Automatic discovery of the sequential accesses from web log data files via a genetic algorithm

Emine Tuğ *, Merve Şakiroğlu *, Ahmet Arslan

Department of Computer Science, Selçuk University, 42300 Konya, Turkey

Received 31 December 2003; accepted 27 October 2005
Available online 27 December 2005

Abstract

This paper is concerned with finding sequential accesses from web log files, using ‘Genetic Algorithm’ (GA). Web log files are independent from servers, and they are ASCII format. Each transaction, whether completed or not, is recorded in the web log files and these files are unstructured for knowledge discovery in database techniques. Data which is stored in web logs have become important for discovering of user behaviors since the using of internet increased rapidly. Analyzing of these log files is one of the important research area of web mining. Especially, with the advent of CRM (Customer Resource Management) issues in business circle, most of the modern firms operating web sites for several purposes are now adopting web-mining as a strategic way of capturing knowledge about potential needs of target customers, future trends in the market and other management factors.

Our work (ALMG—Automatic Log Mining via Genetic) has mined web log files via genetic algorithm. When we search the studies about web mining in literature, it can be seen that, GA is generally used in web content and web structure mining. On the other hand, ALMG is a study about web mining usage. The difference between ALMG and other similar works at literature is this point. As for in another work that we are encountering, GA is used for processing the data between HTML tags which are placed at client PC. But ALMG extracts information from data which is placed at server. It is thought to use log files is an advantage for our purpose. Because, we find the character of requests which is made to the server than detect a single person’s behavior. We developed an application with this purpose. Firstly, the application is analyzed web log files, than found sequential accessed page groups automatically.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Web mining; Genetic algorithm; Knowledge discovery; Sequential access

1. Introduction

The World Wide Web (WWW) is a hypertext body of more than 800 million pages that continues to grow. It exceeds six terabytes of data on roughly three million servers. Almost one million pages get added daily, and typically, several hundred gigabytes are changed every month. Hence, it is clear that the Web currently constitutes one of the largest dynamic data repositories [9]. So, it to be became a need to extract useful knowledge from these raw and dynamic data by knowledge discovery called as web mining on the Internet.

Web mining refers to the analysis and mining of all web-related data, including web content, hyperlink structure, and web access statistics [13]. Analyzing Web Logs for usage and access trends cannot only provide important information to web site developers and administrators, but also help in creating adaptive web sites [10]. Various techniques and methods are used for this reason. Especially, with the advent of CRM (Customer Resource Management) issues in business circle, most of the modern firms operating web sites for several purposes are now adopting web-mining as a strategic way of capturing knowledge about potential needs of target customers, future trends in the market and other management factors.

By this time, there were a lot of studies about web mining using GA. Nasraoui and Rojas investigate a new evolutionary approach based on continuously and dynamically learning the Web access patterns from non-stationary Web usage environments [9]. Kim and Zhang use a genetic algorithm to learn the importance factors of HTML tags which are used to re-rank the documents retrieved by standard weighting schemes for web structure mining [11]. Picarougne et al. present a genetic search strategy for a search engine [5]. Abraham and Ramos propose an ant clustering algorithm to discover Web usage patterns...
In this work, GA based on evolutorial approach for knowledge extraction was used to found optimum solutions for the problem at the minimum time to discover sequential accesses from web log files. Our work can be resembled as a kind of market basket analysis which is performed by using data mining techniques. Marshall [3] tried to find items which are sold together in a market database. In our work, this idea is applied to web mining and tried to find web pages which are visited sequentially on the web site by using a genetic algorithm.

Firstly, Web log files are analyzed and their redundant data deleted. Then web log data is migrated to database. Finally, the GA is used to find sequential access of web pages. At the end of the work, we report result on experiments realized with our new fitness function and a fitness function that is using a support concept known from data mining.

This paper is organized as follows. Section 2 describes web mining. GA is given in Section 3. And, our work is presented in Section 4. The experimental results of the work are given in Section 5. Finally, in Section 6, we present our conclusions and perspectives.

2. Web mining

With the increasing amount of data available on the WWW, it has seemed like a gold mine which is full of useful information by many researchers [4]. We need intelligent data analysis methodologies to discover useful knowledge from web. It has conducted to emerge ‘web mining’ as a new technology. It was first described by Etzioni in 1996. ‘Web Mining is the use of Data Mining techniques to discover and extract information from web documents and services’ [8].

2.1. Taxonomy of web mining

As shown in Fig. 1, web mining is classified roughly into three domains such as web content mining, web structure mining and web usage mining. Web content mining is the process of extracting knowledge from the content of a number of web documents. Web structure mining is the process of inferring knowledge from the organization and links on the web, while web usage mining is automatic discovery of user access patterns from web servers [7].

Web content mining is related with interpreting web contents. Most of the web content data is unstructured text data. Web content mining is often viewed as an intelligent software program that traverses the web to collect information from the web sites Lycos, Alta vista, Web Crawler, etc. which use web content mining technique [2].

As different from the web content mining, web structure mining discovers hyperlink relations between web sites. It is related to links between web pages directly while web content mining is related to web documents.

Web usage mining is discovered as an useful knowledge by using data mining techniques from the secondary web data. Secondary web data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse clicks and scrolls, and any other data generated by the interaction between users and the web. Logs of web access available on most servers are good examples of the data sets used in web usage mining. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on [1]. Since, our goal is to discover the sequential access from web logs, this approach belongs to web usage mining.

2.2. Structure of web logs

Web server log file’s format is ASCII format and independent of server platform. A Web server log is an important source for performing Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the access of a Web site by multiple users. These log files can be stored in various formats such as common log or extended log formats [6].

There are some differences in server softwares, but traditionally there are four types of server logs:

1. Transfer (access) log
2. Error log
3. Referrer log
4. Agent log

The first two types of log files are standard. The referrer and agent logs may or may not be ‘turned on’ at the server or may
be added to the transfer log file to create an ‘extended’ log file format.

Each HTTP protocol transaction, whether completed or not, is recorded in the logs, and some transactions are recorded in more than one log. For example, most (but not all) HTTP errors are recorded in the transfer log and the error log. Now type of information collected in an ‘extended format’ transfer log file is examined.

A single line in a common log file is presented in Fig. 2 and its descriptions are shown in Table 1. This typically displays as one long line of ASCII text, separated by tabs and spaces (useful for importing it into a spreadsheet program).

### 3. Genetic algorithm

Genetic algorithms are probabilistic search methods based on the mechanism of natural selection and genetics. They are generally quite effective for rapid global search to find solutions in non-deterministic problems [11]. This algorithm operates by iteratively updating a pool of individuals, called population. In the context of search, individuals are candidate solutions to a given search problem. On each iteration, all members of population are evaluated according to fitness function. A new population is then generated by probabilistically selecting the most fit individuals from current population. Some of these selected individuals are carried forward into next generation population intact [12]. Genetic operators such as selection, crossover and mutation generate new offspring from the fittest individuals.

The generation of successor in a GA is determined by these operators that recombine and mutate selected members of the current population. They correspond to idealized versions of genetic operations found in biologic evolution. The most common operators are crossover and mutation. The crossover operator produces two new offspring from two parents by copying selected genes from each parent.

The mutation operator produce small random changes to the chromosome by choosing a single bit at random, then changing its value. Mutation is often performed after crossover has been applied [12]. The genetic algorithm used in this work is shown in Fig. 3. The following sections describe several aspects of the proposed algorithm, structure of chromosome and fitness function.

#### 3.1. Structure of chromosome

Structure of chromosome is presented before describing the GA used for this work. Every page, given a page_id value, has been recorded to the database at the step of pre-processing task. These page_id values are used for creating the chromosomes. The length of a chromosome has set as a number of the pages which are wanted to access sequentially. If trio page groups are searched which are accessed orderly, the chromosome’s shape should be as seen in Fig. 4.

![Fig. 3. Flow card of the genetic algorithm.](image-url)
3.2. Genetic operators

The genetic algorithm uses genetic operators to generate the offspring of the existing population. In this section, three operators of genetic algorithms are described: selection, crossover and mutation. They must be adapted to the problem.

3.2.1. Selection

The selection operator chooses an individual in the current population according to fitness function and copies it without changes into the new population. We implement roulette wheel for selection phase, where the individuals are selected randomly from current population. The quality of an individual is determined by the fitness function.

3.2.2. Crossover

The crossover operator produces two new chromosomes from two selected chromosomes by swapping segments of genes and insert to population. A single-point crossover operation with a crossover probability of 30% was used for this work. In the single-point crossover operation, crossover point randomly selected and parent chromosomes are swapped to crossover point and two child chromosomes can be created (Fig. 5).

3.2.3. Mutation

The mutation operator is used for maintaining diversity. During the mutation phase, according to mutation probability value of each gene in each selected chromosome is changed with generating a random number between first page number and last page number (Fig. 5). Mutation probability is selected 30% for this work.

3.3. Fitness function

In this section, our fitness function is described. To formalize the function, three new values are introduced: session, support and similarity rate.

3.3.1. Session

For our work, a session is defined that all request of each visitors made to the server in a piece of time. As is clear from examples of data in Fig. 2 the original log data show sequences of requests by IP address and request time. Firstly, an individual session should have been defined. A session can be defined by a unique IP address and a unique request time. When an IP address is crossed, its Time Stamp is noted. This time is accepted as the beginning of a session. Then a time space is defined as a session period. In our work this time accepted 15 min for each session.

If an individual stayed more than a quarter hour then her/his session is ended and a new session is generated. After creating the sessions, they are separated to subsets. For instance, assume that we looking for sequential trio pages groups are searched, so trio subsets have to create from sessions.

3.3.2. Support

The support is defined as the probability of the individual is being present in all sessions, which is the sum of the gene number compatible with each of the session subset of an individual, divided the total number of the session subsets.

3.3.3. Similarity rate

The gene rate is compatible with session subsets of an individual. This statement is formalized as above.

Lastly, fitness function is formalized in Table 2.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematical description of ALMG parameters</td>
</tr>
<tr>
<td>Support = probability (Σ{records is compatible with session subsets})/ Σ{session subsets})</td>
</tr>
<tr>
<td>Similarity Rate : ( n=\sum_{i=1}^{m} ) compatible gene number with ( n \times 2^n ) subset number in session ( \times ) 100</td>
</tr>
<tr>
<td>Fitness (an individual) = Support × Similarity Rate</td>
</tr>
</tbody>
</table>
Threshold value is selected 0.08 and estimates are made according to this value.

4. The related work and database

The purpose of ALMG is finding web pages which are visited sequentially on the web site. GA is used for this which is based on evolitional approach for knowledge extraction and could have found optimum solutions for the problem. Structure of ALMG is presented in Fig. 6.

The first phase of our application is getting web log files from server as seen in Fig. 6. Structure of web log files is described in Section 2.2. Like other data mining process, web usage mining consists of redundant or irrelevant data which must be removed from the data source. Here, our data source is these web log files. Erasing of noisily records and organizing the web log data by using techniques of pre-processing is a step of knowledge discovery. As shown in Fig. 2, our data is a raw data. Firstly, redundant data are cleaned and then other data transferred to database. The structure of the database is shown in Fig. 7. Here, the record whose extension is .gif, .jpeg, .swf or etc. is accepted as a noise data. After this step, data is ready to be analyzed and interpreted.

Next phase of the ALMG is GA. The genetic algorithm is applied to data for acquiring the trio page groups, here. This algorithm designed for our work consists of the following steps:

**Step 1:** First, 100 individuals are selected randomly from the database and these individuals constitute the population.

**Step 2:** A fitness value is evaluated for each individual in the population. For this a fitness function is used. The function is described in Section 3.3.

**Step 3 (Selection):** Individuals are placed on the roulette wheel according to fitness value and \( N \) individuals are randomly selected. Mating pool is consisting of them.

**Step 4 (Crossover):** Individuals are selected from mating pool according to crossover probability. Crossover probability is selected 30% for this application. Later double individuals are selected from mating pool and crossover with them. One point crossover used for this operation.

**Step 5 (Mutation):** Individuals are selected from new population according to mutation probability. Crossover probability is selected 30% for this application. Later double individuals are selected from mating pool and crossover with them. One point crossover used for this operation.

**Step 6 (Evaluation):** Fitness value of each individual in this new population is evaluated and it is compared to threshold value. Individuals which have not passed the threshold value are applied to algorithm again and the others have passed the threshold value, are taken as result.

![Fig. 6. Structure of ALMG.](image)

![Fig. 7. Structure of the database.](image)
5. The experimental results

In this section, three day’s web access log data at the Selçuk University’s Web site were used for experimentations. Volume of these data is around 720 Mb. An application was developed for this work to analyze this data using genetic algorithm. Two different fitness functions are used here. The first function is described above. Second one uses the support value which is known from data mining.

The experiences incomes from these two functions are compared with each other. The performance comparison of these functions for finding of trio page groups by (a) individual numbers on each generation (b) fittest individual of generation, after certain number of generations is processed.

The performance comparison of ALMG and SGA can be done by examining the graphics. It explain the trio page groups which is found from 100 times running results of ALMG and SGA according to 50% threshold value on data set which is shown in Fig. 8(a). Each of running process represents a generation and also each of trio page groups represents an individual of genetic algorithm. As seen from the graph, ALMG has found same around number and similar page groups according to alternative SGA method. This situation is shown that ALMG’s performance is better than other.

Fig. 8(b) shows the fittest individual of each generation. As can be seen from the graph, ALMG fitness rapidly increases until about the 10th generation although SGA fitness increases until about the 25th generation. So, ALMG is passed the stable state before SGA.

Fig. 9 shows that, trio page groups, is acquired after 100 times running and the frequency of this page groups in the population.

This result set represents the fittest individual that is produced after the ALMG is passed to the stable state. Table 2 shows description of this page groups. The page groups which have a wide slice on the graph, are visited frequently more than the other pages. However, the pages that are not visited frequently but have high visiting rate along the other pages, are placed on the graph. These groups have less part than the others on the graph.

6. Conclusions

With the help of genetic algorithm (ALMG) developed in this work, the information discovery was made on web log data sets. Beside this, there was shown a comparison of genetic algorithm with the alternative one. The genetic algorithm, used as a comparison algorithm, uses usual support value known from data mining. Results of work and comparison of performance are shown with graphs in Section 5. The most coherent and reliable results were obtained with the ALMG.

The data sets used in this work are the web log files of the University of Selçuk. The application implemented in this work can be significant especially for page used in electronic business. The advertisements are placed on the pages where users are spending the most of their time. Or examining sequential page from the user side, special dynamic pages can be provided for customer coming for shopping. Determining by which customer which products were bought, special promotions could be organized. Setting new strategies like this can be used for decision support.

Acknowledgements

The authors acknowledge the support of this study provided by Selçuk University Scientific Research Projects.
References


