JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 10 : 2301~2306 Oct. 2015

# TSK 퍼지 모델 이용한 효율적인 빅 데이터 PCP 예측 알고리즘

김장영*

# An Efficient Algorithm for Big Data Prediction of Pipelining, Concurrency (PCP) and Parallelism based on TSK Fuzzy Model

**Jang-Young Kim***

Department of Computer Science, The University of Suwon, Hwaseong 445-743, Korea

## 요 약

정보가 급증함에 따라 큰 용량의 데이터를 전송해야 할 경우가 있다. 빅 데이터 전송 기술은 큰 용량의 데이터를 전송할 때 필요하다. 본 논문은 빅 데이터를 최적화된 속도로 전송하기 위해 GridFTP의 주된 기능인 PCP를 사용하며 또한 PCP 값을 예측하는 알고리즘을 개발한다. 또한, TSK 퍼지 모델을 적용하여 PCP에 따른 최적화된 전송률을 측정하는데 사용된다. 따라서, 제안된 TSK모델을 이용한 PCP 예측 알고리즘은 본 논문의 우수성을 입증한다.

## ABSTRACT

The time to address the exabytes of data has come as the information age accelerates. Big data transfer technology is essential for processing large amounts of data. This paper posits to transfer big data in the optimal conditions by the proposed algorithm for predicting the optimal combination of Pipelining, Concurrency, and Parallelism (PCP), which are major functions of GridFTP. In addition, the author introduced a simple design process of Takagi-Sugeno-Kang (TSK) fuzzy model and designed a model for predicting transfer throughput with optimal combination of Pipelining, Concurrency and Parallelism. Hence, the author evaluated the model of the proposed algorithm and the TSK model to prove the superiority.

**키워드** : 파이프라이닝, 컨커런시, 패러랠리즘, 빅데이터, TSK 퍼지모델

Key word : Pipelining, Concurrency, Parallelism, Big data, TSK fuzzy model

**Received** 14 August 2015, **Revised** 11 September 2015, **Accepted** 24 September 2015

* **Corresponding Author** Jang-Young Kim (E-mail: jykim77@suwon.ac.kr, Tel: +82-31-229-8345)
Department of Computer Science, The University of Suwon, Hwaseong 445-743, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2015.19.10.2301 print ISSN: 2234-4772 online ISSN: 2288-4165

## Ⅰ. INTRODUCTION

Big data transfer technologies have been in the spotlight with three properties of the velocity, volume and variety.

As these technologies are actively researched, areas of computer network researching data transfer technology are also emerging as important topics. Although FTP is being used in order to allow the transfer of the data among the computers, it may decrease the throughput of data transfer especially in big data transfer. In order to solve this problem, GridFTP [1] is popularly used to achieve an optimal throughput to transfer the large amounts of data by enhancing security, transfer speed and reliability of data transmission. Existing work has tried to find optimal values of Pipelining, Concurrency and Parallelism (PCP) by using historical PCP datasets. Hence, the author is willing to propose an efficient algorithm with Takagi-Sugeno-Kang (TSK) fuzzy model [2, 3] to predict optimal values of PCP and the throughput. PCP includes the following functions.

Pipelining is a method of transmitting files continuously without waiting for a response signal for the previous transmission. Concurrency usually transmits various files via different channels at the same time. Parallelism is a method of transmitting different parts of the same file via multiple parallel data channels at the same time. Therefore, three PCP functions are very useful for large-scale data transmissions.

GridFTP [1] can improve throughput by optimizing PCP values which are manageable major functions depending on file size, number of files, bandwidth, round trip time (RTT) and buffer size. Prior studies [4-10] suggested some algorithms for finding the optimal combination of parallelism. However, a drawback of these studies is that throughput of data transfer may decrease due to overhead, which could occur when the conventional algorithm finds the optimal combination of PCP values by using the information gained by transferring sampling files through network channel.

Hence, this paper suggests an efficient algorithm for predicting the optimal combination of PCP based on fuzzy model; this is appropriate for certain circumstances of data transfer based on the measured abundant experimentation datasets which contain throughput values depending on PCP in different testbed environments. In addition, the author designed a model of predicting the throughput of data transfer under the optimal combination of PCP using the processed experimentation dataset by TSK fuzzy model.

## Ⅱ. BACKGROUND

Abundant PCP experimentation datasets measured in various testbeds contain more than seventeen factors such as file size, number of files, bandwidth, round-trip time, buffer size, pipelining, concurrency, parallelism, throughput, transfer duration and so on. The system will be complex and difficult to interpret because each factor has a non-linear relationship with the throughput and other factors. The PCP background is designed in Fig. 2.

In this paper, the author used clustering, which is a data classification algorithm that identifies the nature of PCP experimentation dataset and specially contains the concept of fuzzy to reflect a more specific characteristic of the data. The author designed a model of predicting the highest throughput of data transfer under the optimal combination of pipelining, concurrency and parallelism based on PCP experimentation dataset. The Takagi-Sugeno-Kang fuzzy model [2, 3] can approximate the very complex non-linear system based on fuzzy rule-based inference.

### 2.1. Takagi-Sugeno-Kang Fuzzy Model

TSK Fuzzy Model is a method of inference based on the fuzzy rule for approximating the complex non-linear system. After creating several rules by dividing input

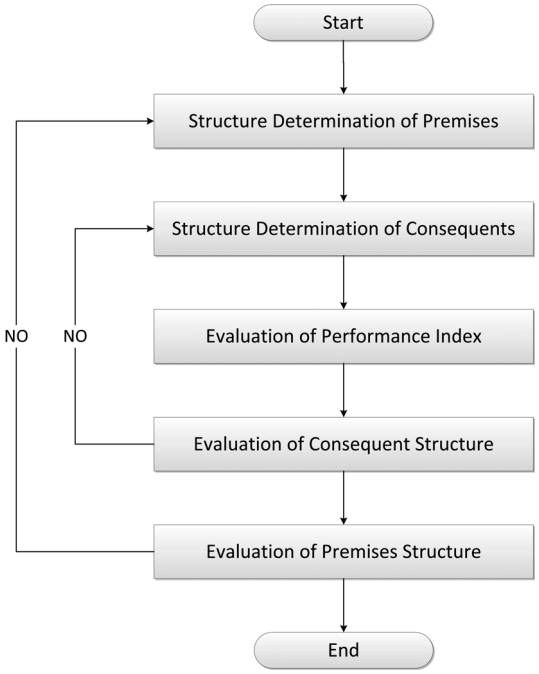


**그림 1.** TSK 퍼지 모델
**Fig. 1** TSK fuzzy model

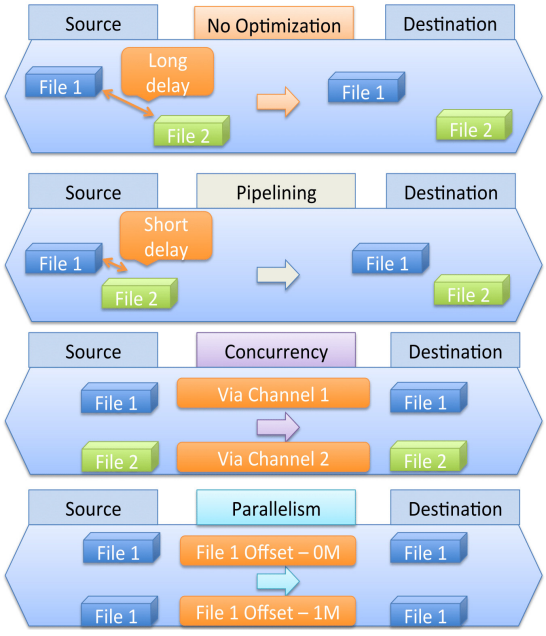


**그림 2.** PCP 지식 배경
**Fig. 2** PCP Background

space into fuzzy area, which infers final output by applying fitness into each rule. Here are the steps of the procedure of TSK Fuzzy Model (Fig. 1).

## Ⅲ. PROPOSED ALGORITHM

The proposed algorithm finds the most similar data by comparing certain input data with the existing abundant datasets. Most similar data is determined by the data that have the lowest sum of error rate by comparing factors of input data with factors of data in the dataset. In addition, the maximum operators were added in order to consider the factor that contains the outlier.

The proposed algorithm uses the following steps to find optimal PCP values. First, each error is calculated by comparing each of the factors of the data in dataset with each factor of the input data. Second, the algorithm calculates the maximum value on each factor. Third, the algorithm finds the error rate by dividing the maximum value of each factor for each factor in error. Fourth, the algorithm calculates the sum of error rate (SumER) for each data. Fifth, the algorithm calculates the maximum value (MaxER) for each of the error rate data. Sixth, the algorithm calculates the final error rate by combining SumER and MaxER using weight coefficient. Finally, the algorithm finds the minimum value out of final error rate and corresponding data number is determined as most similar data with input data.

The above comparison algorithm is a method of comparing each error rate. However, several factors need to be taken into account simultaneously for one data. If the scale of range of each factor is different, then the comparison method cannot determine the most similar characteristics of data. The proposed algorithm is able to compare characteristics of data by compensating for the differences of range for each factor, using rate of error in the case of several factors exist in data.

Hence, they are useful characteristics for comparing data that contains various scale factors. Therefore, the author applied the concept of the maximum operator to error rate to control for the amount of removal of the outlier.

Finally, compared value is calculated by the total amount of the sum of the error rate multiplied by the weight coefficient and maximum value of error rate multiplied by the weight coefficient. If the weight coefficient is close to 1, then it is difficult to remove the outlier because it is a total comparison method. On the other hand, if weight coefficient is close to 0, then it is more likely to remove the outlier because it is a partial comparison method. The weight coefficient can be used by adjusting to appropriately find most similar characteristics of data.

## Ⅳ. EXPERIMENTAL RESULTS

### 4.1. Proposed algorithm for finding optimal combination of PCP values for the highest throughput based on PCP experimentation dataset

The author applied the proposed algorithm to PCP experimentation dataset that contains data transfer throughput depending on file size, number of files, bandwidth, round trip time, buffer size, pipelining, concurrency and parallelism generated from various testbeds. Input data is the data corresponding to certain circumstances of file transmission. The author compared the characteristics of input data with characteristics of data in a PCP experimentation dataset and determined the most similar characteristics of data and determining the PCP values of corresponding data number as an optimal combination of pipelining, concurrency and parallelism for certain circumstance of file transmission.

In this experiment, the author applied the most effective five factors to throughput of data transfer such as file size, number of files, bandwidth, round trip time and buffer size in the proposed algorithm. Also, it is essential to preprocess the PCP experimentation dataset before applying the proposed algorithm to dataset.

PCP experimentation dataset contains many data that have the same values in file size, number of files, bandwidth, round trip time and buffer size. However, PCP values and throughput values are different. Therefore, a preprocessed experimentation dataset is constructed by extracting each data of highest throughput in all possible cases of the values is same in factors such as file size (Bytes), number of files, bandwidth (BW, Mbps), round trip time (RTT, seconds), buffer size (BS, Bytes) and Throughput (Th, Mbps).

**표 1.** 전처리 실험 데이터
**Table. 1** Preprocessed Experimentation Dataset

| No | File Size | Number of Files | BW | RTT | BS | Th |
|---|---|---|---|---|---|---|
| 1 | 262144 | 1000 | 1000 | 0.06 | 131071 | 16.81 |
| 2 | 262144 | 3000 | 1000 | 0.06 | 131071 | 91.69 |
| . . . | . . . | . . . | . . . | . . . | . . . | . . . |
| 78 | 13421772 | 512 | 1500 | 0.002 | 1677721 | 1935.3 |

Testing dataset is randomly constructed within the error of 10% with the processed experimental dataset.

**표 2.** 테스팅 데이터
**Table. 2** Testing Dataset

| No | File Size | Number of Files | BW | RTT | BS | Th |
|---|---|---|---|---|---|---|
| 1 | 280000 | 900 | 1010 | 0.06 | 131071 | 16.81 |
| 2 | 270000 | 3300 | 900 | 0.06 | 131071 | 91.69 |
| . . . | . . . | . . . | . . . | . . . | . . . | . . . |
| 78 | 14421772 | 550 | 1650 | 0.002 | 1677721 | 1935.3 |

**표 3.** 실험결과
**Table. 3** Results of the Experiment

| Number of testing data | Number of processed data | Number of success | Number of failure | Rate of success |
|---|---|---|---|---|
| 78 | 78 | 78 | 0 | 100% |

The result of the experiment shows highly accurate determination capability, though the range of each factor is very different.

### 4.2. TSK fuzzy model for predicting throughput under optimal combination of PCP

A prior preprocessed experimentation dataset is used as training data in design of TSK Fuzzy Model. Also, a testing dataset is constructed by extracting each data with the second highest throughput in the same value in file size, number of files, bandwidth, round trip time and buffer size from original PCP experimentation dataset.

## Ⅴ. CONCLUSION AND FUTURE WORK

In this paper, the author predicted an optimal combination of pipelining, concurrency and parallelism (PCP) for certain circumstances of file transfer such as accruing overhead, network saturation, based on experimental dataset measured by various testbeds. Hence, it would be feasible to transfer large-scale data in optimal throughput with GridFTP by using optimal combination of PCP.

In addition, the author designed a model for predicting throughput of file transfer under the optimal combination of PCP by using TSK fuzzy rule based inference. In future work, the author will optimize weight of error rate in the proposed algorithm based on objective function. Therefore, the new optimization algorithm will provide more efficient way and accomplish to find optimal values of PCP and throughput.

## REFERENCES


[ 1 ] GridFTP, Globus Online “http://www.globus.org”

[ 2 ] T. Takagi and M. Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” *IEEE Trans. Syst., Man, Cybern*., vol. 15. pp. 116-132, Jan. 1985

[ 3 ] M. Sugeno and T. Yasukawa, “A fuzzy-logic-based approach to qualitative modeling,” *IEEE Trans, Fuzzy Syst.*, vol. 1, pp. 7-31, 1993.

[ 4 ] E. Yildirim, J. Kim, and T. Kosar, “Optimizing the sample size for a cloud-hosted data scheduling service,” *Proc. 2nd International Workshop on Cloud Computing and Scientific Applications (CCSA in conjunction with CCGRID12)*, 2012.

[ 5 ] J. Kim, E. Yildirim, and T. Kosar, “A highly-accurate and low-overhead prediction model for transfer throughput optimization,” *Proc. of DISCS Workshop*, November 2012.

[ 6 ] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and S. Tuecke, “Software as a service for data scientists,” *Communications of the ACM*, vol.55:2, pp.81 – 88, 2012.

[ 7 ] E. Yildirim, J. Kim, and T. Kosar, "Modeling Throughput Sampling Size for a Cloud-hosted Data Scheduling and Optimization Service," *In Future Generation Computer Systems (FGCS)*, Vol. 29, No. 7 (2013) pp 1795-1807.

[ 8 ] E. Yildirim, J. Kim, and T. Kosar (Best Paper Award), "How GridFTP Pipelining, Parallelism and Concurrency Work: A Guide for optimizing large dataset transfers," In *Proceedings of IEEE/ACM Supercomputing'12 Workshop on Network-Aware Data Management (NDM 2012)*, Salt Lake City, UT, November 2012.

[ 9 ] E. Yildirim, M. Balman, and T. Kosar, “Data-intensive Distributed Computing: Challenges and Solutions for Large-scale Information Management, ch. Data-aware Distributed Computing, *IGI-Global*, 2012.

[10] E. Yildirim, D. yin, and T. Kosar, “Prediction of optimal parallelism level in wide area data transfers,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 12, pp. 2033-2045, 2011.


## 김장영(Jang-Young Kim)

2005년 2월: 연세대학교 컴퓨터과학 공학사
2010년 5월: Pennsylvania State Univ. 공학석사
2013년 7월: State University of New York 공학박사
2013년 8월: University of South Carolina 조교수
2014년 3월: 수원대학교 컴퓨터학과 조교수
※관심분야 : Big data, Cloud computing, Networks