ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 20, No. 12, pp. 615-622, 2014. 12
http://dx.doi.org/10.5626/KTCP.2014.20.12.615

영화 흥행 실적 예측을 위한 빅데이터 전처리
(Big Data Preprocessing for Predicting Box Office Success)

전 희 국 † 현 근 수 † 임 경 빈 † 이 우 현 † 김 형 주 ††
(Hee-Gook Jun) (Geun-Soo Hyun) (Kyung-Bin Lim) (Woo-Hyun Lee) (Hyoung-Joo Kim)

요 약 국제적 수준으로 성장한 한국의 영화 시장 환경은 더욱 타당한 자료 분석에 근거한 의사 결정 수단을 필요로 하게 되었다. 또한 발전된 정보 환경으로 인해 실시간으로 생성되는 대규모 데이터를 신속히 처리하고 분석하여 보다 정밀한 결과를 예측할 수 있어야 한다. 특히 전처리 작업은 정보 분석 과정 중 가장 많은 시간이 소요 되므로 대규모 데이터 기반 분석 환경에서도 합리적인 시간 내에 처리할 수 있어야 한다. 본 논문에서는 영화 흥행 예측을 위한 대용량 데이터 전처리 방법을 연구하였다. 영화 흥행 데이터의 특성을 분석해 전처리의 각 유형별 처리 방법을 설정했으며 하둡 기반 맵리듀스 프레임워크를 사용하는 방법을 사용하였다. 실험 결과 빅데이터 기법을 사용한 전처리가 기존의 방법보다 더 좋은 수행 결과를 보이는 것을 확인하였다.

키워드: 빅데이터, 하둡, 맵리듀스, 전처리, 영화

Abstract The Korean film market has rapidly achieved an international scale, and this has led to a need for decision-making based on analytical methods that are more precise and appropriate. In this modern era, a highly advanced information environment can provide an overwhelming amount of data that is generated in real time, and this data must be properly handled and analyzed in order to extract useful information. In particular, the preprocessing of large data, which is the most time-consuming step, should be done in a reasonable amount of time. In this paper, we investigated a big data preprocessing method for predicting movie box office success. We analyzed the movie data characteristics for specialized preprocessing methods, and used the Hadoop MapReduce framework. The experimental results showed that the preprocessing methods using big data techniques are more effective than existing methods.

Keywords: big data, hadoop, MapReduce, preprocessing, movie

· 이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0030810)

† 비 회 원 : 서울대학교 컴퓨터공학부(Seoul National Univ.)
hgjun@idb.snu.ac.kr
(Corresponding author임)
gshyun@idb.snu.ac.kr
kblim@idb.snu.ac.kr
whlee@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2014년 7월 29일
(Received 29 July 2014)
논문수정 : 2014년 10월 17일
(Revised 17 October 2014)
심사완료 : 2014년 10월 20일
(Accepted 20 October 2014)

정보과학회 컴퓨팅의 실제 논문지 제20권 제12호(2014. 12)

1. 서 론

한국의 영화 시장 규모는 2013년 기준 140만 달러로 세계순위 6위에 달하는 규모이다[1]. 또한 세계순위 7위에 달하는 영화 제작량을 보이고 있으며[2], 2013년 1조 8,839억원의 매출을 달성했다. 또한 온라인 영화 시장도 지속적으로 증가하여 인터넷 VOD 분야는 729억원의 매출[3]을 기록했다.

이처럼 대규모로 발전한 영화 시장은 지속적인 문화, 경제적 성장을 위해 더욱 타당한 자료 분석에 근거한 의사결정 수단이 필요하다. 이를 위해 영화의 흥행 요인을 파악[4]하고, 통계적 방법[5] 혹은 영화의 유형 판단[6]을 사용한 흥행 예측을 연구하는 등 다양한 연구가 진행되어 왔다.

또한 스마트폰과 같은 현대의 다양한 개인화 정보 기기를 사용한 폭발적 웹 기반 정보의 생성과 공유 활동은 영화 산업 분석에 있어 보다 대중사회의 기호를 반영할 수 있으므로 보다 정밀한 결과를 예측하는 데 도움을 줄 수 있다.

그러나 영화 산업의 성장으로 인한 관련 정보의 증가 및 웹 기반 정보화 시대의 대규모 데이터 생성은 영화 흥행 분석에 있어서 빅데이터적인 문제를 발생시키고 있다. 예를 들어 데이터 규모가 커질수록 다양한 매체로부터 추출한 각종 데이터 양식 통합, 불순 데이터 제거, 자연어 처리, 분석에 영향을 미치는 데이터 선택 등 정보 분석 과정의 전반적인 계산 복잡도가 증가하게 된다.

특히 정보 분석 과정에 있어 실질적인 작업량만을 보았을 때 전체 과정 중 90% 정도의 처리 시간을 요하는 데이터 정제 및 전처리 과정은 대용량 데이터 환경에서는 더욱 긴 처리 시간을 필요로 한다. 따라서 보다 합리적이고 효과적인 영화 흥행 분석을 위해서 효과적인 대용량 데이터 전처리 기술 연구가 필요하다.

본 논문에서는 영화 흥행 분석 과정 중 발생 가능한 대용량 데이터 처리 방법 연구를 진행하였다. 대규모 비정형 데이터를 대상으로 각 데이터 특성 별 발생하는 현상을 정리하고 그에 맞는 빅데이터 전처리 방법을 개발하였다. 또한 빅데이터 전처리 과정을 거친 대규모 영화 데이터는 관계형 데이터베이스와 같은 정형 데이터로 변경되며, 이를 통해 효과적이고 다양한 영화 흥행 분석을 수행할 수 있도록 하였다.

논문의 구성은 다음과 같다. 2절에서 영화 흥행 분석과 빅데이터 전처리에 대한 기존 방법들을 살펴보고 3절에서는 대용량 영화 데이터에 적용한 전처리 방법에 대해 설명한다. 4절에서는 전처리 과정에서 사용된 빅데이터 처리 기술에 대해 설명한다. 5절에서 실험 결과를 제시하고 6절에서 결론 및 향후 연구에 대해 언급한다.

2. 관련연구

2.1 영화 흥행 요소와 예측

영화 흥행 예측은 통계적 데이터 마이닝 방법(다중회귀, 의사결정트리, 인공신경망 등)으로 분석 모델을 개발해 수행이 가능하다[7]. 또한 영화의 모든 흥행 요소를 사용하는 것보다 일부 유의미한 흥행 요소만을 선택해 사용[5]하는 것이 더 높은 흥행 예측 성능을 보일 수 있다.

영화의 흥행 요소는 창조의 영역, 배급의 영역, 홍보의 영역, 경쟁의 영역으로 분류[4]가 가능하다. 표 1은 각 영역에 해당하는 흥행 요소를 나타내고 있다. 흥행 요소 중 총 제작비가 관객동원에 가장 큰 영향을 미치며, 배우와 배급사도 흥행에 큰 영향을 미치는 것으로 확인 되었다. 반면 스크린 수 처럼 나라별 영화 시장마다 영향을 미치는 정도가 다른 특성을 보이는 흥행 요소도 존재하는 것을 확인 하였다. 또한 배우 명성, 감독 명성 등과 같은 주관적 평가 요소는 수상 및 후보로 지명된 회수의 총합으로 수치화하여 객관적인 요소로 변환[6]해 판단 방법으로 사용할 수도 있다.

표 1 영화의 흥행 요소 분류[4]
Table 1 Classification of factors in Box Office success

Sphere	Factors for the box office success
Creativeness	Actor, director, production cost,..
Distribution	Number of screens, …
Marketing	Promotion, rating, …
Competition	Films released at the same time, …

영화의 흥행 요소는 영화의 특성을 기준으로 흥행 예측에 영향을 미치는 정도[6]가 달라질 수 있다. 상업 영화는 스크린 수, 관객 평가, 장르 등과 같은 요소, 예술 영화는 스크린 수, 감독 명성, 상영 등급 평가 등과 같은 요소가 유의미한 영향을 끼치는 요소로 나타났다.

그러나 영화 흥행 예측을 위해 데이터를 사용한 실험은 수행했지만 그 처리 과정이 생략[4-6]되어 있거나, 빅데이터 측면에서는 적용하기 어려운 방법[7]을 제시하고 있어 보다 효과적인 데이터 처리 방법에 대한 논의가 필요하다.

2.2 전처리

전처리는 수집한 데이터를 활용 가능한 형태로 정제를 하는 과정으로, 데이터마이닝 과정 중 가장 많은 시간을 소요하는 단계[8]이다. 이러한 전처리 작업을 거치지 않은 데이터는 크게 세 가지의 문제점[9]을 보이게 된다. 첫 번째는 데이터 중 필요한 속성이 없는 불완전(Incomplete) 문제, 두 번째는 데이터 내에 의미가 맞지 않는 값이 있거나 범위 밖의 이상치가 있는 노이즈(Noisy) 문제, 세 번째는 여러 데이터를 합칠 때 발생할

표 2 데이터 전처리의 다섯 가지 단계
Table 2 Five phases of data preprocessing

Data cleaning	- Fill in missing values - Smooth noisy data - Identify or remove outliers
Data integration	- Integration of multiple databases, data cubes, or files
Data transformation	- Normalization - Aggregation - Summarization
Data reduction	- Obtains reduced representation in volume but produces the same or similar analytical results
Data discretization	- Part of data reduction but with particular importance, especially for numerical data - Convert numerical values to categorical values

수 있는 비일관성(Inconsistent) 문제이다.

전처리 과정은 작업 형태별로 다섯 가지 단계[12]로 구분이 가능하다. 각 단계는 데이터 클리닝(Data Cleaning), 데이터 통합(Data Integration), 데이터 변환(Data Transformation), 데이터 축소(Data Reduction), 데이터 이산화(Data Discretization)로 구성되며 각 과정의 상세한 작업은 표 2와 같이 정리된다[8].

이러한 전처리 과정을 효과적으로 수행하기 위한 일련의 연구가 진행되어 왔다. [10]에서는 웹 사용 마이닝에서 웹사이트 사용자들의 이용 패턴을 파악하는 분석을 위한 웹 서버 로그파일 전처리에 대한 방법론을 제시하였다. 국내 연구로는 유사한 웹 서버 로그 파일 전처리 연구를 국내 도메인 기반으로 수행[11]하였다.

그러나 전처리의 중요성을 강조하고 있는 반면[8-12], 각 전처리 방법에 대한 실질적인 효과를 입증할 수 있는 실험 평가 과정은 제시되어 있지 않다. 보다 실질적으로 연구 활동에 적용할 수 있기 위해서는 재현 가능하고 구체적인 사례를 통한 전처리 방법론의 제시가 필요하다.

3. 영화 빅데이터 전처리

본 절에서는 영화 흥행 예측을 위해 사용될 대용량 영화 데이터를 대상으로 전처리 관점에서의 문제 분석 및 해결 방법을 제시한다.

분석을 위해 3가지 유형의 영화 데이터를 수집하였다. 첫 번째는 영화진흥위원회에서 제공하는 영화 정보(이하 “영화 메타데이터”로 지칭), 두 번째는 영화진흥위원회 제공 일일 흥행 데이터(이하 “일일 흥행 데이터”로 지칭), 세 번째는 웹 검색 사이트인 네이버의 영화 관련 메뉴에 등록된 사용자 댓글 및 평점 데이터(이하 “댓글 데이터”로 지칭)이다.

위 세 데이터를 분석하여 영화 데이터의 전처리적 특성을 찾아내고, 데이터 전처리의 다섯 가지 단계를 활용해 영화 데이터의 전처리를 수행한다.

3.1 데이터 클리닝(Data Cleaning)

3.1.1 댓글 데이터

웹 상에서 추출한 “댓글 데이터”를 CSV(Comma-separated Values) 형식의 반정형 구조로 변경시키기 위해서는 하나의 댓글 내용은 한 줄로 정리해야 한다. 이를 위해 개행 문자 등 특수문자로 인한 줄바꿈 현상을 없애는 작업이 필요하다. 또한 여러 곳에서 대규모 병행 추출하는 과정에 있어 동일한 댓글이 중복 추출될 수 있는 확률이 존재하므로, 댓글을 구분하는 유일키를 사용해 중복된 댓글을 제거하는 작업을 수행한다.

3.1.2 일일 흥행 데이터

영화진흥위원회에서는 사용자가 열람을 원하는 기간의 일일 흥행 데이터를 제공한다. 이 데이터는 엑셀 파일 형태로 제공되나 CSV 파일로 바로 변환할 수 없는 비정형 형태의 데이터 구조를 이루고 있다. 우선 한 행이 한 레코드인 상태가 아니므로 개행 줄이나 의미 없는 내용이 있는 줄을 지우는 작업이 필요하다. 일별 데이터 분석을 위한 날짜 속성이 없는 대신 일주일 단위로 레코드들이 묶어진 상태이므로 합쳐진 레코드를 풀어내어 열 정보로 날짜 속성을 추가한다.

3.1.3 일일 흥행 데이터와 영화 메타데이터

“일일 흥행 데이터”와 “영화 메타데이터” 모두 데이터 중간 비어있는 속성 값이 존재한다. 이런 빈 속성 값으로 인해 비정형 데이터를 CSV 파일로 변환할 때 값 정렬이 깨지는 현상이 발생할 수 있다. 따라서 그림 1과 같이 빈 속성 값 위치에 명시적으로 “NULL”값을 채워 넣어 값 정렬이 변하지 않도록 처리해준다.

78726	The Thieves		Feature film	26 M $	3.5 M Spectators	135 Min.
72054	The Dark Knight Rises	Warner Bros. Pictures	Feature film	16 M $	2 M Spectators	164 Min.

⬇

78726	The Thieves	NULL	Feature film	26 M $	3.5 M Spectators	135 Min.
72054	The Dark Knight Rises	Warner Bros. Pictures	Feature film	16 M $	2 M Spectators	164 Min.

그림 1 데이터 클리닝: 빈 속성 값 처리의 예
Fig. 1 Data cleaning: processing empty values example

3.2 데이터 축소(Data Reduction)

3.2.1 영화 메타데이터

영화 정보를 수집한 “영화 메타데이터”와 “일일 흥행 데이터”간에 중복되는 영화 속성값이 존재하는 현상이 발견되었으며, 중복 정보를 제거하는 작업(그림 2)을 통해 데이터의 일관성을 유지한 채 원본의 63.1% 크기로 데이터를 감소시킬 수 있었다.

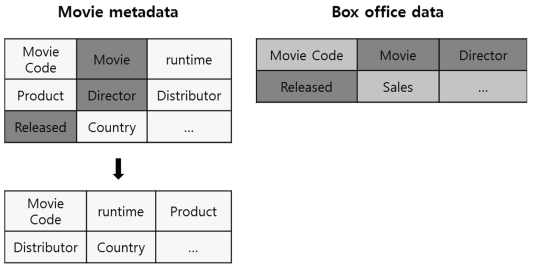

그림 2 중복 속성 값 제거의 예
Fig. 2 Example where duplicate values are eliminated

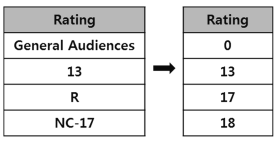

그림 3 속성 값 이산화의 예
Fig. 3 Example of the discretization of values

3.2.2 일일 흥행 데이터

“일일 흥행 데이터”의 관람 등급 속성값에서 통일되지 않은 용어를 사용한 현상이 발견되었다. 예를 들어 “국민학생관람불가”, “중학생 이상 관람가”와 같이 같은 등급을 표현 하는 다른 이름을 가진 경우가 발견되었다. 따라서 정보의 일관성을 지킬 수 있도록 문자로 된 관람 등급 값을 정수 형태의 숫자 나이 값으로 변경하는 작업(그림 3)을 수행하였다.

3.3 데이터 통합(Data Integration)

3.3.1 일일 흥행 데이터와 영화 메타데이터

다양한 출처에서 모은 영화 데이터를 통합하는 과정을 위해 각 영화 정보에 대한 유일한 식별자를 지정해 주어야 한다. 식별자로는 영화 코드나 영화 제목 등 유일함을 판단할 수 있는 영화 속성을 사용해 지정한다.

3.4 데이터 변환(Data Transformation)

3.4.1 일일 흥행 데이터와 영화 메타데이터

“일일 흥행 데이터”와 “영화 메타데이터”에서 하나의 속성에 한 개 이상의 값을 포함하는 속성이 다수 존재한다. 이러한 문제는 영화 데이터 특성상 하나의 영화가 여러 제작사, 장르, 배우 등을 가질 수 있기 때문에 발생한다.

이와 같은 데이터는 관계형 데이터베이스 관점으로 제 1정규화 규정인 데이터 원자화(Atomicity)를 위반하는 것이다. 이 문제를 해결하기 위해 그림 4와 같이 무손실 분해(Lossless Decomposition) 가능한 상태로 각 영화 속성의 값을 분리하는 과정을 거친다.

4. 빅데이터 분석

본 절에서는 3절에서 언급한 전처리 과정을 빅데이터

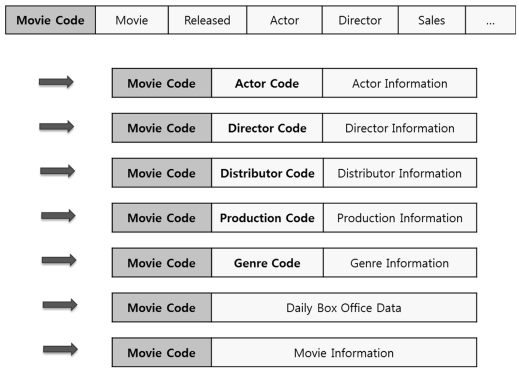

그림 4 데이터 통합: 정규화 처리의 예
Fig. 4 Example where movie data is normalized

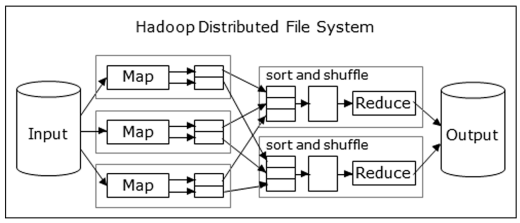

그림 5 하둡 파일시스템에서의 맵리듀스 프레임워크
Fig. 5 MapReduce framework on HDFS

기술을 사용해 처리를 하는 방법에 대해 설명한다.

빅데이터 전처리를 위해 오픈소스인 하둡 맵리듀스 프레임워크를 사용하였다. 맵리듀스는 분산 병렬 시스템에서 대용량 데이터를 처리하기 위해 고안된 프레임워크로써 그림 5처럼 데이터를 특정 함수 및 패턴에 따라 키와 값 집합으로 나누는 맵(Map) 단계와 이 집합을 키 별로 합치는 리듀스(Reduce) 단계로 구성된다. 하둡 파일 시스템(HDFS)은 대용량 데이터를 여러 노드에 분산시켜 병렬적인 맵(Map) 연산 작업을 처리한다.

본 연구에 사용된 “일일 흥행 데이터”, “영화 메타데이터”, “댓글 데이터”는 각각의 영화 관련 속성 정보를 가지고 있으므로 각 속성별 데이터베이스 테이블을 정의해 정형 데이터를 구축하였다. 이 과정 중 대용량 데이터를 대상으로 한 효과적인 데이터 분리 및 정규화 작업을 위해 맵리듀스 알고리즘을 사용해 구현하였다.

4.1 데이터 클리닝

정형화된 영화 데이터베이스를 구축하기 위한 맵리듀스 알고리즘을 수행하기 위해 기본적으로 한 줄에 하나의 영화 정보를 나타내는 형식으로 모든 영화 정보를 정리할 필요가 있다. 이를 위해 입력 받은 영화 원시 데이터(Raw data)를 대상으로 데이터 클리닝 전처리 과정(그림 6)을 수행한다. 입력 데이터에 대해 하나의 영화에 대해 여러 줄로 나뉜 레코드가 있으면 한 줄의 레

```
method Cleaning(string previousLine, string currentLine)
    // Check current line if it is separated
    currentLine ← previousLine + currentLine
    if IsSameMovie(previousLine, currentLine) then
        previousLine ← previousLine + currentLine
        return {previousLine, ""}
    // Remove special characters
    currentLine←Replace(currentLine, '[^0-9a-zA-Z\\s]','')
    // Pad the null value into empty columns
    values ← ColumnSplit(currentLine)
    for all v ∈ values do
        v ← Trim(v)
        if Empty(v) then v ← "NULL"
    end
    return {"", Join(values, "\t")}
```

그림 6 영화 데이터 클리닝 처리
Fig. 6 Movie data cleaning process

코드로 병합하고 특수 문자를 제거하여 줄바꿈 현상을 방지한다. 그 후 속성의 열(Column) 별 값을 확인하여 값이 비어 있는 속성은 “NULL”을 명시적으로 채워 넣어 이후 데이터 처리시 속성 리스트의 열 배치 깨짐 현상을 예방한다.

4.2 데이터 축소

대용량 영화데이터에 대해 맵리듀스 프레임워크로 병렬적인 데이터 축소 및 이산화 작업을 수행할 수 있다. 예를 들어 그림 7은 맵(Map) 단계에서 실행하는 이산화 함수 중 하나로서 다양한 문장으로 표현된 영화 등급을 하나의 정수형 영화 등급으로 변경해준다.

```
method RatingsDiscretization(string rating)
    case rating of
        G: general audiences: rating ← 0
        PG-13: parental guidance suggested: rating ← 13
        12A: suitable for 12 years and over: rating ← 12…
        R-18: adults only: rating ← 18
    end case
    return rating
```

그림 7 영화 등급 이산화
Fig. 7 Discretization of movie ratings

4.3 영화 기본 정보와 흥행 데이터의 정규화

4.1절과 4.2절의 과정을 거친 영화 데이터를 맵(Map) 단계를 통해 각 영화 제목을 기준으로 그룹화하고 리듀스(Reduce) 단계를 통해 배우, 감독, 배급사, 장르, 제작사, 영화 기본 정보, 흥행 통계 속성 별로 분할하여 속성별 데이터베이스 테이블을 구축할 수 있도록 맵리듀스 출력 파일을 생성하였다.

4.3.1 제1정규화

데이터의 중복을 최소화하고 흥행 분석 시 각 속성의 영향력을 용이하게 측정할 수 있는 구조를 유지하기 위해 관계형 데이터베이스 설계의 제1정규형에 따라 각 속성이 오직 하나의 값을 가지는 형태로 구조화하였다. 제1정규형에 따라 영화 별로 1개 이상의 값을 지닌 속성은 개별 테이블로 분할하여 속성 간의 관계를 강화하고 다양한 질의를 테이블 간의 조인을 통해 가능하도록 하였다. 그림 8은 맵리듀스 알고리즘을 사용해 제1 정규화 작업을 하는 과정을 보이고 있다.

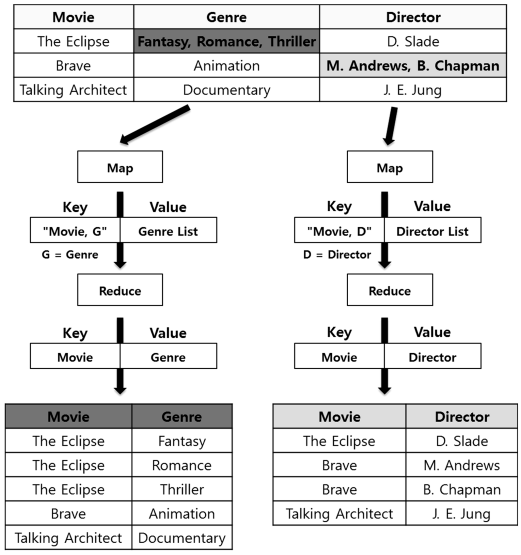

그림 8 맵리듀스 알고리즘을 이용한 제 1정규화 처리
Fig. 8 First normal form using the MapReduce algorithm

4.3.2 제 2정규화

제2정규형은 테이블의 모든 비기본(Non-prime) 속성이 모든 후보 키에 함수 종속일 때 만족된다. “일일 흥행 데이터”의 경우 감독, 장르, 배우 등의 기본 영화 속성들은 ‘영화명’에 함수 종속이다. 반면 당일 매출, 관객수 등 일일 흥행 현황 속성은 ‘영화명, 상영일’ 복합 후보 키에 함수 종속이다. 즉 현재 “일일 흥행 데이터”는 제2정규형을 만족하지 않으므로 영화 정보와 흥행 정보를 분리하여 제2 정규형 규칙을 준수하도록 처리를 한다. 그림 9는 맵리듀스 알고리즘을 사용해 제2 정규화 작업을 하는 과정을 보이고 있다.

4.4 댓글 중복 제거

4.1절의 데이터 클리닝 과정은 댓글 데이터에 대해서도 동일하게 적용 가능하다. 그러나 댓글 데이터는 웹 수집(Crawling) 방법의 특성상 같은 데이터를 중복 수집할 가능성이 있으므로 추가적인 댓글 중복 제거 작업

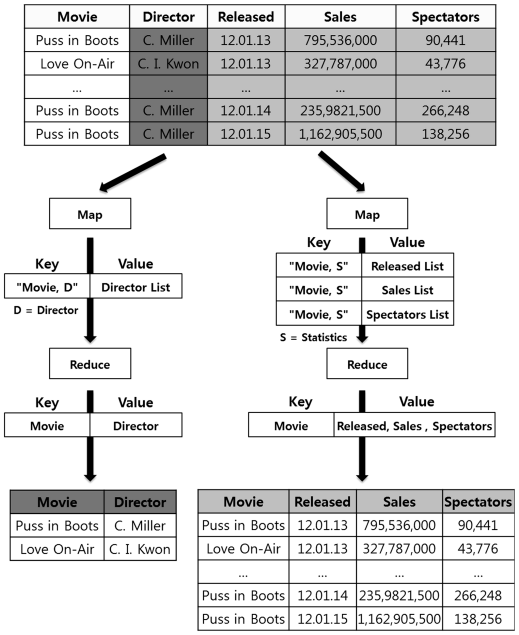

그림 9 맵리듀스 알고리즘을 이용한 제 2정규화 처리
Fig. 9 Second normal form using the MapReduce algorithm

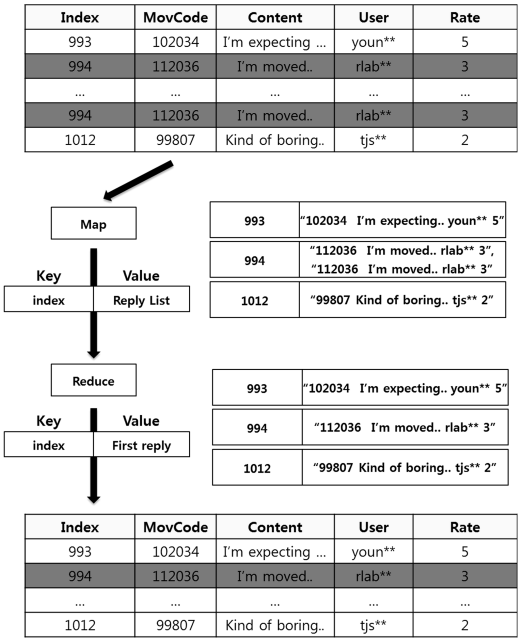

그림 10 맵리듀스 알고리즘을 이용한 댓글 중복 제거
Fig. 10 Elimination of duplicate data using the MapReduce algorithm

이 필요하다. 그림 10은 맵리듀스로 구현한 댓글 중복 제거 알고리즘의 주요 부분을 설명하고 있다. 댓글이 가진 식별 값을 맵리듀스의 키(Key)로 두면 리듀스 단계에서 중복된 값을 한 곳으로 모을 수 있으므로 쉽게 중복 값을 제거할 수 있다.

4.5 정형 데이터베이스 구축

데이터 클리닝 단계를 통해 생성된 반정형 CSV 파일을 읽어 영화 속성의 특성 별로 분리하는 작업과 앞 절에서 설명한 정규화 작업을 거쳐 최종 그림 11과 같은 정형 데이터베이스를 구축하였다.

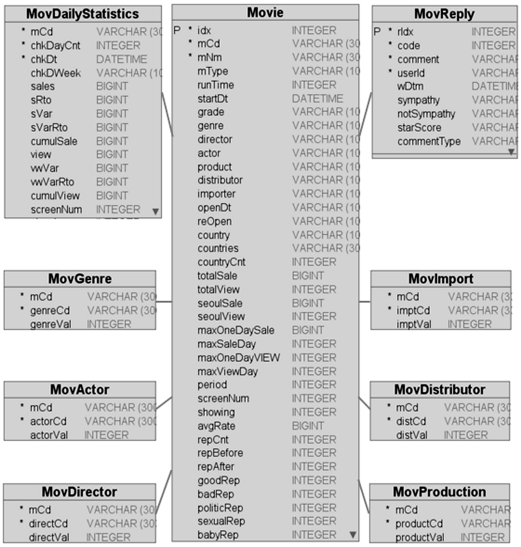

그림 11 전처리 과정 후 구축된 정형 데이터베이스
Fig. 11 Established database form after data preprocessing

5. 실 험

실험을 위해 3절에서 소개한 "영화 메타데이터", "일일 흥행 데이터", "댓글 데이터"를 사용한다. 전체 100개의 엑셀 및 텍스트 파일로 수집되었으며, 데이터 크기는 전체 1 기가, 전체 레코드는 약 100만 레코드로 구성되었다.

그림 12-15는 영화 흥행 예측 분석을 위한 데이터 전처리 과정에 대해 단일 컴퓨터를 사용한 기존의 방법과 빅데이터 기술을 사용한 방법의 성능을 비교한 결과이다. 그림 12는 전처리 과정 중 데이터 클리닝 작업에 대한 성능 비교를 보이고 있다. 그래프의 x 축은 데이터 크기, y축은 수행 시간을 나타낸다. 데이터 크기에 따른 수행 시간 증가폭이 기존의 방법보다 빅데이터 방법이 적은 것을 확인할 수 있다. 그림 13은 데이터 축소 및 속성 값 이산화, 그림 14는 데이터 정규화에 대한 성능 비교이다. 데이터 축소 및 이산화 작업을 비롯하여 데이터 정규화 작업도 빅데이터 방법이 더 좋은 성능을 보이는 것을 알 수 있다. 그림 15는 "댓글 데이터"에서 중복된 컨텐츠를 제거하는 작업에 대한 수행 시간 비교이

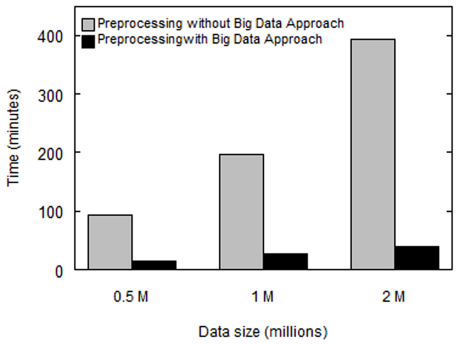

그림 12 데이터 클리닝 시간 비교
Fig. 12 Processing time for data cleaning

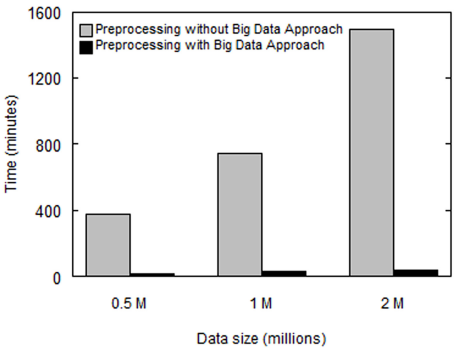

그림 13 데이터 축소 및 이산화 시간 비교
Fig. 13 Processing time for data reduction and discretization

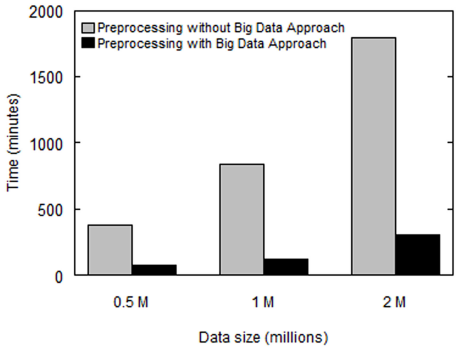

그림 14 데이터 정규화 시간 비교
Fig. 14 Processing time for data normalization

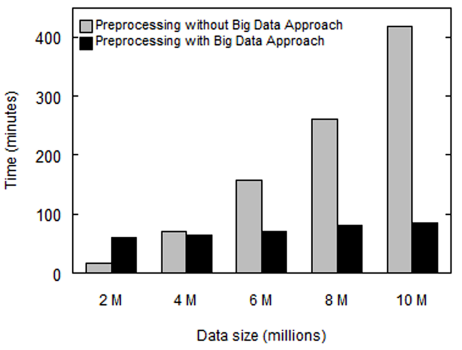

그림 15 댓글 중복 제거 시간 비교
Fig. 15 Processing time for removing duplicate comments

다. 데이터 크기가 작을 때는 단일 기기에서 처리한 시간이 더 빨랐으나 데이터 양에 따른 수행 시간 증가가 큰 폭으로 상승하여 대용량 데이터의 경우는 빅데이터 방법을 사용한 처리가 더 효과적임을 확인할 수 있다.

6. 결 론

본 논문에서는 대규모 영화 시장에서의 보다 효과적인 흥행 예측을 위한 대용량 데이터 전처리 방법을 연구하였다.

정보 분석에서 가장 많은 처리 시간이 소요되는 전처리 과정의 특성상 대용량 데이터 분석 처리 시간을 줄일 수 있는 방법이 필요하다. 따라서 영화 흥행 데이터의 특성을 분석해 전처리의 각 유형별 처리 방법을 지정한 후 하둡 기반 맵리듀스 프레임워크를 사용하는 방법을 사용하였다. 실험 결과 빅데이터 기법을 사용한 전처리가 기존의 방법보다 더 좋은 수행 결과를 보이는 것을 확인하였다.

향후 연구로는 구축한 맵리듀스 알고리즘의 최적화 기법을 연구할 계획이다. 또한 더 정밀한 영화 흥행 예측을 위한 영화 속성 평가 기법에 대해 연구하고자 한다.

References

[1] Motion pictiure association of america, "Theatrical Market Statistics," 2013.

[2] UNESCO Institute for Statistics, "Emerging markets and the digitalization of the film industry," UIS Information Paper No.14, 2013.

[3] Korean Film Council, "2013 Korean Film Industry Report," 2014.

[4] E. M. Kim, "The Determinants of Motion Picture Box Office Performance: Evidence from Movie Exhibited in Korea," *Korean Society for Journalism & Communication Studies*, Vol. 47, No. 2, pp. 190-220, 2003.

[5] H. Y. Jeong, H. J. Yang, "Predicting Financial Success of a Movie Using Multiple Regression Analysis," *Korea Society of Computer & Information Summer Conference*, Vol. 21, No. 2, pp. 275-278, 2013.

[6] S. Y. Kim, "A Comparison Study of the Determinants of Performance of Motion Pictures: Art Film vs. Commercial Film," *The Korea Contents Association*, Vol. 10, No. 2, pp. 381-393, 2010.

[7] S. J. Lee, T. R. Jeon, G. D. Back and S. S. Kim, "A Movie Rating Prediction System Based on Personal Propensity Analysis," *Proc. of kIIS Fall Conference 2008*, Vol. 18, No. 2, pp. 203-206, 2008.
[8] W. Zhang, Available from: http://www.cs.wustl.edu/~zhang/teaching/cs514/Spring11/Data-prep.pdf [Accessed: 17 July 2014]
[9] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, Vol. 17, No. 5-6, pp. 375-381, 2003.
[10] D. Tanasa and B. Trousse, "Advanced data preprocessing for intersites web usage mining," *Intelligent Systems, IEEE*, Vol. 19, No. 2, pp. 59-65, 2004.
[11] W. S. Hyun, "Performance Improvement of Data Preprocessing for Intersite Web Usage Mining," *The Korean Institute of Information Scientists and Engineers Autumn Conference*, Vol. 33, No. 2B, pp. 357-361, 2006.
[12] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *International Journal of Computer Science*, Vol. 1, No. 2, pp. 111-117, 2006.

전 희 국
2004년 동국대학교 컴퓨터공학과 학사. 2007년 고려대학교 전자컴퓨터공학과 석사. 2011년~현재 서울대학교 컴퓨터공학부 박사과정 재학 중. 관심분야는 데이터베이스, 시맨틱 웹, 온톨로지, 빅데이터

현 근 수
2014년 서울대학교 컴퓨터공학과 학사. 2014년 현재 서울대학교 컴퓨터공학부 석사과정 재학 중. 관심분야는 데이터베이스, 시맨틱 웹, 온톨로지, 빅데이터

임 경 빈
2012년 Univ. of Wisconsin-Madison Computer Science 학사. 2013년~현재 서울대학교 컴퓨터공학부 석사과정 재학 중 관심분야는 데이터베이스, 시맨틱 웹, 온톨로지, 빅데이터

이 우 현
2010년 Univ. of Wisconsin-Madison Computer Engineering 학사. 2013년~현재 서울대학교 컴퓨터공학부 석사과정 재학 중. 관심분야는 데이터베이스, 시맨틱 웹, 온톨로지, 빅데이터

김 형 주
1982년 서울대학교 전산학과 학사. 1985년 Univ. of Texas at Austin 석사. 1988년 Univ. of Texas at Austin 박사. 1988년~1990년 Georgia Institute of Technology 조교수. 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이터베이스, 시멘틱웹, 온톨로지, 빅데이터