

Recognize, Categorize, and Retrieve

Kazem Taghva, Thomas A. Nartker, and Julie Borsack

Information Science Research Institute

University of Nevada, Las Vegas

Abstract

A successful text categorization experiment divides a textual collection into pre-defined classes. A true representative for each class is generally obtained during training of the categorizer.

In this paper, we report on our experiments on training and categorization of optically recognized documents. In particular, we will address the issues regarding the effects OCR errors may have on training, dimensionality reduction, and categorization. We further report on ways that categorization may help error correction and retrieval effectiveness.

1 Introduction

Retrieving relevant information from a large textual corpus is a challenging task and entails many manual and automated efforts. Query construction and training of categorizers are two prominent examples of manual efforts while document indexing and clustering would be considered automated. Another expensive but widely used manual effort is the assignment of a controlled vocabulary to pre-defined categories for documents in the corpus. Expert searchers can then use these categories and controlled vocabularies to formulate more compact and effective queries [4].

Although these manual efforts will be needed for the foreseeable future, it is of great interest to establish automated techniques to extract controlled vocabularies and to assign categories to new documents based on similar documents in the corpus. One of the proven approaches is to use an already categorized set of documents to build categorizers for future document classification.

In a large textual repository such as the Licensing Support Network (LSN) being built by the Nuclear Regulatory Commission (NRC), many of the documents are recognized using commercial OCR engines. The OCR errors typically result in an index which is considerably larger than keyed text. Also, for some queries, one may not be able to find cer-

tain relevant documents due to poor quality OCR that generates a high number of errors[13, 14]. Although, for contemporary OCR systems, these errors do not affect average precision and recall, they may produce variations in ranking[12]. But in general, studies show that one can work with OCR text in an information retrieval (IR) environment with few adverse consequences.

In a similar situation, one must know what effects these errors may have on automated text classification. In particular, we want to know what effects (if any) these errors may have on training the classifier to build the categorizer. We would also like to discover if errors cause improper categorization for new documents.

In section 2 of this paper, we give a brief introduction to the work done in the area of OCR and IR. In Section 3, we describe both the Bernoulli and multinomial Bayes categorizers. Sections 4 and 5 report on our categorization experiments in the presence of OCR errors. Finally, in section 6 we give our conclusion and future work.

2 OCR and Information Retrieval

In the early 1990's the use of OCR devices became more widespread for the conversion of printed material to electronic form. At this time, ISRI was heavily involved in OCR system comparison, testing, and research. What became clear was that eventually, this data was to be loaded into an IR system for subsequent retrieval. Initially, it was thought that manual correction was required to bring this OCR generated text to a level of "retrievability." In the case of the LSN, text accuracy was to be no less than 99.8% accurate.

ISRI questioned the necessity of this level of accuracy and the "Noisy Data" experimentation began. Since our first publication on this topic in *JASIS* in 1994[15], the Information Science Research Institute (ISRI) has performed hundreds of experiments involving optically recognized data to discover its ef-

fect on related technologies. Nearly all of our studies have pointed to the same conclusion: In general, using OCR text has little effect on average precision and recall when compared to re-keyed or manually corrected text. This was a consequential result, in particular for collections such as the LSN, because re-keying millions of pages would have been a tedious and expensive task.

With this result in mind, some of our other studies did a little more investigation into the effects of OCR on IR. For example, we found that the index of an OCR collection can be as much as five times the size of a clean data set and that most of this overhead was of no value for retrieval. Further, formulas used to calculate term weights can be affected in several ways by erratic term frequencies in OCR generated text [12, 13]. This in turn affects document ranking. We also found that short documents in particular are affected because of the lack of redundancy in the text. So although our results on average precision and recall holds in nearly every test we performed, there are some considerations when a collection is made up of OCR data.

The categorization experiments we report on here are similar to our IR experiments we have done in the past. First, the data set is derived from documents pertaining to the LSN; the documents tend to be long journal documents with a scientific flavor. Classification software such as *BOW* [8] constructs an index and then uses its categorizers (based on the training sets) as vectors to determine the classification of incoming documents. In this sense, a categorizer is similar to a vector query in the IR domain.

But once a document has been classified properly, even if the OCR is poor, more information about that document is now known. In this case, it may be feasible to apply more specific and exhaustive automatic error correction to documents within a category. We see ample potential in the continued research in OCR, categorization, and other related technologies.

3 Probabilistic Classifiers

There are two distinct approaches to automatic text classification. The first approach is based on machine learning techniques. In this method, the system is given a set of training documents for each category. These documents are used to typically generate a set of propositional Horn clauses that will be used to classify future documents [3, 10, 1]. The second approach is based on traditional IR techniques. In this method, the training documents are used to form an ideal document which represents each category. These ideal documents are known as *categorizers* [8, 7, 6, 4]. The system uses similarity

measures between incoming documents and the categorizers to classify these new documents properly. Our experiments only pertain to the latter approach.

Let $V = \{v_1, v_2, \dots, v_{|V|}\}$ be the set of words in a lexicon. Each document, and each categorizer, can be represented as a vector of the form $(w_1, w_2, \dots, w_{|V|})$, where each component w_t of this vector represents the weight of the term v_t in the document and categorizers. In its simplest form, w_t can be either 0 or 1. In this case, the weight represents the presence or absence of the term v_t in the document. This weight though can carry more information such as the frequency of the term in the document.

Now, let $C = \{c_1, c_2, \dots, c_{|C|}\}$ and $D = \{d_1, d_2, \dots, d_{|D|}\}$ be sets of categories and training documents, respectively. Each category c_j , is represented by a vector of the above form, where the weight, w_t , is calculated from using term frequencies based on the training set of the documents. Using the naive Bayes assumption, that the probability of each word occurring in a document is independent of the occurrence of other words in a document [9], then these weights can be easily calculated. In our experiments, we focus on both the Bernoulli and multinomial methods.

In the Bernoulli method, the frequency of the words do not play any role. Hence, each document is represented by a vector of the form $d_i = (B_{i1}, B_{i2}, \dots, B_{i|V|})$, where each B_{it} is either 1 or 0. In this case, the weight of each component of the categorizer c_j is calculated using the following formula:

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{it}P(c_j|d_i)}{2 + \sum_{i=1}^{|D|} P(c_j|d_i)} \quad (1)$$

In other words, the weight of the term w_t given category c_j is obtained by dividing the number of documents containing the term v_t in category c_j by the total number of documents in category c_j .

Now, the probability of a new document d_i belonging to category c_j is calculated by the following formula:

$$P(d_i|c_j) = \prod_{t=1}^{|V|} (B_{it}P(w_t|c_j) + (1 - B_{it})(1 - P(w_t|c_j))) \quad (2)$$

In the multinomial model, the frequency and the length of the document (i.e. the number of words in the document) play a role. In this setting, a document d_i is represented with a vector of the form $d_i = (N_{i1}, N_{i2}, \dots, N_{i|V|})$, where N_{it} is the frequency of the term v_t in the document d_i . If we use the notation $|d_i|$ for the length of a document, then

the following formulas represent the corresponding calculations for the multinomial model.

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{it}P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is}P(c_j|d_i)} \quad (3)$$

$$P(d_i|c_j) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c_j)^{N_{it}}}{N_{it}!} \quad (4)$$

In practice, there are various methods to decrease the dimension of $|V|$. These methods are known as *dimensionality reduction* techniques which tend to improve the performance of a categorization system. These are further discussed in Section 4.

4 Experimental Environment

Unlike most classification experiments, we do not use a standard categorized collection like *Reuters* [5]. Instead, our testing is dictated by the needs of the Department of Energy and the LSN. They have millions of optically recognized documents and terabytes of email that need to be classified into *Regulatory Guideline 3.69* [2] categories. This guideline determines which documents and email messages are required for the licensing proceedings of the High-Level Radioactive Waste Repository. The documents we use in these experiments are a subset of this collection. Our experimental environment consists of:

BOW text classifier from CMU: a

statistically-based text categorization system applying the probabilistic naive Bayes model [8, 9]. BOW offers several ways to reduce dimensionality of classes. They include:

- Default:** removes no words from the vocabulary.
- Document Count:** removes words that occur in N or fewer documents. In our experiments, $N = 3$.
- Occurrence Count:** removes words that occur less than N times. In our experiments, $N = 10$.
- Information Gain:** removes all but the top N words by selecting words with the highest information gain. We use $N = 10,000$ in our experiments.

DOE documents: studies, reports, plans, correspondence, etc. that may be potentially relevant to the licensing of the High-Level Radioactive Waste Repository. All of these documents are optically recognized.

3.69 Topical Guideline categories: a hierarchical guide of topics that encompass potential licensing issues. Following is the selected categories we use for our experiments:

Collection Statistics	
Document count	138
Number of pages	9015
Average document length (pages)	65
Median document length (pages)	37

Table 1: Experimental collection statistics

- 02.1 The Natural Systems of the Geologic Setting: Geologic Systems
- 02.2 The Natural Systems of the Geologic Setting: Hydrologic Systems
- 02.4 The Natural Systems of the Geologic Setting: Climatological and Meteorological Systems
- 04.1 Engineered Barrier Systems: Waste Package
- 12.1 Geologic Repository Environmental Impact Statement: Environmental
- 12.2 Geologic Repository Environmental Impact Statement: Socioeconomic
- 12.3 Geologic Repository Environmental Impact Statement: Transportation

In Table 1 we give some statistics on our collection.

5 Effects of OCR on Document Classification

Our goal was to formulate experiments that would give us the most insight into what effect OCR errors may have on document classification. Broadly, there are two ways in which errors can influence categorization. First, by introducing errors into the training set, and second, by reducing the ability of incoming documents to get categorized correctly. We report on four experiments that help explain both these possibilities.

Good Training/Bad Test Set: In this experiment, the training set, although uncorrected OCR, was selected for its good quality. The test set was just the opposite; it was selected for its poor OCR quality. These experimental runs are labeled E1.

Mixed Training/Mixed Test Set: This experiment used the same set of documents used in E1, but documents were selected randomly from the complete set for both training and testing. This group of runs we label E2.

Good Training/Auto-Corrected: This experiment is labeled E3. E3 is exactly the same as E1, except that two difficult-to-categorize documents were first run through *MANICURE*, a system we built to improve recognized documents prior to classification or retrieval [16].

Bernoulli	E1	E2	E3	E4
Default	32.35	42.03	32.35	32.35
Document Count	64.71	56.52	64.71	64.71
Information Gain	70.59	59.42	70.59	70.59
Occurrence Count	64.71	57.97	64.71	64.71

Table 2: Average accuracy rates for each dimensionality reduction

Multinomial	E1	E2	E3	E4
Default	94.12	84.06	94.12	94.12
Document Count	97.06	85.51	97.06	97.06
Information Gain	94.12	86.96	97.06	97.06
Occurrence Count	97.06	85.51	97.06	97.06

Table 3: Average accuracy rates for each dimensionality reduction

Good Training/Manually-Corrected: This set of experimental runs, labeled E4, is the same test as E3 except that the two documents in E3 are *manually corrected*.

In each experiment described above, we perform several runs based on both the Bernoulli and multinomial probability models. In addition, each experiment includes a limited vocabulary run that applies the multinomial probability technique. The limited vocabulary “list” consists of several merged dictionaries that include domain specific terms that a general dictionary may have missed in the indexing process. It is comprised of several general dictionaries, geologic and radiologic specific dictionaries, an LSN specific thesauri, and contains 413,216 words. Limiting the vocabulary to pre-defined control terms is a common method of indexing in both retrieval and categorization.

For each run, we apply all the dimensionality reductions described in Section 4. Tables 2, 3, and 4 report the average *accuracy rates* for the various runs. The accuracy rate of a class is the ratio of the *number of correct decisions* made by the system over the *total number of documents in the class*.

5.1 Bernoulli vs. Multinomial

Note first that the Bernoulli results do not compare to either of the multinomial runs. We know from previous research [9] that with longer documents, like the ones we use here, multinomial typi-

Limited Vocabulary	E1	E2	E3	E4
Default	94.12	85.51	94.12	94.12
Document Count	97.06	84.06	97.06	97.06
Information Gain	94.12	85.51	97.06	94.12
Occurrence Count	94.12	82.61	97.06	94.12

Table 4: Average accuracy rates for each dimensionality reduction

Term	Bernoulli Weight	Multinomial Weight
southern	0.750000	0.000138

Table 5: Comparison of weight for OCR error

Dimensionality Reduction	% of Misspellings
Default	48%
Document Count	8%
Information Gain	11%
Occurrence Count	8%

Table 6: Percentage of misspellings for each dimensionality reduction

cally produces better results than Bernoulli. These results mirror this research. We do believe however, based on these results here and results from other experiments we have done [17], that the accuracy rate is particularly poor due to the use of OCR text. Table 5 shows an obvious OCR error in both the Bernoulli run and the multinomial run and its respective probability weights. For Bernoulli, this term is given the highest weight in the category while for Multinomial, this error’s probability is only 1% of the highest ranked term in the category. Examples like this can be found throughout the Bernoulli categories.

5.2 Default vs. Dimensionality Reductions

As with other classification experiments[11], our results show that dimensionality reduction improves categorization. Dimensionality reduction eliminates terms that contribute the least amount of information for the categories. With respect to OCR text, this includes terms that are misrecognized by the device and contribute no value to the category. Examples of obvious OCR errors that were removed due to dimensionality reduction include: `aluminurn`, `tomography`, `sufface`, `therinal`, `requirements`.

We believe that with OCR text, reduction is not an option, it is a requirement. Table 6 reveals the drop in the percentage of misspellings included in the categories when dimensionality reduction is applied.¹ Removal of OCR errors through dimensionality reduction clearly improves the accuracy of categorization.

¹This table excludes the limited vocabulary runs which of course had no misspellings.

Document/Category Changes	
Corrected words	195
Garbage strings removed	19,772
Net improvement to category 02.2	5%

Table 7: Improvements made by MANICURE

5.3 Good Training vs. Mixed Training

Recall that Experiment E1 uses all “good training” documents and a poorly recognized test set and E2 uses a randomly selected “mixed training” and the complement for its test set. Note that in every run except Bernoulli Default, the average accuracy results of E1 are significantly better than in E2. We believe that these improved results are a function of using good quality OCR for training vs. a randomly selected training set from mixed quality documents.

Although more analysis may be required to verify this conclusion, these consistently better accuracy rates point to the fact that although OCR-generated text may have little or no effect in general when incoming documents are being classified, the selection of good quality OCR documents for training is essential.

5.4 Classifying Poor OCR

We discovered in several of our experiments with information retrieval and OCR that some poorly recognized documents were unretrievable without some corrective intervention [15, 14]. This dilemma is paralleled in categorization. In nearly all our experimental runs, there were two poorly recognized documents that just couldn’t seem to get categorized properly. We wanted to see if correcting errors in these documents would help. Experiment E3 applies several algorithms within a single pre-processing system, MANICURE, to see if automated OCR cleanup and error correction could improve classification. In fact, one of the two documents did get categorized correctly after running the documents through MANICURE.

MANICURE (Markup ANd Image-based Correction Using Rapid Editing) [16] applies several algorithms to improve OCR-generated text that not only correct misrecognized terms but also remove “garbage strings” and repetitive text (like headers and footers). The improvements to these two documents after being run through MANICURE appear in Table 7.

The fact that automatic correction helped classify this document correctly is just part of the story. We also report on the improvement to the category itself. Both of these poorly recognized documents belonged to a single category. After MANICURE and

retraining, the percentage of correctly spelled category terms also improved. Of the 118 changed terms, seven more were correct when compared to the non-manicured runs in E1. Although this increase may seem slight, only two documents were run through MANICURE. Additional document processing may prove even more beneficial.

Full manual correction of these documents offered no additional categorization improvement over the automatically MANICURE’d runs.

6 Conclusion and Future Work

Document classification is not an exact science and rarely produces 100% accuracy even with clean textual documents. The results from these experiments show that high accuracy can be attained even when OCR documents are being classified. By comparing experiments using training and test sets with known characteristics, we have identified a few elements that improve categorization.

- Multinomial techniques produce significantly better results for OCR documents than does Bernoulli. We attribute this to the value of weighting based on term frequency in the collection, categories, and incoming documents.
- Good optically recognized documents are essential for training. The difference between using good OCR and randomly selected documents from the full set was pronounced. Category term selection and weighting is heavily influenced by statistics both in the collection and in the training documents. This influence manifests itself in the accuracy of incoming document classification.
- Dimensionality reduction is highly recommended. Reduction rids the categories of insignificant terms, which in this case, includes hundreds of misspellings and garbage strings produced by the OCR. As with IR, these terms have no value. For categorization, these words are a detriment to proper document placement. Of course, in general, reduction techniques will improve categorization. Several should be tried so that the accuracy is maximized.
- Unless a controlled vocabulary shows marked improvement over free text categorization, we do not view it as highly beneficial. Even though our dictionary was quite extensive and specific to our collection’s domain, none of our experiments showed improved results for these runs. This can undoubtedly be attributed to important terms and proper names that are not included in the dictionary.

- If a document is poorly recognized because of OCR errors, it may never get classified properly. This was an issue for IR as well. In some cases, if enough of the errors are corrected, proper classification is the result. Some pre-processing may be required for certain poorly recognized documents.

Most of what we learned through our experimentation is that by applying good classification techniques, improvement in results can be expected. But more than that, for OCR text, if these techniques are not applied, results will be inferior.

References

- [1] William W. Cohen and Haym Hirsh. Joins that generalize: text classification using WHIRL. In Rakesh Agrawal, Paul E. Stolorz, and Gregory Piatetsky-Shapiro, editors, *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining*, pages 169–173, New York, 1998. AAAI Press, Menlo Park.
- [2] Nuclear Regulatory Commission. Regulatory guide 3.69. <http://www.nrc.gov/NRC/RG/03/03-069.html>, 1996.
- [3] P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt. TCS: a shell for content-based text categorization. In *Proc. of CAIA-90, 6th IEEE Conf. on Artificial Intelligence Applications*, pages 320–326, Santa Barbara, CA, 1990.
- [4] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proc. of ECML-98, 10th European Conf. on Machine Learning*, pages 4–15, Chemnitz, Germany, 1998.
- [5] David D. Lewis. Reuters-21578 text categorization test collection, distribution 1.0. September 1997.
- [6] M. E. Maron. Automatic indexing: An experimental inquiry. *Journal of the ACM*, 8:404–417, 1961.
- [7] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.
- [8] Andrew McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/~mccallum/bow>, 1996.
- [9] Andrew McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.
- [10] Isabelle Moulinier and Jean-Gabriel Ganascia. Applying an existing machine learning algorithm to text categorization. In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors, *Connectionist, statistical, and symbolic approaches to learning for natural language processing*, pages 343–354, Heidelberg, DE, 1996. Springer Verlag.
- [11] Fabrizio Sebastiani. Machine learning in automated text categorisation. *ACM Computing Surveys*, 2001. to appear.
- [12] Kazem Taghva, Julie Borsack, and Allen Condit. Results of applying probabilistic IR to OCR text. In *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pages 202–211, Dublin, Ireland, July 1994.
- [13] Kazem Taghva, Julie Borsack, and Allen Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Inf. Proc. and Management*, 32(3):317–327, 1996.
- [14] Kazem Taghva, Julie Borsack, and Allen Condit. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transactions on Information Systems*, 14(1):64–93, January 1996.
- [15] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. The effects of noisy data on text retrieval. *J. American Soc. for Inf. Sci.*, 45(1):50–58, January 1994.
- [16] Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth. The MANICURE document processing system. In *Proc. IS&T/SPIE 1998 Intl. Symp. on Electronic Imaging Science and Technology*, San Jose, CA, January 1998.
- [17] Kazem Taghva, Tom Nartker, Julie Borsack, Steve Lumos, Allen Condit, and Ron Young. Evaluating text categorization in the presence of ocr errors. In *Proc. IS&T/SPIE 2001 Intl. Symp. on Electronic Imaging Science and Technology*, pages 68–74, San Jose, CA, January 2001.