해시태그 분석 알고리즘의 자료구조별 성능 분석

황영석
성균관대학교 전자전기공학부
hhyun3032@g.skku.edu

***Abstract*—본 논문에서는 해시태그 분석 알고리즘의 자료구조별 성능 분석에 대해 연구한다. 연결 리스트, 이진 탐색트리와 같은 기존의 단일 자료구조뿐만 아니라 여러 단일 자료구조를 결합한 새로운 자료구조를 통해 성능을 확인하고 분석한다. 또한 해당 분석 결과는 다양한 자료 구조의 적절한 결합을 통한 사용이 성능의 향상에 중요한 역할을 할 수 있음을 시사한다.**

I. INTRODUCTION

현대 사회에서 전 세계 인구의 60% 이상이 인스타그램, 페이스북 및 다양한 소셜 네트워크 서비스(Social Network Service, SNS)를 사용하고 있다 [1]. 이들은 해시태그(Hashtags)를 사용하여 글이나 영상에 주제를 포함시킴으로써 다른 사용자들과 관련 주제를 공유하고 공감을 표현한다. 전 세계적으로 널리 사용되는 이 해시태그들은 특정 주제나 사건에 관심이 있는 사람들을 연결하기 때문에, 이를 분석하는 것은 현재 사람들 사이의 트렌드와 관심사를 파악하고 사회 현상을 이해하는 데 핵심적인 역할을 할 수 있다.

하지만, 매일 1 억 2 천만개 이상의 해시태그가 생성되는데 [2], 매일 이렇게 생성되는 방대한 양의 해시태그 데이터를 분석하는 것은 쉬운 일이 아니다. 따라서 이를 효율적으로 관리하고 분석하기 위한 적절한 자료구조의 선택은 매우 중요하다. 이 연구는 이와 같은 방대한 해시태그 데이터를 효율적으로 처리할 수 있는 자료구조를 탐색하고 분석하는 것을 목표로 한다. 해시태그 정보 분석을 통해 현대 사회의 트렌드 파악과 사회 동향 이해가 가능하기 때문에, 트렌드 분석이 필요한 기업이나 각종 기관들에게 큰 도움이 될 것이다. 따라서 이 연구를 통해 이에 기여하고자 한다.

II. DESIGN AND IMPLEMENTATION

A. System Design

1) 데이터 구성

해시태그의 정보는 “#socialmedia, United States, 2023.12.02”와 같이 해시태그명, 국가명, 날짜로 구성되어있다. 해당 정보를 가공하여 어떤 날짜별로 어떤 해시태그가 많이 사용되었는지, 또는 어떤 국가에서 어떤 해시태그가 많이 사용되었는지와 같은 복합적인 정보를 얻을 수 있다.

2) 자료 구조

1. 3-Dimensional Array

기본적인 3 차원 배열이며, 첫 번째 차원은 국가명, 두 번째 차원은 해시태그명, 세 번째 차원은 날짜를 나타낸다. 해시태그 데이터의 특성상 3 차원 배열로 데이터를 효과적으로 저장하고 나타낼 수 있다.

2. Linked list

연결 리스트이며, 각 노드는 해시태그 정보의 구조체이다. 또한 각 노드는 다음 노드를 가리키는 포인터를 포함한다. 새로운 데이터가 추가될 때 리스트를 순회하며 기존에 존재하는 데이터와 일치하는지 확인하고, 이미 존재하는 경우에는 빈도수를 업데이트한다.

3. 이진 탐색 트리(BST)

Linked list 와 마찬가지로 트리의 각 요소는 해시태그 정보의 구조체이다. 새로운 데이터가 추가될 때, 트리를 순회하며 적절한 위치를 찾고, 이미 존재하는 데이터는 빈도수를 업데이트한다.

4. 2D Array with linked list

2 차원 배열과 linked list 의 특성을 결합한 새로운 자료구조이다. 2 차원 배열의 행은 해시태그명을, 열은 국가명을 나타낸다. 배열의 각 요소는 날짜 데이터를 저장하는 연결 리스트의 헤드 포인터를 가지고 있다. 새로운 날짜 데이터가 들어오면 노드를 추가하고, 이미 존재하는 경우에는 빈도수를 업데이트한다.

5. Multiple Queue with linked list

다중 큐와 linked list 의 특성을 결합한 새로운 자료구조이다. 큐 포인터를 담은 배열은 큐를 가리키는데, 각 큐에는 같은 해시태그명을 가진 요소들이 있다. 또한, 큐의 요소는 국가를 기준으로 추가 / 삭제 되며 4 번의 자료구조와 비슷하게 각 큐의 요소는 날짜 데이터를 저장하는 연결 리스트의 헤드 포인터를 가지고 있다.

B. Implementation in Detail

웹크롤링을 통해 SNS 상의 실제 해시태그 정보를 추출하는 방식으로 구현하는 것이 좋지만, 해당 작업은 시간이 아주 오래 걸릴 것으로 판단되어 텍스트 파일에 해당 해시태그 정보를 입력해두고 텍스트 파일 안에 있는 해시태그 정보를 분석하는 방식으로 구현하였다. 텍스트 파일 데이터셋은 해시태그의 정보에 맞게 직접 작성하였다. 총 150 종류의 해시태그 조합이 존재하도록 데이터셋을 구성하였다. 분석 알고리즘 프로그램은 각 자료구조에 맞는 구조체를 선언하고, 해시태그의 빈도수를 측정하여 해시태그명별, 국가별 해시태그의 빈도수에 대한 분석 결과를 출력하도록 하였다.

III. Performance Evaluation

A. Experiment Setup

실행 시간을 측정하기 위해 리눅스의 gpof 컴파일러를 활용하여 시간을 측정하였다. 측정 환경 및 측정 기준은 아래와 같다.

OS : AMD Ryzen 5 5625U with Radeon Graphics
CPU 클럭 : 2.3 GHz
사용 txt 파일 용량 : 340,065 KB
시행 횟수 : 3 회
결과값 기준 : 3 회 실행 평균 속도

B. Analysis

실험 결과는 아래의 Fig 1 에 나타난다. 그래프의 가로축은 각각 순서대로 3 차원배열, Linked list, 이진 탐색 트리, 2D Array with linked list, Multiple Queue with linked list 에 해당한다. 3 차원 배열은 8.32 초로 가장 긴 평균 실행 시간을 보여주었고, Multiple Queue with linked list 구조는 0.27 초로 가장 짧은 시간을 기록했다. 이진 탐색 트리(BST)와 2D Array with linked list 도 Multiple Queue with linked list 와 유사한 짧은 실행 시간을 나타냈다.

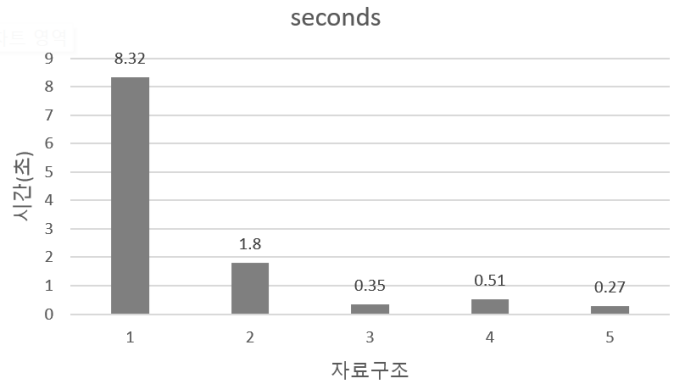

Fig. 1. 자료구조별 실행 시간

이러한 결과는 복잡한 데이터 세트를 효율적으로 처리하기 위해서는 단일 자료구조보다 여러 자료구조의 결합이 유리할 수 있음을 시사한다. 3 차원 배열은 데이터의 저장 및 접근에 있어서 직관성과 구조적 단순성에서 장점을 가질 수 있지만, 데이터의 크기가 크고 방대할 경우에는 검색 및 갱신 과정에서 비효율적일 수 있다. 반면, 다중 큐, 2 차원 배열과 연결 리스트의 결합은 새로운 데이터의 삽입이나 데이터 검색 과정에서 더 빠른 성능을 보여줄 수 있다. 특히, Multiple Queue with linked list 구조는 해시태그에 따라 데이터를 분류하고 각 큐에서 국가 및 날짜 별 데이터를 효율적으로 관리할 수 있어, 대규모 데이터에 대한 빠른 처리가 가능하다. 추가적으로, 이진 탐색 트리도 많은 데이터를 처리하는데 좋은 성능을 보인다는 것을 실제로 확인할 수 있었다.

IV. Conclusions

이 연구는 복잡하고 방대한 데이터를 처리하는 데 있어 여러 자료구조의 결합이 단일 자료구조 사용보다 효과적일 수 있음을 시사한다. 특히, 다중 큐와 linked list 의 결합은 해시태그 데이터와 같은 대규모 정보를 효율적으로 처리하는 데 매우 유용함을 보여주었다. 또한, 이진 탐색 트리와 같은 특정 자료 구조들도 뛰어난 효율성을 제공함을 확인했다. 따라서, 데이터의 특성을 고려하여 적절한 자료구조를 선택하고 필요에 따라 여러개의 자료구조를 현명하게 결합한다면, 해시태그 분석뿐만 아니라 대량의 데이터를 처리해야하는 여러 분야에서도 효율적으로 데이터를 관리하고, 만족스러운 결과를 도출할 수 있을 것으로 기대된다.

Reference

[1] Mehdi Houshmand Sarkhoosh et al. "Soccer on Social Media," *Oslo Metropolitan University*. Norway, 10.48550/arXiv.2310.12328.
[2] Matt Stevens, “As the Hashtag Celebrates Its 10th Birthday, Are We #Blessed?,” The New York Times, Aug. 23, 2017, https://www.nytimes.com/2017/08/23/business/hashtag-anniversary-twitter.html.