-Going Research Projects

- **Web Ranking/Classification Related**
  - **MatchSim**: link-based web page similarity measurements (WI’07)
  - **Diffusion rank**: ranking web pages based on heat diffusion equations (SIGIR’07)
  - Web text classification (WWW’07)

- **Social Networks and Web/Opinion Mining**
  - **Recommender system**: accurate recommendation based on sparse matrix (SIGIR’07)
  - **Feature-opinion association** for sentiment analysis: "The quality is low" vs. "The noise level is low"
  - **User credibility analysis**: "the enemy of my enemy is my friend!" phenomena, etc.
  - **Expert finding** in DBLP bibliography database

- **Machine Learning**
  - Transductive Support Vector Machine (NIPS’07)
  - Global and local learning (ICML’04, JMLR’04)
On-Going Research Projects

- **Web Ranking/Classification Related**
  - **MatchSim**: link-based web page similarity measurements (WI’07)
  - **Diffusion rank**: ranking web pages based on heat diffusion equations (SIGIR’07)
  - Web text classification (WWW’07)

- **Social Networks and Web/Opinion Mining**
  - **Recommender system**: accurate recommendation based on sparse matrix (SIGIR’07)
  - **Feature-opinion association** for sentiment analysis: ”The quality is low” vs. ”The noise level is low”
  - **User credibility analysis**: ”the enemy of my enemy is my friend!” phenomena, etc.
  - **Expert finding** in DBLP bibliography database

- **Machine Learning**
  - Transductive Support Vector Machine (NIPS’07)
  - Global and local learning (ICML’04, JMLR’04)
On-Going Research Projects

- **Web Ranking/Classification Related**
  - **MatchSim**: link-based web page similarity measurements (WI’07)
  - **Diffusion rank**: ranking web pages based on heat diffusion equations (SIGIR’07)
  - Web text classification (WWW’07)

- **Social Networks and Web/Opinion Mining**
  - **Recommender system**: accurate recommendation based on sparse matrix (SIGIR’07)
  - **Feature-opinion association** for sentiment analysis: "The quality is low" vs. "The noise level is low"
  - **User credibility analysis**: "the enemy of my enemy is my friend!" phenomena, etc.
  - **Expert finding** in DBLP bibliography database

- **Machine Learning**
  - Transductive Support Vector Machine (NIPS’07)
  - Global and local learning (ICML’04, JMLR’04)
Acknowledgments

The Chinese University Plagiarism IDentification Engine (CUPIDE) Project
- Prof. Jimmy Lee
- Patrick Lau, Research Assistant
- Wei Wei, Ph.D.
- Final year students
There will be a poster with demo tomorrow!

The Social Web team in the Intelligent Data Engineering And Learning (IDEAL) Lab
- Prof. Jimmy Lee
- Prof. Michael Lyu
- Thomas K.T. Chan
- Hongbo Deng, Ph.D.
- Haixuan Yang (Postdoc Fellow)
- Zhenjiang Lin, Ph.D.
- Hoa Ma, Ph.D.
- Wei Wei, Ph.D.
- Kaizhu Huang (Postdoc Fellow)
- Haiqin Yang, Ph.D.
- Zenglin Xu, Ph.D.
Acknowledgments

The Chinese University Plagiarism IDentification Engine (CUPIDE) Project
- **Prof. Jimmy Lee**
- **Patrick Lau**, Research Assistant
- Wei Wei, Ph.D.
- Final year students

There will be a poster with demo tomorrow!

The Social Web team in the Intelligent Data Engineering And Learning (IDEAL) Lab
- **Prof. Jimmy Lee**
- Prof. Michael Lyu
- Thomas K.T. Chan
- Hongbo Deng, Ph.D.
- Haixuan Yang (Postdoc Fellow)
- Zhenjiang Lin, Ph.D.
- Hoa Ma, Ph.D.
- Wei Wei, Ph.D.
- Kaizhu Huang (Postdoc Fellow)
- Haiqin Yang, Ph.D.
- Zenglin Xu, Ph.D.
Acknowledgments

The Chinese University Plagiarism IDentification Engine (CUPIDE) Project

- Prof. Jimmy Lee
- Patrick Lau, Research Assistant
- Wei Wei, Ph.D.
- Final year students

There will be a poster with demo tomorrow!

The Social Web team in the Intelligent Data Engineering And Learning (IDEAL) Lab

- Prof. Jimmy Lee
- Prof. Michael Lyu
- Thomas K.T. Chan
- Hongbo Deng, Ph.D.
- Haixuan Yang (Postdoc Fellow)
- Zhenjiang Lin, Ph.D.
- Hoa Ma, Ph.D.
- Wei Wei, Ph.D.
- Kaizhu Huang (Postdoc Fellow)
- Haiqin Yang, Ph.D.
- Zenglin Xu, Ph.D.