Document Image Retrieval Techniques for Chinese

Yuen-Hsien Tseng
Dept. of Library and Information Science
Fu Jen Catholic University
510 Chung Cheng Road, HsinChuang,
Taipei, Taiwan, R.O.C. 242

Douglas W. Oard
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD
20742, USA

Abstract:

In this paper we present experiment results for retrieval from a collection of scanned article clippings from Chinese newspapers. The test collection consists of 8,438 articles from China, Taiwan and Hong Kong in a mix of traditional and simplified Chinese. A commercial OCR system was used to produce errorful text. Exhaustive relevance assessment was performed over the entire collection for 30 Chinese queries by multiple judges. Indexing a combination of unigrams and overlapping bigrams was found to outperform overlapping bigram indexing alone, and byte length normalization was found to outperform cosine normalization. No improvement resulted from the addition of query expansion using blind relevance feedback on the same collection.

1. Introduction

The advent of the World Wide Web has made access to digital information easier than ever before. Many information providers have therefore been inspired to digitize existing paper materials to enable access through networked information services. A number of approaches for this purpose are possible, including: (1) manual re-keying of the text; (2) creation of metadata; (3) creation of document images through scanning; and (4) layout analysis and optical character recognition (OCR) of document images [1, 2]. Many current systems have combined approaches (2) and (3), using metadata to support search and document images to support electronic document delivery. Although manual creation of metadata can be much more economical than manually re-keying the full text for each document, it still involves considerable cost in time and human effort. Furthermore, manually produced metadata can only support searches based on information needs that could be anticipated when the metadata was created. Combining approaches (3) and (4) offers complementary strengths, using OCR to produce searchable (although sometime erroneous) full-text representations, and document images as a basis for electronic document delivery. When sufficiently accurate OCR is possible, this can provide relatively inexpensive support for searches based on the full vocabulary used by document authors. Combining approaches (2), (3) and (4) can provide even richer support for information access. For example, full-text search can be used to find documents that address previously unforeseen topics, and metadata can be used to limit the search to a range of creation dates that is appropriate to the user's task. In this paper, we focus on the use of approach (4) to support the process of searching a collection of printed Chinese documents.

Recently there has been considerable interest in the application of approach (4) to historical newspaper materials. The Pathfinder Library System in Grand Junction, Colorado, has started a Library Services and Technology Act (LSTA) project that will explore digitization, indexing, copyright, and other issues relating to providing access to historical Colorado newspapers over the Internet to students, researchers, and other potential users [3]. This "proof-of-concept" pilot project intends to scan the daily Aspen Times for the year 1887, perform OCR on the resulting document images to generate text files for full-text indexing, create a searchable database of content indexes with links to the newspaper images and text files, and develop a prototype Web site for the project.

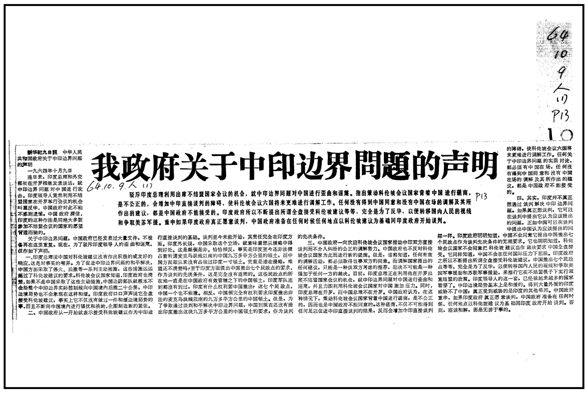
Research in languages other than English clearly indicates that effective support for information access using approach (4) requires some degree of language-specific processing. For example, researchers working with historical Greek newspapers in a project at the Lambrakis Press have developed techniques that account for changes in the Greek languages over time [4]. The experiments reported in this paper are motivated by a similar project at the Socio-Cultural Research Center (SCRC) at Fu Jen Catholic University in Taiwan, which has scanned 600,000 of the 800,000 newspaper clippings that they have collected from Mainland China, Hong Kong, and Taiwan over the past 50 years with the ultimate goal of providing access to the collection over the Internet. In this case, each clipping has been separately scanned to create 300 dot per inch (dpi) TIFF images, so segmentation is less of an issue. But the inventory of Chinese characters is far larger than for Western languages, and the lack of explicit word boundaries makes both OCR and retrieval more challenging.

Information retrieval is an experimental science in which test collections provide the basis for tuning systems for optimum retrieval effectiveness. For the Text Retrieval Conference, OCR results were simulated by applying a confusion model trained on actual OCR output to an existing information retrieval test collection. While this approach can provide some degree of insight into the sensitivity of a technique to OCR errors, evaluations based on actual scanned document images are generally preferred by OCR researchers because OCR accuracy depends on a wide array of situations (e.g., bleedthrough in two-sided printing or spurious marks on historical materials) that might not be modeled with sufficient fidelity. For this reason, the Fritz Kutter-Fonds Foundation in Zurich sponsored an evaluation of automatic cataloguing and free text searching in 1999 that was based on 500 books in four European languages that were published between 1770 and 1970 [5, 6, 7]. Book pages on which the cataloguing was to be based were scanned to produce document images that were then converted to text and Xerox Xdoc layout description files using layout analysis and OCR. This paper complements that work describing what we believe is the first information retrieval test collection for an Asian language based on scanned documents images. Experiment results obtained using the collection are presented for a variety of retrieval techniques

The remainder of the paper is organized as follows. The next section introduces the test collection, and Section 3 then briefly surveys previous work on document image retrieval. Our techniques and experiment results are then presented in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper.

2. The SCRC Chinese Document Image Retrieval Test Collection

An information retrieval test collection contains a set of documents, a set of topic descriptions from which queries can be constructed, and a set of relevance judgments that identifies the relevant documents for each topic. From the SCRC news clippings, we selected a 11,108 document images to create the test collection. The selected stories focus mostly on diplomatic and military developments in Mainland China between 1950 and 1976. Stories from 30 news agencies are represented in the collection, from Mainland China (where simplified Chinese is used), Hong Kong, and Taiwan (where traditional Chinese is used). Most documents are in simplified Chinese, but some are traditional Chinese. A sample image is shown in Figure 1.



64
10
9
人
17
P13
10

我政府关于中印边界問題的声明

64.10.9.人 (17)

驳斥印度总理利用出席不结盟国家会议的机会，就中印边界问题对中国进行歪曲和诬蔑。指出策动科伦坡会议国家背着中国进行磋商，是不公正的，会增加中印直接谈判的障碍，使科伦坡会议六国将来更难进行调解工作。任何没有得到中国同意和没有中国在场的调解及其所作出的建议，都是中国政府不能接受的。印度政府所以不断提出所谓全盘接受科伦坡建议等等，完全是为了反华，以便转移国内人民的视线和争取美苏军援。重申如果印度政府真正愿意谈判，中国政府准备在任何时候任何地点以科伦坡建议为基础同印度政府开始谈判。

P13

新华社九日訊　中华人民共和国政府关于中印边界问题的声明

一九六四年十月九日

连日来，印度总理和外交部长在开罗相继发表谈话，就中印边界问题对中国进行攻击。印度领导人竟然利用不结盟国家在开罗举行会议的机会叫嚣反华，中国政府对此不能不感到遗憾。中国政府深信，印度的这种作法是同绝大多数参加不结盟会议的国家的愿望背道而驰的。

关于中印边界问题，中国政府已经发表过大量文件，不准备再在这里重复。现在，为了驳斥印度领导人的歪曲和诬蔑，仅作如下声明：

一、印度总理说中国对科伦坡建议没有作出积极的或友好的响应，这是对事实的嘲弄。为了促进中印边界问题的和平解决，中国方面采取了停火、后撤等一系列主动措施，这些措施远远超过了科伦坡建议的要求。科伦坡会议国家知道，印度政府也清楚，如果不是中国采取了这些主动措施，中国边防部队就根本不会沿整个中印边界实际控制线向中国境内后撤二十公里，中印边境局势也不会象现在这样和缓。印度政府口口声声说它全盘接受科伦坡建议，事实上它不仅没有做过一件和缓边境局势的事，而且不断向中国境内进行骚扰和挑衅，企图制造新的紧张。

二、中国政府从一开始就表示接受科伦坡建议作为中印进行直接谈判的基础。谈判至今未能开始，其责任完全在印度方面。印度外长说，中国采取这个立场，就意味着想从侵略中得到好处。这是颠倒是非。恰恰相反，事实是印度至今还非法侵占着所谓麦克马洪线以南的中国九万多平方公里的领土，而中国方面却从来没有占领过印度一寸领土。究竟是谁在侵略，难道还不清楚吗？至于印度方面提出中国撤出七个民政点的要求，作为谈判的先决条件，这是完全没有道理的。这些民政点的所在地一直是在中国政府有效管辖之下的中国领土，印度军队连到都没有到过，印度有什么权利要中国撤出？这七个民政点，中国一个也不能撤。相反，中国倒完全有权利要求印度撤出非法的麦克马洪线以南的九万多平方公里的中国领土。但是，为了争取通过谈判和平解决中印边界问题，中国到现在还没有提出印度撤出这块九万多平方公里的中国领土的要求，作为谈判的先决条件。

三、中国政府一向欢迎科伦坡会议国家推动中印双方直接谈判而不介入纠纷的公正的调解努力。中国政府也不反对科伦坡会议国家为此而进行新的磋商。但是，谁都知道，任何有效的调解活动，都必须取得当事双方的同意。而调解国家提出的任何建议，只能是一种供双方考虑的推荐，而决不可能是一种强加于任何一方的裁决。目前，印度总理正在利用他在开罗出席不结盟国家会议的机会，就中印边界问题对中国进行歪曲和诬蔑，并且力图利用科伦坡会议国家对中国施加压力。同时，印度总理在开罗，而中国总理不在开罗。中国政府认为，在这种情况下，策动科伦坡会议国家背着中国进行磋商，是不公正的，因而也是中国政府不能同意的。这种磋商，不仅不可能得到任何足以促进中印直接谈判的结果，反而会增加中印直接谈判的障碍，使科伦坡会议六国将来更难进行调解工作。任何关于中印边界问题的实质讨论，都必须有中国在场，任何没有得到中国同意和没有中国在场的调解及其所作出的建议，都是中国政府不能接受的。

四、其实，印度并不眞正想通过谈判解决中印边界问题。如果眞正想谈判，它可以在谈判中提出它认为应该提出的问题，正如中国可以在谈判中提出中国认为应该提出的问题一样。印度政府明明知道，中国不会同意它提出中国撤出七个民政点作为谈判先决条件的无理要求。它也明明知道，科伦坡会议国家不会同意把科伦坡建议当作裁决要求中国全盘接受。它同样知道，中国不会在任何国际压力下屈服。印度政府之所以不断提出所谓全盘接受科伦坡建议，中国撤出七个民政点等等，完全是为了反华，以便转移国内人民的视线和争取美国军事援助和苏联军事援助，来推行它在不结盟幌子下实行双重结盟的政策。印度领导人的这一套，已经被越来越多的国家看穿了。中印边境局势基本上是和缓的。得到大量外援的印度威胁不了中国，眞正受到威胁的是印度的其他邻邦。中国政府重申，如果印度政府眞正愿意谈判，中国政府准备在任何时候、任何地点以科伦坡建议为基础同印度政府开始谈判。否则，空谈和解，那是无济于事的。

Figure 1. A sample newspaper clipping image.

The 11,108 images were converted to text by a commercial OCR system, yielding 8,438 valid text documents. Others are rejected by the OCR system due to low image quality or other limitations to the recognition capability of the system. To get an idea of the OCR quality, we tabulated the system-reported character recognition rates for a 1,300 document sample. We found that the average recognition success rate was only 0.69 (with a standard deviation is 0.124). The distribution of the recognition rates is shown in Figure 2. Compared to the figures claimed by the OCR vendor, where a recognition rate of over 0.9 or even 0.95can be expected for ordinary printed materials, the low OCR rate in this case might be due to low print quality of the aging clippings. Although these statistics represent system-generated estimates rather than character accuracy based on ground truth data, they do provide an initial basis for characterizing the difficulty of the recognition challenge for these materials.

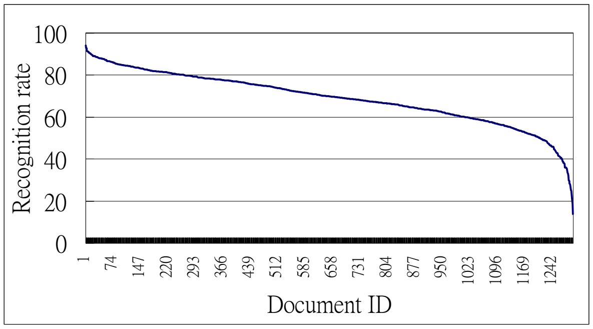


Figure 2. The recognition rate distribution, sorted by descending order.

As to the topic set, it would be best to assemble it from real searchers' information needs. However, SCRC's research library does not record the nature of users' requests for reference assistance. We therefore gathered possible query topics from various journal articles published at about the same time as the news stories. This was based on our belief that if some issue was being written about in a journal article, there may have some information needs related to that issue. From 100 paper titles, 30 were selected and rewritten as formalized information need statements (topics) in Chinese using a format similar to that of the Text Retrieval Conference (TREC) topic descriptions. These topics have also been translated into English by SCRC's social science researchers to support possible cross-lingual (English-Chinese) retrieval experiments in the future. Figure 3 shows an example of an English translation of a Chinese topic.

<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>

Figure 3. A sample topic in English.

The degree of relevance for each document with respect to each topic was judged by three assessors (two of whom majored in history, with the other having majored in library science). Three levels of relevance could be specified. Complete (i.e., exhaustive) relevance assessments were performed, with each document image (not the possible erroneous OCR text) being examined for relevance to any topic. We used exhaustive assessment because the alternative, a sampling strategy known as "pooled relevance assessment" would have required the participation of multiple teams using different techniques in a coordinated evaluation. A level of 0 was assigned to irrelevant documents, 1 for partially relevant documents, and 2 for fully relevant documents. The ability to specify the degree of relevance may allow assessors to express relationships in a more natural way than binary relevance judgments would. Each assessor required an average of 4 minutes to judge the relevance of one document to 30 queries, so a total of (4 x 8438 x 3) = 101,256 minutes was invested over two months to perform the (8438 x 30 x 3) = 759,420 relevance judgments. The relevance levels for each topic-document pair were then summed over the 3 assessors to produce a value between 0 and 6 that could serve as the ground truth degree of relevance for a document. Retrieval performance is often expressed in terms of precision and recall, where precision is the ratio of relevant documents in the retrieved set to the total size of the retrieved set, and recall is the ratio of relevant documents retrieved over all relevant documents in the collection. Such measures require binary-valued relevance judgments, which can be produced by applying whatever threshold to this value that the experimenter believes would best represent the retrieval task that they seek to model. In the experiments reported below, we treat a document as relevant for purposes of evaluation if it has a non-zero value, and irrelevant otherwise.

The 8,438 images were converted to text in BIG-5 code by the OCR software mentioned above. For the convenience of researchers with tools optimized for the GB code that is in common use in Mainland China, standardized GB versions of the recognized text were produced using the "ConvertZ" freeware utility (http://www.speednet.net/~shing/). In the experiments reported below, we used only the GB representation.

3. Previous Work

A number of researchers have done studies of automatic retrieval using degraded text produced by (or modeling) OCR. In this section we summarize the results with respect to the faceted classification of approaches summarized in Table 1. A fuller description of each study can be found in [1]. Although no study that we are aware of has yet explored retrieval based on Chinese OCR results, we found that the experience of others in working with degraded text shed considerable light on the directions that we could take in our work.

Taghva et al. did a series of studies to identify the effects of OCR errors on text retrieval using different models. In [8], they used a Boolean logic retrieval system, finding that the effect of OCR errors was insignificant for a small collection or relatively long documents (38 pages per document). The same group did two other studies [9, 10], one using the InQuery system [11], which uses a probabilistic retrieval model, the other using the SMART system [12], which uses a vector space retrieval model. Unlike the Boolean model, both of these retrieval produce a ranked list in which the documents most likely to be relevant to a query are listed first. Results obtained using both the probabilistic and the vector space retrieval models showed that although no statistically significant differences were found between the mean average precision of the OCR and the manually corrected collection the results for individual queries can be greatly affected. They attributed this effect to unreliable term frequency statistics derived from the noisy OCR text. The term frequency statistics greatly affect the term weighting measure, on which the probabilistic retrieval model is based to calculate the query-document relationship. Additional findings with the vector space retrieval model was that *cosine normalization* had a negative effect when compared to the unnormalized inner product, and that *relevance feedback* could not be used to compensate for OCR errors caused by badly degraded documents. Relevance feedback is an automatic process that uses information derived from known relevant and non-relevant documents to reformulate queries. It has been consistently shown by various experiments that relevance feedback is an effective approach to improve performance for ordinary clean text [13, 14], so this was a surprising result.

Indexing method	Word-based indexing. N-gram indexing (fixed or variable length).
Retrieval model	Boolean logic positional model. Vector space model. Probabilistic model. Approximate string matching.
Test collection	Direct OCR output. Simulated OCR output.
Evaluation measure	Percentage of documents returned from the OCR set Mean Average precision (over recall levels and topics). Document ranking fluctuation (mean, variance, or correlation).
Performance comparison	Compared with original clean text. Compared with manually corrected text. Compared among different levels of simulated OCR error.
Specific strategy	No strategy: rely on information redundancy. Long query: sort of document relevance feedback. Query expansion: term expansion based on the original query. Preprocessing: automatic correction of OCR errors. Interaction with users.

Table 1. Faceted classification of OCR text retrieval approaches.

The experiments done by Taghva et al. showed that some widely used weighting schemes that are known to be effective for ordinary text might lead to more unstable results for OCR degraded text. Singhal et al. [15] analyzed this phenomenon closely using the SMART system and a simulation of the expected degradation from OCR in the large (742,202 documents) TREC collection. In their research, Singhal et al., found that an erroneous term like "systom" that might be produced by a recognition error could have a large inverse document frequency (*idf*) value, thus incorrectly affecting the weights of the index terms if a mutually dependent normalization like the cosine is used. They found that instead using a byte size normalization scheme could mitigate this source of error. For a document, their *byte size normalization factor* is computed as:

$$(\mathit{byte\ size})^{0.375}$$

Singhal et al. found that bite size normalization produced a higher mean average precision and was more robust across topics than cosine normalization for both OCR output and ordinary text.

Another attempt to seek robust weighting methods was made by Mittendorf et al. [16]. They used expected term frequency (*tf*) and expected *idf* for term weighting under a probabilistic model instead of the direct term frequency statistics from the OCR collection. Eight hundred library catalog cards were scanned, OCR was performed with 67% word accuracy, and the result was split into training and test sets of equal size. Manual re-entry of the same data was done to derive the actual *tf* and *idf*, and the training set was used to estimate this actual *tf* and *idf* based on observed values. This resulted in a 23% relative improvement in the average number of relevant documents found in the first position of the ranked list in a known-item retrieval evaluation. In this case the documents were quite short (averaging 23 terms). Although this research suggests that parameter estimation can be helpful, the required training documents and their associated ground truth may not be available for other cases.

Lopresti and Zhou [17] examined the effects of varying the degree of degradation on the effectiveness of Boolean, fuzzy Boolean, vector space, extended Boolean, fuzzy extended Boolean, proximity Boolean, and fuzzy proximity Boolean retrieval models. One thousand news articles were collected from the Internet and corrupted to varying degrees using a model of OCR effects. An analysis of rank correlation coefficients within a technique for varying degrees of degradation showed that fuzzy retrieval models based on approximate string matching appear to be generally more robust than their traditional counterparts. The approximate string matching techniques used in the study were quite inefficient, however, raising questions about the practicality of the technique in large-scale applications.

Character n-grams offer a more efficient way of achieving some degree of approximate string matching. Pearce and Nicholas [18] applied fixed-length (overlapping) n-gram indexing to index both OCR results and ordinary text, exploring alternative normalization functions. Their most robust normalization function, which they called *similarity link*, was unfortunately computationally intractable. Harding et al. experimented with simultaneous use of multiple n-gram lengths for OCR-based retrieval [19], finding that this improved retrieval performance over word-based indexing at 10% or greater OCR degradation. N-gram indices are, however, larger than word indices, and their efforts to find more efficient n-gram indexing produced adverse effects on retrieval effectiveness

From this survey of prior work, we can observe fairly clear agreement on the following factors:

- OCR errors have relatively little negative effect on retrieval effectiveness for long documents. Redundancy is often beneficial in information retrieval applications, and longer documents naturally offer more scope for redundancy.
- Byte length normalization results in better retrieval effectiveness than cosine normalization when using the vector space model.
- N-gram indexing can result in better retrieval effectiveness than word-based indexing if OCR errors are relatively common.
- There can be a tradeoff between retrieval effectiveness and retrieval efficiency when developing techniques to search in the presence of OCR errors.

In the next section we apply these observations to the design of techniques for OCR-based retrieval of Chinese document images.

4. Experiment Design

There are thousands of Chinese characters, about 2,000 to 3,000 of which are in common use. Chinese words vary in length from a single character (almost every Chinese character has meaning as a word on its own) to nine or more characters, with an average of about two characters (for contrast, the average length of an English word is about 5 characters). In many cases, longer "words" are actually better thought of as compound terms, since some speakers of the language could segment them into shorter words and recover the same meaning. From the perspective of information retrieval, Chinese differs from English in two important ways. First, Chinese retrieval is more challenging because written Chinese includes no delimiters between words and available automatic segmentation techniques are imperfect. Second, Chinese retrieval is made somewhat easier by the relative absence of morphological variants, thus obviating the need for stemming.

Comparative studies have established that n-gram indexing (usually with n=2, for bigrams) works about as well for Chinese retrieval as word-based indexing, both for ordinary text [20] and for text produced by automatic speech recognition [21]. As we have seen above, n-gram indexing is also known to be relatively robust in the presence of OCR errors. We therefore chose to use n-grams as the basis for our experiments. This raises the question of how to choose the optimum value for n. Longer n-grams might match character sequences that are recognized without error fairly well, but shorter n-grams might help mitigate the effect of OCR errors. We therefore tried n=1, n=2, and a combination of the two n-gram lengths. The large inventory of Chinese characters results in excessively large indices for n>2, so we did not try larger values of n.

Previous studies have shown an interaction between the retrieval model and the term weighting and normalization techniques. We therefore ran experiments with InQuery, a widely used system based on a probabilistic retrieval model, and a locally developed vector space retrieval system called Crystal. InQuery provides a standard reference implementation which has been extensively debugged, while Crystal provides complete access to the internal features of the system. We used the default term weighting and "weighted sum" normalization techniques of the InQuery system. In contrast, weighting schemes in vector space model may vary quite differently. With Cyrstal, we experimented with both byte length normalization and cosine normalization. For query n-gram weights, we emphasized longer n-grams over shorter n-grams as follows:

$$q_k = \frac{tf_k(3w_k - 1)}{\sqrt{\sum_{i=1}^{t} tf_i(3w_i - 1)}}$$

where q_k is the weight of n-gram k from a query, tf_k is its term frequency, w_k is the number of characters in n-gram k, and t is the total number of n-grams in the query. With this formula, single bigram match (with weight 5) is given more emphasis than two unigram matches (each with weight 2), but less than three unigram matches.

To explore whether retrieval effectiveness could be further improved, we also ran a small set of blind relevance feedback experiments using the same collection and Rocchio's method:

$$W_{new} = \alpha W_{old} + \beta \frac{1}{|R|} \sum_{X \in R} X - \gamma \frac{1}{|T-R|} \sum_{X \in T-R} X$$

where W_{old} is the initial query weight vector, R is the set of relevant document vectors, T is the set of all document vectors, and α, β, and γ are coefficients controlling the contribution of each factor. For blind relevance feedback (i.e., without manual relevance judgments), the top N documents in the initial result set are assumed to be relevant.

Run ID	System	Field	Indexing Method	Weighting scheme	Ave. P
I1s	Inquery	title only	1-gram		0.3612
I2s	Inquery	title only	2-gram		0.4083
Ins	Inquery	title only	1-gram and 2-gram		0.4397
I1	Inquery	all fields	1-gram		0.3472
I2	Inquery	all fields	2-gram		0.4621
In	Inquery	all fields	1-gram and 2-gram		0.4692
Gb1s	Crystal	title only	1-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	0.3509
Gc1s	Crystal	title only	1-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	0.3059
Gb2s	Crystal	title only	2-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	0.4044
Gc2s	Crystal	title only	2-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	0.4000
Gbs	Crystal	title only	1-gram and 2-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	0.4164
Gcs	Crystal	title only	1-gram and 2-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	0.4098
Gb1	Crystal	all fields	1-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	0.3963
Gc1	Crystal	all fields	1-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	0.3157
Gb2	Crystal	all fields	2-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	0.4582
Gc2	Crystal	all fields	2-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	0.4344
Gb	Crystal	all fields	1-gram and 2-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	0.4757
Gc	Crystal	all fields	1-gram and 2-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	0.4459

Table 2. Experiment results using different retrieval models, query sets, indexing methods, and weighting schemes.

Run ID	System	Field	Indexing Method	Weighting scheme	Feedback parameters	Ave. P
Gb	Crystal	all fields	1-gram and 2-gram	Best in basic strategies	(No expansion)	0.4757
Gbrf	Crystal	all fields	1-gram and 2-gram	Same as above	Top N=1, $\alpha = \beta = 1$	0.4117
Gbrf5	Crystal	all fields	1-gram and 2-gram	Same as above	Top N=5, $\alpha = 5$, $\beta = 1$	0.4255

Table 3. Experiment results using Rocchio's relevance feedback formula.

5. Results

The results obtained are listed in Table 2 and 3. In those tables, the retrieval effectiveness is characterized by the mean (over topics) of the uninterpolated average precision, a commonly used evaluation measure computed by the trec_eval program (available at ftp://ftp.cs.cornell.edu/pub/smart/). As can be seen, the combination of unigrams and overlapping bigrams consistently performs better than that overlapping bigrams alone, which in turns consistently outperforms unigrams alone. This is true for the probabilistic and vector space retrieval models, and for both long (all topic fields) and short (title field only) queries. Another consistent result is that byte length normalization performs better than cosine normalization for different indexing methods and for different query lengths.

Several studies of Web searching behavior have shown that searchers typically use only a few query terms, perhaps because it is easier to understand the behavior of the system when only a few terms are used. The title field in each topic description typically contains between two and five Chinese characters that are chosen as terms that a Web searcher might issue as a query. The results show that long queries (using all fields) generally perform better than short queries (using only the title field), achieving a 6.7% relative improvement with InQuery and a 14.2% relative improvement with Crystal. Viewed another way, the two retrieval models (probabilistic and vector space) perform about equally well with long queries, but InQuery's probabilistic model achieved a 5.6% relative improvement over Crystal's vector space model with short queries.

Table 3 shows the results from our blind relevance feedback experiments. With the best retrieval strategy, no performance improvement was observed using the Rocchio parameter values that we tried. Although these results should be interpreted as extremely preliminary because we have yet to systematically explore the space of possible parameter values, they do suggest that correctly tuning blind relevance feedback parameters without a test collection of the type we have developed would be impractical. In fact, it is not yet clear that blind relevance feedback using documents that contain OCR errors will be helpful. Results obtained by Singhal et al. in a spoken document retrieval application [22] suggest that effective blind relevance feedback may actually require comparable (i.e., topically similar) text that is free of OCR errors.

6. Conclusions and Future Work

OCR text provides the cheapest and fastest way to make full-text images searchable, but optimizing retrieval effectiveness under these conditions requires that the retrieval technique be adapted to mitigate the effect of OCR errors. Previous work has shown that OCR errors have little effect on retrieval at low OCR error rates, but that relatively short documents with poor image quality can produce severe adverse effects. Unfortunately, this is often the case in real-world applications such as retrieval from the SCRC collection of Chinese newspaper clippings. We have described a new Chinese document image retrieval test collection and a set of retrieval experiments that we performed with that collection to explore the effect of different retrieval models, query lengths, n-gram lengths, and normalization schemes. The results show that n-gram length has the greatest effect on retrieval effectiveness, with a combination of unigrams and overlapping bigrams producing the best results. Query length was also found to have a substantial effect, as has been seen in other retrieval evaluations.

The test collection that we have developed opens the door to a number of interesting questions that we are interested in exploring. Because we have already translated the topic descriptions into English, cross-language document image retrieval is a logical next step. We already have some experience with cross-language spoken document retrieval between English and Chinese, so much of the required infrastructure for such an experiment is already in place. A second interesting direction for exploration is differential handling for documents with many recognition errors. For each document in the test collection, we know the estimated error rate reported by the OCR system. It might prove useful to develop corpus-trained correction algorithms even if such algorithms are computationally expensive, because we might productively apply those algorithms to only the most severely degraded documents. As a first step in this direction, we have begun to create manually corrected text for a portion of the test collection. Finally, we believe that it would be interesting to explore additional normalization techniques, particularly those that are optimized for short queries. InQuery's excellent performance with short queries suggests that there is likely some room for improvement in this regard. We expect that others will find additional uses for the same test collection. With multi-level judgments from multiple judges, topic descriptions in two languages, and manually corrected text for some of the documents, we believe that it is a rich resource that can support important research on document image retrieval in the years to come. Researchers interested in using the collection should contact the first author.

Acknowledgements

This work is supported in part by National Science Council, Taiwan, under the grant numbers: NSC 88-2418-H-001-011-B8908 and NSC 88-2418-H-001-011-B9003 and by the U.S. Defense Advanced Research Projects Agency under contract N6600197C8540 and under cooperative agreement N660010028910. The authors will like to thank researchers in the Socio-Cultural Research Center, Fu Jen Catholic University, for helping create the OCR test collection, the University of Massachusetts for the use of InQuery, and Tapas Kanungo for his helpful suggestions.

References

[1] Yuen-Hsien Tseng, "Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text", Journal of American Society for Information Science and Technology (JASIST), Vol. 52, No. 5, March, 2001, pp. 378-390.

[2] Yuen-Hsien Tseng, "An Approach to Retrieval of OCR Degraded Text", Journal of Library Science, National Taiwan University, No. 13, Dec. 1998, pp.153-168.

[3] http://www.colosys.net/pathfinder/AboutPathfinder/HistoricNewspapers.htm

[4] S.L. Mantzaris, B. Gatos, N. Gouraros and S.J. Perantonis, "Linking Article Parts for the Creation of a Newspaper Digital Library", *Content-Based Multimedia Information Access International Conference (RIAO2000*), Paris.

[5] "Contest 1999 - Automatic Cataloguing and Searching Contest", Fritz Kutter-Fonds, http://www.kutter-fonds.ethz.ch/contest99.html

[6] Andreas Myka, "QDOC'99 - Querying Document Bases," Final report submitted to the Fritz Kutter-Contest 1999, ETH Zurich.

[7] Yuen-Hsien Tseng, "Report for Automatic Cataloguing and Searching Contest," Final report submitted to the Fritz Kutter-Contest 1999, ETH Zurich.

[8] K. Taghva, J. Borsack, A. Condit, S. Erva, “The Effects of Noisy Data on Text Retrieval,” Journal of the American Society for Information Science, Vo.45. No. 1, 1994, pp.50-58.

[9] Kazem Taghva, Julie Borsack and Allen Condit, "Results of applying probabilistic IR to OCR text," Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval July 3 - 6, 1994, Dublin Ireland, pp. 202-211.

[10] K. Taghva, J. Borsack, and A. Condit, “Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model,” Information Processing and Management, Vol. 32, No.3, 1996, pp. 317-327.

[11] J. P. Callan, W. B. Croft, and S. M. Harding, "The INQUERY retrieval system," Proceedings of the 3rd International Conference on Database and Expert Systems, Springer-Verlag, New York, 1992, pp.78-83.

[12] Gerard Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Retrieval*, Englewood Cliffs, NJ, 1971, Prentice Hall Inc.

[13] William B. Frakes and Ricardo Baeza-Yates, *Infomation Retrieval: Data Structure and Algorithms*, Prentice Hall, 1992.

[14] Harman, D. “Overview of the third Text REtrieval Conference (TREC-3)” Proceedings of the Third Text Retrieval Conference, 1994, pp.1-19.

[15] Amit Singhal, Gerard Salton, and Chris Buckley, “Length Normalization in Degraded Text Collections,” Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 149-162.

[16] Elke Mittendorf, Peter Schäuble and Páraic Sheridan, "Applying probabilistic term weighting to OCR text in the case of a large alphabetic library catalogue", Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval July 9 - 13, 1995, Seattle, WA USA, pp. 328-335.

[17] Daniel Lopresti and Jiangying Zhou, “Retrieval Strategies for Noisy Text,” Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 255-269.

[18] Claudia Pearce and Charles Nicholas, “TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data,” Journal of the American Society for Information Science, 47(4), 1996, pp.263-275.

[19] Harding, W. B. Croft, and C. Weir, “Probabilistic Retrieval of OCR Degraded Text Using N-Grams,” in Research and Advanced Technology for Digital Libraries, Carol Peters and Costantino Thanos, Editors, 1997. pp. 345-359. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir-115.ps.gz

[20] Ross Wilkinson, “Chinese Document Retrieval at TREC-6,” in The Sixth Text REtrieval Conference (TREC-6), edited by D. K. Harman, Nov., http://trec.nist.gov/, 1997.

[21] Helen Meng, Berlin Chen, Erika Grams, Wai-Kit Lo, Gina-Anne Levow, Douglas Oard, Patrick Schone, Karen Tang and Jian Qiang Wang, “Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval,” Technical Report, Johns Hopkins University, Oct. 2000, http://www.clsp.jhu.edu/ws2000/groups/mei/.

[22] Amit Singhal and Fernando Pereira, “Document expansion for speech retrieval”, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval August 15 - 19, 1999, Berkeley, CA USA, pp. 34 – 41.