SCRC Test Collection for Evaluation of Chinese OCR Text Retrieval

Yuen-Hsien Tseng
Fu Jen Catholic University,
Taipei, Taiwan, R.O.C.
tseng@blue.lins.fju.edu.tw
Retrospective Information Access Approaches

• Full-text search based on manually re-keying the text
  – Prohibitively expensive at large scale

• Search based on bibliographic metadata
  – May be difficult to adequately describe the materials.

• Full text based on Optical Character Recognition (OCR)
  – Inexpensive and relatively rapid
  – Sensitive to OCR accuracy
  – Effectiveness evaluation is needed to justify the cost advantage, **which requires a test collection.**
Socio-Cultural Research Center (SCRC) Collection

- 800,000 newspaper clippings from 1950-1976
  - Scanned over 300,000 at 300 dpi
- 30 China, Hong Kong, and Taiwan news agencies
  - Mostly simplified Chinese, some traditional Chinese
- Selected images focus on diplomatic and military activities

---

The Chinese government's statement on the Sino-Indian border issue.

- The Chinese government has issued a statement on the Sino-Indian border issue. The statement points out that the Chinese government has always adhered to the principle of non-interference in internal affairs, and will continue to uphold this principle.

---

The practices and strategies of the Chinese government in dealing with international disputes.

- In terms of international disputes, the Chinese government has always adhered to the principle of non-interference in internal affairs, and will continue to uphold this principle. This principle has not only been practiced in the Sino-Indian border issue, but also in other international disputes.

---

The future direction of the Chinese government in international relations.

- The Chinese government will continue to adhere to the principle of non-interference in internal affairs in international relations. This principle will guide the Chinese government in future disputes.
SCRC Test Collection for Chinese OCR Text Retrieval Evaluation

• Document Set
  – 8438 images and OCR texts

• Query Topics
  – 30 topics in TREC-like format

• Relevance Judgment
  – Exhaustive judgment by 3 persons
Document Preparation

- Selected 11,108 scanned document images
- OCR yielded 8,438 valid documents
  - OCR software package: Presto! OCR Pro, Big-5
  - Average valid document had a 69% system-reported “recognition rate”
    - Computed on a sample of 1,300 documents
- Second version prepared using Big-5 to GB conversion
Topic Preparation

• Based on contemporaneous Chinese journal articles
  – From 100 paper titles, 30 were selected and rewritten as Chinese topics

• Made English translations for cross-language experiments
  – Translated by native speakers of Chinese

<top>
<num> 12
<title>Anti-Chinese Movements
<description>
  Activities related to the anti-Chinese movements in Indonesia
<narrative>
  Articles must deal with activities related to the anti-Chinese movement in Indonesia; case reports or articles dealing with PRC's criticism of the Anti-Chinese movement will be considered partly relevant.
</top>
Relevance Judgments

- Exhaustive tri-state relevance judgments
  - Irrelevant (=0), partially relevant (=1), fully relevant (=2)
- Every topic-document pair judged by 3 assessors
  - 2 majored in history, 1 majored in library science
  - Averaged 4 minutes per document image (for all 30 topics)
- Sum of the judgments provides a final estimate
  - 0=not relevant, 1…5=partially relevant, 6=fully relevant
  - Threshold as desired to reflect the intended application
    - In our experiments, any score > 0 is treated as “relevant”

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Doc. ID</th>
<th>1st Assessor</th>
<th>2nd Assessor</th>
<th>3rd Assessor</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0053487</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>01</td>
<td>0053489</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>0054425</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>02</td>
<td>0054452</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How to Get the Test Collection

Please visit
for details