한국빅데이터학회
제1권 제2호, 2016, pp. 1-8

대규모 로그를 사용한 유저 행동모델 분석 방법론*

The Analysis Framework for User Behavior Model using Massive Transaction Log Data

이종서 · 김성국†

연세대학교 글로벌융합공학부

요 약

사용자로그는 많은 숨겨진 정보를 포함하고 있지만 데이터 정형화가 이루어지지 않았고, 데이터 크기도 너무 방대하여 처리하기 까다로워서 아직 밝혀져야 할 부분들을 많이 내포하고 있다. 특히 행동마다의 모든 시간정보를 포함하고 있어서 이를 응용하여 많은 부분을 밝혀낼 수 있다. 하지만 로그데이터 자체를 바로 분석으로 사용할 수는 없다. 유저 행동 모델 분석을 위해서는 별도의 프레임워크를 통한 변환과정들이 필요하다. 이 때문에 유저 행동모델 분석 프레임워크를 먼저 파악을 하고 데이터에 접근해야 한다.
이 논문에서는, 우리는 유저 행동모델을 효과적으로 분석하기 위한 프레임워크 모델을 제안한다. 본 모델은 대규모 데이터를 빨리 처리하기 위한 분산환경에서의 MapReduce 프로세스와 유저별 행동분석을 위한 데이터 구조 설계에 대한 부분을 포함한다. 또한 실제 온라인 서비스 로그의 구조를 바탕으로 어떤 방식으로 MapReduce 를 처리하고 어떤 방식으로 유저행동모델을 분석을 위해 데이터 구조를 어떤식으로 변형할지 설명하고, 이를 통해 어떤 방식의 모델 분석으로 이어질지에 대해 상세히 설명한다. 이를 통해 대규모 로그 처리 방법과 분석모델 설계에 대한 기초를 다질 수 있을 것이다.

■ 중심어 : 사용자 행동모델분석 프레임워크, 대규모 로그

Abstract

User activity log includes lots of hidden information, however it is not structured and too massive to process data, so there are lots of parts uncovered yet. Especially, it includes time series data. We can reveal lots of parts using it. But we cannot use log data directly to analyze users' behaviors. In order to analyze user activity model, it needs transformation process through extra framework. Due to these things, we need to figure out user activity model analysis framework first and access to data. In this paper, we suggest a novel framework model in order to analyze user activity model effectively. This model includes MapReduce process for analyzing massive data quickly in the distributed environment and data architecture design for analyzing user activity model. Also we explained data model in detail based on real online service log design. Through this process, we describe which analysis model is fit for specific data model. It raises understanding of processing massive log and designing analysis model.

■ Keyword : User Behavior Model Analysis Framework, Massive Transaction Log Data

2016년 6월 30일 접수; 2016년 7월 14일 수정본 접수; 2016년 7월 20일 게재 확정.
* 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 ICT명품인재양성사업의 연구결과로 수행되었음(IITP-R0346-16-1008).
† 교신저자 songkuk@yonsei.ac.kr

Ⅰ. 서 론

사람들은 인터넷의 편리함을 바탕으로 통해 쇼핑, 검색 등 많은 일을 수행하고 있다. 최근에 사용자들이 이렇게 온라인상에서 많은 서비스를 이용하고, 이를 통해 정보를 획득하고 있다. 이러한 사용자들의 활동은 “사용자로그”로 서버에 기록된다. 사용자 행동 로그는 높은 잠재적 연구 가치를 가지고 있다. 음악 스트리밍 청취와 같은 서비스 행태의 데이터 처리방식에서 유저 행동에 대한 기록들은 획일화된 데이터베이스가 아닌 수많은 사용자 로그에 기반하는데, 정보 구조적 체계성이 부족할 뿐만 아니라 데이터 양이 거대하기 때문에 분석을 위해 상당한 시간과 자원이 필요하다. 이러한 이유로 아직 사람들에게 알려지지 않은 많은 부분이 존재한다. 기존의 데이터 분석 툴인 SPSS, SAS, Excel 등은 데이터를 분석하기 위해서 굉장히 편하고 좋은 기능들을 제공하고 있지만, 사용자 로그와 같은 대규모 데이터를 처리해서 사용자를 분석하기에는 몇 가지 제약 사항들이 존재한다. 첫째로 이러한 툴들은 데이터가 정제되어 컬럼화 되어있지 않으면 원하는 항목을 제대로 분석할 수 없다. 둘째로 시스템 자체가 대규모 데이터에 적합하지 않다. 일반적으로 실제 사용자 로그는 월별 Gigabyte 단위로 구성되는데 기존 일반 소프트웨어로는 이 데이터를 업로드 조차할 수 없다. 마지막으로 업로드가 완료가 된 상황이라도 분석하는데 시간이 상당히 소모될 수 있다. 이 연구에서, 우리는 대규모의 로그를 사용하여 사용자의 행동모델을 분석하기 위한 분석프레임워크를 제시한다. 본 모델을 통해 사용자로그와 대규모 데이터를 Data Set, Bag, Session 형태로 전처리하고 자신이 분석하고자 하는 유사도분석과 같은 유저분석법에 연결시킬 수 있는 인사이트를 얻을 수 있다. 이 논문의 이후 부분에서, 우리는 가장 먼저 대규모 로그 분석을 수행한 연구, 로그를 통해 사용자 측면에서 프레임워크를 제시한 연구들을 정리하였다. 다음으로 로그기반 사용자 행동모델 분석을 위한 프레임워크에 대해 소개한다. 프레임워크를 통해 적합한 데이터구조를 어떻게 설명하고 이 구조가 어떤 분석으로 이어질지에 대해 상세하게 다룬다. 이를 위해 기초가 되는 로그데이터와 Hadoop과 MapReduce 처리 방식에 대해서도 함께 서술한다. 이어서 제시한 사용자 행동분석 프레임워크의 제약점 등에 대한 토론 세션을 설명한다. 마지막으로 논문에 대한 결론 및 앞으로의 미래 연구방향에 대해 서술한다.

Ⅱ. 관련연구

2.1 로그를 통한 대규모 데이터 분석연구

로그는 사용자의 많은 행동을 내포하기 때문에 많은 학자들이 로그를 통해 사용자들의 행동을 분석해왔다. Domais et al.[3]은 로그분석을 통한 유저 활동에 대한 이해라는 주제로 HCI 도메인에서의 기존 연구 상황들을 정리하고, 데이터 수집, 정제 실험에 이르기까지 다양한 상황을 정리하였다. 이상준, 이동훈[1]은 로그를 이용해서 실시간 예측분석시스템에 대한 연구를 수행하여, 분석 시스템을 직접디자인하고 구현까지 완성시켰다. 검색활동의 로그분석연구도 활발히 이루어졌는데 Park et al.[12]은 네이버의 트랜잭션 로그를 분석해서 사용자들의 검색 활동을, Ke et al.[8]은 대만의 Eslevier ScienceDirect.com을 분석하여 사용자들의 저널 탐색활동을 연구하였다. 이밖에도 음악 스트리밍과 같은 온라인 스트리밍 서비스의 스트리밍 로그 분석도 연구가 많이 진행되고 있다. Lee[9, 10]은 상황정보를 결합하여 음악을 추천하기 위한 모델을 제안하는 연구를 수행하였다. Levy and Bosteels[11]는 Last.fm으로부터 음악 스트리밍데이터를 수집하고 롱테일 기준으로 유사도를 계산하여 추천하는 모델을 제안하였다.

2.2 분산환경에서의 데이터 분석 프레임워크연구

로그와 같은 비정형 데이터를 효율적으로 처리하고 분석하기 위한 프레임워크에 대한 연구도 활발히 진행되어 왔다. Hussain et al.[6]은 로그분석 및 스트리밍 데이터 마이닝을 위한 고성능의 프레임워크를 통해 이벤트를 잘 수집하고 실시간으로 계산하는 모델을 선보였다. Sharma and Busch[13]는 분산계층구조상에서의 유전 알고리즘을 사용해서 다이나믹하게 계산을 수행해주는 분석 프레임워크를 제안하였다. Jiang et al.[7]은 Hadoop 상에서 모바일 인터넷 접근 로그를 가지고 분산된 데이터 마이닝 분석 프레임워크를 제안하여 대규모 데이터를 빠르게 처리하는 모습을 보여주었다. 대부분의 기존 연구에서는 분산환경을 이용하여 빠르게 처리하는 성능적 개선 연구를 수행하였다. 본 연구는 기존 연구와는 다르게 사용자의 행동 패턴을 효율적으로 분석하기 위한 데이터 조작적 측면에서의 프레임워크 연구를 다루고 있다.

Ⅲ. 방법론

3.1 온라인 서비스 사용자 활동 로그

사용자 활동 로그란 인터넷상에서 사용자의 각 활동(클릭)에 따른 데이터 처리의 시간 순서 기록을 말한다. 보통 로그의 크기는 웹 사이트의 사용자 수와 활성화 정도, 그리고 개발자의 로그파일 설계에 따라 그 크기가 달라질 수 있다. 수십만의 회원이 있는 음악 스트리밍 서비스 로그의 경우 한 달에 단순 음악 청취기록만 약 4 Gigabyte에 달한다. 사용자 활동 로그에 대한 구성정보는 다음 식 (1)과 같다.

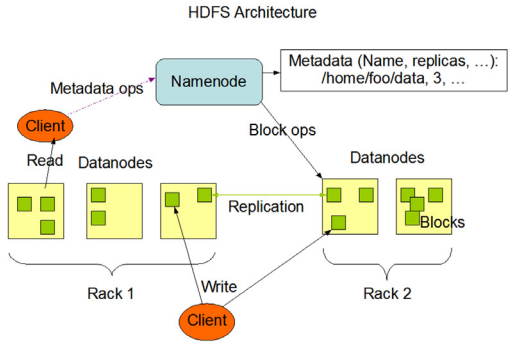
$$Object_{ID} + TimeStamp + AccessRoute + Customer_{ID} \quad (1)$$

Object_ID에는 실제 어떤 데이터에 접속했는지에 대한 기록이다. 음악 스트리밍 서비스에 경우에서는 Song_ID처럼 특정곡이 해당된다. Timestamp는 시간에 대한 기록으로 사용자가 언제 행위를 했는지 파악할 수 있게 한다. Timestamp는 년도, 월, 일, 시, 분, 초의 정보를 포함하고 있다. Access_route는 모바일이나 웹 등의 접속기기를 말하고 마지막으로 Customer_ID는 행위를 특정 사용자 개인을 식별해주는 ID를 말한다. 위와 같은 항목이 특정 기간 동안 수백 만 개, 수천 만 개가 모여서 한 달간의 수 Gigabyte에 달하는 사용자 활동 로그를 이룬다.

3.2 Hadoop 시스템, MapReduce, Key-Value 구조

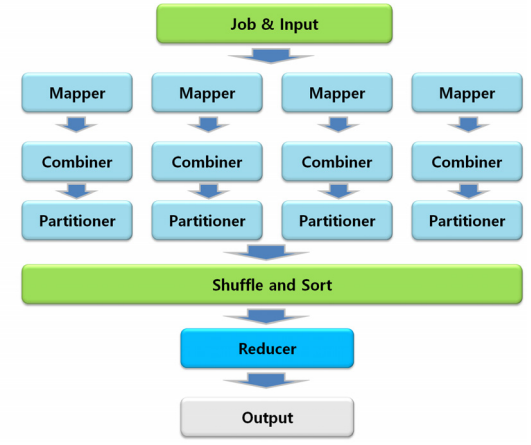
Hadoop 분산 파일 시스템(HDFS)은 하드웨어 여러 대를 사용하여 동작을 수행하기 위해 디자인 되었다[5, 14]. 기존 분산시스템들과 매우 흡사하지만 오류를 스스로 복구할 수 있고 저비용으로 분산환경을 구축할 수 있도록 디자인 되었다. Hadoop 분산 파일 시스템은 처리량이 매우 우수하여 대규모 데이터에 적합하다. 아래 <그림 1>처럼 하나의 Namenode가 복수의 Datanodes들을 관리하여 Job을 수행하고, 결과를 취합한다.

이러한 처리과정은 MapReduce라는 메커니즘에 의해 구동된다[2]. MapReduce는 수행하는



〈그림 1〉 HDFS의 구조[11]

Job에 대하여 Mapper라는 과정을 통해 여러 개의 분산환경에서 나눠서 작업을 하고 Reducer라는 합치는 과정을 통해 결과를 빠르게 도출해 낼 수 있다. 이러한 작업을 통하면 워드카운트, 키 인덱스 작업등 많으면서 단순한 작업들을 많은 컴퓨터에서 분산시켜 빠르게 작업할 수 있다. 아래 <그림 2>는 MapReduce의 Job 수행과정을 나타낸 것이다. 특정 Job에 대해 여러 개의 Mapper로 일을 분산 시키면서 최적화를 위하여 합칠 수 있는 것은 미리 합쳐서 계산을 진행하는 Combiner 과정과 key가 같은 것끼리 모아서 보내주는 Partitioner 과정을 거쳐서 Reducer를 통해 Key에 따라 필요한 작업들을 최종적으로 합쳐서 Output을 출력하는 작업을 보여준다.



〈그림 2〉 MapReduce의 Job 수행과정

여기서 중요하게 봐야할 것이 Key-Value라는 데이터 구조이다. MapReduce를 처리 시 상당수의 작업이 Key-Value 자료구조기반으로 수행된다. 대규모의 작업을 분산환경을 통해 나눠서 수행하고 Reducer로 관련 있는 것끼리 취합할 때의 기준이 바로 Key이다. 아래 <그림 3>은 실제 상단의 로그기록과 같은 데이터가 MapReduce의 분산 처리환경을 거쳐서 Customer_ID를 Key로 하여 관련 유저의 행동기록(Value)끼리 연결하여 출력되는 과정을, <그림 4>는 Reducer에서 취합된 유저별 데이터를 분석하기 편하게 변형하여 출력한 최종결과 예시를 보여주고 있다.

```
323424155 20160401000014 M 132344445
545434345 20160401000334 M 121111145
311111115 20160401000637 M 121111145
555524155 20160401000014 W 122222325
555524155 20160401000014 M 234578233
555524155 20160401000014 W 143455555
311111115 20160401000956 M 121111145
```

```
121111145 545434345_20160401000334
121111145 311111115_20160401000637
121111145 311111115_20160401000956
```

〈그림 3〉 사용자 로그 기록 원본(상)과 MapReduce 처리후의 사용자별 취합된 행동 기록(하)

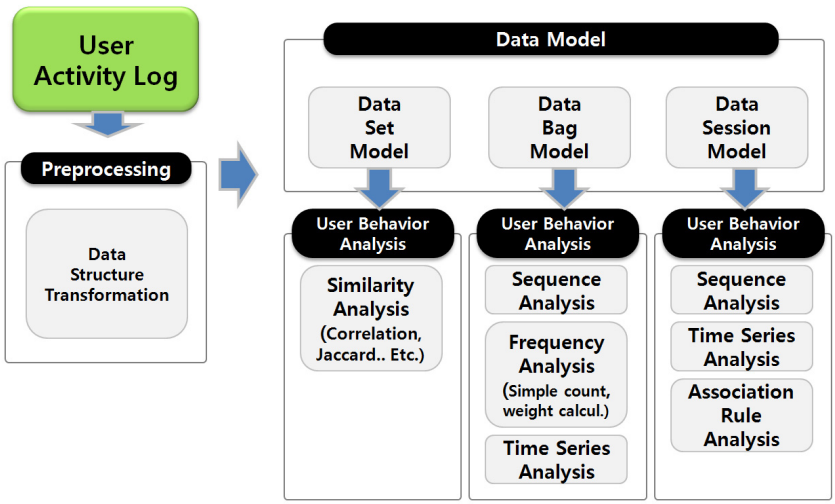
〈그림 4〉 MapReduce에서 Reducer의 Input, Output에 해당하는 Key-Value 구조와 예시

3.3 유저 행동모델 분석 프레임워크

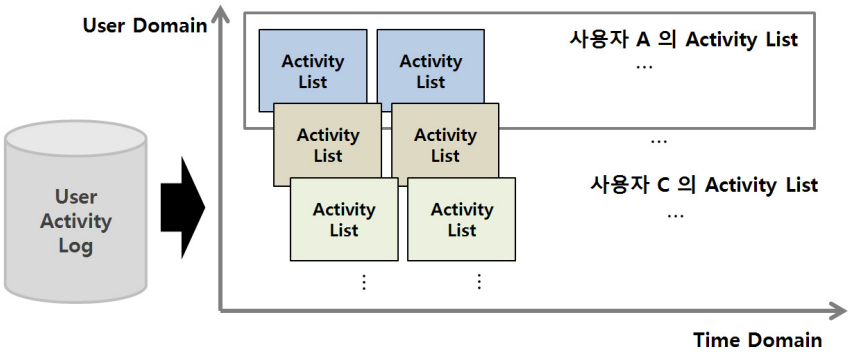
3.3.1 프레임워크 전반적 개요

사용자 행동을 분석할 수 있는 여러 가지 모델들이 존재하는데 각각 효과적으로 분석하기 위해서 요구되는 데이터 구조들이 다르다. 사용자 활동 로그는 비정형 데이터로 전처리 과정을 거쳐서 각각의 분석모델에 적합한 데이터 구조인 Data Set, Data Bag, Data Session 형태로 바뀌어서 분석을 수행하면 효과적으로 결과를 얻어낼 수 있다. 우리가 제시하는 로그기반 유저 행동 모델 분석 프레임워크는 <그림 5>와 같다.

먼저 사용자 활동 로그를 전처리 과정을 통해 어떻게 변형 하는지에 대해 알아보고, 다음으로 각각의 Data 모델들의 상세한 구조와 이것이 어



〈그림 5〉 사용자 행동모델 분석 프레임워크



〈그림 6〉 사용자 활동 분석을 위한 Data Model 추출과정

떤 모델 분석으로 이어질 수 있는지에 대해 설명하고자 한다.

<그림 4>와 같이 사용자 활동 로그에 대해 사용자 도메인과 시간 도메인으로 나누어 데이터를 정리하여 보면 <그림 6>과 같이 사용자별 순차적 상황에 대한 행동 리스트들을 도출해낼 수 있다. 이것을 중복을 허용하지 않고 합치는 것을 <그림 5>의 Data Set Model이고, 중복을 허용하여 뒤로 계속 시간순서대로 이어 붙이는 것을 Data Bag Model, 마지막으로 연달아 한 행동이지만 Timestamp 상에서 이어지는 행동과 Break time이 포함된 후에 이어지는 리스트가 있을 수 있다. 이것을 고려해서 나열한 것이 Data Session Model이다.

3.3.2 데이터 구조 상세설명 및 적합한 분석모델

이 섹션에서 설명할 데이터들의 구조는 전부 <그림 4>에서 설명한 Key-Value 구조에 기반하고 있다. Data Set, Bag, Session Model 에 대해 상세히 알아보고 어떤 분석으로 바로 이어질 수 있는지 설명한다.

3.3.2.1 Data Set Model

먼저 Data Set Model은 <그림 7>과 같이 하나의 Key인 User_ID와 거기에 Object_ID들의 조합으로 구성되는데 여기서 모든 Object들은 Unique 하다. Data Set Model의 장점은 사용자별로 어떤 구성원소들이 있는지 파악이 바로 되기 때문에

사용자간의 유사도 파악이 굉장히 용이하다. Jaccard Similarity나 Pearson Correlation Coefficient 같은 Common data를 기반으로 한 유사도 계산을 쉽게 진행하여 수많은 유저 중에 특정유저와 비슷한 유저를 계산해서 도출해 내는 상황에 적합하게 사용될 수 있다. 음악 스트리밍 서비스의 경우로 예를 들면 Data Set Model을 가지고 서로 전체 노래 중에 얼마나 많은 같은 노래를 들었는지를 가지고 상호간의 유사도를 측정할 수 있다.

- Data Set Model

User_ID Object_ID1| Object_ID2 | Object_ID3 | Object_ID4 | Object_ID5 |.....

스트리밍 청취 행동 예시
Customer_ID Song_ID7| Song_ID32| Song_ID44| Song_ID55| Song_ID11|.....

〈그림 7〉 Data Set Model Key-Value 구조 및 상세 예시

3.3.2.2 Data Bag Model

다음으로 Data Bag Model은 아래 <그림 8>과 같이 Data Set Model처럼 하나의 Key인 User_ID와 거기에 Object_ID들의 조합으로 구성되는데 여기서 모든 Object들은 Repetitive하다. 사용자들은 모든 독립적인 행동만 하는 것이 아니라 자신이 선호하는 몇 개의 소수의 행동을 반복적으로 수행하게 된다. 이러한 점을 고려해 Data Bag Model은 사용자가 하는 행동을 중복유무에 상관없이 순차적으로 이어서 붙여나간다.

- Data Bag Model

User_ID Object_ID1| Object_ID2 | Object_ID1 | Object_ID1 | Object_ID1 |.....

스트리밍 청취 행동 예시
Customer_ID Song_ID7| Song_ID32| Song_ID7| Song_ID7| Song_ID7|.....

〈그림 8〉 Data Bag Model Key-Value 구조 및 상세 예시

Data Bag Model의 장점은 사용자별로 수행한 모든 행동이 나열 되게 때문에 행위 간의 Frequency Analysis, Sequence Analysis, Time Series Analysis 등이 가능하고, 또한 이를 사용하여 Object별 Weight도 계산할 수 있다. 음악 스트리밍 서비스에서는 이 데이터 구조를 통해 각 유저가 들은 노래별로 Frequency 기반으로 전체 청취수 대비 노래에 대한 weight를 부여할 수 있다.

3.3.2.3 Data Session Model

마지막으로 Data Session Model은 아래 <그림 9>와 같이 하나의 Key인 User_ID와 거기에 Object_ID들의 조합으로 된 여러 개의 중복이 허용된 Data Bag 리스트들을 포함한다. 중요한 건 여러 리스트가 한 사람에게 연결된다는 것이다. 그리고 각 리스트 간에는 시간상의 공백이 존재한다. 즉 Timestamp에서 연달아 한 것은 하나의 리스트에 포함되고 break time이 존재한 후 한 행동에 대해서는 다른 리스트로 구분된다는 것이다.

- Data Session Model

User_ID1 Object_ID1| Object_ID2 | Object_ID3 | Object_ID4 | Object_ID5
User_ID1 Object_ID7| Object_ID8 | Object_ID9
User_ID1 Object_ID11| Object_ID12 | Object_ID11 | Object_ID11

스트리밍 청취 행동 예시
Customer_ID1 Song_ID7| Song_ID32| Song_ID44| Song_ID55| Song_ID11
Customer_ID1 Song_ID3| Song_ID4| Song_ID3| Song_ID4

〈그림 9〉 Data Session Model Key-Value 구조 및 상세 예시

Data Session Model은 사용자가 순차적 행동들이 리스트로 존재한다는 점이 특징이다. 이는 쇼핑의 경우로 예를 들면 같이 구매한 상품이 뭔지를 파악할 수 있는 형태이다. 즉 일시적 기점으로 동시에 한 행동들의 패턴을 분석하는 Association Rule Analysis가 가능하다. 또한 리스트간의 순차가 존재하기 때문에 Data Bag Model과 마찬가지로 Sequence Analysis, Time- Series Analysis도 가능하다.

Ⅳ. 토의

제안된 프레임워크는 로그가 Time-Series 정

보를 포함하고 있는 것을 핵심으로 잡고 설명을 하고 있다. 따라서 로그자체에 Timestamp 정보가 포함되어 있지 않다면 Data Bag Model이나 DATA Session Model에서 Time-Series 분석이나 Sequence 분석을 제대로 수행할 수 없다.

다음으로, 사용자 행동모델 분석프레임워크를 설명함으로써 하나의 로그가 하나의 행동만을 포함한다는 가정하에 설명을 진행하였다. 하지만 개발자의 로그 설계에 따라 이는 바뀔 수도 있다. 다양한 행동을 하나의 로그에 전부 기록한다고 한다면 먼저 행위 별로 로그를 분류하는 작업을 먼저 수행한 후에 제시된 프레임워크를 통해 분석을 하면 효과적으로 복합 로그도 다룰 수 있다.

또한 사용자 행동모델 분석 프레임워크를 설명하면서 하나의 대규모 로그에서만 추출하는 방식으로 설명을 진행하였는데, 이 프레임워크는 데이터가 확장이 가능하기 때문에, 기간이 다른 여러 로그파일에서 필요한 Data Model을 추출 후 비교 및 활용이 용이 하다.

마지막으로, 본 모델은 프레임워크에 대한 성능테스트를 실시하지 않았다. 추후 전수모델을 사용한 방법과 본 모델을 사용한 방법의 효율성 개선 측면을 추가적으로 상세하게 연구할 계획이다.

Ⅴ. 결론

우리는 대규모의 로그를 사용하여 사용자의 행동모델을 분석하기 위한 분석프레임워크를 제시하였다. 본 모델을 통해 사용자로그와 대규모 데이터를 Data Set, Bag, Session 형태로 전처리하고 자신이 분석하고자 하는 유사도분석과 같은 유저분석법에 연결시킬 수 있는 인사이트를 얻을 수 있다. 기존에 분석툴들은 사용자 로그와 같은 대규모 데이터를 다루기에 적합하지 않고 효율적인 분석을 위해 접근하는 방법도 용이하지 않았다. 결론적으로, 사용자 행동모델 분석 프레임워크는 이런 기존의 상황을 개선할 수 있다. 로그들을 Hadoop 시스템의 MapReduce 전처리 과정을 통해 제시한 Data Set, Bag, Session Model로 변환하고, 해당 모델에 적합한 분석을 수행하면 효과적으로 작동할 것이다. 본 연구진은 이 프레임워크를 바탕으로 사용자 유사도와 시퀀스 분석 같은 연구를 더 진행해 나갈 예정이다.

참 고 문 헌

[1] 이상준, 이동훈, “빅데이터 로그를 이용한 실시간 예측분석시스템 설계 및 구현”, 정보보호학회논문지, 제25권, pp.1399-1410, 2015.

[2] Dean, J. and S. Ghemawat, “MapReduce: simplified data processing on large clusters”, *Communications of the ACM*, Vol.51, pp.107-113, 2008.

[3] Dumais, S., R. Jeffries, D.M. Russell, D. Tang, and J. Teevan, “Understanding user behavior through log data and analysis”, in Ways of Knowing in HCI, ed: Springer, pp.349-372, 2014.

[4] Flexer, A. and D. Schnitzer, “Effects of album and artist filters in audio similarity computed for very large music databases”, *Computer Music Journal*, Vol.34, pp.20-28, 2010.

[5] Foundation, A.H., “Architecture of Hadoop Distributed File System”, 2013.

[6] Hussain, A.R., M.A. Hameed, and S. Fatima, “A Proposal: High-Throughput Robust Architecture for Log Analysis and Data Stream Mining”, in *Innovations in Computer Science and Engineering*, ed: Springer, pp.305-314, 2016.

[7] Jiang, Y., J. Yang, L. Tang, Y. Liu, X. Zhao, and X. Hao, “A Distributed Data Mining System Framework for Mobile Internet Access Log Based on Hadoop”, in *Transactions on Edutain-*

ment XI, ed: Springer, pp.243-252, 2015.

[8] Ke, H.-R., R. Kwakkelaar, Y.-M. Tai, and L.-C. Chen, "Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan", *Library & Information Science Research*, Vol.24, pp.265-291, 2002.

[9] Lee, J.S. and J. C. Lee, "Music for my mood: A music recommendation system based on context reasoning", in *Smart sensing and context*, ed: Springer, pp.190-203, 2006.

[10] Lee, J.S. and J.C. Lee, "Context awareness by case-based reasoning in a music recommendation system", in *Ubiquitous Computing Systems*, ed: Springer, pp.45-58, 2007.

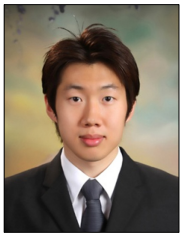
[11] Levy, M. and K. Bosteels, "Music recommendation and the long tail", in 1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain, 2010.

[12] Park, S., J.H. Lee, and H.J. Bae, "End user searching: A Web log analysis of NAVER, a Korean Web search engine", *Library & Information Science Research*, Vol.27, pp.203-221, 2005.

[13] Sharma, G. and C. Busch, "An analysis framework for distributed hierarchical directories", *Algorithmica*, Vol.71, pp.377-408, 2015.

[14] Shvachko, K., H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system", in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium*, pp.1-10, 2010.

저 자 소 개



이 종 서(Jongseo Lee)

- 2010년 : 아주대학교 경영대학 e-Business 학과전공, 경영학과 복수전공 (e-Business, 경영학사)
- 2012년 : 연세대학교 공과대학 정보산업공학과 (공학석사)
- 2012년～현재 : 연세대학교 공과대학 글로벌융합공학부 박사과정
- 관심분야 : 대규모 데이터 분석, 검색 모델 설계

김 성 국(Songkuk Kim)

- 서울대학교 컴퓨터공학과 (공학학사, 공학석사)
- 2005년 : University of Michigan, Computer Science (공학박사)
- 2007년~2011년 : Google U.S. (Software Engineer)
- 2011년～현재 : 연세대학교 공과대학 글로벌융합공학부 교수
- 관심분야 : 네트워크, 대규모 데이터 분석, 클라우드 컴퓨팅