미니 PC 기반의 하둡 클러스터를 이용한 트렌드 분석 서비스

전영호, 김은상, 박효주, 이기훈[1)]
광운대학교 컴퓨터공학과
e-mail: kihoonlee@kw.ac.kr

A Trend Analysis Service Using a Hadoop Cluster of Mini PCs

Young-Ho Jeon, Eun-Sang Kim, Hyo-Ju Park, and Ki-Hoon Lee
Dept. of Computer Engineering, Kwangwoon University

요 약

IT 산업의 발전에 따라 생성되는 데이터의 양이 폭발적으로 증가하고 있다. 이러한 빅 데이터는 여러 대의 컴퓨터로 구성한 하둡 클러스터를 이용하면 상당히 빠른 속도로 처리할 수 있으나, 일반적으로 하둡 클러스터를 구성하기 위해 많은 비용과 공간이 소요되는 단점이 있다. 본 논문에서는 저가의 미니 PC로 하둡 클러스터를 구성하여 비용 및 공간적 문제점을 해결하고, 구축한 하둡 클러스터를 이용한 트렌드 분석 서비스를 제안하였다. 실험 결과 미니 PC로 이루어진 하둡 클러스터가 고가의 서버보다 트렌드 분석에 더 좋은 처리 성능을 보였다.

1. 서론

IT산업의 발전에 따라 생성되는 데이터의 양이 폭발적으로 증가하고 있다. 이러한 빅데이터는 여러 대의 컴퓨터로 구성한 하둡 클러스터를 이용하면 상당히 빠른 속도로 처리할 수 있다. 하지만 일반적으로 하둡 클러스터를 구성하기 위해 큰 공간이 필요하고 많은 비용이 든다는 단점이 있다. 본 논문에서는 저가의 미니 PC를 이용하여 하둡 클러스터를 구성함으로써 비용 및 공간적 문제점을 해결하고, 구축한 하둡 클러스터를 이용한 트렌드 분석 서비스를 제안한다.

트렌드 분석 서비스를 위해 먼저 인터넷 뉴스 사이트에서 뉴스 기사를 크롤링하여 수집한다. 다음으로 형태소 분석기를 이용하여 수집한 뉴스에서 단어만을 추출한 후에 하둡 클러스터를 이용해 단어의 빈도수를 계산한다. 계산한 단어 빈도수를 데이터베이스에 저장하여 사용자가 트렌드를 검색하고 확인할 수 있도록 한다.

실험을 통해 저가의 미니 PC로 구성한 하둡 클러스터가 고가의 서버에 비해 단어 빈도수 계산 성능에 있어서 우수함을 보였다.

2. 관련 연구

SNS와 MapReduce를 이용한 SNS 트렌드 예측 시스템에 대한 연구[1]가 진행되었다. 하둡을 이용하여 트렌드를 분석한다는 점은 유사하나, 본 논문에서는 SNS보다 정확한 정보를 가지는 뉴스 데이터를 사용하는 점과 미니 PC로 하둡 클러스터를 구축하여 비용 및 공간적 문제점을 해결한다는 점에서 차이가 있다.

3. 제안하는 트렌드 분석 서비스

제안하는 트렌드 분석 서비스에서는 그림 1과 같이 검색하고자 하는 년, 월, 카테고리를 설정한 후에 확인을 클릭하면 그림 2와 같은 표 형식으로 해당 년, 월, 카테고리의 키워드들이 빈도수의 내림차순으로 정렬되어 출력된다.

트렌드 검색				
2013 ▼	년	1 ▼	월	확인
카테고리		정치 ▼		
Top		100 ▼		

(그림 1) 기간 및 카테고리 별 검색 서비스

2013년 1월 정치 분야 트렌드		
키워드	빈도	순위
인수	6305	1
당선	4567	2
박근혜	3331	3
민주	3110	4
이동흡	1996	5
김용준	1984	6

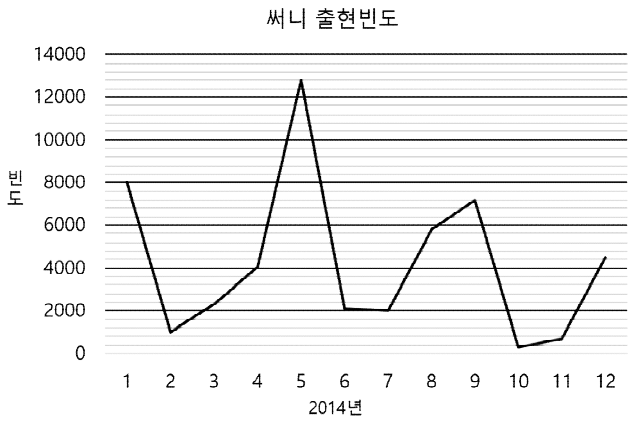
(그림 2) 기간 및 카테고리 별 검색 결과

1) 교신저자

기간별, 카테고리별 검색뿐만 아니라 키워드를 이용한 검색도 가능하다. 그림 3과 같이 원하는 키워드를 입력하고 확인 버튼을 누르면 그림 4와 같이 기간별로 키워드 빈도수가 어떻게 변화하는지를 차트로 시각화하여 보여준다.

키워드 검색	
써니	확인

(그림 3) 키워드 검색 서비스



(그림 4) 키워드 검색 결과

4. 실험 및 평가

4.1. 실험 방법

Mini-ITX 규격(17cm * 17cm)의 PC로 구성된 하둡 클러스터와 서버의 처리 성능과 전력 소모량을 비교한다. 미니 PC와 서버의 사양은 표 1과 같다. 하둡 클러스터는 미니 PC 4대로 구성되는데 1대는 Name Node로 사용하고 나머지 3대는 Data Node로 사용한다.

형태소 분석이 끝난 8.5GB 크기의 뉴스기사 텍스트 파일에 대해 하둡으로 word count 프로그램[2]을 실행하여 처리 시간과 전력 소모량을 측정한다. 하둡의 각종 설정 값은 미니 PC와 서버 각각에 대해 최적값을 찾아서 표 2와 같이 설정한다.

<표 1> 미니 PC 및 서버 사양

구분	미니 PC	서버
사양	Intel Core i5-4690, DDR3 4GB RAM, Samsung 840 PRO SSD 256GB	Intel Xeon Processor E5-2620, DDR3 24GB RAM, Samsung 840 PRO SSD 256GB
가격	280만원 (70만원 * 4대)	420만원

<표 2> 실험에서 사용한 하둡 설정 값

구분	미니 PC	서버
Data Node 당 최대 병렬 수행 Map 태스크 수	4	11
Data Node 당 최대 병렬 수행 Reduce 태스크 수	3	11
블록 복제수	2	2
신규 생성 파일의 기본 블록 크기	128MB	128MB

4.2. 실험 결과

처리 성능의 실험 결과를 보면 표 3과 같이 미니 PC로 구축한 하둡 클러스터가 서버보다 2.21배 더 빠르다. 표 4의 전력 소모량을 보면 하둡 클러스터가 서버와 거의 비슷한 전력을 소모한다.

<표 3> word count 프로그램 처리 시간

구분	미니 PC	서버
처리 시간	4m 15s	9m 23s

<표 4> word count 프로그램 수행 시 전력 소모량

구분	미니 PC	서버
전력 소모량	236.3W * 255초 = 60,257Ws	113.5W * 563초 = 63,901Ws

5. 결론

본 논문에서는 미니 PC를 이용하여 하둡 클러스터의 비용적·공간적 문제점을 해결할 수 있는 방안을 제시하였으며, 뉴스기사를 이용한 트렌트 분석 서비스를 제안하였다. 미니 PC 4대로 구축한 하둡 클러스터와 서버 1대 간의 비교 실험을 통해 미니 PC 하둡 클러스터가 서버보다 빅데이터 처리 성능이 좋다는 것을 확인할 수 있었다. 또한 전력 소모량은 비슷하다는 것을 확인하였다.

감사의 글

본 연구는 중소기업청에서 지원하는 2014년도 산학연협력 기술개발사업(No. C0187264)의 연구수행으로 인한 결과물임을 밝힙니다.

참고문헌

[1] 이현진, 박석천, 김종현, “빅데이터 기반 맵리듀스를 이용한 트렌드 예측 시스템 설계”, *한국인터넷정보학회 춘계학술대회 논문집*, Vol. 15, No. 1, pp. 159-160, 2014년 10월.

[2] 한기용, 직접 해보는 하둡 프로그래밍, 이지스퍼블리싱, 2013년.