



WHITE PAPER

Accurate Document Categorization of OCR-Generated Text

By Robert J. Price

Content Analyst Company, LLC
11720 Sunrise Valley Drive
Reston, VA 2019,1 U.S.A.

rprice@contentanalyst.com

Abstract

The purpose of this research was to examine the application of the Latent Semantic Indexing (LSI) algorithm to the categorization of Optical Character Recognition (OCR) -generated documents (or text). LSI is a robust dimensionality reduction technique for the processing of textual data. The technique can be applied to collections of documents independent of subject matter or language. Given a collection of documents, LSI indexing can be employed to create a vector space in which both the documents and their constituent terms can be represented. In practice, vector spaces of several hundred dimensions are typically employed. The resulting vector space possesses some unique properties that make it well suited to a range of information-processing problems. Of particular interest to the document conversion community is the fact that the technique is highly resistant to noise in text that is generated by the OCR conversion process.

Noise in OCR-generated documents nominally takes the form of missing characters or improperly interpreted characters that result in word misspellings. Normally, human operators are employed to perform corrective post processing of noisy OCR-generated text prior to categorization or other document workflow processes that require highly accurate text (e.g., Boolean searches). A technology such as LSI that could eliminate human review and editing of OCR-generated text, while maintaining highly accurate categorization or information retrieval rates, would result in a dramatic increase in workflow throughput with substantial labor savings.

Introduction

Previous work [1] has demonstrated the high performance of Latent Semantic Indexing (LSI) [5] on the Reuters-21578 [9] test set. In this paper, we have examined the ability of LSI to categorize documents that contain corrupted or noisy text (e.g., misspellings, transliterations differences, OCR errors). In an earlier case study for a U.S. Intelligence Agency related to a pending FOIA release, utilizing documents derived via OCR, indicated the LSI technology could possibly provide good retrieval performance on OCR-generated text. Attempts at using other technologies to examine the document set for potential “mosaic effect” connections had failed.

A key feature of LSI that makes it attractive for categorizing noisy text is its capability of automatically extracting the conceptual content of text items [5, 6]. With knowledge of their content, these items then can be treated in an intelligent manner. For example, documents can be routed to individuals based on their job responsibilities. Similarly, e-mails (or documents with noisy text) can be filtered accurately. Information retrieval operations can be carried out based on the conceptual content of documents, not on the specific words that they contain (or variants thereof generated by noise sources, e.g., OCR). Of particular note is that the LSI algorithm does not employ any auxiliary data structures as part of its processing (e.g., thesauri, grammars, taxonomies or ontologies).

Discussion

Numerous pattern-matching technologies have been studied and applied towards finding relevant search patterns in noisy text [2,4]. If the image conversion process is poor (i.e., originals are of poor quality) and the characteristics of the OCR engine are well understood, a model of the overall conversion process can be created and employed to train a pattern-matching engine for searching the OCR text. Such models are too expensive and time consuming to produce in practice. If the image conversion process is relatively clean (i.e., originals are of good quality, image enhancement software and production quality scanning equipment is used), then pattern-matching technologies may be able to overcome some inconsistencies in the OCR text and achieve some reasonable confidence in the precision/recall of the search results. Often the real world is somewhere in between [11].

This problem is also very akin to the conversion of text from speech and subsequent retrieval of the text by search [3]. Techniques described in this paper are also applicable to this domain.

Approach

To formally examine the robustness of the LSI algorithm under conditions of noisy text, a categorization-style testing approach was selected since the authors had previous experience testing the LSI algorithm using this mode of operation. The Reuters categorization test sets were selected due to their availability and active use in other categorization testing efforts by other researchers. The Reuters categorization test sets come in two versions (Reuters-21578 [9] and Reuters RCV1-V2 [10]) and consist of document training and document test sets. The training set provides a knowledge base to train the algorithm undergoing examination and the test documents are used in evaluating the categorization accuracy of the subject algorithm. In the case of the LSI algorithm, the training documents were employed to train the LSI vector space. The makeup of the Reuter’s corpuses is discussed in the following section.

In the case of LSI, the training documents are used to create a matrix that relates the documents and the words that occur in them. The rows of the matrix correspond to terms that occur in the documents. The columns correspond to individual documents. The number entered in the i th row and j th column of the matrix corresponds to the number of times that the i th term appears in the j th document. The matrix produced in this manner can be very large. In practical applications,

it can involve hundreds of thousands of terms and even larger numbers of documents. Fortunately, the matrix is very sparse and is amenable to dimensionality reduction.

A powerful mathematical technique, known as singular value decomposition (SVD), is used to reduce this matrix to a product of three matrices. One of these matrices has non-zero values only on the diagonal. Small values on this diagonal, and their corresponding rows and columns in the other two matrices are then deleted. This truncation process generates a matrix of greatly reduced dimensionality. For any given dimensionality, this technique can be shown to produce an optimal approximation of the original matrix. The columns of the associated matrices can be used to create a vector space in which both terms and documents are represented. The dimensionality of this vector space can be chosen to work well in a particular application. Typically, LSI spaces with a dimensionality of several hundred are employed. The dimensionality reduction has the effect of extracting semantic information that is latent in the processed text, hence the name latent semantic indexing. In some cases unlabeled background data is used to enhance the knowledge base of the training set [7]. This approach was not utilized in this work.

To test the robustness of LSI in conditions of increasing document degradation, a procedure was developed that degraded percentages of text in the test documents by inserting, deleting and substituting characters randomly at specified error rates based on experience with numerous OCR engines and OCR degradation models published in the literature [4]. In our model, for any given document, 66 percent of the errors were random alphanumeric substitutions, 17 percent were deletions, 12 percent were insertions of random alphanumeric characters and 5 percent were random insertions of spaces. Although true OCR errors are more systematic, the intent here was to show to what extent the text of the documents could be degraded and still retain useful categorization results. Appendix A illustrates the impact of the degradation algorithm on one test document at various stages of degradation.

The collections of degraded test documents were then presented to the LSI algorithm for categorization and the accuracy results recorded.

Test Corpus and Performance Measures

To understand the performance of LSI on degraded text, the ModApte version of the Reuters-21578 test set [9] was utilized. A second test using the larger, more recent RCV1-V2 test set was also performed to examine the impacts of a larger training set and much larger test set.

The ModApte version has been used in a wide number of studies [8] due to the fact that unlabeled documents have been eliminated and categories have at least one document in the training set and the test set. We followed the ModApte split defined in the Reuters-21578 data set in which 71% of the articles (6552 articles) are used as labeled training documents and 29% of the articles (2581 articles) are used to test the accuracy of category assignments.

The larger Reuters RCV1-V2 [10] consists of 804,414 documents broken up into 23,149 training documents and 781,265 test documents. We employed the Topic category orientation of the test set, which consists of 103 Topic categories.

For evaluating the effectiveness of category assignments to documents by LSI, we adopted the break-even point (the arithmetic average of precision and recall) as reported in [12], and the total ("micro-averaged") precision P and recall R (defined in [13]).



Results

The categorization accuracy of LSI on the Reuters-21578 test set was reported in detail in [1]. Figure 1 shows the performance of the LSI algorithm on the Reuters-21578 test set with regards to categorization accuracy for various percentages of document degradation. As can be observed in Figure 1, the categorization accuracy falls off at a very slow rate for an introduced character error rate of 0 to 30 percent.

To compare results across the two Reuters data sets, the authors plotted baseline categorization accuracy versus the degradation levels of the degraded text. Figure 2 (Reuters-21578) and Figure 3 (Reuters RCV1-V2) show the plots of the overall categorization accuracy compared to the baseline versus the introduced character-error rate.

It is interesting to note that the two curves are very similar up to an introduced character-error rate of 15 percent. Examination of Figure 3 revealed an interesting artifact in the Reuters RCV1-V2 curve. After an introduced character-error rate of 15 percent, the curve for the Reuters RCV1-V2 test set falls off much faster than the curve for the Reuters-21578 test set. As noted previously, the Reuters RCV1-V2 test set contains 718,265 test documents versus 2,581 test documents for the Reuters-21578 test set. One explanation for the differences (and of continuing author investigation) is the larger number of new unique word combinations generated by degrading the RCV1-V2. There is a 33:1 ratio for the number of test documents versus training documents in the Reuters RCV1-V2 document set. This same ratio is 2.5:1 for the Reuters-21578 document collection (ModApte variation). It could be possible that a larger training set (than the one supplied with the RCV1-V2 test collection) is required at higher levels of document degradation.

Figures 4 and 5 show plots of the baseline categorization accuracy versus the percentage of corrupted words. The interest in these plots stemmed from the observation that at a 20% introduced character-error rate, roughly three-quarters of the words in a test document are corrupted. It can be observed that the LSI algorithm maintains a very

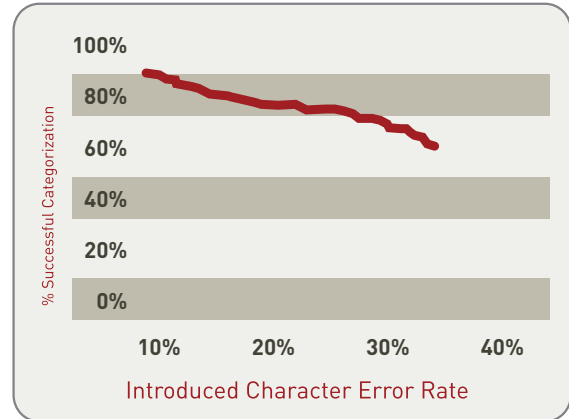


Figure 1. Reuters-21578 document categorization accuracy versus introduced character-error rate.

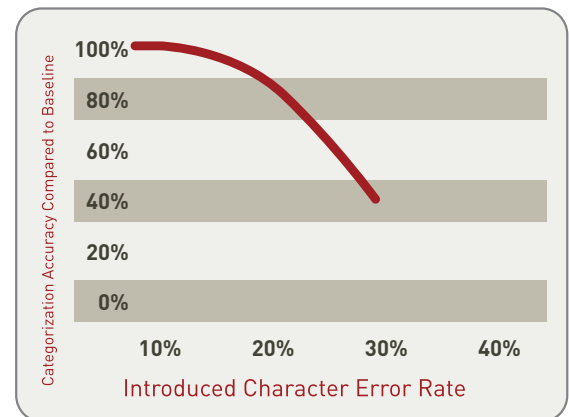


Figure 2. Reuters-21578 categorization accuracy compared to baseline versus introduced character-error rate.

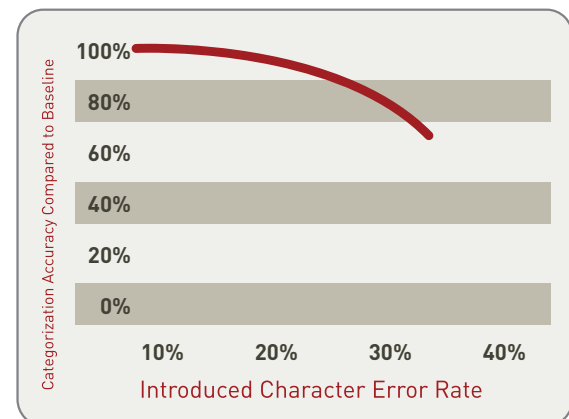


Figure 3. Reuters RCV1-V2 categorization accuracy compared to baseline versus introduced character-error rate.

flat curve out to a point where approximately 60% of the words are corrupted in a test document.

To the authors' knowledge, this level of information retrieval on noisy text has never been achieved using other technologies and is worthy of further investigation. Further examination of Figures 4 and 5 also shows the more rapid falloff at the tail of the curve between the two document collections as observed in Figures 2 and 3.

Conclusions

The authors find these results very encouraging. The evidence that the accuracy of the LSI algorithm falls off very slowly in categorization testing, even at high levels of text errors, indicates LSI could be an excellent solution to the general indexing of noisy text. In most applications where OCR text is targeted for information retrieval, the text conversion process most often requires a human-in-the-loop to either (1) extract document meta data which is indexed (rather than the OCR text) or (2) read the OCR text and provide human-intensive review and correction prior to full-text indexing. The results of this study indicate that reasonably accurate OCR conversion processes combined with LSI indexing could eliminate the human review process normally associated with the indexing of OCR text.

Future research will address the application of LSI in other areas, e.g., indexing of machine readable text as a result of speech-to-text conversion, the use of LSI-based text summarization of noisy text to replace human-generated meta data, and the potential of using LSI to mitigate the effects of noisy data on machine translation systems.

Acknowledgements

We thank Roger Bradford and Dr. Janusz Wnek for their useful comments when reviewing this paper

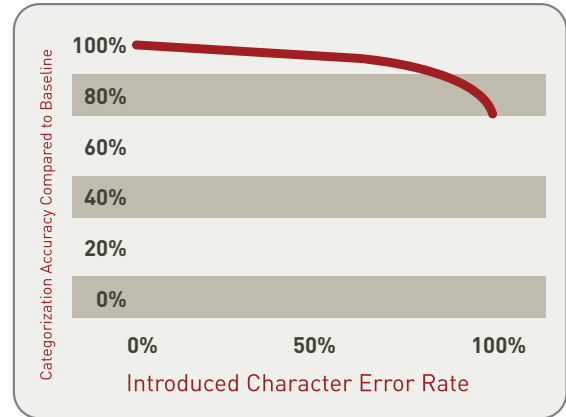


Figure 4. Reuters-21578 categorization accuracy compared to baseline versus the percentage of words corrupted.

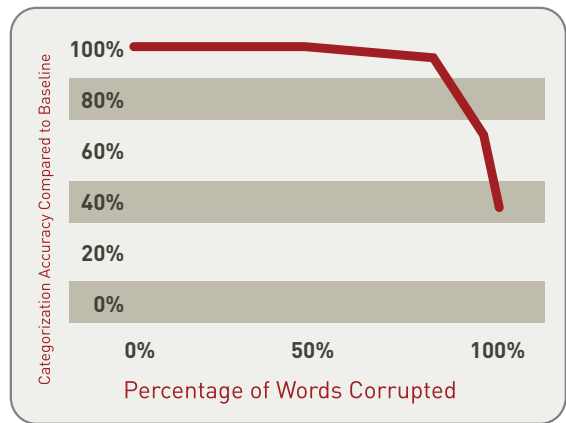


Figure 5. Reuters RCV1-V2 categorization accuracy compared to baseline versus the percentage of words corrupted.

References

- [1] A. Zukas and R. Price, Document Categorization Using Latent Semantic Indexing, In Proceedings: 2003 Symposium on Document Image Understanding Technology, Greenbelt, MD, April 2003 87-91.
- [2] J. Jeuring, Polytypic Pattern Matching, In Proceedings of the Seventh International Conference on Functional Programming Languages and Computer Architecture, ACM (1995) 238-248.
- [3] A. Singhal and F. Pereira, Document Expansion for Speech Retrieval, Proceedings SIGIR (1999) 34-41.
- [4] R. Baeza-Yates and G. Navarro, A Practical Index for Text Retrieval Allowing Errors, In CLEI, Volume 1 (1997) 273-282.
- [5] S. Deerwester et al., Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41(6), pp. 391-407.
- [6] T. Landauer and D. Lanham, Learning Human-like Knowledge by Singular Value Decomposition: a Progress Report, Advances in Neural Information Processing Systems 10 (1998), Cambridge: MIT Press, pp. 45-51.
- [7] K. Nigam, Using unlabeled data to improve text classification, PhD. Thesis, Carnegie Mellon University, May 2001.
- [8] S. Rice, F. Jenkins, T. Nartker, The Fifth Annual Test of OCR Accuracy, Technical Report TR-96-01, Information Science Research Institute, University of Nevada, Las Vegas, NV, 1996.
- [9] D. Lewis. Reuters-21578 Text Categorization Test Collection, Distribution 1.0, README file (Version 1.2), Manuscript, September 26, 1997, <http://www.daviddlewis.com/resources/testcollection/reuters21578/readme.html>.
- [10] D. Lewis, Y. Yang, T. Rose, F. Li, RCV1: A New Benchmark Collection for Text Categorization Research, Journal of Machine Learning Research, 5(2004):361-397, 2004, <http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf>.
- [11] H. Baird, Document Image Models and Their Uses, ICDAR, Proc. Intl. Conf. Document Anal. Recog, (1992) 62-67.
- [12] Y. Yang and X. Liu, A Re-examination of Text Categorization Methods, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), (1999) 42-49.
- [13] Y. Yang, An Evaluation of Statistical Approaches to Text Categorization, Journal of Information Retrieval, Volume 1, No. 1/2 (1999) 67-88.

Appendix A

China wheat purchases reach 80 pct of 1996 target. China's total state purchases of wheat reached 18.53 million tonnes by the end of August, accounting for 80.26 percent of the target for 1996, the official People's Daily said on Sunday. Of the total, purchases of wheat at state-fixed prices were 13.11 million tonnes, more than 90 percent of the year's target, the newspaper said. The state also buys at market prices. The state bought 1.95 million tonnes of rapeseed, fulfilling 66.21 percent of state plans, it said. China's grain output was expected to be a record 475 million tonnes in 1996, the Xinhua news agency said on Sunday.

Figure A-1. Sample document with no errors.

China wheat purchases reach 80 pct of 1996 target. China's total state purchases of wheat reached 18.53 million tonnes by the end of August, accounting for 80.26 percent of the target for 1996, the official People's Daily said on Sunday. Of the total, purchases of wheat at state-fixed prices were 13.11 million tonnes, more than 90 percent of the year's target, the newspaper said. The state also buys at market prices. The state bought 1.95 million tonnes of rapeseed, fulfilling 66.21 percent of state plans, it said. China's grain output was expected to be a record 475 million tonnes in 1996, the Xinhua news agency said on Sunday.

Figure A-2. Sample document with 5% errors.

China wheat purchases reach 80 percent of 1996 target. China's total state purchases of wheat reached 18.53 million tonnes by the end of August, accounting for 80.26 percent of the target for 1996, the official People's Daily said on Sunday. Of the total, purchases of wheat at state-fixed prices were 13.11 million tonnes, more than 90 percent of the year's target, the newspaper said. The state also buys at market prices. The state bought 1.95 million tonnes of rapeseed, fulfilling 66.21 percent of state plans, it said. China's grain output was expected to be a record 475 million tonnes in 1996, the Xinhua news agency said on Sunday.

Figure A-3. Sample document with 15% errors.

China wheat purchases reach 80 percent of 1996 target. China's total state purchases of wheat reached 18.53 million tonnes by the end of August, accounting for 80.26 percent of the target for 1996, the official People's Daily said on Sunday. Of the total, purchases of wheat at state-fixed prices were 13.11 million tonnes, more than 90 percent of the year's target, the newspaper said. The state also buys at market prices. The state bought 1.95 million tonnes of rapeseed, fulfilling 66.21 percent of state plans, it said. China's grain output was expected to be a record 475 million tonnes in 1996, the Xinhua news agency said on Sunday.

Figure A-4. Sample document with 30% errors.