Characterizing Large Text Corpora Using a Maximum Variation Sampling Genetic Algorithm

Robert M. Patton
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN USA 37831
Ph: 1-865-576-3832

pattonrm@ornl.gov

Thomas E. Potok
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN USA 37831
Ph: 1-865-574-0834

potokte@ornl.gov

ABSTRACT

There exists an enormous amount of information available via the Internet. Much of this data is in the form of text-based documents. These documents cover a variety of topics that are vitally important to the scientific, business, and defense/security communities. Currently, there are a many techniques for processing and analyzing such data. However, the ability to quickly characterize a large set of documents still proves challenging. Previous work has successfully demonstrated the use of a genetic algorithm for providing a representative subset for text documents via adaptive sampling. In this work, we further expand and explore this approach on much larger data sets using a parallel Genetic Algorithm (GA) with adaptive parameter control. Experimental results are presented and discussed.

Categories and Subject Descriptors

I.7.0 [Document and Text Processing]: General

General Terms

Algorithms, Performance, Design, Experimentation

Keywords

Text analysis, parallel genetic algorithm, intelligent agents.

1. INTRODUCTION

The focus of this work is on analysis of textual data. To characterize effectively a large and streaming set of news articles, the following goals are proposed in order to create an algorithm that provides a useful result to a human analyst, it must:

1. Be capable of sufficiently reducing the data to manageable levels

2. Be able to provide a fast and accurate processing of massive amounts of data

3. Efficiently and effectively deal with duplicate data

4. Be able to work with streaming data

5. Not require prior knowledge concerning the data set

To address the five goals identified, this work describes a genetic algorithm that performs an adaptive, maximum variation sampling technique. It is well known that a genetic algorithm performs very well for large search spaces and is easily scalable to the size of the data set. In addition, GA's are also particularly suited for parallelization [1][2]. To better understand the need for scalability and the size of the search space in this problem domain, consider a document set with only 10,000 news articles in it. Now, suppose an analyst needs to reduce this data set to 200 representative articles (only 2% of the entire data set). In that case, there are approximately 1.7 x 10424 different combinations of documents that could be used to create a single sample. Clearly, a brute force approach is unacceptable. In addition, many of the combinations would consist of duplicate data, which would lower the quality of the result for the analysts. Ultimately, an intelligent and scalable approach such as a genetic algorithm is needed to help address goals 1 and 2. As demonstrated by Mutalik [3], a parallel genetic algorithm is well suited to a combinatorial optimization problem.

The remainder of the goals is addressed via the MVS technique. Since this technique is searching for data points that maximize diversity, this approach will avoid duplicate data from being included in the results. In addition, it does not require that all duplicate data be first identified. This is a tremendous advantage since duplicate data can often be a significant portion of the data set. Furthermore, the MVS technique does not require the data set to remain static, but a dynamic set is easily handled. Finally, the MVS technique is a sampling technique and therefore does not require prior knowledge of the data set, and will naturally reduce the data set to the appropriate size as determined by the analysts.

The following sections will discuss the details of our work, the test data used and experiments performed, and the results obtained.

2. MVS-GA DESIGN

Two of the most critical components of implementing a GA are the encoding of the problem domain into the GA population and the fitness function to be used for evaluating individuals in the population. To encode the data for this particular problem domain, each individual in the population represents one sample of size N. Each individual consists of N genes where each gene represents one document (each document is given a unique numeric identifier) in the sample. For example, if the sample size is 15, each individual would represent one possible sample and consist of 15 genes that represent 15 different documents. This representation is shown in Figure 1.
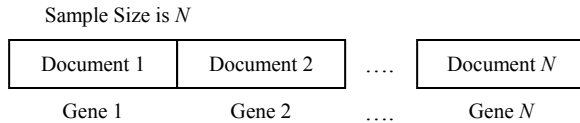
Sample Size is N

Document 1	Document 2	….	Document N
Gene 1	Gene 2	….	Gene N

Figure 1. Genetic representation of each individual

The fitness function evaluates each individual according to some predefined set of constraints or goals. In this particular application, the goal was to achieve an ideal sample that represents the maximum variation of the data set without applying clustering techniques or without prior knowledge of what the categories of the population are. To measure the variation (or diversity) of our samples, the summation of the similarity between the vector-space models of each document (or gene) in the sample is calculated as shown in Equation 1.

$$Fitness(i) = \sum_{j=0}^{N} \sum_{k=j+1}^{N} Similarity(Gene(i,j), Gene(i,k))$$

Equation 1. Fitness function

In Equation 1, the Similarity function calculates the distance between the vector space models of gene j and k of the individual i. This distance value ranges between 0 and 1 with 1 meaning that the two documents are identical and 0 meaning they are completely different in terms of the words used in that document. Therefore, in order to find a sample with the maximum variation, Equation 1 must be minimized. In this fitness function, there will be (N2 – N) / 2 comparisons for each sample to be evaluated.

The defined fitness function can be computationally intensive for large sample sizes or for data sets with lengthy news articles. To compensate for this, the GA developed for this work was designed as a global population parallel GA. For this particular work, the selection process used an "above average" measure for the selection. For each generation, an average fitness value is calculated for the population. Individuals with fitness values that are above this average are selected as parents, while the other individuals are discarded. The crossover and mutation operators are 1-point operators. The crossover rate was set to 0.6. The mutation rate was set to 0.01.

3. TESTS
The data set used for the tests described previously was the Reuters-21578 Distribution 1.0 document collection [4]. This corpus consists of 21,578 Reuters news articles from 1987, and was specifically developed for categorization research purposes. As a result, this corpus includes additional information concerning the documents in the set. This corpus was chosen due to its availability, its size and for the additional information (e.g., category information) for each document, which will be used for future comparisons and research.

To evaluate the performance of this implementation, several tests were conducted, and are briefly summarized in the following table.

Table 1. List of tests performed

Test Num.	Corpus Size	Sample Size	Known Duplicates
1-3	1,000	15	No
4-6	9,494	135	No
7-9	21,578	200	No
10-12	1,000	15	Yes
13-15	9,494	135	Yes

For each test, ten runs were performed with a population size of 100 and 100 generations. However, on test 7 – 9, only 3 runs of 400 generations each with a population size of 100 were performed due to time constraints.

4. CONCLUSIONS
After conducting the defined test and analyzing the results, several interesting observations are evident. The hypothesis that the MVS-GA would be "immune" to duplicate data or take advantage of it did appear to hold true. There is a very slight decrease in fitness values as duplicates are added. While this is not as big of a decrease as was expected, it still supports the hypothesis that the MVS-GA is not dramatically affected by duplicate data. In addition, this approach successfully reduces massive data amounts to manageable levels.

Finally, while the results demonstrated several significant relationships and behaviors, future work will be needed to further understand these relationships and to develop improved parameter control functions.

5. REFERENCES

[1] H. Muehlenbein, "Parallel Genetic Algorithms, Population Genetics, and Combinatorial Optimization", Proc. of the Third International Conference on Genetic Algorithms, Morgan Kaufmann, 1989.

[2] R. Tanese, Distributed Genetic Algorithms for Function Optimization, Ph.D. thesis, University of Michigan, 1989, Computer Science and Engineering.

[3] P.P. Mutalik, et al., "Solving Combinatorial Optimization Problems Using Parallel Simulated Annealing and Parallel Genetic Algorithms", Proceedings of the 1992 ACM/SIGAPP symposium on Applied computing: technological challenges of the 1990's, 1992, pp 1031 – 1038.

[4] Reuters-21578 Distribution 1.0, http://kdd.ics.uci.edu/databases/reuters21578/