

RESEARCH STATEMENT

MURAT ALI BAYIR

1 Overview

My research vision is user-centered, socially beneficial and system-oriented. I like working on real application projects which has potential social impact in our daily lives. Before attacking any research problem, I always ask myself which problem I am trying to solve and what are the benefits of solving it. For the last seven years of my graduate study (MS + PhD) and industrial employment period, I have been lucky to have several collaborators in both industry and academia. I was also fortunate to produce MS Thesis which led to industrial research project supported by National Science Foundation of Turkey (300\$K industrial project fund) and 2 private companies.

In my research, I have focused on the areas of Data Mining & Machine Learning, Ubiquitous Computing, Information Retrieval and Optimization Algorithms.

Below, I will describe my research in each direction:

2 Data Mining and Machine Learning

I have been working in Machine Learning and Data Mining for more than 7 years and both of my thesis studies (MS and PhD) include applications of these two areas. During my Master Thesis, my work focused on analyzing navigation behavior of web users. This work led to an industrial project named "A Web Analytics Tool for Improving Structure of Commercial Web Sites", and it was funded by TUBITAK (National Science Foundation of Turkey) with 300K \$ support. As an application of my thesis, I developed link recommendation and decision support systems for improving inner and intra page structures of commercial web sites. My work in PhD Thesis focused on using Machine Learning and Data Mining algorithms for analyzing mobility behaviors of cell phone users. My analysis over 9 month data of MIT Reality Mining group including 100 subjects resulted in a new model for representing human mobility in city wide level. Based on my findings, I developed applications for predicting future locations of cell phone users and an observation based self learning distributed routing protocol for human networks (carrying cell phones). Below, I will give brief summary of our approaches and contributions in each project:

2.1 Mobility Profiler: A Framework for Discovering Mobility Profiles of Cell Phone Users

Problem: The location logs that can be collected from cell phones are low level data units (either collected in cellular or GPS enabled environment). These preliminary data units makes difficult to access mobility profiles of the cell phone users for several applications. To make mobility data more accessible to cell phone applications, higher level data abstractions are needed. In order to achieve

this, we focus on the problem of discovering spatiotemporal mobility profiles from cell phone-based location logs.

Approach: In order to capture the mobility behaviors of cell phone users at a level of suitable abstraction, we introduce formal definitions for the concepts of end locations (corresponds to locations where users spend significant amount of time), mobility path (denoting a user’s travel from one end-location to another), mobility pattern (denoting a popular paths that users traveled frequently), and mobility profile (providing a synopsis of a user’s mobility behavior by integrating the frequent mobility patterns with time contextual data (2D location component of profiles), and end locations with time distribution data (1D component of the profiles)). These four different mobility concepts correspond to abstraction of mobility information in different levels listed from lower to the upper ones. Here, we design and implement a complete framework, the Mobility Profiler [1, 2], for discovering mobility profiles from raw cell tower connection data by producing all of the data type abstractions intermediate levels described above.

Contributions: Our analysis of the cell phone users’ mobility behaviors yields important lessons for researchers interested in testing systems with human mobility. As also identified in a recent studies, we found that users spend approximately 85% of their time in 3 to 5 favorite locations, e.g., home, work, shopping. However, our analysis has exposed a more interesting phenomena for the distribution of the remaining 15% of the users’ time. We identify a significant long tail in a user’s location-time distribution: Approximately a total of 15% of cell phone user’s time is spent in locations that each appear with less than 1% of total time.

Our another finding is that significant amount of human mobility (85%) exhibits spatial and temporal regularity where users move between their top-k locations. The regularity property of these mobility profiles with time contextual information (days of week, and hours of day domain) enables to develop different smartphone applications such as early warning systems and route prediction applications etc.. By coupling the time-context with the mobility paths, these mobility profiles may be useful for the purposes of synthetic mobility scenario generation for evaluation of different network algorithms.

2.2 Smart Miner: A New Framework for Mining Large Scale Web Usage Data

Problem: Web usage mining which can be defined as the application of data mining techniques to web log data in order to discover user access patterns. Web usage mining has various applications such as link prediction, site reorganization and web personalization. The success of all of these applications is significantly related to the outcomes of web usage mining process which includes session construction and frequent navigation pattern discovery phases. Among these phases, session reconstruction is the most important part of web usage mining since the performance of any application using frequent patterns or sessions are significantly affected by session reconstruction phase. Previous approaches for session construction didn’t consider forward link information between pages which is very precious. The problem we attack here is to develop web usage mining framework with session construction and pattern discovery phases which considers link information between pages during the web user navigation. Another issue is the scalability, popular web sites may have significant amount of web page access daily which can only be handled by scalable distributed systems.

Approach: In order to solve the problem above, we propose a novel framework called Smart-Miner [3] for web usage mining problem which uses link information for producing accurate user sessions and frequent navigation patterns. Unlike the simple session concepts in previous methods, where sessions are sequences of web pages requested from the server or viewed in the browser, Smart Miner sessions are set of paths traversed in the web graph that corresponds to users' navigations among web pages. We have modeled session construction as a new graph problem and utilized a new algorithm, Smart-SRA [4,5], to solve this problem efficiently. For the pattern discovery phase, we have developed an efficient version of the Apriori-All technique which uses the structure of web graph to increase the performance. In order to solve the scalability issues, we have implemented distributed version of the Smart Miner framework by employing Map/Reduce Paradigm on Hadoop framework.

Contributions: Using both simulated and real data, we have showed that in terms of our accuracy measure the maximal frequent patterns discovered by Smart-Miner framework is at least 30% much better than the ones obtained by web usage mining techniques utilizing previous session construction methods. We also showed that frequent patterns generated by Smart Miner framework has better performance for web usage mining applications such as link prediction and decision support systems. Our results show that we have a significant performance improvement in the distributed version of the Smart-Miner. Our Experimental results imply that the run time performance of distributed Smart-Miner increases linearly and it can be scaled-up easily to process any size of data. We conclude that we can efficiently process terabytes of web server logs belonging to multiple web sites by our scalable framework.

2.3 Identifying Breakpoints in Public Opinion

Problem: In this work [6], we consider the problem of identifying breakpoints in public opinion about certain topic. While traditional poll applications achieve this by providing a snapshot of public opinion, they can neither track temporal opinion changes nor capture opinions that are not asked in the questionnaire. In order to overcome difficulties of these applications, we propose to use Micro-blogs to capture changes in public opinion.

Approach: In order to detect opinion changes, we integrate tweet based emotion vector representation of time periods in vector space model and document based representation of time periods. We show that the set space model can be used to eliminate false positive opinion changes that vector space model finds on a time domain. For representing events that caused changes in public opinion, we propose a new scoring function to discover popular terms by considering temporal dimension.

Contributions: By successfully combining these methods, we identified the time intervals that include a change in public opinion over two case studies. From the experimental results we found that using emotion corpus based method and set space model methods together eliminates false positives and improves the accuracy of breakpoint detection. We also create a customized news tracking application that can notify users without flooding them with every new entry. In this aspect, our application is superior to other services such as Google Alert because we notify users only for significant events.

2.4 Web Analytics Tool Project

Problem: As the information provided by the Web is increasing at a fast rate every day, web sites become structurally more complex, thus making efficient information access a seriously difficult task. The way with which a user browses through the content of a web site is heavily dependent on her needs, knowledge and interests. However, these requirements may differ significantly from the ones considered by the web designer when constructing the site. Consequently, it is essential that a site structure and content should be optimized with respect to the users preferences. This type of commercial requirements increase the need for intelligent web analytics tools which is designed for continuously mining site usage logs and extracting site improvement decisions both semantically and structurally.

Approach: In order to solve the problem above, we have developed an intelligent web analysis system, which includes modules for traffic analysis, usage pattern mining and decision support system. Our system runs web usage mining methods and carries out certain analysis on usage log of each web site. In addition, our system will also have decision support system which analyzes usability of complex structures of web sites like goal paths to show improvements on site structure. In order to increase data processing performance and meet the demand of increasing user logs from multiple web sites our system collects usage data from multiple web sites in its central storage cluster, and a distributed data processing architecture is implemented for running web usage mining methods over usage logs of multiple web sites.

Contributions: Over the usage logs of medium size web site with more than 3K web pages and 200K daily page request, we showed that our web analytics tool has capable of improving commercial web sites both semantically and structurally. Our software framework provides feedback and detailed analysis of commercial web site with several components from web usage mining system to decision support system. Our decision support system utilizes navigation paths with different usage and temporal dimension for optimizing link structure of web site [4]. We also modeled user actions over the web site as a semantic event objects and utilize relationship between these semantic event objects in order to produce different reports for content optimization.

2.5 Integration of topological measures for eliminating non-specific interactions in protein interaction networks

Problem: High-throughput protein interaction assays aim to provide a comprehensive list of interactions that govern the biological processes in a cell. These large-scale sets of interactions, represented as protein-protein interaction networks, are often analyzed by computational methods for detailed biological interpretation. However, as a result of the tradeoff between speed and accuracy, the interactions reported by high-throughput techniques occasionally include false-positive interactions. Unfortunately, many computational methods are sensitive to noise in protein interaction networks; and therefore they are not able to make biologically accurate inferences.

Approach: In this work [7], we propose a novel technique based on integration of topological measures for removing non-specific interactions in a large-scale protein-protein interaction network. After transforming a given protein interaction network using line graph transformation, we compute clustering coefficient and betweenness centrality measures for all the edges in the network. Motivated by the modular organization of specific protein interactions in a cell, we remove edges

with low clustering coefficient and high betweenness centrality values. We also utilize confidence estimates that are provided by probabilistic interaction prediction techniques. We validate our proposed method by comparing the results of a molecular complex detection algorithm (MCODE) to a ground truth set of known *Saccharomyces cerevisiae* complexes in the MIPS complex catalogue database.

Contributions: Our results show that by removing false-positive links in the protein interaction networks, we can significantly increase the biological accuracy of the complexes reported by MCODE.

3 Ubiquitous Computing

We are currently moving to the era of ubiquitous computing and recent development in hardware technology paved the way to small and portable devices such as wireless sensors, PDAs, iPods etc.. Among these technologies, cell phones have been integrated into our life faster than any other one: as of 2009, the number of cell phone subscribers has exceeded 3.3 billion users. The rate of innovation in this field has also been head-spinning. Nokia, Google, Microsoft, and Apple have all introduced cell phone operating systems (Symbian, Android, Windows Mobile, iPhoneOS) and provided APIs for enabling open application development on the cell phones. These modern cell phones with PDA capability, which are dubbed as smartphones, enable location-aware applications as well as empowering the users to generate and access multimedia content.

Mobility information of cell phone users plays an important role in a wide range of cell phone applications, such as context-based search and advertising, early warning systems, traffic planning, route prediction, and air pollution exposure estimation. However the mobility information captured in the cell phone environment are low level data units and they should be converted into suitable format (mobility profile) in order to benefit these applications.

Another gap in the area of Ubiquitous Computing is that despite the availability of the sensor and smartphone devices to fulfill the ubiquitous computing vision, there is no concrete infrastructure exist to task/utilize these devices for collaboration and coordination.

My research projects in the area of ubiquitous computing is related to enhancing smartphone applications with high level mobility profiles and developing infrastructure for utilizing ubiquitous devices for collaboration and coordination. Below, I will give brief summary of our approaches and contributions in each project:

3.1 A Web Based Personalized Mobility Service for Smartphone Applications

Problem: Nowadays, most of the basic web services use instant location information for providing suitable content to the smartphone users. However, more intelligent smartphone applications such as context-based search and advertising, early warning systems, city-wide sensing applications may require additional information about smartphone users such as their mobility profiles. In order to meet more personalized demand of these applications more personalized web services needed.

Approach: In order to support different smartphone applications with personalized mobility information of cell phone users, we propose TRACK ME [8]: A new web based framework for smart-

phone applications with personalized lightweight mobility service as well as location tracking and mobility profile construction. We showed that our personalized mobility service support different smartphone applications and it is lightweight enough to provide fast access to the mobility profiles of smartphone users. Apart from personalized mobility service, our framework also provides solution for location tracking and mobility profile construction problem by employing Map/Reduce architecture. The proposed framework separate heavy processing of mobility profile construction from mobility profile access (which is extremely fast operation).

Contributions: We showed that our personalized mobility services support multiple applications such as location prediction and air pollution exposure risk estimation. Here, we propose an on-line solution to location prediction applications where it is possible to predict future locations of smartphone user instantly by using query interface that is provided by our mobility service. We illustrate that this application is easily used for solving early warning problems mentioned in the example scenario above. For the air pollution exposure risk estimation, we have showed that it is possible to obtain more accurate risk estimation by using our mobility service than residential based approach. We illustrate the example case study for air pollution risk estimation by using Reality Mining data set.

3.2 PRO: A Profile-based Routing Protocol for Pocket Switched Networks

Problem: Delay Tolerant Networks (DTNs), which are also known as intermittently connected networks, or opportunistic, store and-forward networks investigate routing techniques in the environment where the connectivity is not exist all the time. Recently Pocket Switched Networks (PSNs) have been formulated as a subfield of DTNs where each node represents a person with a communication device. Unlike the general DTNs, human factor plays very important role in PSNs. The nature of human mobility and the structure of social networks emerge as important factors in in Pocket Switched Networks, while DTN routing algorithms have been oblivious to them. Therefore, more context aware (mobility profile and social network aware) routing protocols should be developed for PSNs.

Approach: In this work, we are motivated by the observation that using smartphones it is possible to maintain more detailed contextual information about the nodes in the network, and hence design faster and more lightweight routing protocols than the existing work on PSNs. Here, we propose a fast (low-delivery-latency) and efficient (low-message-overhead) routing protocol for PSNs, based on the regularity of human mobility profiles and of intercontact events. Our protocol, namely PRO [9,10], (profile-based routing protocol), is simple yet general enough to be easily instantiated to solve the smartphone search applications.

Contributions: In a break from previous routing protocols, we showed that our protocol treats node encounters as periodic patterns and exploit them to predict times of future encounters. Our profile based estimation of intercontacts yields an accurate ranking of the potential forwarding nodes as to their ability to deliver the message earlier to the destination. We showed that PRO routing protocol is completely decentralized and local to the nodes. PRO runs in an adhoc manner and does not depend on any central infrastructure or third party like Telephone Service Providers.

Using the Reality Mining dataset, we compare the performance of our protocol with most popular previous approaches over both cell based mobility data (coarse granularity) and Bluetooth con-

nection data (fine granularity). Our experimental results showed that PRO routing outperforms previous approaches in terms of end to end delay and communication cost. Finally, we measure the performance of PRO on smartphone queries described above and show that PRO achieves similar query performance with Epidemic routing (in terms of delay and success) while using significantly less communication cost.

3.3 Crowd-Sourced Sensing and Collaboration Using Twitter

Problem: Despite the availability of the sensor and smartphone devices to fulfill the ubiquitous computing vision, the state-of-the-art has gap due to lack of infrastructure to task/utilize these devices for collaboration and coordination. We propose that Twitter can provide an open publish-subscribe infrastructure for sensors and smartphones, and pave the way for ubiquitous crowd-sourced sensing and collaboration applications.

Approach: We design and implement a crowd-sourced sensing and collaboration over Twitter [11], and showcase our system in the context of two applications: a crowd-sourced weather radar, and a participatory noise-mapping application. Our system is composed of three components namely Askweet, Sensweet and Twitter clients. Sensweet is a smartphone application that publishes real-time readings from the integrated-sensors to Twitter. Askweet is a program that listens to its Twitter account for questions and processes the questions and aggregates the replies it receives to these questions from Sensweet and the Twitter clients.

Contributions: We present an analysis of our real-world Twitter experiments to give insights for the feasibility of our approach. We find that although we do not offer the user any incentives to reply, our queries receive at least 15% reply ratios. Surprisingly, 50% of the total replies arrive within the first 10 minutes of our query, and 80% of the replies arrives within the first 2 hours, enabling low latency operations for crowd-sourcing applications. Our experiments also found that consistently the majority of replies come from users that access Twitter from their mobile phones.

3.4 Ongoing Work

3.4.1 Smartphone Applications for Health Care

We are currently working on the air pollutant exposure estimation problem as an application of smartphone based sensing. The previous modelling approaches for estimating air pollutant exposures of the individual use the residential address. The main problem with these methods is that they do not consider time activity data, such as locations that users spend their time. As we have shown in our previous work [1] it is not easy to generalize time activity behaviour of people due to large tail in the location distribution. This implies that by using the residential based methods it is infeasible to reach 100% location coverage, as these approaches capture only the top-k locations, which make up only about 80%-85% of total time. In order to solve this problem, we propose a web-based framework, iMAP [12], which integrates pollution estimation models with time based location data of individuals collected from cell phones in order to estimate air pollution exposure risk.

4 Information Retrieval

During my employment at AGMLAB Information Technologies in Turkey, I have worked on the development of first national search engine called Bilgi.com. Bilgi.com uses Apache Hadoop and Lucene technologies and indexes Turkish content on the Internet periodically.

My first role in Bilgi.com project was to improve the link based page importance calculation methods for Turkish web. To this end, I have extended page rank algorithm with new features such as assigning different weights to the links by using extra information (domain type, existence in DMOZ category). This twist leads to calculation of better link scores to web pages that has incoming links from more trustable and popular domains (based on number of visitors). For performance evaluation we fixed the text based score by feeding same downloaded web pages to search engine and asked a group of users to test query top keywords in Turkish language and assigning scores to the search results. The experimental results showed that we significantly improve the query results for the top Turkish language query terms by tuning the pagerank algorithm.

My first second role in Bilgi.com project was detecting spam farms by using links graph of the Turkish Web. Here, We employed a group of users to manually detect spam urls. Then, we construct spam clusters from this possibly disconnected spam graph by using hierarchical graph clustering methods. After that, we constructed bigger and dense spam clusters by adding new urls to initial clusters by checking outgoing links from Turkish web graph. Using this method we identified more than 50K spam urls automatically by starting with manually determined 3K spam urls.

5 Optimization Algorithms

5.1 Genetic Algorithm for the Multiple-Query Optimization Problem

Problem: Producing answers to a set of queries with common tasks efficiently is known as the multiple-query optimization (MQO) problem. Each query can have several alternative evaluation plans, each with a different set of tasks. Therefore, the goal of MQO is to choose the right set of plans for queries which minimizes the total execution time by performing common tasks only once. Since MQO is an NP-hard problem, several, mostly heuristics based, solutions have been proposed for solving it. However previous approaches such as A* algorithm for MQO problem has scalability problems.

Approach: One of the most popular meta heuristic techniques used for solving complex optimization problems is the genetic algorithms (GA). Since its introduction in early 80s, the GA has attracted a lot of interest from the computer science community, and it has been successfully used for developing efficient solutions too many NP-complete problems.

Here we propose GA approach to the MQO problem [13] since MQO can easily be modeled for a GA. 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. Each gene of a chromosome corresponds to a query. The value of the gene is the plan selected for the evaluation of the corresponding query. To select the chromosomes

for the next generation, the quality of the solution represented by the chromosome is used. This quality is represented by the fitness function, which is simply the inverse of the total execution time of all the tasks in the selected plans for the queries. Under this modeling, MQO is also very suitable for genetic operations to solve the MQO problem.

Contributions: Our work presents evolutionary techniques for solving the MQO problem. Successful A* heuristics have been used for finding optimal solutions to the moderately small (up to ten queries) sized MQO problems. Our work shows that GA is scalable and it is very good candidate for solving bigger MQO problems including more than ten queries. We also showed that it is not practical to solve complex MQO problems including large amount of queries with NP-Complete approaches like A*.

References

- [1] Murat Ali Bayir, Murat Demirbas, and Nathan Eagle. Mobility profiler: A framework for discovering mobility profiles of cell phone users. *To Appear at Pervasive and Mobile Computing Journal, Elsevier*, 2010.
- [2] Murat Ali Bayir, Murat Demirbas, and Nathan Eagle. Discovering spatiotemporal mobility profiles of cell phone users. In *WOWMOM, (acceptance rate=24%)*, pages 1–9, 2009.
- [3] Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar, and Guven Fidan. Smart miner: a new framework for mining large scale web usage data. In *WWW, (acceptance rate=12%)*, pages 161–170, 2009.
- [4] Murat Ali Bayir and Ismail Hakki Toroslu. Link based session model for improving structure of commercial websites. *Submitted to VLDB 2010*, 2010.
- [5] Murat Ali Bayir. A new reactive method for processing web usage data. Master’s thesis, Middle East Technical University, June 2006.
- [6] Cuneyt Gurcan Akcora, Murat Ali Bayir, Murat Demirbas, and Hakan Ferhatosmanoglu. Identifying breakpoints in public opinion. In *SOMA 2010, SIGKDD Workshop on Social Media Analytics (full talk acceptance rate=20.6%)*.
- [7] Murat Ali Bayir, Tacettin Dogacan Guney, and Tolga Can. Integration of topological measures for eliminating non-specific interactions in protein interaction networks. *Discrete Applied Mathematics*, 157(10):2416–2424, 2009.
- [8] Murat Ali Bayir, Murat Demirbas, and Ahmet Cosar. A web based personalized mobility service for smartphone applications. *To Appear at The Computer Journal, Oxford University Press, The British Computer Society*, Special Issue on best papers of ISCIS 2009 (ranked in top 5% among 240 submissions), 2010.
- [9] Murat Ali Bayir and Murat Demirbas. Pro: A profile based routing for pocket switched networks. *Submitted to GLOBECOM 2010*.
- [10] Murat Ali Bayir and Murat Demirbas. On the fly learning of mobility profiles for intelligent routing in pocket switched networks. *Technical Report, 2009-03, Department of Computer Science and Engineering, University at Buffalo*.
- [11] Murat Demirbas, Murat Ali Bayir, Cuneyt Gurcan Akcora, Yavuz Selim Yilmaz, and Hakan Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. In *WOWMOM, (acceptance rate=21%)*, 2010.
- [12] Murat Demirbas, Carole Rudra, Atri Rudra, and Murat Ali Bayir. imap: Indirect measurement of air pollution with cellphones. In *PerCom Workshops*, pages 1–6, 2009.
- [13] Murat Ali Bayir, Ismail Hakki Toroslu, and Ahmet Cosar. Genetic algorithm for the multiple-query optimization problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 37(1):147–153, 2007.