論文の要旨

題目：**Studies on MapReduce based Area Skyline Query and Parameter Estimation of Queueing Systems**

(MapReduceによるエリアスカイライン問合せ及びキューイングシステムのパラメータ推定に関する研究)

氏名 **CHEN LI**

## Abstract

In this dissertation, the author discusses two kinds of research works, MapReduce based computation of area skyline query for location recommendation, and parameter estimation of queueing systems with utilization data, respectively.

With the rapid development of computer technology and network technology, a vast amount of available information may cause information overload problem. The exponential growth of data makes customers inundate with various choices. Since the ability to resolve the ``information overload'' problem, recommendation systems are widely accepted in E-commerce in recent years. The basic premise of recommenders is to reduce noise and filter out information, which is not relevant to customers' tastes. In the past decades, many researchers worked on generating good recommenders. Recommendation systems can help enterprises improve economies. For example, Amazon.com claimed that product sales grew 35% from the recommendation system. Moreover, Netflix.com claimed that 66% of movies were rented by using recommendation systems. Google News Recommendations generated 38% more click-throughs.

Location recommendations are the main part of recommendation systems. For example, the selection of good locations is essential in the service filed, which may help customers find implicit preferable places in an unfamiliar environment. Moreover, it is also critical for a businessperson to find a suitable location to build a company.

Skyline query is a well-known information retrieval approach for recommendation systems. Skyline query can select a small number of representative objects from an extensive database, which already has been applied in the location selection problem. For example, a traveler wants to find a hotel from several candidate hotels. In this case, we can use the skyline query to help the traveler select a preferable hotel. However, in some real-world scenarios, the candidate points are not given for the

location recommendation problem. For example, the businessman looks for a good vacant area to build a hotel at any cost. In such a situation, candidate points do not exist. Besides, the two-dimensional area is much complicated than a candidate point, which makes the problem of good location selection more difficult and challenging. However, most of the skyline algorithms to solve the problem of location selection assume that the candidate points always exist. To generate the spatial relationship between objects, we consider solving such problem by using our team's previous work, called the area skyline query.

Area skyline query is a new skyline query for selecting spatial area objects on a map. Grid-based Area Skyline (GASKY) algorithm is an efficient and practical algorithm by dividing data structure into grids for retrieving interesting areas. Assume a businessman wants to find an excellent area to build a hotel. The area should be close to some preferable facilities, such as bus/train stations and sightseeing spots, and is far from some unpreferable facilities, such as open landfills and noisy places. In such a situation, GASKY first divides a square region on a map into several girds. Then, we calculate the minimum and maximum distance from an area to the closest preferable facility and the farthest unpreferable facility of each facility type. After the calculation of the minimum and maximum distance of every grid, the problem of the area skyline query can be transformed into the conventional skyline query. However, the computational cost of the time complexity for GASKY algorithm is much higher than the conventional skyline query. Furthermore, the average processing time increases linearly with the growth of facilities.

To resolve the poor performance problem of GAKSY and to handle ``big data'' well, we consider a distributed algorithm for computing the GASKY. MapReduce is a programming model and an associated implementation for processing big data. Mainly, a MapReduce program is composed of a Map function and a Reduce function. The Map function takes a set of data and converts it into another collection of data, where individual elements are broken down into tuples. The Reduce function takes the output from the Map function as an input and combines those data tuples into a smaller set of tuples. In recent years, the MapReduce framework is widely used for computing skyline query. In general, there are three main MapReduce based algorithms for skyline query processing, called MR-BNL, MR-SFS, and MR-Bitmap. Unfortunately, only a few algorithms focus on spatial skyline query in such parallel way. The GASKY algorithm is a new skyline query for selecting spatial area objects. In this work, we propose a novel algorithm to improve the performance of GASKY to handle large-scale database.

In the second work, the author discusses the parameter estimation problem of queueing systems.

Performance evaluation plays a vital role in the computer design phase. Performance evaluation can help computer designers determine the optimal system configuration, such as the size of memory, the number of CPU, and the storage capacity. As the complexity of system increases, the new system architecture tends to integrate a collection of independent systems, called a system of systems (SoS). To configure such kind of an SoS, performance evaluation becomes more challenging and more critical.

In general, there are three main methods for performance evaluation of computer systems, measurement-based performance evaluation, simulation-based performance evaluation, and model-based performance evaluation, respectively. If the system already exists, we usually use the measurement method to evaluate performance. When the system is under designing, we select to use simulation-based method and model-based method for performance evaluation. Simulation-based performance evaluation is more flexible, accurate, and credible, but require more time to derive models. Model-based performance evaluation constructs a systematic mathematical model and quantitatively analyzes the model. For example, Markov chains, queuing systems, and queuing networks are conventional techniques of model-based methods, which are usually used in telecommunication systems. Although to some extent, the model-based approach is an approximation, the amount of calculation is small.

Queueing systems for model-based performance evaluation are very common. Typically, a single queue has three main components: input, queue, and server. The input is a stream of customers or jobs. They wait for service in the queue and leave the system after the service. More specifically, we usually define a queueing system by symbol A/S/m/k. A is the type of the arrival process, and S is the distribution of the service time of customers. m is the number of servers, and k is the queue capacity. For example, M/M/1/K means that the customers arrive at the service facility via the Poisson process, and the service time follows an exponential distribution.

Moreover, there is only one server in the system with a capacity of k. In other words, the input processes and service distributions are two necessary components for a queueing system. Unfortunately, in practice, it is not easy to model the input process and service distribution. In general, we must observe the arrival process and service time in a fixed time, and then we can estimate the arrival rate and service rate. In most queueing systems, the arrival process and service time is observable. In other words, we can know the exact job arrival time and service time by a period of observation. However, the

parameter estimation of the arrival processes in computer systems, such as CPU utilization, are quite limited.

Moreover, we usually assume that the arrival rate is constant for a queueing system. However, in the real world, the fixed arrival rate cannot model the arrival process very well. For example, the customers' arrival rates in a convenient store are always varying dynamically. The arrival rate of morning and evening rush hours may be higher than the rate of other periods.

Non-Homogeneous Poisson Process (NHPP) is a Poisson process over a non-linear time scale. In this work, the author considers an NHPP for the job arrival process of a computer system, and generate an Mt /M/1/K queueing system. The author aims to estimate the arrival rate, which varies as a function of time by using utilization data.

To summarize, this dissertation introduces two kinds of research works. The outline of this dissertation is organized as follows.

Chapter 1 discusses the motivations of the two research works and gives the organization of this dissertation.

In Chapter 2, the author introduces the first research work in detail. In this chapter, we first introduce the definitions of skyline query, spatial skyline query, and area skyline query. Secondly, we review the related works. And then, we propose our novel MapReduce based algorithm for the calculation of area skyline query. Finally, we conduct two numerical experiments on both the synthetic dataset and real dataset. The experimental results confirm that our proposed model is efficient and effective for handling ``big data.''

In Chapter 3, the author presents the detail of the other research work: parameter estimation of queueing systems with dynamic arrival process from utilization data. In this chapter, we first give some definitions of queueing system, utilization data, and Non-Homogeneous Poisson Process. Secondly, we review the related works. And then, we introduce a parameter estimation method, which is called maximum likelihood estimates. Also, we propose to use the Expectation-Maximization algorithm to overcome the incomplete information problem. Finally, we conduct two experiments on both synthetic utilization data and real CPU utilization data to discuss the performance of our proposed model.

Finally, a concluding discussion of the contributions of the two research works and conclusions are discussed in Chapter 4.