데이터 관리 효율성을 위한 데이터 마이닝 기반 데이터 접근 주소 분석

이현섭*
백석대학교 컴퓨터공학부 교수

Data Mining-Based Data Access Address Analysis for Data Management fficiency

Hyun-Seob Lee*
Professor, Division of Computer Engineering, Baekseok University

요 약 스토리지 기술의 급속한 발전으로 플래시 메모리 기반 스토리지가 기존의 하드 디스크 드라이브를 대체하며 고성능 스토리지 솔루션으로 부상하고 있다. 특히 SSD는 빠른 접근 속도와 낮은 전력 소비 때문에 데이터 센터, 클라우드 컴퓨팅, 모바일 장치에서 널리 채택되고 있다. 그러나 SSD의 미디어는 비대칭의 읽고 쓰기 성능과 데이터를 쓰고 지워야 하는 특성 등 플래시 메모리의 독특한 특성이 있다. 따라서 SSD의 성능과 수명을 극대화하기 위해서는 효율적인 데이터 관리가 중요하며, 이를 위해서는 내부 데이터 구조와 접근 패턴에 대한 이해가 필요하다. 특히, 랜덤하게 사용되는 데이터에서 빈번하게 접근되는 데이터 패턴을 파악하면 스토리지 시스템 성능을 크게 향상하고, SSD 수명을 연장하며, 효과적인 데이터 관리 전략을 수립할 수 있을 것으로 기대한다. 따라서 본 연구는 데이터 관리 효율성을 높이기 위해 SSD의 접근 패턴을 분석하는 것을 목표로 한다. 이를 위해 데이터 마이닝 기법을 사용해 대규모 SSD 접근 주소 데이터 세트에서 의미 있는 패턴을 추출하고 분석한다. 이 연구는 먼저 SSD 접근 주소를 시각화하여 접근 패턴을 파악하고, 빈번한 접근의 빈도와 분포를 분석하며, 성능 및 데이터 관리 전략을 최적화하기 위한 예측 모델을 제안한다. 제안된 방법론은 데이터 마이닝을 통해 SSD 데이터 구조와 접근 패턴을 분석하는 체계적인 접근 방식을 제공하여 다양한 스토리지에 적용할 수 있으며, 분석된 연구 결과는 실제 시스템에 적용되어 스토리지를 최적화하고 효과적인 데이터 관리 솔루션 개발을 지원할 수 있을 것으로 기대한다.

주제어 : 데이너 분석, 데이터 마이닝, 접근 빈도, 데이터 패턴, 플래시 메모리

Abstract With rapid advances in storage technology, flash memory-based storage is replacing traditional hard disk drives and emerging as a high-performance storage solution. SSD, in particular, are widely adopted in data centers, cloud computing, and mobile devices because of their fast access speeds and low power consumption. However, the media in SSD has unique characteristics of flash memory, including asymmetrical read and write performance and the need to write and erase data. Therefore, efficient data management is critical to maximize the performance and lifespan of SSD, which requires an understanding of their internal data structures and access patterns. In particular, identifying frequently accessed data patterns in randomly used data is expected to significantly improve storage system performance, extend SSD lifetime, and enable effective data management strategies. Therefore, this research aims to analyze the access patterns of SSD to improve data management efficiency. To achieve this, we use data mining techniques to extract and analyze meaningful patterns from a large dataset of SSD access addresses. The study first visualizes SSD access addresses to identify access patterns, analyzes the frequency and distribution of frequent accesses, and proposes a predictive model to optimize performance and data management strategies. The proposed methodology provides a systematic approach to analyze SSD data structure and access patterns through data mining, which can be applied to a wide range of storage, and the results of the study can be applied to real-world systems to optimize storage and support the development of effective data management solutions.

Key Words : Data Analysis, Data Mining, Access Frequency, Data Pattern, Flash Memory

*This paper was supported by 2025 Baekseok University Research Fund
*교신저자 : 이현섭(hyunseob@bu.ac.kr)
접수일: 2025년 02월 13일 수정일 2025년 03월 31일 심사완료일 2025년 04월 10일

1. 서론

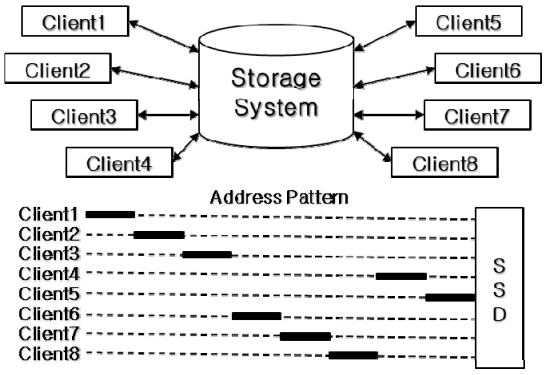
최근 저장장치 기술의 급속한 발전으로 플래시 메모리 기반 저장장치가 전통적인 HDD(hard disk drive)를 대체하며 고성능 스토리지 솔루션으로 자리 잡았다. 특히, SSD(solid state drive)는 빠른 접근 속도와 저전력 소비의 장점 때문에 데이터 센터, 클라우드 컴퓨팅, 모바일 기기 등 다양한 분야에서 사용되고 있다[1-4]. 그러나 SSD의 미디어인 플래시 메모리는 쓰기 전 저장 영역을 지워야 하고(erase before write) 읽고 쓰는 단위와 지우는 단위의 메모리 영역 크기가 다른 특성 때문에 효율적인 데이터 관리가 필요하다[5-9]. 이러한 배경에서 SSD의 데이터 관리 효율성을 극대화하기 위해서는 내부 데이터 구조와 접근 패턴을 이해하는 것은 중요하다. 특히, 랜덤하게 접근되는 데이터로부터 빈번하게 접근하는 데이터의 구조와 접근 패턴을 찾아내면 저장시스템의 성능과 효율성을 크게 개선하고, 수명을 연장하며, 데이터 관리 전략을 수립하는 데 있어 중요한 역할을 할 수 있을 것으로 기대된다.

데이터마이닝[10-13] 기술은 대용량 데이터에서 유의미한 패턴을 추출하기 위한 기법으로 활용하고 있다. 따라서 본 연구에서는 데이터 마이닝 기법을 이용하여 SSD에서 접근된 데이터의 주소로부터 성능 및 데이터 관리에 영향을 줄 수 있는 패턴을 분석하는 것은 스토리지 시스템의 효율성을 크게 향상할 것으로 기대된다. 이를 위해 먼저 SSD의 접근 주소를 시각화하여 주소의 패턴을 분석한다. 그리고 빈번하게 접근되는 데이터의 빈도와 분포를 분석한다. 또한, 분석 결과를 바탕으로 스토리지 시스템의 성능 최적화 및 데이터 관리 전략에 도움을 줄 수 있는 예측 모델을 제안한다. 데이터 마이닝 기법을 통해 SSD의 데이터 구조와 접근 패턴을 분석하는 방법은 다양한 저장시스템에 활용될 수 있고, 분석을 통해 각 저장시스템의 성능과 효율성을 최적화하는 연구에 크게 이바지할 것이다. 따라서 이 연구를 통해 SSD를 효율적으로 활용할 수 있는 데이터 패턴 분석의 기반을 마련하고, 실제 시스템에 적용 가능한 솔루션을 제시함으로써 스토리지 기술의 발전에 기여할 것으로 기대한다.

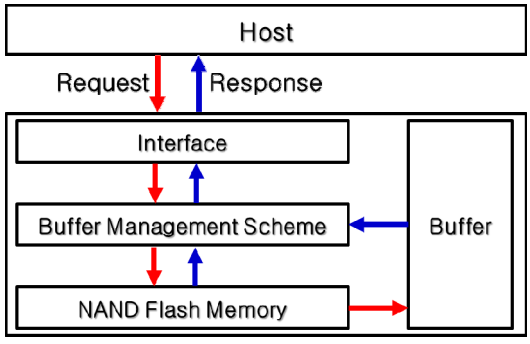
2. 배경

2.1 저장시스템의 접근 및 사용 패턴

Fig. 1은 단일 저장시스템에 복수의 클라이언트가 데이터를 사용하는 일반적인 데이터 접근 및 사용 환경을 보여주고 있다. 그림과 같이 하나의 저장시스템은 직접 연결되거나 네트워크를 통해 연결된 여러 개의 클라이언트의 데이터를 저장 및 관리한다. 또한, 데이터의 접근 및 순서는 정해져 있지 않다. 따라서 각 클라이언트가 고유한 패턴의 데이터에 접근해도 저장시스템의 관점에서는 다수 클라이언트로부터 전달되는 다양한 패턴의 데이터를 처리해야 해서 데이터 접근 패턴은 랜덤한 패턴이 된다. 그림의 예제에서는 8개의 클라이언트가 하나의 저장시스템에 접근하여 SSD의 데이터를 사용했고, 각기 다른 패턴으로 다른 주소 영역의 데이터에 접근했다.



[Fig. 1] Access Pattern on Storage System



[Fig. 2] Buffer Management System

Fig. 2는 SSD에서 버퍼 관리 시스템의 예를 보여주고 있다. 그림과 같이 저장장치는 호스트의 요청으로 데이터를 플래시 메모리에서 읽어서 호스트로 전송한다. 이때 버퍼 관리 정책에 따라 일부 데이터는 고속으로 데이터를 읽을 수 있는 버퍼에 유지한다. 이 방법은 호스트에 빠른 속도로 응답할 수 있는 장점이 있다[14]. 그러나 버퍼의 공간은 제한되어 있어서 일정 크기의 데이터만 유지할 수 있다. 따라서 버퍼 관리 정책의 알고리즘에 따라

버퍼에 적재된 데이터를 처리할 때에만 성능이 향상된다. 그러나, 버퍼 관리 정책은 사용이 예측된 데이터를 버퍼에 유지하기 때문에 예측이 어려운 랜덤 데이터 패턴에는 효율이 떨어지는 문제가 있다.

2.2 데이터 마이닝을 이용한 주소 분석 연구

데이터 마이닝(Data Mining)은 대량의 데이터에서 의미 있는 패턴, 규칙, 그리고 관계를 추출하는 기술로, 다양한 분야에서 데이터 분석 및 의사결정을 지원하는 핵심 도구로 활용되고 있다. 특히, 저장 매체의 데이터 접근 패턴과 지역성(Locality)을 분석하는 데 있어 데이터 마이닝은 중요한 역할을 한다. 데이터의 지역성은 시간적 지역성(temporal locality)과 공간적 지역성(spatial locality)으로 구분되며, 이는 데이터 접근 빈도와 물리적 저장 위치 간의 관계를 통해 파악할 수 있다.

SSD는 기존 HDD와 비교하여 빠른 데이터 접근 속도와 낮은 전력 소비 때문에 널리 사용되고 있지만, 내부적으로는 플래시 메모리의 특성상 데이터가 블록(block)과 페이지(page) 단위로 관리되며, 웨어 레벨링(wear leveling), 가비지 컬렉션(garbage collection)과 같은 고유의 알고리즘이 동작한다. 이러한 특성은 SSD의 데이터 주소(logical block address)와 물리적 주소(physical block address) 간의 매핑 관계를 복잡하게 만들며, 데이터의 지역성과 접근 패턴을 분석하는 데 있어 새로운 분석이 필요하다.

데이터 마이닝 기술을 활용하여 SSD의 데이터 주소를 분석하면, 데이터의 지역성과 접근 패턴을 정량적으로 파악할 수 있다. 이를 통해 데이터 관리 전략을 최적화하고, SSD의 성능을 향상하는 전략을 세우는 데 영향을 줄 수 있다. 또한, 데이터의 지역성을 기반으로 캐시 메모리의 효율성을 높이고, 웨어 레벨링이나 가비지 컬렉션 알고리즘의 성능을 개선하는 데 활용할 수 있다.

2.3 데이터 마이닝을 활용한 데이터 지역성 분석

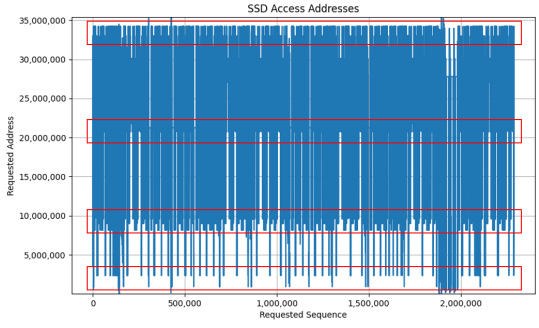
데이터 지역성은 데이터 접근 패턴을 분석하는 데 있어 핵심 개념으로, 시간적 지역성과 공간적 지역성으로 구분된다. 시간적 지역성은 특정 데이터가 짧은 시간 간격으로 반복적으로 접근되는 현상을 의미하며, 공간적 지역성은 특정 데이터 주변의 데이터가 함께 접근되는 현상을 의미한다. 데이터 마이닝 기술은 이러한 지역성을 분석하기 위해 다양한 알고리즘을 활용한다. 예를 들어, 클러스터링(clustering) 알고리즘은 데이터 접근 패턴을 그룹화하여 지역성을 파악하는 데 활용된다. K-means 클러스터링은 데이터 접근 빈도와 위치를 기반으로 유사한 패턴을 보이는 데이터를 그룹화하며, 이를 통해 데이터의 지역성을 정량적으로 분석할 수 있다. 또한, 시퀀스 패턴 분석(sequence pattern analysis)은 데이터 접근 순서를 분석하여 시간적 지역성을 파악하는 데 활용된다. HDD와 같은 전통적인 저장 매체에서 데이터 지역성 분석은 캐시 메모리 및 디스크 스케줄링 알고리즘의 성능을 최적화하는 데 활용되었다. 예를 들어, 데이터 접근 패턴을 분석하여 LRU(least recently used) 캐시 교체 알고리즘의 효율성을 높이거나, 디스크 I/O 성능을 개선하는 연구가 있다.

본 연구는 데이터 마이닝 기술을 활용하여 SSD의 데이터 주소를 분석함으로써, 데이터의 지역성과 접근 패턴을 파악하고, 데이터 관리에 영향을 미치는 특징을 분석한다.

3. 시각화를 통한 주소 분석

3.1 라인 그래프를 이용한 시각화 분석

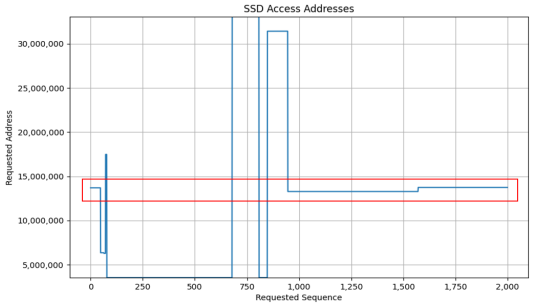
본 논문에서는 데이터마이닝 기술의 한 분야인 시각화를 통해 SSD에서 사용된 데이터의 주소를 분석하였다. 분석을 위해 엔터프라이즈 서버 중 특정 볼륨의 SSD에서 수집된 트레이스로부터 주소 데이터를 수집하였다. 트레이스로부터 추출된 데이터는 사용된 순서대로 512B 단위의 논리적 페이지 주소를 추출하였다. 추출한 페이지 주소 데이터 세트는 총 2,289,161개의 주소이고, 주소의 범위는 27,448번 페이지부터 35,404,287번 페이지이다.



[Fig. 3] Line Graph for Analysis

Fig. 3은 읽기 데이터에 대한 주소 분석을 위해 요청된 순서대로 호스트로부터 요청된 주소를 선으로 연결한

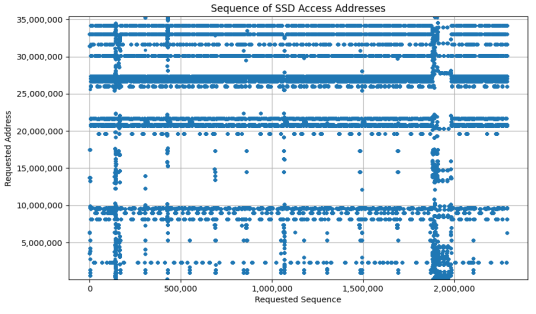
라인 그래프다. 그림에서 x축은 호스트로부터 요청된 데이터의 순서다. 그리고 y축은 요청된 데이터의 주소다. 그림과 같이 읽기 동작이 수행된 약 200만 개의 데이터에 대해, 저장된 주소가 27,448번 페이지부터 35,505,287번 주소까지 랜덤하게 저장되었다. 전체적으로 읽기 요청 데이터가 랜덤하게 처리되었으나, 그림에 표시한 4개의 상자와 같이 집중적이고 빈번하게 읽기 요청이 발생하는 데이터가 있음을 확인하였다. 이러한 데이터를 분석하기 위해서는 더 정밀한 시각화가 필요하다.



[Fig. 4] Analytic Graph between 0 and 2,000

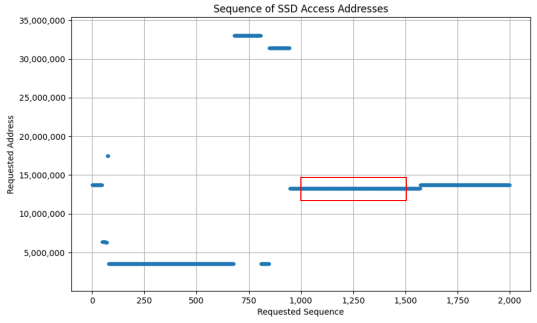
Fig. 4는 분석을 위해 전체 데이터 중 약 0.001%의 일부 데이터를 샘플링 하여 시각화한 그래프다. 샘플링한 데이터는 초기 0에서 2,000번 사이 데이터이다. 그림에서 x축은 요청된 데이터의 순서이고, y축은 요청된 데이터의 주소이다. 그림의 상자로 표시한 영역 내에 직선 그래프 선과 같이 데이터의 약 50% 이상은 특정 주소 영역에 접근하고 있는 것을 확인하였다. 따라서 데이터 사용이 집중적으로 발생한 영역을 분석하는 것은 데이터를 관리에 유의미한 분석이 될 것으로 판단된다.

3.2 도트 그래프를 이용한 시각화 분석



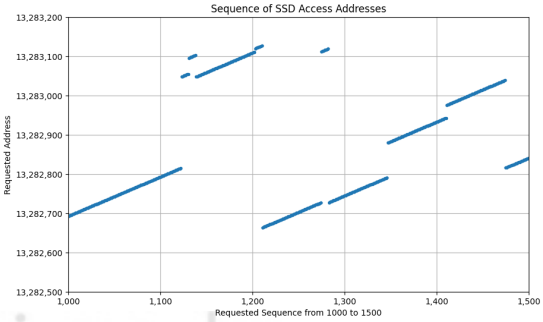
[Fig. 5] Dot Graph for Analysis

Fig. 5는 분석 패턴을 정교하게 파악하기 위해 읽기 데이터를 도트로 시각화한 그래프다. 그래프에서 x축은 데이터의 순서고, y축은 데이터의 주소다. 그림에서 보는 것과 같이 데이터의 표시를 최소화했을 때 특정 주소 영역을 빈번하고 집중적으로 접근하는 것을 더 명확하게 확인할 수 있다.



[Fig. 6] Dot Graph between 0 and 2,000

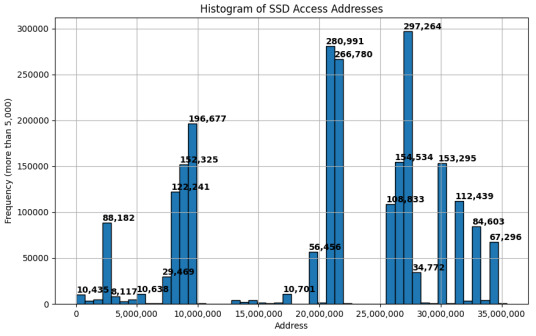
Fig. 6은 정교한 분석을 위해 트레이스 데이터의 초반 2,000개의 주소를 샘플링하여 시각화한 도트 그래프다. 그래프의 x축과 y축은 각각 요청된 데이터의 순서와 요청된 데이터의 주소다. 그림과 같이 도트를 이용하여 시각화한 그래프를 통해 더 정확한 패턴을 확인할 수 있다. 특히 선명한 두 구간의 패턴은 78번 데이터에서 678번 데이터 구간과, 946번 데이터에서 2,000번 데이터 구간이며, 각각 3,545,968블록에서 3,546,207블록 사이 주소에 대한 읽기 요청과, 13,283,112블록에서 13,743,004블록에 사이 주소에 대한 읽기 요청이었다. 이 패턴의 범위는 239개의 블록 주소 범위와 459,892개의 블록 주소 범위에서 발생했고 2,000개의 샘플 데이터 중 82.5%를 차지하고 있고 전체 약 35,376,834개의 주소 범위에서 약 1.3% 범위에 해당한다. 즉 샘플 데이터를 이용하여 통계를 내었을 때 전체 데이터의 약 82%의 데이터가 1.3% 영역의 구간에 집중된 것을 예상할 수 있다.



[Fig. 7] Magnified Graph from 1,000 to 1,500

Fig. 7은 Fig. 6에서 데이터 주소가 집중되는 구간 중 박스 구간인 1,000에서 1,500번째 블록의 주소의 패턴을 확대 시각화한 결과이다. 그림에서 x축은 1,000에서 1,500번 블록까지 읽기 요청이 발생한 순서이고, y축은 각 요청에 해당하는 주소다. 분석을 위해 확대한 그림의 결과와 같이 데이터 패턴은 여러 개의 짧고 순차적인 7개의 패턴이다. 이 분석 구간의 각 패턴에서는 약 100개(50KB) 블록 미만의 데이터가 처리된 것을 확인할 수 있다.

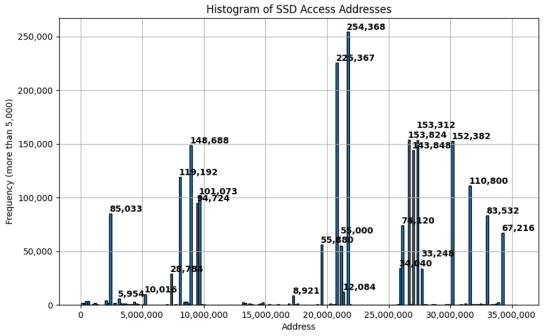
3.3 히스토그램 그래프를 이용한 시각화 분석



[Fig. 8] Histogram for Requested Address

Fig. 8은 트레이스의 주소 구간별로 요청된 데이터 블록의 수를 히스토그램으로 시각화한 결과이다. 그림에서 x축은 읽기 요청된 데이터 구간이고, 트레이스의 주소 데이터 범위를 50개의 구간으로 분할 하였다. 한 구간별 범위는 블록의 페이지 주소 기준으로 약 707,357개의 블록이다. 그리고 y축은 구간별 요청한 읽기 페이지의 누적 횟수이다. 그림에서는 일정 개수 이상 사용된 구간에서 유의미한 분석을 위해 구간별로 5,000회 이상 읽기 요청이 있었던 구간에만 블록의 개수를 표시하였다. 표시된 블록의 총합은 2,246,048개의 블록으로 전체 트레이스의 약 98.12%에 해당한다. 이 구간은 전체 주소의 약 40%에 해당한다. 그런데 여기서 만약 100,000번 이상 읽기 요청이 발생한 구간으로 조건을 변경하면 요청된 블록의 개수는 1,845,379개이고, 약 80.61%의 데이터이다. 이 구간은 전체 범위 중 약 10%에 해당한다.

Fig. 9는 전체 주소 영역을 200개로 분할하고, 정밀하게 데이터 빈도를 시각화한 그래프다. 이 분석 그래프에서는 5,000회 이상 읽기 요청이 발생한 구간은 전체 주소 범위의 약 12%에 해당한다. 또한, 총 2,211,406개의 블록이 처리되었고, 이 개수는 전체 데이터의 약 96.6%에 해당한다. 만약 요청 범위를 50,000회 이상 읽기 요청이 처리된 구간으로 변경할 경우, 범위는 전체 주소 영역의 8.5%에 해당하고 데이터는 총 데이터 대비 약 90.79%에 해당한다. 이 결과는 전체 저장공간의 약 10% 이내에서 전체 데이터 중 90%가 처리되는 것을 의미한다. 따라서 SSD 기반 저장장치를 위한 관리 시스템에서 전체 구간을 관리하는 것보다 패턴에 따라 집중적으로 사용되는 10%이 이내의 구간을 관리하는 것이 SSD의 버퍼 관리 정책, 쓰레기 수집, 자원 관리 등, 데이터를 관리하는 기술 분야에서 자원과 비용을 최적화할 수 있을 것으로 예상한다.



[Fig. 9] Precisely Segmented Histogram

4. 결론

본 논문에서는 데이터 관리 효율성을 높이기 위해 SSD의 접근 패턴을 분석하였다. 패턴 분석을 위해 데이터마이닝의 시각화 기법을 이용하였고, 랜덤한 패턴 안에서 데이터가 집중되는 약 10%의 영역을 확인하였다. 또한, 10% 영역 안에서 처리되는 데이터가 전체 데이터양의 약 90%에 해당하는 것을 확인하였다. 이 결과를 통해 SSD 기반 저장장치를 관리하기 위한 자원 활용을 최적화할 수 있을 것으로 예상한다. 향후에는 추가적인 데이터 마이닝 기법을 적용하여 더 세밀하고 유의미한 패턴의 정보를 찾는 방법과 찾은 패턴을 이용하여 저장장치 관리 기술의 효율과 성능을 올리는 방법을 연구할 예정이다.

REFERENCES

[1] H.S.Lee, "A Design of SSD Dedicated RAID System for Efficient Resource Management," *Journal of Internet of Things and Convergence*, Vol.10, No.2, pp.109-114, 2024.

[2] H.S.Lee, "An Efficient Resource Optimization Method for Provisioning on Flash Memory-Based Storage," *Journal of Internet of Things and Convergence*, Vol.9, No.4, pp.9-14, 2023.

[3] H.S.Lee, "A Memory Mapping Technique to Reduce Data Retrieval Cost in the Storage Consisting of Multi Memories," *Journal of Internet of Things and Convergence,* Vol.9, No.1, pp.19-24, 2023.

[4] H.S.Lee, "A Study on the Performance Measurement and Analysis on the Virtual Memory based FTL Policy through the Changing Map Data Resource," *Journal of Internet of Things and Convergence,* Vol.9, No.1, pp.71-76, 2023.

[5] H.S.Lee, "A Design of Temperature Management System for Preventing High Temperature Failures on Mobility Dedicated Storage," *Journal of Internet of Things and Convergence*, Vol.10, No.2, pp.125-130, 2024.

[6] H.S.Lee, "An Efficient SLC Transition Method for Improving Defect Rate and Longer Lifetime on Flash Memory," *Journal of Internet of Things and Convergence*, Vol.9, No.3, pp.81-86, 2023.

[7] H.S.Lee, "A Study on Characteristics and Techniques that Affect Data Integrity for Digital Forensic on Flash Memory-Based Storage Devices," *Journal of Internet of Things and Convergence*, Vol.9, No.3, pp.7-12, 2023.

[8] H.S.Lee, "Performance analysis and prediction through various over-provision on NAND flash memory based storage," *Journal of Digital Convergence*, Vol.20, No.3, pp.343-348, 2022.

[9] H.S.Lee, "High Efficiency Life Prediction and Exception Processing Method of NAND Flash Memory-based Storage using Gradient Descent Method," *Journal of Internet of Things and Convergence*, Vol.11, No.11, pp.44-50, 2021.

[10] Y.Bai, M.Zhao, R.Li and P.Xin, "A new data mining method for time series in visual analysis of regional economy," *Information Processing & Management*, Vol.59, No.1, 2022.

[11] M.Ali, M.W.Jones, X.Xie and M.Williams, "Visual Analytics Approach for Temporal Pattern Discovery in Large-Scale Data Mining," *The Visual Computer*, Vol.35, pp.1013-1026, 2019.

[12] M.Ali, Ali.Alqahtani, M.W.Jones and X.Xie, "Clustering and Classification for Time Series Data in Visual Analytics: A Survey," *IEEE Access*, Vol.7, pp.181314-181338, 2019.

[13] D.A.Keim, "Information Visualization and Visual Data Mining,", *IEEE Transactions on Visualization and Computer Graphic*s, Vol.8 pp.100-107, 2002.

[14] H.S.Lee, "A Prediction-Based Data Read Ahead Policy using Decision Tree for improving the performance of NAND flash memory based storage devices," *Journal of Internet of Things and Convergence*, Vol.8, No.4, pp.9-15, 2022.

[15] SNIA, http://iotta.snia.org/traces/block-io/388.

이 현 섭(Hyun-Seob Lee) [종신회원]

- 2013년 2월 : 한양대학교 컴퓨터공학과 (공학 박사)
- 2012년 3월 ~ 2021년 2월 : 삼성전자 책임연구원
- 2021년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 조교수

〈관심분야〉
인공지능, 저장시스템, 임베디드 시스템