圖 書 館 學 刊
第十三期（民國87年12月）：頁153-168
NATIONAL TAIWAN UNIVERSITY
JOURNAL OF LIBRARY SCIENCE
NO.13, DECEMBER 1998, pp.153-168
ISSN 1018-3817

OCR雜訊文件之檢索

An Approach to Retrieval of OCR Degraded Text

曾元顯*
Yuen-Hsien Tseng

Abstract

The major problem with retrieval of OCR text is the unpredictable distortion of characters due to recognition errors. Because users have no ideas of such distortion, the terms they query can hardly match the terms stored in the OCR text exactly. Thus retrieval effectiveness is significantly reduced, especially for low-quality input. To reduce the losses from retrieving such noisy OCR text, a fault-tolerant retrieval strategy based on automatic keyword extraction and fuzzy matching is proposed. In this strategy, terms, correct or not, and their term frequencies are extracted from the noisy text and presented for browsing and selection in response to users' initial queries. With the understanding of the real terms stored in the noisy text and of their estimated frequency distributions, users may then choose appropriate terms for a more effective searching. A text retrieval system based on this strategy has been built. Examples to show the effectiveness are demonstrated. Finally, some OCR issues for further enhancing retrieval effectiveness are discussed.

Keywords: Optical character recognition, information retrieval, fault-tolerant retrieval, keyword extraction, fuzzy matching

*Department of Library and Information Science, Fu Jen Catholic University 242, Taipei, Taiwan, R.O.C.
E-mail: tseng@blue.lins.fju.edu.tw

1. Introduction

The advent of the Internet and the World Wide Web has made publication, dissemination, and access of information more easily. Such accessibility of networked resources has inspired more and more information providers to digitize their data for networked information services. Although future information is likely to be present in full digital form, the digitization of data on published paper materials, however, is not an easy task. A number of possible approaches include: (1) manual re-entry of the data; (2) creating data indexes or bibliographic data for ease of locating the source materials; (3) full-text image scanning; (4) optical character recognition of full-text images. Most current systems have taken the approach (2) and (3) as a model for digitization, indexing, searching and accessing. That is, paper data are scanned into full-text images for ease of accessing, while data indexes or bibliographic data are created for searching in the large volume of the scanned files. However, indexes or bibliographic data have their limitation: They require large amounts of human effort in analyzing the source files to obtain descriptive metadata and such metadata, whether subject headings or keywords, tend to be insufficient in describing the source materials. This limitation often leads to less effective retrieval of specific data in a large database. For example, a scholarly article is often described by only 5 keywords. Other important concepts contained in the article, but not in the keyword list, can not be used to locate this article. If full texts are available, retrieval effectiveness can be greatly enhanced with current advanced information retrieval technologies. However, full texts obtained from manual re-entry of data is seldom available due to the prohibitive amount of time and cost. It seems that text obtained from automatic optical character recognition (OCR) become alternative full-text sources for the digitized information.

However, the major problem with retrieval of OCR text from image data is the inevitable corruption of characters which results from even the best OCR system. Until now, only a few researches have tackled this problem. It is shown that OCR errors have little effect on retrieval with good quality input; effectiveness is significantly reduced in short texts with poor image or scanning quality (1,2). Harding, et al (3), reported an approach to retrieving OCR degraded English text using n-gram matching. The results show that direct retrieval of documents using n-gram leads to improved

performance over standard (word based) method on the same data when a level of 10 percent degradation or worse was achieved. In this paper, we report our preliminary results for reducing the losses from retrieving corrupted, or "noisy", Chinese OCR text through the proposed fault-tolerant information retrieval techniques.

In a project launched by Socio-Cultural Research Center, Fu Jen Catholic University, we have to make hundreds of thousands of newspaper clippings over the past 40 years accessible and searchable through the World Wide Web. By the end of this summer, over 100,000 clippings will have been scanned into image files in TIFF format. In the meantime, commercial OCR software will be used to convert them into text as soon as possible. A search engine is built to deal with these OCR-degraded texts.

The distinct features of the search engine are that it has a multilingual keyword extraction module without using dictionaries and that it provides fuzzy matching for searching. Terms, correct or not, and their term frequencies from the noisy texts are extracted to form a keyword database. With this collection-specific keyword resource, users can search the text database directly or search the keyword database first and then the text database indirectly. By first searching the keyword database with fuzzy matching, the system prompts a ranked list of related terms and their estimated term frequencies. With the understanding of the real terms stored in the text database and of their frequency distributions, users may choose appropriate terms for a more effective searching.

This retrieval system is highly fault-tolerant in that the search in the keyword database with fuzzy matching results in candidate terms used in the text database and that the candidate terms are selected interactively by users, thus reducing the vocabulary mismatch between the terms submitted and the terms used in the noisy text documents. Moreover, the retrieval system is highly automatic in that keyword extraction and document indexing require no manual involvement, both are handled by the software we developed. The overall effect is that the process of document digitization and indexing can be processed by hardware devices and software automatically without sacrificing much retrieval effectiveness. This model of digitization, indexing, searching, and accessing will greatly reduce the time and cost needed in providing world-wide information access and service.

The rest of the paper is organized as follows: In the next section, the indexing and

retrieval models used in the OCR text retrieval application is introduced. This is followed by a brief description of the keyword extraction for interactive term selection in Section 3. Section 4 illustrates a number of fault-tolerant retrieval examples. Finally Section 5 concludes the paper.

2. Fuzzy Matching and Document Ranking

The information retrieval system we have been developing is based on the n-gram indexing model (4), the vector space retrieval model (5), and a self-developed keyword extraction algorithm (6,7). It supports a number of advanced search functions such as natural language-like queries, fuzzy matching, document ranking, term suggestion and term relevance feedback. Below we introduce the ideas of the n-gram indexing model and the vector space retrieval model. These two models have the effect of enabling fuzzy matching and document ranking, which are required features in a fault-tolerant retrieval system.

In the vector space retrieval model, a document d is represented by an n-dimensional vector in the form of: $d = (w_{d1}, w_{d2}, ..., w_{dn})$ where w_{di} denotes the weight of the term t_i in the document d. Ideally each term represents a concept mentioned in the document. In practice, the terms are usually extracted from the document directly. In Chinese documents, for example, a term could be a meaningful word, a single character, or any consecutive string found in the document. The weight w_{di} represents the importance of the term t_i to that document. If a term is absent from the document, its weight is set to zero; otherwise, a weighting scheme, usually based on term frequency, is applied to determine its value.

Similarly, user queries are converted to query vectors in the same way as document vectors. A query q is represented by a vector of the form: $q = (w_{q1}, w_{q2}, ..., w_{qn})$, where w_{qi} denotes the weight of the term t_i in the query q. With the query and the documents transformed into the same vector space, the similarity between the query q and each document d is calculated by the inner product of the corresponding vectors as follows:

$$Sim(d, q) = \sum_{i=1}^{n} (w_{di} \times w_{qi})$$

Thus the higher the degree of the similarity between the query q and the document d, the higher the value of their inner product. The retrieval results of the query are a

list of documents ranked with their similarity values in descending order, thus the name vector space retrieval model.

While the retrieval model determines how each document is related to a user query, the indexing model determines how each index term is represented as a basis for retrieval. Different term representations may lead to different retrieval performance even with the same retrieval model. In the above vector space model, for example, if a word is submitted as a query and that word does not match any index terms used to form the query vector, the query with that word will retrieve nothing even if there are documents containing that word. A case like this is known as "vocabulary mismatch". To reduce vocabulary mismatch, an indexing model known as n-gram indexing is used in most recent Chinese information retrieval systems (8,9,10).

In n-gram indexing, any consecutive n characters in a document are used as index terms. For example, the string "WORDS" is indexed with the terms "WO", "OR", "RD", and "DS" with n equal to 2. Because n-gram indexing may results in excessive terms for large n, the value n is usually set to 1 and 2 for Chinese documents. It can be shown that for Chinese text retrieval 1-gram and 2-gram indexing with the vector space model (or with other similar ranking retrieval models) is sufficient to support most syntactically approximate matching, or fuzzy matching in short. As an example, the term "文學史" (literature history) can match "文學的歷史" (history of literature) with 4 common n-grams (namely, "文", "學", "史", and "文學") out of a total of 10 different n-grams (the other six are "學史" from the first term and "的", "歷", "學的", "的歷", and "歷史" from the second term).

Document ranking sorts retrieved documents according to the similarity measurement between queries and documents. High-similarity documents are placed at the top of the result set in an attempt to help users locate relevant documents more easily. Such ranking is enabled by the fuzzy matching which retrieves documents containing string patterns syntactically approximate to the query string. Fuzzy matching is a powerful tool to free users from worrying about search failure due to the vocabulary problem. However, it is also the main cause that baffles most end users for its obscure similarity measurement. As we will show later, fuzzy matching alone may produce excessive results. It is better used in the first-half phase of interactive term selection, a topic discussed in the next section.

3. Automatic Keyword Extraction for Interactive Term Selection

Since OCR text contains corrupted characters and terms, the vocabulary mismatch problem is likely to occurs for any queries. This is because users have no ideas about whether terms in OCR texts are recognized correctly or not. The only thing that users can do is to submit correct terms and then to expect that fuzzy matching can retrieve those documents containing exactly the same terms and those documents containing the corrupted terms as well. Although fuzzy matching may alleviate the problem of vocabulary mismatch, those relevant documents with lower similarity scores may be interfered with other irrelevant ones with higher similarity scores. The low-score relevant documents may be placed far away from the top of the result set, thus making identification of relevant documents difficult.

To tackle this problem, we propose a solution which relies on automatic extraction of significant terms, whether recognized correctly or not, from the OCR text to make up a keyword database for term selection. By searching the keyword database before searching the text database, syntactically approximate terms are present for user selection. Since these terms are extracted from the noisy OCR text, there will be no discrepancy between the term submitted and the term contained in the text database. Besides, irrelevant but high-similarity documents due to fuzzy matching can be filtered out by this term selection approach.

Basically, there are three approaches to automatic keyword extraction. The first and perhaps the most simplest way is the use of a dictionary or a lexicon to match against the input text. Text terms appear in the dictionary are selected as system-generated keywords. While this method is common in most Chinese OPAC systems in Taiwan, it suffers from the cost of maintaining such dictionaries. Besides, the quality of the keywords depends on the chosen dictionaries, and ordinary dictionaries do not cover the names of individuals, institutes, places, or emerging buzzwords of various fields, let alone many more unpredictable erroneous terms in OCR text.

Another approach is based on the parser developed from natural language processing techniques (11,12,13,14,15). Noun phrases are identified and then filtered to yield representative terms. Such parser is language dependent and often requires resources such as manually tagged corpora. Another limitation is that it requires the input text be confined to a well-defined grammar. Thus documents such as bibliographic titles

(usually noun clauses) and text from OCR or speech recognition (usually contains erroneous words) do not satisfy the requirement.

The third approach is based on some statistical features of the input text (16,17,18,19,20). Although the statistical approach might result in illegal terms and miss those terms with insufficiently statistical features, it does not have the disadvantages as those mentioned above. Names of individuals, institutes, places, or emerging buzzwords can be extracted. Degraded or noisy text from OCR or speech recognition fits this method.

Our approach to keyword extraction is an algorithm similar to the statistical approach. It is based on an assumption that keywords are repeated patterns because documents concentrating on a topic tend to mention a set of words in a specific sequence a number of times. The proposed algorithm has some distinct features: it requires no extra resources such as lexicons, corpora, or NLP parsers; the time complexity are linear in average case; the threshold of term frequency, the only parameter in this algorithm, is easily tuned (usually set to value 1 or 2); key phrases of any length can be identified; when used in character level, single words or word stems can be identified as well as multiple-word phrases (6,7).

Once keywords are extracted, they can be applied in the term selection application. Term selection can take the forms of term relevance feedback (TRF) and term suggestion (TS). In TRF, terms extracted from retrieved documents are shown in certain order of listing for user selection. Users then inspect the term list and select those terms related to their goals for another round of searching. In TS, collection-specific terms are extracted from the text collection. These terms and their frequencies constitute the keyword database for suggesting terms in response to users' queries. The difference between TS and TRF is that TS suggests users with terms similar to their query before searching the text database, while TRF provides terms not necessarily syntactically similar to the query terms after an initial search of the text database. These term selection approaches and the foregoing fuzzy matching provide the fault-tolerant retrieval demonstrated in the next section.

4. Examples of Fault-Tolerant Retrieval

The fault-tolerant techniques have been applied to three text databases: a biblio-

graphic database, a full-text news database, and a noisy OCR text database. From these three databases, three corresponding keyword databases have been built with the keyword extraction algorithm. A web page is available at http://xlib.fju.edu.tw/demo/index.html for demonstrating the retrieval of ordinary text documents as well as OCR noisy documents. These databases are briefly described as follows.

1. FJU OPAC: the OPAC (Online Public Access Catalog) system of the library of Fu Jen Catholic University. There are currently 356,000 bibliographic records.

2. Keywords of FJU OPAC : contains 118,430 Chinese and English keywords extracted from the titles of the bibliographic records. The threshold of term frequency is set to 1 so that any keywords appear more than once are collected in the database.

3. News Articles: contains over 13000 Chinese news articles from the web sites of two news agencies, Central Daily News and China Times, from July 1997 to December 1997.

4. Keywords of News Articles: contains 66,500 keywords extracted from the news articles. Since these news articles are full-text documents, the threshold of term frequency is set to 2 to filter out most illegal terms and retain as many useful terms as possible.

5. Newspaper Clippings: currently contains 71 scanned images and their OCR noisy text from the newspaper clippings provided by Socio-Cultural Research Center, Fu Jen Catholic University. This database will expand to include over 100,000 scanned images and their OCR text in the future.

6. Keywords of News Clippings: contains 3466 terms extracted from the OCR noisy text. Since keywords in the OCR text might contains erroneous characters due to recognition error, the threshold of term frequency is set to the lowest possible value 1 so that any string patterns occurs more than once are extracted as candidate terms for user selection.

An example of searching the keyword database of Newspaper Clippings is illustrated in Fig. 1. The query term is "江澤民" (Jiang Zemin) and 2 records are retrieved with fuzzy matching. The search results are sorted by match score and then by term frequency. Although the keyword database does not have a record exactly matching the query term, it returns the term "江滯民" (Jiang Ze-min), which contains an incorrectly recognized character by OCR. This term is morphologically similar to the query term (and happens to be similar to the query term phonetically). Note the term is

retrieved with very low match score. This is only possible for such keyword databases without worrying excessive results. By selecting the exact term used in the OCR text as the actual query term, the vocabulary mismatch problem is avoided. Figure 2 shows the search results. Note the OCR text is used only for enhancing retrieval performance, the actual documents accessed by users are the full-text images provided in the hyper-links within the search results, as shown in Figure 2.

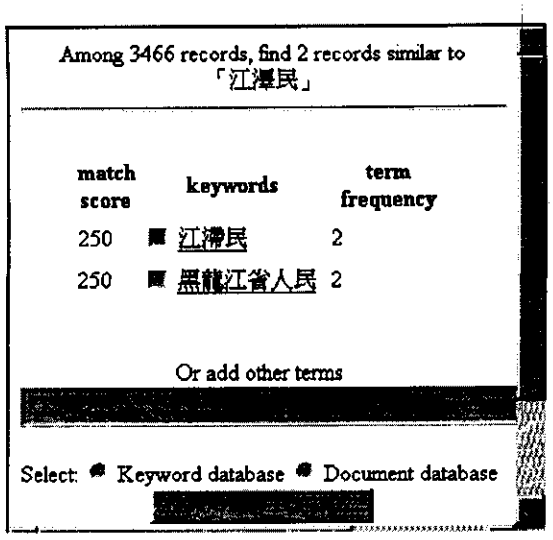


Figure 1. An example from the Keywords of Newspaper Clippings for term suggestion.

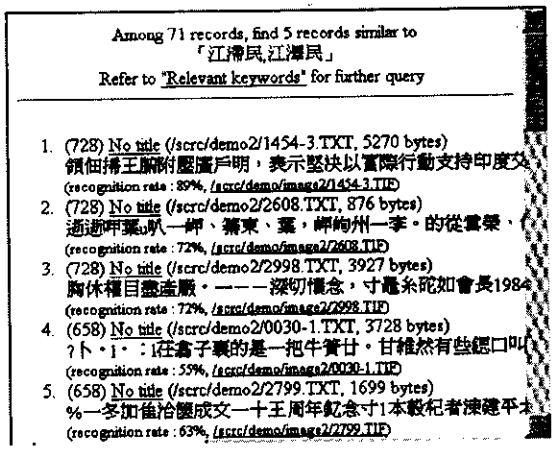
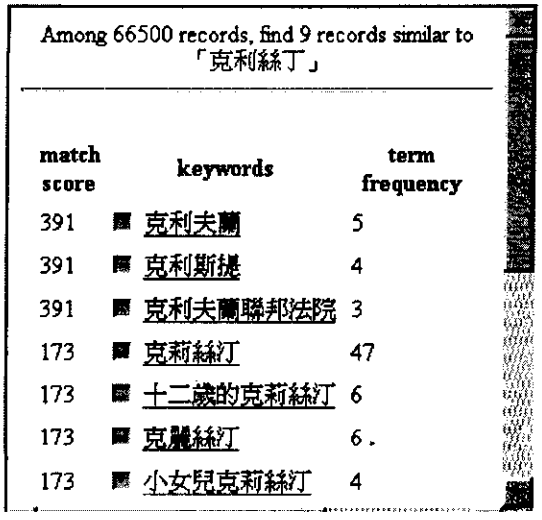


Figure 2. An example of search results from the Newspaper Clippings.

Since the current collection of Newspaper Clippings contains too few documents for further demonstration, the second example is taken from the News Articles. In Figure 3, the query term is the name of a South African little girl who behaved bravely in a kidnap event. Note the various translation of an English name might be. The query gives one translation "克利絲丁" for "Christine", the database records two translations "克莉絲汀" and "克麗絲汀" from two news agencies, or from different reporters. This case is similar to those OCR text which might contain different corrupted versions of the same term such that users do not know the exact forms of the term before their query submission. Note the low-score fuzzy matching helps locate the correct terms. If the original query is directly submitted to the text database, incorrect documents (namely those about "克利夫蘭" (Cleveland) and "克利斯提" (Christie)) are retrieved at the top of the ranked results.

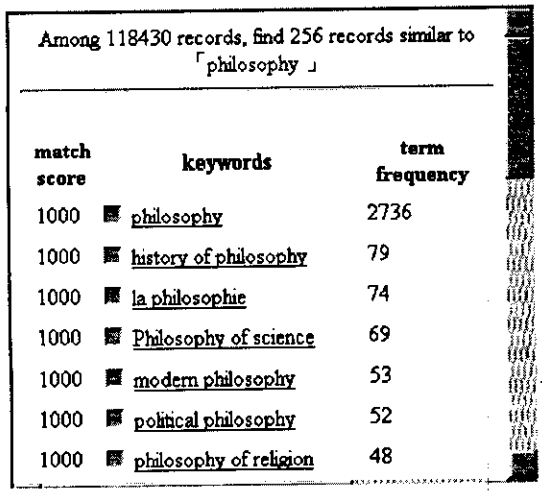


Among 66500 records, find 9 records similar to 「克利絲丁」

match score	keywords	term frequency
391	克利夫蘭	5
391	克利斯提	4
391	克利夫蘭聯邦法院	3
173	克莉絲汀	47
173	十二歲的克莉絲汀	6
173	克麗絲汀	6
173	小女兒克莉絲汀	4

Figure 3. An example from the Keywords of News Articles for term suggestion.

These keyword databases have an additional effect that if the query term contains a broad meaning, the query results function as a dynamic directory. That is, more specific terms under the category specified by the query are present in the results. From this dynamic directory, users are able to understand the distribution of the collection by inspecting the term frequencies and then to choose more specific terms for searching. This would avoid excessive outcomes resulting from submitting the original query to the document database. The example from the Keywords of FJU OPAC in Figure 4 with the query term "philosophy" illustrates this effect.



Among 118430 records, find 256 records similar to 「philosophy」

match score	keywords	term frequency
1000	philosophy	2736
1000	history of philosophy	79
1000	la philosophie	74
1000	Philosophy of science	69
1000	modern philosophy	53
1000	political philosophy	52
1000	philosophy of religion	48

Figure 4. An example illustrating the effect of "dynamic directory".

The next examples demonstrate the effects of term relevance feedback. In Figure 5, a number of keywords is extracted for feedback from the results of querying the bibliographic database with the term "image processing". Another example from the full-text News database is in Figure 6, where the query term is "拜耳撤資案" ("The recall of Bayer's investment"). This kind of local TRF provides users more search terms which are semantically related to (but not necessarily syntactically similar to) the query terms under the same topic. An example is the term "computer vision" in Figure 5. More examples are in Figure 6, such as "外商" (foreign company), "投資" (investment), "憂心" (worry) and "公投" (referendum) (Bayer is reluctant to go through a referendum held by a local government for his investment). Thus TRF has the effects of expanding users' search vocabulary and of guiding users in search directions closer to their goals.

5. Conclusions and Future Work

Most full-text image systems rely on manual preparation of bibliographic metadata for supporting information retrieval. Due to the large amount of time and efforts of human indexers required in this process, and due to the limitation of relatively few searchable data provided by such metadata, this paper proposes using OCR text as an

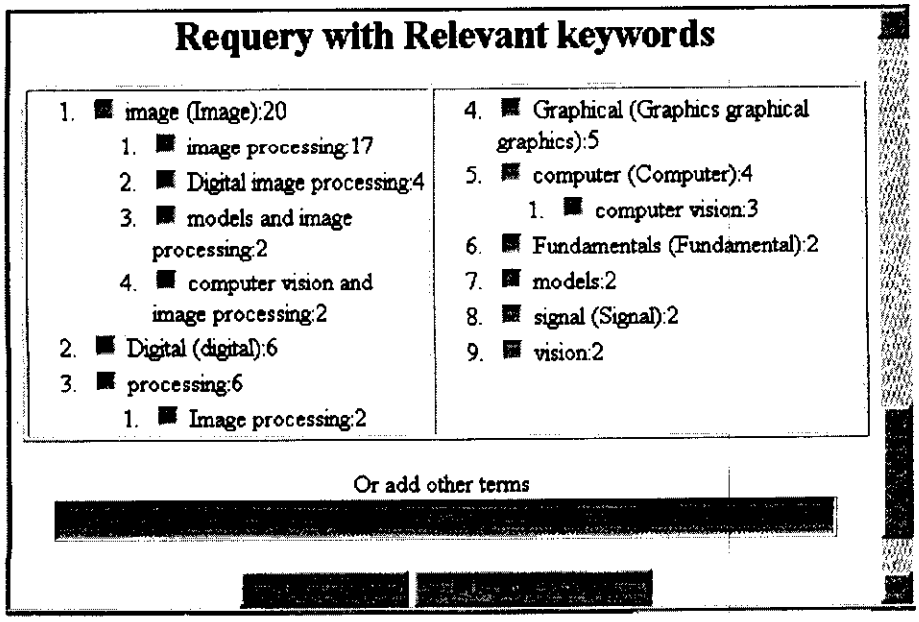


Figure 5. An example from the bibliographic database for term relevance feedback.

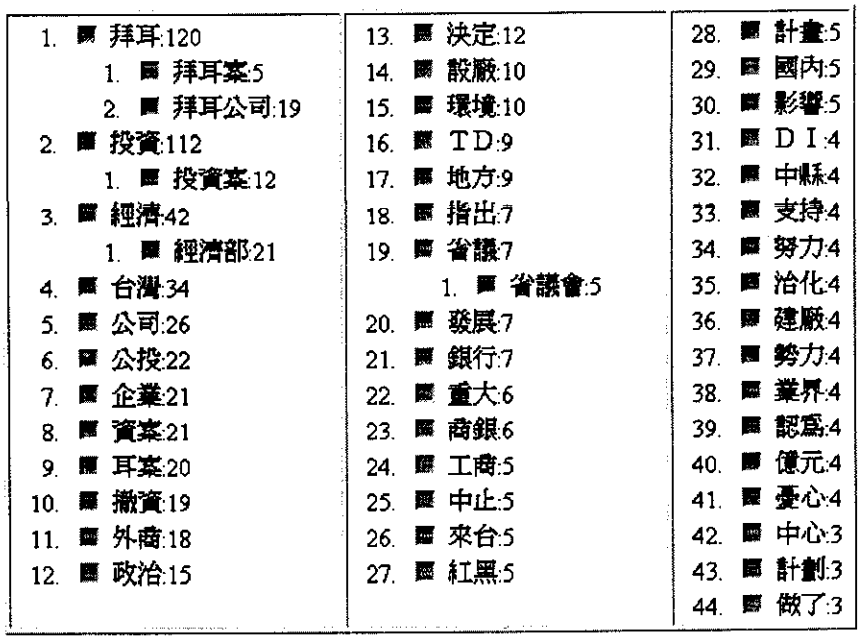


Figure 6. An example from the News Articles for term relevance feedback.

alternative or supplementary source for retrieval of full-text images. However, because optical character recognition can not guarantee 100% correction rate, the resultant text may contain erroneous characters. To support retrieval of such database, a retrieval strategy based on automatic keyword extraction and fuzzy matching is proposed. Terms, correct or not, and their estimated frequencies are extracted for term selection. Such a term listing allows users understand the real terms stored in the database so as to choose appropriate terms for a precise searching.

The term selection retrieval is applied in two forms: one is term suggestion, the other is term relevance feedback. These two retrieval approaches together with direct fuzzy matching have been evaluated for ordinary electronic text (21). The results show that term suggestion and term relevance feedback increase performance by 38.2% and 29.1%, respectively, over direct retrieval of documents using fuzzy matching in the bibliographic database. A similar performance increment of 18.2% and 7.3% is observed for term suggestion and term relevance feedback, respectively, in the full-text database of news articles. These results show that such interactive retrieval is able to provide effective retrieval for ordinary text. As this paper demonstrates, it is also able to yield favorable performance for noisy OCR text. The future work is to provide a more rigorous performance evaluation for noisy OCR text.

In our experience with OCR text retrieval, there are some problems that cannot be solved without the help from OCR technologies. The first is the low recognition rates for newspaper titles due to their different fonts and sizes from most of the other text. Because titles are often the most important information for newspaper articles, the absent of them in the OCR documents affect the presentation of search results and the retrieval performance expected by users. Thus selective recognition for special sections such as titles, captions, and emphasized blocks in an article is highly desirable. The second problem is the random corruption of the same term observed in most OCR systems. Since the automatic keyword extraction algorithm relies on the repetition of strings for extraction, random erroneous patterns of the same term make such extraction more difficult. If mutual character statistics instead of single-character statistics is considered in the recognition process, random corruption can be reduced such that a term is recognized into a fixed pattern. If this desirable effect can be achieved, the keyword extraction algorithm can extract those terms that occur more than once, no matter they are correct or not. The documents containing these terms can then be locat-

ed by the fault-tolerant retrieval approach. The third problem is the need for batch processing of the OCR task. Since a large number of data is to be digitized, automatic recognition of most paper data will speed up the conversion process from digital images to OCR text. In addition, the recognition rate of each image should be reported as a result of the batch processing, so that text of low-recognition rate can be tuned manually to reduce the recognition errors. All these problems can only be solved by OCR systems. Solutions to the problems will greatly enhance the effectiveness and possibility of retrieving full-text images with OCR text.

The analysis and processing of noisy text are rarely discussed in the past information retrieval researches. It is a core technology for information digitization, indexing, searching, and accessing in digital libraries. The results in this paper can not only be applied to retrieval of noisy OCR text, but also existing electronic documents, or even other types of noisy text, such as those from speech or video recognition results.

References

1. S. M. Croft, K. Taghva, Harding, and J. Borsack, "An Evaluation of Information Retrieval Accuracy with Simulated OCR Output," Symposium of Document Analysis and Information Retrieval, 1994.
2. J. Taghva, A. Borsack, S. Erva, Condit, "The Effects of Noisy Data on Text Retrieval," In UNLV Information Science Research Institute Annual Report, (place: publisher, 1993), 71-80.
3. W. Harding, B. Croft, and C. Weir, "Probabilistic Retrieval of OCR Degraded Text Using N-Grams," 1995, in Research and Advanced Technology for Digital Libraries, Carol Peters and Costantino Thanos, (Place: publisher, 1997), 345-359. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir-115.ps.gz
4. William B. Cavnar, "Using An N-Gram Based Document Representation With A Vector Processing Retrieval Model," in Proceedings of the Third Text Retrieval Conference (TREC-3), (Place: publiser, 1995), 269-277.
5. Gerard Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer (Place: Addison-Wesley, 1989).
6. Yuen-Hsien Tseng, "Multilingual Keyword Extraction for Term Suggestion," in 21st International ACM SIGIR Conference on Research and Development in Informa-

tion Retrieval - SIGIR '98, Aug. 24-28, Australia, 1998, pp.337-378

7. Yuen-Hsien Tseng, "Fast Keyword Extraction of Chinese Documents in a Web Environment," International Workshop on Information Retrieval with Asian Languages - 1997, Oct. 8-9, Japan, pp.81-87
8. GAIS亞太 WWW 資源搜尋引擎，http://gais.cs.ccu.edu.tw/www2-adv.html
9. Csmart:網路中文資源檢索系統，http://csmart.iis.sinica.edu.tw/
10. "Online Public Access Catalog, Library of Fu Jen Catholic University," http://xlib.fju.edu.tw/
11. Timonthy C. Craven, "An Experiment in the Use of Tools for Computer-Assisted Abstracting" ASIS 1996 Annual Conference Proceedings, Oct. 19-24, 1996. Also available at http://www.asis.org/annual-96/ElectronicProceedings/craven.html
12. Zimin Wu and Gwyneth Tseng, "ACTS: An Automatic Chinese Text Segmentation System for Full Text Retrieval," Journal of American Society for Information Science, 4:6 (1995): pp.83-96.
13. Antti Arppe, "Term Extraction from Unrestricted Text," http://www.lingsoft.fi/doc/nptool/term-extraction.html, 1995.
14. Mathis H. C. Chen, Tsong-Yi Tseng, Jason J. S. Chang, "Automatic Generation of Indices for Chinese Books," http://nlplab.cs.nthu.edu.tw/~mathis/own/html/PAPER/JNL/96/cpcol/BookIdx.htm
15. Jean Godby, "Two Techniques for the Identification of Phrases in Full Text, "http://www.oclc.org/oclc/research/publications/review94/part1/twotech.htm .
16. Jonathan D. Cohen, "Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting", Journal of the American Society for Information Science, 46(3), 1995: pp.162-174.
17. Craig G. Nevill-Manning, Ian H. Witten and Gordon W. Payner, "Browsing in Digital Libraries: A Phrase-Based Approach" Proceedings of Digital Library 97, Philadelphia PA, USA, 1997, pp.230-236.
18. Richard Sproat, Chilin Shih, William Gale, and Nancy Chang, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese" Computational Linguistics, 22:3 (1996): pp. 376-404
19. Fagan, J. L. "The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval," Journal of American Society for Information Science, 40:2 (1989): pp.115-132.

20. Jones, L. P., Gassie, E. W., & Radhakrishnan, S. "INDEX: The Statistical Basis for an Automatic Conceptual Phrase-indexing System," Journal of American Society for Information Science, 41:2 (1990): pp. 87-98

21. Yuen-Hsien Tseng "Solving Vocabulary Problems with Interactive Query Expansion", Journal of Library & Information Science, 24: 1(1998):pp.1-18