Greedy 알고리즘을 사용한 데이터 분할 시스템

김민우[O*], 김세준[*], 이병준[*], 김경태[*], 윤희용[**]
[O*]성균관대학교 정보통신대학 전자전기컴퓨터공학과
[**]성균관대학교 소프트웨어대학 소프트웨어학과
e-mail: {kimmw95, ksj105, byungjun}@skku.edu[O*], kyungtaekim76@gmail.com[*], youn7147@skku.edu[**]

Data Segmentation System using Greedy Algorithm

Min-Woo Kim[O], Se-Jun Kim[*], Byung-Jun Lee[*], Kyung-Tae Kim[*], Hee-Yong Youn[**]
[O*]Dept. of Electrical and Computer Engineering, Sungkyunkwan University
[**]Dept. of Software, Sungkyunkwan University

● 요 약 ●

머신 러닝 환경에서 많은 양의 데이터를 한꺼번에 학습하게 되면 데이터 트래픽이 증가함에 따라 흐름 정체가 발생하고 학습 품질이 저하되며 학습속도 지연 등의 문제가 발생한다. 본 연구는 머신러닝 환경에서 빅 데이터 학습 데이터 분할을 위한 핵심 목표인 Greedy 알고리즘에 대해 설명하고 간단한 Greedy 알고리즘을 사용하여 각각의 데이터 파티션을 생성하여 학습 속도의 효율성을 향상시키는 방법을 제안한다.

키워드: 학습 데이터(training data), 빅 데이터(big data), 분할(segmentation), 분산 처리(distributed processing)

I. Introduction

정보화 사회로 접어들며 데이터의 사용이 많아지고 데이터의 크기 또한 커지게 되었다. 이러한 데이터들을 계산하기위해서는 오랜 시간이 소요되며 데이터 트래픽으로 인해 지연이 발생하게 되면 데이터의 손실을 야기 할 수 있다. 빅 데이터의 활용은 계속 증가하고 있으며 인공지능의 학습을 기반으로 빅 데이터가 사용되어지면 학습 데이터의 처리를 위해 많은 시간이 소요되는 문제점에 직면하게 된다. 본 논문에서는 Greedy 알고리즘의 관점에서 문제에 접근하여 학습 소요시간을 단축시키는 해결책을 제안한다. 2장에서는 해외의 분산 처리 시스템과 관련된 기존의 연구에 대한 내용들을 서술하였고 3장에서는 Greedy 알고리즘과 분산 처리 시스템에 대한 간단한 설명과 프로세스에 대해 기술하였다[1][2]. 4장에서는 본 논문에 대한 결론과 향후 연구 방향에 대하여 서술하였다.

II. Preliminaries

1. Related works

해외의 분산 처리 시스템의 동향을 살펴보면 미국을 선두로 독일, 영국 등 유럽에서 관심을 보이고 있으며 기상, 인공지능, 수학, 생명공학, 암호학 등 다양한 분야에서 사용되어진다. 빅 데이터의 방대한 양으로 인해 데이터를 분할 시켜 처리하는 분산 처리와 여러 프로세서들이 병렬로 처리하는 병렬 처리 등의 방법이 제안되었다. Dremeld[3]의 논문을 기반으로 제안 되어 만들어진 Cloudera[4], Apache Tez[5] 등 과 Facebook의 Presto[6] 등이 병렬, 분산처리를 위해 사용되었다.

III. The Proposed Scheme

3.1. Greedy 알고리즘/분산 처리 시스템

Greedy 알고리즘은 최적의 해를 구할 때 사용되는 방법 중 하나로 순간마다 최적의 결과 값을 선택하는 방식으로 최종 결과에 도달한다. 본 논문에서 제안하는 Greedy 알고리즘을 적용하기 위해서는 먼저 분산 처리 시스템에 대해 간단히 알아야 한다. Fig 1은 분산 처리 시스템의 기본적인 구조이다.

Fig 1의 구조에서 볼 수 있듯이 여러 프로세서들의 처리 능력을 이용하여 Greedy 알고리즘을 통해 분할된 빅 데이터들을 효율적으로 학습할 수 있다.

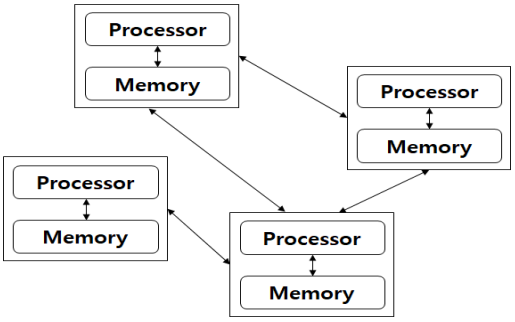

Fig. 1. 분산 처리 시스템 구조

3.2. Greedy 알고리즘 프로세스

본 논문에서 제안하고자 하는 메인 아이디어는 데이터를 N개의 값으로 분할할 때 Greedy 알고리즘의 특성상 전체적인 결과가 최적이라는 보장이 없다는 문제에 대해 결과 값이 평균에 미치지 못하거나 최소인 블록에 최대 결과를 낼 수 있는 항목을 추가하여 분할한다. 그 후 N개의 파티션으로 나누어 전체 목표를 점진적으로 개선하여 결과 값을 높인다. Fig 2는 Greedy 알고리즘을 사용한 데이터의 분할 처리 과정을 간단하게 보여준다.

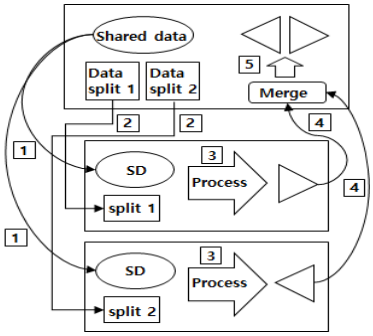

Fig. 2. Greedy 알고리즘 프로세스

동시에 공유되어 주어지게 되는 빅 데이터로부터 개별적으로 계산 또는 학습될 데이터의 분할들은 서로 의존적이지 않아야 하며 계산 후에 병합을 통해 완전한 결과를 만들어 낼 수 있어야한다. 이와 같이 분할된 데이터에 대해 연산을 수행하는 부분은 하위 프로세스로 구현해야 한다.

IV. Conclusions

본 연구를 통해 Greedy 알고리즘을 사용하여 데이터 분할의 효율성에 따른 가능성을 제안하였다. 향후 연구로서는 Greedy 알고리즘을 사용하여 분할된 데이터를 처리하고 계산 등의 연산을 수행하는 하위 프로세스를 구현할 것이며 분할된 데이터를 집계하는 과정에 대한 분산 컴퓨팅에 대하여 연구할 것이다. 또한 분산 처리에 적합한 서브 모듈러에 최적으로 배포하는 방법에 대한 연구도 진행할 예정이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 정보통신·방송연구 개발 사업(No. 2016-0-00133, 초연결 IoT 노드의 군집 지능화를 통한 Edge Computing 핵심 기술 연구), SW중심대학지원사업(2015-0-00914), 한국연구재단 기초연구사업(No.2016R1A6A3A11931385, 실시간 공공안전 서비스를 위한 소프트웨어 정의 무선 센서 네트워크 핵심기술 연구, 2017R1A2B2009095, 실시간 스트림 데이터 처리 및 Multi-connectivity를 지원하는 SDN 기반 WSN 핵심 기술 연구), 삼성전자, BK21PLUS 사업의 일환으로 수행되었음.

REFERENCES

[1] L. Zhao, H. Nagamochi, T. Ibaraki, “On generalized greedy splitting algorithms for multiway partition problems.”, Discrete applied mathematics, pp. 130-143, 2004.

[2] K. Wei, R. Iyer, J. Bilmes, “Submodularity in data subset selection and active learning.” In ICML, 2015.

[3] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis, “Dremel: Interactive Analysis of Web-Scale Datasets”, Proceedings of the VLDB Endowment, Vol. 3, No. 1, pp. 330-339, 2010

[4] Cloudera. https://www.cloudera.com/

[5] Apache Tez. http://tez.apache.org

[6] Presto. https://prestodb.io/