Research Statement

1. Overview

My primary research topics are data and web mining, information retrieval, query optimization, query processing and simulation systems. My research in web mining is mainly related to web usage mining. I have implemented a new technique for constructing meaningful web user sessions, which can be used for several applications including pre-fetching, web personalization and recommendation systems. In data mining area, I have been also working on pattern discovery methods on web user sessions data with various constraints. In the area of information retrieval, I have participated in the development of the first search engine designed specifically for Turkish language. In information retrieval area, I have implemented “focused crawl” for crawling specific topics in the web and preserving refreshes of determined set of pages. I have also designed different scoring systems as a combination of usage aware page ranking, page-rank with various constraints and OPIC scoring system. In the area of query optimization, I have worked on query processing and multiple query optimization problems. In this work, we have implemented a genetic algorithm technique for multiple query optimization problem and proved that it is possible to obtain optimum or close to optimum solutions for multiple query optimization problem by using genetic algorithm. In the simulation system area, I worked in the development small-scale combat simulation system, which is the first agent-based programming simulation of Turkish Armed Forces. In this project, I have worked on modeling semi-automated individual commando behavior and perception modeling.

Below, I will briefly talk about my researches in each direction.

2. Research

Data&Web Mining

A New Reactive Method for Processing Web Usage Data is one of my main research projects that I have studied during my MS study [1, 2]. In this work, we added link information to web usage processing methods in order to construct web user sessions. We have improved previous session reconstruction heuristics with additional time limitations. In our work, we have modeled web domain as a web graph [2]. By this way, the session reconstruction problem becomes a graph problem in which we try to determine all connected maximal sequences from initial web user page access sequences. All pages in connected maximal sequences satisfy link topology of corresponding domain. That means there has to be web page link between each consecutive item in the sequence. We called our session reconstruction technique Smart-SRA. Smart-SRA [1, 2] is a reactive session reconstruction technique, which uses web log data and the site topology. In order to compare Smart-SRA with previous reactive methods, a web agent simulator has been developed. Our agent simulator models behaviour of web users and generates web user navigations as well as the log data kept by the web server. In this way, the actual user sessions will be known and the successes of different techniques can be compared. In our studies [1, 2], we have proved that the sessions generated by Smart-SRA are more accurate than the sessions constructed by previous heuristics. Although Smart-SRA is better than previous session reconstruction heuristics, it has some deficiencies and these deficiencies are removed by the new method called ISSRA (Improved Smart-SRA). As we stated in [7], Smart-SRA misses some actually possible user sessions since it does not consider sub-sequences in the navigation paths of the users with edges corresponding to forward edges in the topology graph. Our improved technique, ISSRA captures all possible user behaviors in the maximal sessions constructed. The first phase of ISSRA is the same as Smart-SRA. The second phase of ISSRA constructs all maximal sub-sessions from the candidate sessions generated at the first phase of the algorithm corresponding to the paths in the web site topology graph. The main idea of ISSRA is to determine all possible paths in the web topology captured in the server logs. These paths are assumed to represent potential user navigation behaviors. In our research [7], this idea of ISSRA is shown to produce more accurate sessions than previous heuristics. We have also started to investigate the effect of the results of our session reconstruction technique in frequent pattern discovery phase. We proved that ISSRA produces more accurate sessions than previous heuristics and Smart-SRA. We have also showed that ISSRA runs faster than Smart-SRA in construction time of per session.
Like in ordinary Data Mining, Discovering Frequent Web User Access Patterns is the next step of my research after working on session reconstruction heuristics. Session reconstruction phase is the first stage of web usage mining. Pattern discovery phase is the second step which takes reconstructed sessions and produce frequent web usage patterns. As in the classical data mining pattern discovery it is the main issue in web usage mining. Many algorithms have been proposed on capturing frequent user navigation patterns. Finding desired patterns is quite challenging in very large data. The search space increases exponentially as pattern lengths increase. Also, discovered patterns must be interpreted and understandable knowledge must be extracted from them. In our recent research [4], we have used time oriented heuristics on web log data for constructing user sessions. After that step, we apply different pattern discovery methods on the constructed sessions. These algorithms are SPADE, GSP, Breadth First Search and Depth First Search algorithms. Their comparisons and performance analysis is studied in [4]. In our study, we have showed that SPADE is the best method for finding patterns from sessions constructed by time oriented heuristics, because SPADE works on prefix-based equivalence classes, which has a much smaller search-space. Another important result is that SPADE uses session-id timestamp list structure which prevents unneccessary database scans for evaluating frequency of intermediate patterns in search space.

Our another research project on pattern discovery methods is related to discovering more accurate web usage patterns [5]. In this work, we have applied pattern discovery methods on sessions constructed by Smart-SRA. We used Sequential Apriori technique which uses link information to produce frequent patterns. In our method, unlike the ordinary large item set discovery problem, consecutive pages in the discovered pattern should also appear in consecutive positions in the reconstructed sessions supporting the pattern. Our final pattern obeys string-matching relation with the reconstructed sessions that they appear. In this work, we used agent simulator for producing web user access data logs and constructed sessions. We measure the accuracy of first phase by comparing agent simulators session and Smart-SRA’s sessions constructed from agent simulator’s web log data. Then, we measure the accuracy of frequent patterns by comparing two sets of patterns coming from sequential Apriori on Smart-SRA’s sessions and sequential Apriori on agent simulators sessions. We concluded that the accuracy of pattern discovery phase is greater than the accuracy of session reconstruction phase. We are currently working on alternatives of sequential Apriori technique for pattern discovery phase after session reconstruction phase with ISSRA.

In the area of Web usage mining, we have started a project for combining both our session reconstruction methods and pattern discovery methods to implement complete Web Usage Miner System. We think that our web usage miner system will have various application areas such as web pre-fetching, link prediction, site reorganization and web personalization. Also this system can be a subcomponent of bigger systems like Web Analytics tools or Adaptive web site systems. Adaptive web site shows dynamic structure, which changes with respect to user interests. The main purpose of these systems are decreasing the length in web navigation pathways and increasing page-stay time of web user on a particular domain. It is impossible to achieve this without strong web usage miner systems. We think that our web usage miner system have great potential for being a part of this type of dynamic systems. In addition, our web usage mining plays important role in the detection of PPC Fraud in e-commerce web sites. Many web analytics tools have an integrated web usage miner system to analyze web users’ behaviors. In this perspective, our web usage miner system has a potential of being part of this kind of web analytic tools.

**Information Retrieval**

In the area of Information retrieval, I have studied on Search Engine Technologies. I have worked in the development of first Turkish search engine, which is bilgi.com. Search engines are document retrieval systems designed to find information stored on a computer system, such as on the World Wide Web, inside a corporate or proprietary network, or in a personal computer. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria. Crawler is a data collection component of search engine; they fetch information from www and store them in web databases. In our project bilgi.com, I have implemented a focused crawler for preserving refreshness of the determined set of pages. For preserving refreshness of the determined set of pages, our crawling systems find changing period of each page. The focus crawler I have implemented fetches web pages in time period less than the smallest changing period of the page in our set. We have kept a signature key of web pages which is obtained from hashing content of the page. The signature changes shows that corresponding page is updated, and then our system updates the corresponding page with
the fresher one. In crawling part, I have implemented another focused crawler for fetching information about specific topic in the web. In this project I used an initial word set for representing particular semantic domain. This word set is modeled as unit vector representing a given domain. Each web site is represented by a topic vector containing frequency of each element in the unit vector of a corresponding domain. The focused crawler is given a similarity threshold \( q (0<q<1) \) as an input. Then program checks the cosine of the angle between unit vector of that domain and topic vector of that page. If that value is greater than \( q \), corresponding web page is marked as related to that semantic domain; other web pages not satisfying the rule are pruned. Web crawling process continues depth by depth with pruning search space in this manner until enough data is collected. In search engine project, I have also designed a different scoring system combination of usage aware page ranking and OPIC scoring system. For implementing OPIC scores system, we have crawl database containing all the pages crawled in common time-periods. In this database, we keep page importance score of each web page. Each web page contains two importance score called total and cache scores. Cache score is the score that the page obtained after the most recent fetch. Our systems distribute cache points of each page to pages which current page refers. Since we crawl depth by depth; in each depth, fetched pages distribute current cache scores to pages, which it links out. In this way, important pages of current depth distribute higher scores to newly discovered pages from that page at that level. This scoring system keeps score of web pages coming from link analysis. We have also structure called usage-score database, which keeps usage statistics of clicks for links in the search engine result. At indexing phase, we calculated overall score as linear combination of OPIC score and usage score. At query time, query results are sorted with respect overall score and relevance to the query terms.

**Database Systems and Query Optimization**

Genetic Algorithm for Multiple Query Optimization Problem is also my other research topic in the area of Database Systems and Query optimization. Multiple Query Optimization (MQO) problem is a well-known database research problem and the database community has studied it since 1980s. The goal of MQO is to reduce the execution cost of a set of queries by performing their common tasks only once. A query may have more than one solution plan. Traditional query optimization considers a single query at a time, and tries to generate the most efficient solution by examining and finding cheaper alternative plans. In our work [6] related to MQO problem, we showed that MQO problem could easily be modeled by using Genetic Algorithms. In the context of MQO, a chromosome corresponds to a solution instance for the set of queries of the MQO problem. In a chromosome, each gene of a chromosome represents a plan to the corresponding query. Under this modeling, MQO is also very suitable for genetic operations. Crossover and mutation operations can easily be defined to produce new valid solution instances. We have applied different selection and crossover operations for finding optimal solution for MQO problem. We have compared performances of various combinations of genetic methods. After that, we have picked up the best genetic method determined by the selection type and crossover type. The best genetic method is compared with \( A^* \) algorithm which finds the optimum solution for MQO problem by exploring all search space. We have proved that Genetic algorithm finds solutions which is very near to the optimum solution in a very short time, however \( A^* \) finds optimum solution in exponential time. We have also showed that it is not possible to use \( A^* \) algorithm for more than 10 queries since it takes too long time to get the solution. Under these conditions, genetic algorithms are very suitable for solving MQO problems.

**Simulation Systems**

My research on simulations system is related to project small-scale combat simulation system [3], which I participated in. Small-scale combat simulation system is implemented for simulating and analyzing commando and patrol defense operations with the help of agent-based modeling and HLA-based distributed simulation technologies. I have joined the project in the improvement and the maintenance phase. The project has two core components, called agent and perception module. I studied on both agent and perception modules and improved important parts of them. We have multi-tiered agent architecture in this project and our agent structure has three layers. First one is the top layer for simulating predefined external basic agent behaviors for planned tasks. Second one is the middle layer for simulating predefined and dynamic external basic agent behaviors for user-defined tasks. The final layer is the bottom layer for simulating predefined external basic autonomous agent behaviors. I have improved the behaviors in the bottom layer and hierarchically change behaviors in upper layers, which use changed behavior in the bottom layer. In the perception module I have studied on visual and voice perception. For calculation of voice perception the Voice and Radio Communication Federate...
implemented. It keeps track of the locations of all agents and the locations of all events that produce sound. I have studied on the effects of ground height in voice perception. In the visual perception, I have changed the format of visual perception packets for processing efficiently in other federates in the system.

References:


[7] Ismail H. Toroslu, Murat Ali Bayir, "Graph Theoretic Approach for Session Reconstruction Problem". (Under Submission)