UNIVERSITI PUTRA MALAYSIA

GENETIC ALGORITHM FOR WEB DATA MINING

KEVIN LOO TEOW AIK

FSKTM 2001 19

GENETIC ALGORITHM FOR WEB DATA MINING

By

KEVIN LOO TEOW AIK

Project Paper Submitted in Partial Fulfillment of the Requirement
for the degree of Master of Science (Information Technology)
in the Faculty of Computer Science and Information Technology
Universiti Putra Malaysia

April 2001

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my project supervisor, Dr. Md. Nasir Sulaiman for his guidance and constructive comments throughout the study.

Next, the appreciation goes to the external examiners, especially Dr. Mohamed Othman, for their time and comments.

I would like to extend a special appreciation my beloved parents, sisters and brother who provide me emotional support throughout my studies. Last but not least, my appreciation to friends who had directly or indirectly helped and motivated me from the beginning until the end of this study.

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

The use of various search engines could influence the number of search results in the World Wide Web. Therefore, this study attempted to discover any association between the word types or the information types used to search through the World Wide Web using the available search engines. By doing so, it could assist the process of data mining for information in the World Wide Web.

This study used a prototype program based on genetic algorithm to manipulate the initial set of data. Three sets of inputs were used to generate new populations based on the individual fitness. New strains of individuals from a new population were used to test the results obtained from the World Wide Web. Eight search engines used for this study were tested with two groups of words. All the eight words were used as keyword search in all the eight search engines, and the numbers of web pages returned by each search engines were collected. The total web pages based on the selected new individuals were calculated and tabulated. In order to find any association between the search word and the search engines combinations, the individuals were ranked based on the most web pages to the least according to each of the eight words.

Results obtained through the creation of new populations by the prototype program showed that the average fitness of each population improves as new populations were created and new strains of individuals were created through this evolution process. The test on results obtained from the Internet showed that certain class of words could be associated by certain combination of search engines.

ABSTRAK

Penggunaan pelbagai enjin pencari boleh mempengaruhi keputusan pencarian di dalam “World Wide Web”. Dengan itu, kajian ini cuba menyelidik sebarang hubungan antara jenis maklumat yang di cari dengan penggunaan enjin pencari yang sedia ada. Kajian inin mungkin dapat membantu proses pancarian maklumat baru (“data mining”) di dalam “World Wide Web.”

Kajian ini menggunakan program prototaip yang berasaskan algoritma genetik untuk memanipulasikan set data asas. Tiga set data telah digunakan untuk mencipta populasi baru berasaskan nilai individu. Individu baru yang terhasil digunakan untuk menguji maklumat yang diperolehi dari “World Wide Web.” Lapan enjin pencari telah digunakan untuk mengkaji dua jenis golongan perkataan. Kelapan-lapan perkataan tersebut digunakan sebagai katakunci dalam setiap enjin pencari dan jumlah lamanan web yang diperolehi dikumpul. Jumlah lamanan web berasaskan individu baru dikumpul dan dijadualkan. Untuk mencari sebarang hubungan antara maklumat yang dicari dengan kombinasi enjin pencari, individu- individu terlibat disusun megikut jumlah lamanan web yang tertinggi hingga terrendah berdasarkan kelapan-lapan perkataan yang digunakan.

Keputusan yang diperolehi melalui program prototaip itu menunjukkan bahawa nilai purata setiap populasi meningkat mengikut populasi baru yang dihasilkan dan individu baru terhasil melalui proses evolusi ini. Keputusan kajian berasaskan maklumat dari Internet menunjukkan terdapat hubungan antara jenis maklumat yang dicari dengan kombinasi enjin pencari yang digunakan.

CHAPTER I

INTRODUCTION

This chapter consists of insights to the background and identification of the research problem. It proceeds on with the research questions, hypotheses, purpose of the study and definition of terms. The significance of the study, and assumptions and limitations of the study conclude this chapter.

Background to the Study

The World Wide Web has grown into a mass resource of information, both useful and least useful information. Determining the exact size of the Web is almost impossible, but according to a company, BrightPlanet (Liedtke, 2000) estimated that there are now about 550 billion documents stored on the Web, and 95 percent could be accessed by the public. The figure keeps increasing rapidly as the amount of information stored multiplied every second. The many search engines available online cover only the surface of the Web's vast information reservoir. According to a research paper by CompletePlanet (Liedtke, 2000), the World Wide Web is 500 times larger than the maps provided by the available search engines. However, most of the information gathered by these search engines was not up to expectation and need further filtering by users. This is even more troublesome in searching for local contents in the Web. In a survey of 33, 000 search engine users, NDP New Media Services found that 81 percent of the respondents said they find what they are looking for all or most of the time, an improvement from 77 percent in 1999.

Identification of the Research Problem

Users tend to use more than one search engine when finding information in the World Wide Web. The task of searching for information from the Web is a tasking job, and to get the right information is another bigger problem. This study assumed that there might be some relevancy between the word class and the combination of search engines used. Therefore, by implementing genetic algorithm concept and the data mining techniques, certain group or type of information can best be located from the Web based on the search engines combinations or even certain search criteria such as the number of keywords used, information type or phrase length used. This study classified the information type as word class. For example, the word class for "Artificial Intelligence" consists of neural network, genetic algorithm, fuzzy logic, etc, and the word class for "local fruits" consists of durian, rambutan, etc. The sample result could be used to associate the information type with the search combinations. By pinpointing the best solutions provided, future search for the same category of information can be easily and accurately applied based on the prior results.

Research Questions

Research questions have been constructed to provide a more focused area of study. Three main questions have been formed:

1. How will the data and results be represented in genetic algorithm form?
2. To what extent has the genetic algorithm provide a better result?

3. Is there any association between search engines combination and the word class used for web searching?

Hypothesis

Based on the research problem and questions, a hypothesis has been formed: There are some associations between the search engines combination and the type of information in the Web, while genetic algorithm helps to provide a better set of search engine combinations.

Purpose of the Study

This study will try to implement the use genetic algorithm concept to help determine the combinations of search engines that provide the optimum search results in the World Wide Web. This technique will try to predict the association between the type of information search through the Web and the combination of search engines used. By determining the most suitable combinations of search engines for certain types of information search, the search process will be more focused and fast. This study helps to pave the way for future study related to web data mining and promote the use of different techniques such as genetic algorithm in improving the data mining process.

Definition of Terms

Since this study deals with a special branch of terminology, it is important to define a few key terms and concepts.

Genetic Algorithm

Genetic algorithms originated from the studies of cellular automata conducted by John Holland and his colleagues at the University of Michigan. The methods were first widely distributed by means of a book by Holland (1992) entitled Adaptation in Natural and Artificial Systems. Genetic algorithms are relatively new paradigm for search in Artificial Intelligence, which are based on the principles of natural selection (Petry and Buckles, 1992).

According to Raggett and Bains (1992), genetic algorithms are a type of learning algorithm, a specific version of an empiricist algorithm, which learns through trial and error. They are often used in rules-based systems of the production-rule type. Each rule has a certain associated probability. At each cycle of the program, all the rules whose left-hand sides match the actual conditions at the time are collected. Then one of them is 'fired'. Rules with higher associated probabilities are fired more often. If the result of that rule firing matches the ideal solution, then the probability of its firing the next time increases. Thus after several cycles the rules that produce good answers become the ones that almost always 'fire'. Most rules are formed by recombining old ones (i.e. taking bits of different rules and splicing them together); some are formed by changing bits of a rule at random in a 'mutation'.

Gene

Gene is the smallest unit in genetic algorithm. The gene represents the smallest unit of information in the problem domain and can be thought of as the basic building block for a possible solution. If the problem context were, for example, the creation of a well-balanced investment portfolio, a gene might represent the number of shares of a particular security to purchase (Marakas, 1999).

Chromosome

Chromosome is a series of genes that represent the components of one possible solution to the problem. The chromosome is represented in computer memory as a bit string of binary digits that can be "decoded" by the genetic algorithm to determine how good a particular chromosome's gene pool solution is for a given problem. The decoding process simply informs the genetic algorithm what the various genes within the chromosome represent (Marakas, 1999). Along this study, these chromosomes will be called individuals.

Population

A population is a pool of individuals (chromosomes) that will be sampled for selection and evaluation. The performance of each individual will be computed and a new population will be reproduced using standard genetic operators (Muller et al, 1995).

Reproduction

Reproduction is the process of creating new individuals called offspring from the parents' population. These new population will be evaluated again to select the desired results. Reproduction is done basically using two genetic operators: crossover and mutation. However, according to Muller et al (1995), the genetic operators used can vary from model to model, there are a few standard or canonical operators: crossover and recombination of genetic material contained in different parent chromosomes, random mutation of data in individual chromosomes, and domain-specific operations, such as migration of genes.

Crossover

Crossover involves the exchange of gene information between two selected chromosomes. The purpose of the crossover operation is to allow the genetic algorithm to create new chromosomes that shares positive characteristics while simultaneously reducing the prevalence of negative characteristics in an otherwise reasonably fit solution (Marakas, 1999).

Mutation

Mutation is another refinement step that randomly changes the value of a gene from its current setting to a completely different one. The majority of the mutations formed by this process are, as is often the case in nature, less fit than more so. Occasionally, however, a highly superior and beneficial mutation will occur. Mutation provides the genetic algorithm with the opportunity to create chromosomes and information genes that can explore previously uncharted areas of the solution space, thus increasing the chances for the discovery of an optimal solution. However, mutation is normally set to a very low frequency of occurrence and is primarily used to ensure that the probability of searching any point within the solution is never zero (Marakas, 1999).

Data Mining

Data mining is the set of activities used to find new, hidden, or unexpected patterns in data. Using information contained within the data warehouse, data mining can provide answers to questions about organization that a decision maker had previously not thought to ask. An increasingly common synonym for data mining techniques is knowledge data discovery (Marakas, 1999).

Chik (1997) added that data mining combines of techniques including statistical analysis, visualization, induction, and neural networks to explore large amount of data and discover relationships and patterns that shed light on business problems. Tai (1997) stated that data mining is the data-driven extraction of information from large databases. It is the process of automated presentation of patterns, rules or functions to a knowledgeable user for review and examination.

Significance of the Study

This study holds the key to the problem faced by individuals and organizations trying to mine data or search information from the Internet. The concept of genetic algorithm is relatively new in data mining. Thus this study hopes to encourage the use genetic algorithm concept in data mining process. In addition, this research will act as a platform for future studies on the related topic. With the scarcity of such kind of research, this study may contribute additional information to the other existing studies.

Assumptions and Limitations

Throughout the progress of this study, the research is conducted by assuming that there are some associations between the search engines combinations and the word class used for searching, such as a class of local fruits. Next, it is assumed that certain combinations search engines have a better link to certain type of search category than the others. The measure used for this study is solely based on the number of web pages or results returned from each different search engines based on a common keyword or word class. This measure will be used a fitness measure in the genetic algorithm implementation.

However, this study is limited to its purpose only, which is to implement genetic algorithm in web data mining. The results will be solely based on the search result of a few words. The research results will only limited to the examples of words used as the concept fundamentals and tests.

CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter looks into some of the related work and findings done by other individuals, which are crucial to the foundation of this study.

The Concept and Functions of Genetic Algorithm

Originally proposed by Holland (1992) while working at MIT in the 1940s, the term genetic algorithm applies to a set of adaptive procedures used in a computer system that are based on Darwin's theory of natural selection and survival of the fittest. Following Darwin's suggestion that species adapt to changes in their environment in an effort to become more dominant, Genetic Algorithms "reproduce" themselves in various recombination in an effort to find a new recombinant that is better adapted than its predecessors.

Koza ((1992) stated that in nature, the evolutionary process occurs when the following four conditions are satisfied:

- An entity has the ability to reproduce itself.
- There is a population of such self-reproducing entities.
- There is some variety among the self-reproducing entities.
- Some differences in ability to survive in the environment is associated with the variety.

In nature, variety is manifested as variation in the chromosomes of the entities in the population. This variation is translated into variation in both the structure and the behaviour of the entities in their environment. Variation in structure and behaviour is in turn reflected by differences in the rate of survival and reproduction. Entities that are better able to perform tasks in their environment (i.e. fitter individuals) survive and reproduce at a higher rate; less fit entities survive and reproduce, if at all, at al lower rate. This concept of survival of the fittest and natural selection was described by Charles Darwin in *On the Origin of Species by Means of Natural Selection* (1859). Muhlenbein and Schlierkamp-Voosen (1994) added that evolution of natural organisms is based on three major components – reproduction, variation, and selection. Some reproductions of natural organisms occur with "failures" called mutations. A more systematic variation of the genetic material happens in sexual reproduction. Variation is necessary to allow selection to work. Selection in nature is very difficult to define precisely.

According to Koza, John Holland's pioneering book *Adaptation in Natural and Artificial Systems* (1975) provided a general framework for viewing all adaptive systems and then showed how the evolutionary process can be applied to artificial systems. Any problem in adaptation can generally be formulated in genetic terms. Once formulated in those terms, such a problem can often solved by what we now call the Genetic Algorithm.

The genetic algorithm simulates Darwinian evolutionary process and naturally occurring genetic operations on chromosomes. In nature, chromosomes are character strings in nature's base-4 alphabet. The four nucleotide bases that appear along the length of the DNA molecule are adenine (A), cytosine (C), guanine (G), and thymine (T). This sequence of nucleotide bases constitutes the chromosomes string or the genome of a biological individual. Molecules of DNA are capable of accurate self-replication. Moreover, sub-strings containing a thousand or so nucleotide bases from the DNA molecule are translated, using the so-called genetic code, into the proteins and enzymes that create structure and control bahaviour in biological cells. The structures and behaviours thus created enable an individual to perform tasks in its environment, to survive, and to reproduce at differing rates. The chromosomes of offspring contain strings of nucleotide bases from their parent or parents so that the strings of nucleotide bases that lead to superior performance are passed along to future generations of the population at higher rates. Occasionally, mutations occur in the chromosomes.

The genetic algorithm is a highly parallel mathematical algorithm that transforms a set (population) of individual mathematical objects (typically fixed-length character strings patterned after chromosome strings), each with an associated fitness value, into a new population (i.e. the next generation) using operations patterned after the Darwinian principle of reproduction and survival of the fittest and after naturally occurring genetic operations.

To illustrate a simple example of optimization problem, Koza (1992) gave an example of the hamburger restaurant problem. A businessman needs to find the best business strategy for a chain of four hamburger restaurants. The strategy of running the restaurants will consist of making three binary decisions:

- Price – Should the price of hamburger be 50 cents or $10?
- Drink – Should wine or cola be served with the hamburger?
- Speed of service – Should the restaurant provide slow, leisurely service by waiters or fast service through counter?

The goal is to find the combination of these three decisions that produces the highest profit. Since there are three variables, each of which can assume one of two possible values, it would be very natural to represent each possible business strategy as a character string of length L=3 over an alphabet of size K=2. For each decision variable, a value of 0 or 1 is assigned to one of the two possible choices. The search space for this problem consists of $2^3 = 8$ possible business strategies. Identification of a suitable representation scheme is the first step in preparing to solve this problem. Table 1 shows four of the eight possible business strategies expressed in the representation scheme.

Table 1: Representation Scheme for the Hamburger Restaurant Problem

Restaurant	**Price**	**Drink**	**Speed**	**Binary representation**
1	High	Cola	Fast	011
2	High	Wine	Fast	001
3	Low	Cola	Leisurely	110
4	High	Cola	Leisurely	010

According to Koza (1992), the businessman knows nothing about the environment he is facing, so he might reasonably decide to test a different initial random strategy in each of his four restaurants for one week. The businessman can expect that this random approach will achieve a payoff approximately equal to the average payoff available in the search space as a whole. In fact, the businessman is proceeding in the same way as the genetic algorithm. Execution of the genetic algorithm begins with an effort to learn something about the environment by testing a number of randomly selected points in the search space. Using the four different strategies in Table 1, the genetic algorithm begins at generation 0 (the initial random generation), with a population size equal to four. For each generation for which the genetic algorithm is run, each individual in the population is tested against the unknown environment in order to ascertain its fitness in the environment. Fitness may be called profit, or it may be called payoff, utility, goodness, benefit, value of the objective function, score, or some other domain-specific name. Table 2 shows the fitness associated with each of the four individuals in the initial random population and the mating pool created after reproduction.

Table 2: Fitness of Initial Random Population and The Mating Pool Created After Reproduction

	Generation			**Mating pool created after reproduction**	
i	String (X_i)	Fitness $f(X_i)$	$\frac{f(X_i)}{\sum f(X_i)}$	Mating pool	$f(X_i)$
1	011	3	0.25	011	3
2	001	1	0.08	110	6
3	110	6	0.50	110	6
4	010	2	0.17	010	2
Total		12			17
Worst		1			2

Average	3.00		4.25
Best	6		6

The fitness (profit) for each business strategy has, for simplicity, been made equal to the decimal equivalent of the binary chromosome string. From the initial population, the businessman learnt that the strategy 110 produces a profit of $6 (the global optimum is $7) for the week. This strategy is the best-of-generation individual in the population for generation 0. The strategy 001 produces a profit of only $1 per week, making it the worst-of-generation individual. The only information used in the execution of the genetic algorithm is the observed values of the fitness measure of the individuals actually present in the population. The genetic algorithm transforms one population of individuals and their associated fitness values into a new population of individuals using operations patterned after the Darwinian principle of reproduction and survival of the fittest and naturally occurring genetic operations.

The operation of fitness-proportionate reproduction is performed by copying individuals in the current population into the next generation with a probability proportional to their fitness. The sum fitness for all the four individuals in the population is 12. The best-of-generation individual in the current population has fitness 6. Therefore, the fraction of the fitness of the population attributed to individual 110 is 1/2. In fitness-proportionate selection, individual 110 is given a probability of 1/2 of being selected for each of the four positions in the new population. Thus, it is expected that string 110 will occupy two of the four positions in the new population. Since the genetic algorithm is probabilistic, there is a possibility that string 110 will appear three times or one time in the new population.

There is even a small possibility that it will appear four times or not at all. Similarly, individual 011 has a probability of 1/4 of being selected for each of the four positions in the new population. Thus, it is expected that 011 to appear in one of the four positions in the new population. The strategy 010 has probability of 1/6 of being selected, whereas strategy 001 has only a probability 1/12 of being selected. Thus, it is expected that 010 to appear once in the new population, and 001 to be absent from the new population. Table 2 shows one possible mating pool resulting from applying the operation of fitness-proportionate reproduction to the initial random population.

The genetic operation of crossover (sexual recombination) allows new individuals to be created. It allows new points in the search space to be tested. The operation of crossover starts with two parents. The individuals participating in the crossover operation are selected proportionate to fitness. The crossover operation produces two offspring, and the two offspring are usually different from their two parents and different from each other. Each offspring contains some genetic material from each of its parents.

The crossover operation begins by randomly selecting a number between 1 and L-1 using a uniform probability distribution. There are L-1 = 2 interstitial locations lying between the positions of a string of length L = 3. Suppose the interstitial location 2 is selected, this location becomes the crossover point. Each parent is then split at this crossover point into a crossover fragment and a remainder. The first two individuals from the mating pool are shown in Table 3. The crossover fragments of parents 1 and 2 are shown in Table 4, while the remainders of parent 1 and 2 are shown in Table 5.

Table 3: Two Parents Selected Proportionate to Fitness

Parent 1	Parent 2
011	110

Table 4: Crossover Fragments from the Two Parents

Crossover fragment 1	Crossover fragment 2
01-	11-

Table 5: Remainders from the Two Parents

Remainder 1	Remainder 2
--1	--0

The remainder 1 will be combined with crossover fragment 2 to create offspring 1, while remainder 2 will be combined with crossover fragment 1 to create offspring 2 as shown in Table 6.

Table 6: Two Offspring Produced by Crossover

Offspring 1	Offspring 2
111	010

The one possible outcome of applying reproduction and crossover operations is shown in Table 7. Compare the new population of generation 1 as a whole against