하둡을 이용한 개인화 영화 추천 시스템

김세준*, 박두순**, 홍민**
*순천향대학교 컴퓨터학과, **순천향대학교 컴퓨터소프트웨어공학과
e-mail : kalkal47@hotmail.com

A Personalized Movie Recommender Systems using Hadoop

Se-jun Kim*, Doo-soon Park**, Hong Min**
*Dept of Computer Science, Sooncheonhyang Univ.
**Dept of Computer Software Engineering, Sooncheonhyang Univ.

요 약

인터넷의 발달함에 따라 데이터가 기존에 비해 기하급수적으로 늘어나게 되는 이른바 빅데이터 시대를 맞이하게 되었다. 이러한 빅데이터는 기존의 시스템으로 처리하기가 쉽지 않아 이를 처리하기 위해 하둡이 개발되었다. 하둡은 분산파일 시스템으로 기존의 시스템에 비해 빅데이터를 처리하는데 적합하며 이를 이용한 다양한 오픈 소스들이 등장하게 된다. 그중 기계학습 알고리즘을 구현한 오픈소스 Mahout은 추천 시스템을 구현하는데 적합하다. 이를 이용하여 기존에 구현한 개인화 영화 추천 시스템을 하둡 시스템으로 구현하고 기존의 XLMiner로 구현한 시스템과 결과를 비교해 본다.

1. 서론

기술발달로 시스템의 성능이 향상되고 인터넷의 발달로 전 세계 정보량이 기하급수적으로 증가하고 있다. 이는 시스템의 성능이 향상되어 데이터를 처리, 저장 하는 능력이 크게 증가하게 되었기 때문이다. 기존에는 발생한 정보를 시스템의 처리, 저장 등의 성능 문제로 많은 데이터를 축적할 수 없었으며 그 중에서도 일부 데이터만을 활용해 왔다. 그러나 시스템 성능이 발달하고 이와 더불어 인터넷의 보급은 정보량 자체의 증가에 크게 기여하였으며 특히 스마트폰을 필두로 모바일 스마트 기기의 확산과 소셜미디어의 증가는 공적인 정보뿐만 아니라 사적인 정보까지 교류함으로써 빅데이터의 서막을 알리는 계기가 되었다[1]. 전 세계적인 정보량의 증가는 2011년 약 1.8ZB 이며 2020년에는 관리해야할 정보의 양이 50배 이상 증가 할 것이라 예측하고 있다[2]. 이런 빅데이터 시대에 축적된 정보를 처리하기에는 기존의 시스템으로는 힘든 면이 있었고 이를 기존의 시스템보다 효율적인 처리를 위한 하둡 시스템이 새롭게 부각되고 있다.

하둡은 구글의 GFS(Google File System), 맵리듀스(MapReduce)를 바탕으로 만들어진 것으로 HDFS(Hadoop Distributed File System)과 맵리듀스를 구현한 것이다. HDFS와 맵리듀스는 분산 파일 관리와 분산 배치 처리를 각각 담당하고 있다. 하둡은 빅데이터를 효과적으로 저장, 처리가 가능하도록 하였으며 대표적으로 트위터, 야후, 에버노트 등의 회사들이 하둡을 이용하여 빅데이터 처리를 위해 사용하고 있다. 하둡은 오픈소스를 기반으로 한 다양한 서브 프로젝트 등이 있고 그 중 아파치 Mahout은 오픈소스 기계학습 라이브러리로 전통적인 기계학습 알고리즘을 하둡에서 동작할 수 있도록 알고리즘을 개선하여 제공하고 있다. 기계학습 중에서 추천엔진과 군집 분류를 주로 처리하여 추천 시스템을 구성하는데 적합하다. 이를 바탕으로 기존에 개발한 영화 추천 시스템을 아파치 Mahout을 이용하여 하둡 시스템상에서 구현해보고 동작해보고 기존의 XLMiner로 구현한 시스템과 비교해 본다.

2. 아파치 Mahout을 이용한 추천 시스템

본 연구는 기존 연구에서 사용자들이 평가한 평가치가 충분할 경우 기존의 알고리즘인 협업 필터링을 이용하고, 평가 데이터가 충분치 않은 희박성 문제가 발생한다면 명시적으로 개인화 요인을 입력받아 K-means Clustering 기법으로 군집군을 형성하였다. 입력받은 개인화 요인 6개를 이용한 63가지 방법을 통하여 최적의 개인화 성향으로 나이, 선호장르, 성격을 제안하였으며 데이터는 MovieLens의 데이터를 이용하였다[3]. K-means Clustering을 사용하여 군집화 하였으며 거리 측정에는 유클리드 거리 측정법을 사용하였다.

(그림 1)은 벡터 변화를 용이하기 위해 구분자(,)를 사용하여 입력 데이터를 변화 시킨 것의 일부이다.

```
750,28,16,8
751,24,5,16
752,60,8,6
753,56,8,14
754,59,8,16
755,44,8,4
756,30,1,10
757,26,1,6
758,27,8,13
759,20,1,3
760,35,14,14
761,17,8,13
762,32,14,6
```

(그림 1) 입력 데이터(일부)

유저 번호, 나이, 선호장르, 성격 순으로 구성되었으며 Mahout이 입력으로 받는 벡터 처리를 용이하기 위해 구분자(,)를 사용하여 입력 데이터를 변경 하였다. Mahout에서 벡터를 처리하기 위한 3개의 클래스 DenseVector,RandomAccessSparseVector,SequentialAccessSparseVector 가 있으며 이중 입력 데이터를 보면 double 타입의 배열로 배열의 크기가 데이터의 특성의 수인 DenseVector를 사용하였다. Mahout에는 군집 알고리즘인 K-means Clustering을 KmeansClusterer 또는 KmeansDriver 클래스를 제공하며 KmeansClusterer는 인메모리형식으로 군집화를 실행하며, KmeansDriver는 맵리듀스 작업으로 실행한다. KmeansDriver로 맵 리듀스 작업으로 실행하였으며 거리 측정 방법에는 기존의 시스템에서 사용한 유클리드 거리 측정 방법을 사용하였다. (그림 2)는 Mahout을 이용한 군집화 결과의 일부이다.

```
21:0.3786941542319488: [26.000, 16.000, 5.000]cluster : 3
22:0.30661437540008984: [25.000, 5.000, 7.000]cluster : 2
23:0.37157082809447517: [30.000, 8.000, 3.000]cluster : 4
24:0.28303590793452527: [21.000, 8.000, 10.000]cluster : 2
25:0.28520250838214634: [39.000, 5.000, 5.000]cluster : 5
26:0.33589775456911675: [49.000, 8.000, 9.000]cluster : 5
27:0.25625933781922083: [40.000, 16.000, 7.000]cluster : 5
28:0.24300735955798403: [32.000, 16.000, 12.000]cluster : 3
29:0.2585881566070373: [41.000, 16.000, 5.000]cluster : 5
30:0.2646761624402919: [7.000, 5.000, 9.000]cluster : 2
31:0.2731025549222964: [24.000, 8.000, 8.000]cluster : 2
32:0.3099915782154941: [28.000, 1.000, 13.000]cluster : 0
```

(그림 2) 군집화 결과(일부)

맵리듀스의 (key, value)쌍에서 cluster가 key가 되고 앞에 출력된 값들이 value이며 클러스터는 0부터 시작하게 된다.

Mahout을 이용하여 기존의 논문[3]에서 제안한 최적의 개인화 요인들을 고려한 추천 목록과 기존의 시스템에서 추천된 추천 목록은 <표1>, <표2>와 같다.

영화 추천 결과 10개 중 5개인 "Wrong Trousers, The (1993)", "Close Shave, A (1995)", Casablanca (1942), Wallace &Gromit: The Best of Aardman Animation (1996),One Flew Over the Cuckoo's Nest (1975)가 일치하는 것을 확인하였다. 10개의 50%인 5개 일치함을 확인하였다. 이는 군집군의 개수를 정하는 k값의 차이 인 것으로 추측된다.

<표 1> Mahout 영화 추천 결과

"Saint of Fort Washington, The (1993)"
Someone Else's America (1995)
"Wrong Trousers, The (1993)"
"Close Shave, A (1995)"
Casablanca (1942)
Maya Lin: A Strong Clear Vision (1994)
Wallace &Gromit: The Best of Aardman Animation (1996)
Rear Window (1954)
12 Angry Men (1957)
One Flew Over the Cuckoo's Nest (1975)

<표 2> 기존 시스템 추천 결과

'Casablanca (1942)'
'"Close Shave, A (1995)"'
'"Wrong Trousers, The (1993)"'
'Rear Window (1954)'
'Wallace & Gromit: The Best of Aardman Animation (1996)'
'"Third Man, The (1949)"'
'"Manchurian Candidate, The (1962)"'
'Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1963)'
'Citizen Kane (1941)'
'One Flew Over the Cuckoo's Nest (1975)'

4. 결론

본 논문에서는 기존에 구현하였던 영화 추천 시스템을 하둡에서 동작하는 기계학습 알고리즘을 제공하는 Mahout 오픈소스를 활용, K-means Clustring 군집 알고리즘을 사용하였으며 그 결과 기존의 시스템에서 추천한 영화와 5개가 일치하며 이는 50%만 일치하고 나머지 50%는 일치 하지 않는 것을 확인하였다. 추후 이러한 결과의 차이가 발생하게 된 이유를 찾아 개선한다면 더 좋은 시스템이 될 것이다.

참고문헌

[1] '빅데이터(Big Data)' 활용에 대한 기대와 우려 p.29

[2] IDE & EMC, 'Digital Universe Study 2011'

[3] Woon-hae Jeong, Se-jun Kim, Doo-soon Park and Jin Kwak, "Performance Improvement of a Movie Recommendation System based on Personal Propensity and Secure Collaborative Filtering", Journal of Information Processing Systems Volume 9, Number 1, 2013