Toward a Multi-Sensor Neural Net Approach to Automatic Text Classification*

Venu Dasigi

Department of Computer Science and information Technology
Sacred Heart University
Fairfield, CT.

and

Reinhold Mann

Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Presentation To Be Made At:

IFIP'96 World Conference on Advanced IT Tools, Intelligent Systems Track

Canberra, Australia
September 2-6, 1996

***Oak Ridge National Laboratory, managed by Lockheed Martin Energy Research Corp. for the U.S. Department of Energy under contract number DE-AC05-96OR22464.**

Submitted to IFIP-96

Toward a Multi-Sensor Neural Net Approach to Automatic Text Classification

Venu Dasigi*
Department of Computer Science and Information Technology
Sacred Heart University
Fairfield, CT 06432-1000
e-mail: dasigi@shu.sacredheart.edu
and
Reinhold Mann
Intelligent Systems Section
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6364
e-mail: mannrc@ornl.gov

January 26, 1996

Abstract

Many automatic text indexing and retrieval methods use a term-document matrix that is automatically derived from the text in question. Latent Semantic Indexing, a recent method for approximating large term-document matrices, appears to be quite useful in the problem of text information *retrieval*, rather than text *classification* [Deerwester, et al., 90]. Here we outline a method that attempts to combine the strength of the LSI method with that of neural networks, in addressing the problem of text classification. In doing so, we also indicate ways to improve performance by adding additional "logical sensors" to the neural network, something that is hard to do with the LSI method when employed by itself. Preliminary results are summarized, but much work remains to be done.

Broad Classification: Information Retrieval, Neural Networks

*Send correspondence to this author.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Toward a Multi-Sensor Neural Net Approach to Automatic Text Classification

1 Introduction & Background

Most contemporary approaches to information retrieval use terms contained in a text document directly as indexes into the document. "Vector-based" approaches view documents as vectors of such terms. Thus, a "library" of documents is represented as a term-document matrix, where the entries represent the frequency of each term in each document. Such term-document matrices tend to be very large and sparse.

Latent Semantic Indexing (LSI) is a recent method that captures the "latent semantic structure" of documents, as indicated in a term-document matrix[Deerwester, et al., 90]. The large sparse matrix is reduced into three relatively small matrices by singular value decomposition (SVD), whose product approximates the original sparse matrix. Our work is an initial effort to combine the valuable ideas OF LSI with the powerful pattern-matching and learning capabilities of neural networks. A major stumbling block in applying neural networks to most IR applications has been that the size of a typical IR problem results in impractically large neural networks. An LSI-based approach may be used to address the issue.

2 A Neural Net Approach with Multiple Input Sensors

Specifically, in this initial effort, we focused on two main goals. First, create input to a neural network that is LSI-based, so that the size of the neural net will be practical. Further, a second goal is to see if additional sensors can be added easily to the neural net input, to give improved results. The relationship between the LSI component and the neural network is symbiotic. The LSI-based input compresses the input to the neural network to a much smaller size. Further, LSI is based on a solid mathematical theory, adding strength to the resulting system. For its part, the neural network adds trainability to the LSI-based method, and also makes it possible to integrate other sensors to supplement the LSI-based input.

A straightforward, but simple-minded input vector for the neural network would be a document represented as a vector of *all-possible* term frequencies, which generally number in the thousands. LSI work suggests a way to represent a document using around a hundred "factors", derived from the much longer term vector and the SVD of a "reference matrix"[1]. The developers of LSI indicate that a query may be viewed as a pseudo-document and may be represented by a vector of a chosen number of factors [Deerwester, et al., 90]. First a reference term-document sparse matrix X is derived from the library of documents that are of interest. This matrix is split into three matrices by SVD, so that

$$X = T.S.D'$$

Here, X is a txd matrix, where t is the numbers of distinct terms (word roots) and d is the number of documents in the reference collection. The order of T is txk, that of S, which is a *diagonal* matrix, is kxk, and that of D is dxk, where k is the chosen number of factors. Now, the 1xk pseudo-document vector D_Q corresponding to a 1xt query vector Q may be derived simply as:

$$D_Q = Q.T.S^{-1}$$

In this work, we use this same idea to squash any *regular* document vector into a 1xk vector that serves as input to the neural network. *The only care that must be exercised is to make sure that the reference term-*

[1] A *reference matrix* is the term-document matrix of a *reference library/collection* of documents. A reference library is simply the collection of documents that "adequately" represents all concepts of interest.

document matrix that is used as the starting point is one that "adequately" represents all concepts of interest. Note that this requirement is no more stringent than would be required in the standard LSI approach.

Figure 1 shows a high level view of how the proposed method works. The input to the system is an individual document that needs to be classified into one of several categories. Different logical sensors are applied to the document, constituting different kinds of preprocessing to derive salient features. The first such sensor compresses the 1xt input term vector into a 1xk vector as explained above. The output is an indication of the category to which the document belongs.

3 Experiments and Results

Our initial focus was on hundreds of AP news wire stories. Here, some example output categories are: accidents, crime, business and finance, culture, politics and government, weather, obituary, etc. A major stumbling block in this work was the manual categorization of documents for training and testing purposes. Within the time frame of this initial effort, we could manually categorize only 480 news stories, which is what we worked with.

Below, we describe the results obtained using two different neural networks, one using just LSI-based inputs and a second one using both LSI-based inputs and those derived from a simple second sensor. This second sensor is based on simple profiles of the output categories. In this simple-minded sensor, each category profile is simply a set of keywords characterizing that particular category. There is one output from this logical sensor corresponding to each category. Each output simply represents what fraction of the terms in the given document match the category profile.

Of the 480 documents, the first 380 were used as the "reference library". That is the term-document matrix used in *all* LSI/SVD operations has 380 columns. The number of SVD "factors" used in this work was 112. The neural net inputs were also based on this matrix. The neural net was a simple feedforward net with back propagation, and used the delta rule for learning and the *tanh* transfer function. It was tested in two configurations: one with 112 inputs (just based on LSI alone) and another with 112 LSI-based inputs *plus* another 10 inputs based on simple category profiles. Both configurations used 10 output units, one for each category.

We compared the multi-sensor neural net approach against an LSI-based classification, which is done by first identifying the best matching reference document and then looking up its category. When the LSI method was used by itself to classify the 100 documents that are outside the reference library, the results were somewhat surprising. LSI classified the reference library documents with a perfect 100% accuracy, the percentage of correct results when the 100 new documents were used dropped to 54%.

Table 1 summarizes the performance of the 112-input single sensor neural net on each of the dozen different data sets, and Table 2 is for the 122-input, two sensor neural net. The percentages of correct results shown in the tables represent the peak performance that did not get any better with more iterations. For testing the neural nets, inputs were created for all the 480 documents, including a correct answer for each case. This "answer" was to be used either for training the neural net or for comparison, in the case of a testing. The data were cross validated by generating a dozen pairs of training and test inputs with a 90%-10% split.

It may be seen that in the case of individual data sets, there was an improvement in performance with 10 out of the 12 data sets in the two sensor neural net, compared to the single sensor version, with marginal decrease of performance in the other two cases. But in one of these two cases (data set 2), the superficially

Data Set No.	No. of Iterations	Percent Correct with Test Data	Percent Correct with Training Data
1	48K	58	80.70
2	16K	72	77.67
3	64K	72	76.98
4	64K	62	77.21
5	32K	62	76.05
6	48K	62	76.98
7	80K	62	80.47
8	48K	58	78.14
9	64K	68	77.44
10	48K	60	78.14
11	48K	64	78.84
12	32K	68	76.05

Table 1: Results of Classification with the Single Sensor Neural Net

Data Set No.	No. of Iterations	Percent Correct with Test Data	Percent Correct with Training Data
1	48K	62	85.11
2	48K	70	84.42
3	32K	76	84.19
4	64K	64	83.72
5	32K	64	84.88
6	16K	66	83.26
7	32K	66	84.88
8	16K	62	84.19
9	32K	70	83.49
10	32K	58	83.02
11	16K	64	80.47
12	32K	72	83.02

Table 2: Results of Classification with the Two Sensor Neural Net

better performance of the single sensor neural net decreased a few percent with more training. These anomalies can perhaps be attributed to the simple-mindedness of the second sensor that was employed.

With the neural nets, although the need for more training inputs is obvious, there is a clear improvement of classification results compared to the LSI method by itself. And the two sensor version, even with its simple-minded second sensor, performs better in most cases than the single sensor version. Other researchers appear to have evaluated LSI-based approaches, too [Schuetze, Hull and Pederson, 95]. Our approach differs from theirs in using a reference library, and in employing multiple sensors. It may be noted that SVD is computationally very expensive, but our approach performs an SVD just once - on the reference collection.

Ongoing and future work plans include further training of the nets, developing other more informative sensors, possibly using natural language techniques (for faster training), strengthening the category profiles in the second sensor by various means, etc.

References

[Deerwester, et al., 90] Deerwester, S., S. Dumais, G. Furnas, T. Landauer and R. Harshman, Indexing by Latent Semantic Analysis *Journal of the American Society for Information Science*, 41(6), pp. 391-407, 1990.

[Schuetze, Hull and Pederson, 95] Schuetze, H., D. Hull and J, Pedersen, A Comparison of Classifiers and Document Representations for the Routing Problem, *Proc. of SIGIR-95*, pp. 229-237, 1995.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Multi-Sensor/Neural Net Intelligent Information Retrieval

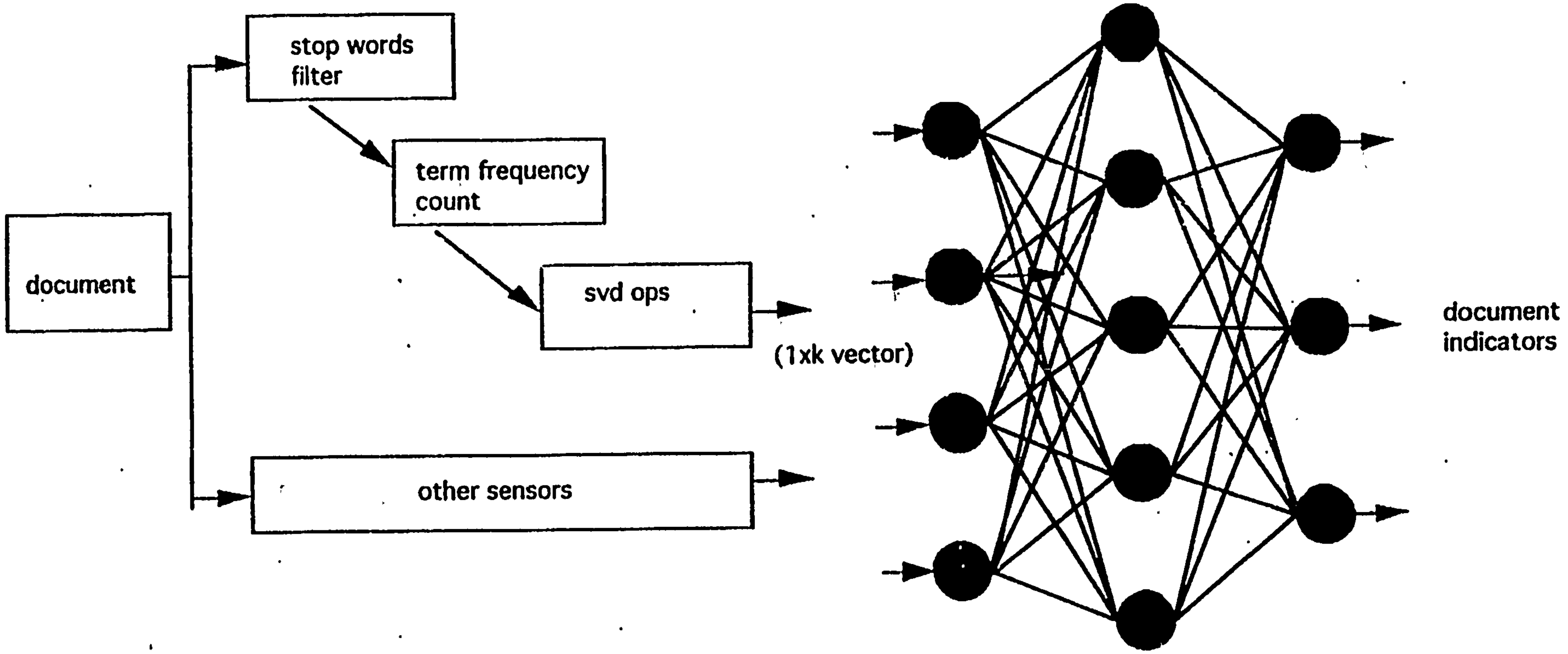

Figure 1. Schematic of multi-sensor/neural network approach to intelligent text retrieval and document classification.