A Survey of Retrieval Strategies for OCR Text Collections

Steven M. Beitzel, Eric C. Jensen, David A. Grossman
Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
{steve,ej,grossman}@ir.iit.edu

Abstract

The importance of effectively retrieving OCR text has grown significantly in recent years. We provide a brief overview of work done to improve the effectiveness of retrieval of OCR text.

Introduction

As electronic media becomes more and more prevalent, the need for transferring older documents to the electronic domain grows. Optical Character Recognition (OCR) works by scanning source documents and performing character analysis on the resulting images, giving a translation to ASCII text, which can then be stored and manipulated electronically like any standard electronic document. Unfortunately, the character recognition process is not perfect, and errors often occur. These errors have an adverse effect on the effectiveness of information retrieval algorithms that are based on exact matches of query terms and document terms. Searching OCR data is essentially a search of "noisy" or error-filled text.

We briefly survey approaches to searching OCR text. These include: defining an IR model for OCR text collections, using an OCR-error aware processing component as part of a text categorization system, auto-correction of errors introduced by the OCR process, improving string matching for noisy data, and issues in cross-language retrieval of image data. We also note that an excellent survey of the more general area of indexing and retrieving document images can be found in [Doer98]. Other recent papers surveying current approaches in OCR Information Retrieval appeared in the 2002 SIGIR workshop on OCR Information Retrieval[1]. The following sections of this paper will examine each of these approaches, and give a summary of the progress made in each area.

IR models for OCR text

Most work in the field of OCR Information Retrieval has been relatively recent. A primary reason for this is that, especially in early years, obtaining a sufficient quantity of data on which to test was very difficult. To get around this problem, Croft and colleagues published one of the first studies of the effects of OCR data on IR by using simulated OCR collections [Crof94]. This study found that for high quality OCR conversion, not much degradation in retrieval effectiveness was encountered, however the retrieval of

[1] http://www.info.uta.fi/sigir2002/html/ws2.htm

short and low-quality documents was adversely affected. Lopresti and Zhou examined the performance of several models of IR on simulated OCR data, and were able to show that it was plausible to use modified approaches (they made use of fuzzy logic and approximate string matching) to increase effectiveness on noisy data [Lopr96]. These conclusions led to the development of models of IR designed specifically for operating on a body of OCR text. Some of the initial work on developing a model of Information Retrieval specifically suited for operating on a collection of OCR text was done by [Mitt95]. Their study involved the development of a term weighting scheme for the probabilistic model of information retrieval. The motivations for this work arose from conclusions reached in [Crof94, Tagh94a, Tagh94b, Tagh96] (and later explored in depth by Mittendorf and colleagues in [Mitt96]), which show that, although retrieval performance is not generally adversely affected by errors in an OCR text collection, performance degradation is often observed in cases where there are few documents in the collection, or the documents in question are very short. It was also observed that the presence of errors in OCR text tends to lead to unstable and unpredictable retrieval performance. In an effort to develop a retrieval model that circumvents these limitations, they incorporate the occurrence probabilities of several kinds of typical OCR errors into their term-weighting schemes. In a set of experiments on a collection of very short documents, they achieved a 23-30% improvement in retrieval effectiveness by using a form of the probabilistic model. Their general equation for the Retrieval Status Value (RSV) of a document is given in Equation 1 below:

$$RSV(q,d_j) \equiv \sum ff(\varepsilon_i,q)\,ff(\varepsilon_i,d_j)/\lambda_j$$

Where:

q	= query
d_j	= document
$ff(\varepsilon_i,d_j)$	= feature frequency in document
$ff(\varepsilon_i,q)$	= feature frequency in query
λ_j	= number of occurrences of feature frequency in document

Equation 1 – Retrieval Status Value for OCR-IR

The referenced work presents a number of methods for estimating various feature frequencies and the probabilities of their occurrence.

Further work has been done in adapting the probabilistic model of IR to handle error-ridden OCR text. Ohta enhanced the probabilistic model to take advantage of expected errors in OCR text [Ohta97]. Specific character transformations and character occurrence bigrams were used to generate candidate terms for each “true”search term. Documents retrieved by each candidate term are then evaluated for inclusion into the final result set. This resulted in minor improvements in recall for moderate quality OCR documents. Another study was performed in [Hard97]. This study makes use of n-grams to overcome the problems introduced by OCR text, which is a quite fitting solution, when considering that a large percentage of typical OCR errors involve mis-identifying a sequence of one or two characters within a given word. Their basic approach defines a

set of "binding operators" over the constituent n-grams of a word. The goal of these operators is to be strict enough to exclude noise and non-relevant information from the top documents, while lenient enough to prevent elimination of relevant data when a particular constituent n-gram is missing. The experimentation performed in the study was designed to discover which operators would maximize retrieval performance, and it was found that the "passage5" operator, which ranks documents containing n-gram components within windows of five word positions, while allowing the windows to cross word boundaries was the most effective approach. It is theorized that this is to be expected because of the extremely common OCR error wherein spaces are added to the text in improper locations. An example of this can be seen with the word "environmental", which is broken up into the ngram components (en env envi ironm onm ment tal al). The length of this word makes it a very likely candidate for incorrect identification during the OCR process via the insertion of one, or several spaces, however, the passage5 operator binds these components together loosely over a large range that may cross word boundaries, helping to mitigate this problem. In their experiments, operating on four databases that were randomly degraded using data developed at UNLV [Rice93], an improvement ranging from 5-38% was observed, depending on the test collection. The authors used a similar approach to aid in query term expansion. They used n-grams to discover candidate expansion terms that were a match or a near-match to terms in the original query. As expected, a further improvement of 9-18% was observed when using n-grams for query expansion. Very recently, work has been done that takes common OCR errors into account when generating a language model [Jin02]. This language model can then be used to approximate an "uncorrupted" version of a particular document, and it can be used for retrieval in a language modeling approach. The authors found significant improvement when using this approach over other approaches that explicitly correct each error found in the source documents.

Processing OCR Text for Categorization

In addition to document retrieval, there are other areas of IR that must effectively handle OCR text. Once such area is Text Categorization, wherein a group of documents are examined and assigned to a set of categories, typically to aid in document browsing or visualization. Some of the early work in this area is presented in [Hoch94]. This study does not deal directly with mitigating problems introduced by OCR Text, rather, they describe an approach to the development of an automatic indexing and classification system that uses advanced morphological analysis of the text, along with term frequency analysis, index term weighting, and training of the classifier based on a previously-defined document model [Deng92]. The results of this study indicate that classification performance was inhibited by the degradation of index terms from the OCR process. No solution to the problem is given, although it is reasonable to assume that the incorporation of a specialized term weighting scheme for OCR documents, such as the ones described above, would help to improve performance. Evidence of this assumption is presented in [Cavn94] and [Junk97], wherein advanced techniques such as n-gram processing and morphological analysis are used to aid in reducing the effect of imperfections introduced by the OCR process on retrieval effectiveness.

A survey of common techniques used to enhance effectiveness in text categorization can be found in [Seba02]

Auto-Correction of OCR errors

One valuable area of research involves post-processing systems that work to correct the errors that are introduced in the OCR process. This has far-reaching implications, as being able to efficiently compensate for OCR errors allows conventional IR techniques to be used without experiencing degradation in effectiveness. Some early work in this area is discussed in [Liu91]. This study examines and classifies each type of error that can be introduced by the OCR process. Furthermore, it identifies which errors are the most typical and most likely to introduce confusion in the resulting documents. Several techniques for correcting the errors are discussed as well. Dictionary lookups on candidate terms are one of the most course-grained techniques used, and they help to identify words that have likely been corrupted. In addition, terms in the source text are broken up into digrams, and a frequency matrix is kept to help identify which character sequences are indicative of errors. Based on this, adaptive character conversion maps are constructed that allow the system to identify a character sequence likely to be in error, and automatically correct it by performing a lookup in the map and replacing it with the corrected version. These automated techniques were performed in concert with user interaction, and resulted in a significantly improved final text that did not require nearly as much user intervention as prior corrective approaches.

Another study involving the use of a post processing system for OCR error correction was performed in [Tagh94a, Tagh94b]. They used techniques similar to the ones mentioned above, with the addition of a clustering technique that grouped collections of misspellings in with their correctly spelled target term. After frequency analysis on the term spelling clusters eliminated unlikely candidates, each misspelled term was replaced with its correctly spelled counterpart. A more complete overview of error correction techniques in post processing systems can be found in [Kuki92].

Improved String Matching on Noisy Data

For applications where it is desirable to find all occurrences of a particular term, there is the notion of exact string matching. When the data is noisy or corrupted, as is the case with OCR text, exact string matching becomes difficult. This problem has been approached by training language models to recognize terms that are improperly spelled, as done in [Coll01]. A more detailed survey of string matching approaches can be found in [Nava01].

OCR Issues in Cross-Language Retrieval

Performing Information Retrieval on OCR documents in non-English languages provides some interesting and unique challenges. Oard and colleagues have implemented a full system for cross-language retrieval, and have extended that system to support text from document images [Oard99]. Basically, this system proposes the use of character-confusion statistics and character-class recognition algorithms that are specific to the target language in order to mitigate the ambiguities and errors introduced into a text by the OCR process. More recently, Darwish and Oard have focused on retrieval of OCR'd

Arabic text, and have found that using a combination of light-stemming approaches and character n-grams for the selection of candidate index terms generally provides the most improvement in retrieval effectiveness, and is robust over a variety of OCR errors [Darw01, Darw02].

Summary

We have briefly surveyed current techniques in use to facilitate information retrieval on collections of OCR text. There are a large number of proposed techniques and models available for use, and hopefully in the future a generalized solution that takes on aspects of all available techniques will be available to members of the Information Retrieval community.

References

[Cavn94] W. Cavnar and J. Trenkle. N-Gram based text categorization. In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161-175, Las Vegas, NV, 1994.

[Crof94] W. B. Croft, S. M. Harding, K. Taghva and J. Borsack. An Evaluation of Information Retrieval Accuracy with Simulated OCR Output. Proceedings of the Symposium on Document Analysis and Information Retrieval, 1994.

[Coll01] Kevyn Collins-Thompson and Charles Schweizer and Susan Dumais. Improved string matching under noisy channel conditions. Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM), 2001.

[Darw01] Darwish, Kareem, D. Doermann, R. Jones, D. Oard, and M. Rautiainen. TREC-10 Experiments at Maryland: CLIR and Video. TREC-2001, 2001.

[Darw02] Kareem Darwish and Douglas W. Oard. Term Selection for Searching Printed Arabic. Proceedings of the Twenty-Fifth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2002.

[Deng92] A. Dengel, R. Bleisinger, R. Hoch, F. Fein, F, Hones. From Paper to Office Document Standard Representation. IEEE Computer, vol. 25, no. 7, 1992, pp. 63-67.

[Doer98] David Doermann. The Indexing and Retrieval of Document Images: A Survey, The Journal of Computer Vision and Image Understanding: CVIU, Volume 70, Number 3, 1998.

[Fras01] Paolo Frasconi and Giovanni Soda and Alessandro Vullo. Categorization for multi-page documents: A Hybrid Naive Bayes HMM Approach. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, 2001.

[Hard97] Stephen M. Harding and W. Bruce Croft and C. Weir. Probabilistic Retrieval of {OCR} Degraded Text Using N-Grams. Proceedings of the European Conference on Digital Libraries (ECDL), 1997.

[Hoch94] Rainer Hoch. Using IR techniques for text classification in document analysis. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994.

[Junk97] M. Junker and R. Hoch. Evaluating OCR and non-OCR text representations for learning document classifiers. Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 1997.

[Kuki92] Karen Kukich. Technique for automatically correcting words in text, ACM Computing Surveys, Vol 24, No. 4, 1992.

[Liu91] Lon-Mu Liu and Yair M. Babad and Wei Sun and Ki-Kan Chan. Adaptive post-processing of OCR text via knowledge acquisition. Proceedings of the 19th Annual Conference on Computer Science, 1991.

[Lopr96] D. Lopresti and J. Zhou. Retrieval Strategies for Noisy Text. Proceedings of the Symposium on Document Analysis and Information Retrieval, 1996.

[Mitt95] Elke Mittendorf and 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, 1995.

[Mitt96] E. Mittendorf and P. Schauble. Measuring the effects of data corruption on information retrieval. Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR), 1996.

[Nava01] Gonzalo Navarro. A guided tour to approximate string matching, ACM Computing Surveys, vol 33, no. 1, 2001.

[Oard99] Douglas W. Oard. Issues in Cross-Language Retrieval from Document Image Collections. In Proceedings of the 1999 Symposium on Document Image and Understanding Technology, 1999.

[Ohta97] M. Ohta, A. Takasu, and J. Adachi. Retrieval Methods for English text with misrecognized OCR characters. Proceedings of the International Conference on Document Analysis and Recognition, 1997.

[Rice93] S. Rice, J. Kanai, and T. Nartker. An Evaluation of Information Retrieval Accuracy. In UNLV Information Science Research Institute Annual Report (1993), 9-20.

[Seba02] Fabrizio Sebastiani. Machine learning in automated text categorization, ACM Computing Surveys (CSUR), 34:1, 2002.

[Tagh94a] Kazem Taghva and 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, 1994.

[Tagh94b] Kazem Taghva, Julie Borsack and Allen Condit. An Expert System for Automatically Correcting OCR Output, in Proceedings of the SPIE – Document Recognition, 1994.

[Tagh96] Kazem Taghva and Julie Borsack and Allen Condit. Evaluation of model-based retrieval effectiveness with OCR text. ACM Transactions on Information Systems (TOIS), 14:1, 1996.