Journal of KIIT. Vol. 16, No. 12, pp. 1-11, Dec. 31, 2018. pISSN 1598-8619, eISSN 2093-7571 1
http://dx.doi.org/10.14801/jkiit.2018.16.12.1

실시간 웹로그 스트림데이터를 이용한 고객행동평가시스템 구현

이한주*, 박홍규**, 이원석***

Implementation of Customer Behavior Evaluation System Using Real-time Web Log Stream Data

Hanjoo Lee*, Hongkyu Park**, and Wonsuk Lee***

"이 논문은 2018년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원과 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2017R1A2B4005344, No.2015-0-00579[빅데이터 환경에서 비식별화 기법을 이용한 개인정보보호 기술 개발])."

요 약

최근 온라인 쇼핑 유통시장의 규모는 지속적이고 빠르게 성장하고 있기 때문에 고객 행동평가분석을 통한 맞춤형 쇼핑서비스가 매우 중요해지고 있다. 하지만 기존의 분석 방식은 소비자의 프로파일 및 행동에 대한 분석 데이터만을 제공하고, 디스크기반 마이닝 탐사로 인해 실시간 분석의 한계가 존재했다. 그러므로 실시간 처리 및 분석이 필요한 웹 서비스와 같은 분야에 기존 방식을 적용하기에는 정확성의 문제와 시스템 성능 문제가 존재한다. 본 연구에서는 실시간으로 발생되는 웹 클릭 로그 스트림을 분석하고 특정 상품에 대한 집중도를 분석하여 상품 구매 의지가 있는 관심고객을 찾아내며, 이를 바탕으로 전체 고객 대상이 아닌 관심고객 중심의 상품 프로모션을 진행할 수 있는 시스템을 구현하고 이들의 효율성과 정확성을 검증한다.

Abstract

Recently, the volume of online shopping market continues to be fast-growing, that is important to provide customized service based on customer behavior evaluation analysis. The existing systems only provide analysis data on the profiles and behaviors of the consumers, and there is a limit to the processing in real time due to disk based mining. There are problems of accuracy and system performance problems to apply existing systems to web services that require real-time processing and analysis. Therefore, The system proposed in this paper analyzes the web click log streams generated in real time to calculate the concentration level of specific products and finds interested customers which are likely to purchase the products, and provides and intensive promotions to interested customers. And we verify the efficiency and accuracy of the proposed system.

Keywords

customer behavior evaluation, web logs, stream mining, big data stream, data analysis, target promotion

* 연세대학교 컴퓨터공학과
- ORCID: https://orcid.org/0000-0001-5192-4973
** 동양미래대학교 컴퓨터소프트웨어공학과
- ORCID: https://orcid.org/0000-0002-6446-9081
*** 연세대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4341-3839

· Received: Sep. 17, 2018, Revised: Dec. 04, 2018, Accepted: Dec. 07, 2018
· Corresponding Author: Wonsuk Lee
Dept of Computer Science, Yonsei University, Seoul, Korea,
Tel. +82-2-2123-2716, Email: leewo2001@gmail.com

Ⅰ. 서 론

최근 생활 수준의 상향 평준화와 인터넷의 발달로 온라인 여행상품 판매 수요의 증가로 웹 로그 분석을 통한 여행상품 시장규모[1]는 2016년 11,890억달러 규모에서 2017년 12,350억 달러규모로 매년 평균적으로 4%의 성장률을 보이고 있다. 무한히 발생되는 데이터 스트림을 모두 분석하여 실시간 상품 추천 마케팅 서비스를 제공하기 위해서는 서비스 제공자와 고객 입장에서 정확한 판단을 할 수 있는 척도를 제공해야 하는데, 소위 정보의 홍수 현상으로 인해 정확한 고객 맞춤 서비스 제공의 어려운 점이 존재한다. 다양한 환경에서 발생되는 로그 데이터에 대한 연구[2][3]는 과거에서부터 현재까지 지속적으로 진행되고 있다. 이러한 상황에서 온라인 상품 구매에 대한 실시간 마케팅 서비스에 관한 연구[4]-[7]는 많이 진행되어오고 있다. 상품 구매와 관련된 고객들의 행동을 이해하고 고객들이 구매할 것인지 확인하는 것은 많은 연구에서 중요한 관심사이기도 하다. 이를 위해 고객의 행동을 파악하기 위해 직접 고객의 온라인 설문조사를 통한 고객 구매 성향 분석을 진행하기도 했고[8], 특정 브랜드나 상품 카테고리를 분석하기위해 해당 고객의 소셜 미디어상에서의 패턴을 분석하기도 했다[9, 10]. 또한 상품의 판매율을 예측하기 위해 C2C형태의 고객 간의 의사교환을 진행하는 온라인 커뮤니티 의사 교환 패턴에 대해서도 분석하여 고객의 행동을 추적하기도 했다[11]. 클릭 스트림을 이용한 사용자의 웹 페이지 방문기록 로그에 존재하는 다양한 속성을 이용하여 고객의 구매를 예측하기도 했다[4, 5]. 하지만, 기존 연구들은 고객의 정확한 행동을 파악 하지 못하고, 수동적으로 참여하는 온라인 설문조사[8]에 의존적이던가, 일부 고객에 대한 분석을 통해 전체 고객에 대하여 일반화된 판단[7]을 하는 한계점이 존재했다. 본 연구에서는 실시간으로 발생되는 웹 클릭 로그 스트림을 이용하여, 특정 고객의 상품 검색 집중도와 실제 구매 내역을 분석하였으며 고객의 특정상품에 대한 집중도를 점수화하여 관심고객을 분류하고, 관심고객으로 분류된 고객들의 행동 분석 결과를 확인할 수 있는 시스템을 제안한다. 본 연구에 사용된 데이터는 2016년 1월부터 2017년 01월까지 수집된 약 63만개의 웹로그 데이터와 약 37만개의 여행상품 조회 로그 및 해당 기간 중 발생한 1930건의 구매 로그를 활용하여, 관심고객을 분류하고, 전체 고객의 구매율과 관심고객의 구매율을 분석하여, 행동 집중도 기반 관심 고객 집계를 통해 전체 고객 대비 더 높은 구매율을 보일 수 있는 점을 확인하였다. 이를 통해 향후 연구에서 상품 구매 임박도가 높은 고객을 선별하여 최종적인 구매율을 향상 시킬 수 있는 객관적인 지표로 활용이 가능할 것으로 예상된다.

본 연구에서 제안하는 방법은 온라인에서 발생되는 웹 클릭 로그 스트림 데이터로부터 현재까지 추출된 고객 행동에 대한 점수를 토대로 상품 조회수 및 행동 집중도를 이용하여 실시간으로 고객 행동을 평가하고, 집중도 점수가 높은 고객들을 구매에 임박한 고객으로 분류하여 해당 관심고객에 대한 프로모션 진행을 통해 효율적인 마케팅을 수행할 수 있다. 분석가 입장에서 고객행동 분류를 위해 여러 개의 점수화 기준을 분류하여 각 단계별 고객의 평가 및 구매가능성을 산출하고, 관심 정보의 실시간 시각화를 통해 분석가가 즉시적인 대응을 할 수 있도록 한다.

Ⅱ. 관련 연구

웹 로그 데이터 마이닝은 데이터 분석가가 유용한 정보를 발견하기 위한 지식 탐사 기술로 다양한 분야에서 연구[12][13]되어 오고 있으며 실제 분석 도구[14]를 활용하여 모바일과 PC에서 발생되는 웹 로그를 분석한 소비자 행동분석 연구[13]도 진행되고 있다. 웹로그 데이터 분석 관련 연구[12]는 웹 페이지 사용자별 행동을 분석하여 사용자의 관심에 따른 웹페이지 최적화 방법을 제시하며, 불필요한 자원 낭비를 최소화하여 검색시간의 단축과 데이터 전송량의 감소 및 데이터 가용성을 향상 시켰다. 때문에 웹로그 분석은 데이터 마이닝 분야에서 중요한 부분을 차지한다.

다른 웹 로그 분석을 통한 고객 행동평가 연구[15]는 발생된 모든 클릭 로그 데이터를 디스크에 적재한 후 분석하는 연구로 디스크 공간에 읽고 쓰는 지연시간의 발생으로 대용량의 데이터가 적재되

어 있는 상황에서 실시간 분석이 어려운 점이 존재했다. 무한히 발생되는 클릭 로그 스트림 환경을 고려하여 해당 데이터에 대한 내재된 지식을 실시간으로 추출하기 위해서는 다음과 같은 조건을 고려해야한다. 첫번째, 데이터 스트림의 각각의 트랜잭션이 발생될 때 단 한번의 스캔을 통해 마이닝 결과를 생성해야 한다. 두번째, 데이터 스트림은 무한히 생성되는 특성을 고려하기 위해 물리적으로 한정된 메모리 공간 내에서 처리해야한다. 세번째, 새롭게 생성된 트랜잭션마다 가능한 빠르게 마이닝 탐사를 수행해야 하고, 해당 스트림에 대한 최신의 결과는 즉시적으로 제공해야한다.

이러한 제한 사항을 충족하려고 하다보면 마이닝 결과에 오차가 포함된다. 사용자의 구매 성향을 파악하기 위해 제안되었던 기존 연구[7]는 클릭 로그 데이터를 활용하여 다양한 유형의 예측 변수를 고려한 구매행동에 영향을 주는 요인들을 고려하여 연구하였고, 방문자의 일반적인 클릭 스트림 행동과 고객 인구 통계 등을 고려한 온라인 구매 행동을 예측하는데 있어서 다양한 범주의 변수를 사용하였지만, 한정적인 데이터 세트의 크기로 인해, 고객 분류의 객관성이 떨어지며, 일부 고객의 행동통계 결과를 이용해 모든 고객들을 일반화 하였다는 한계점이 존재했다.

이러한 한계점으로 인해, 일부 데이터를 이용한 잠재적인 고객에 대한 예측을 수행할 수 있지만, 새로운 고객의 행동 분석이 소매 업체 데이터에 국한되어 있어서 정교한 분석 결과 제공은 어렵다는 점이 존재했다. 상품의 판매율을 예측하기 위해 고객간(C2C)의 커뮤니케이션 행위 추적 및 분석을 위한 연구[11]는 온라인 커뮤니티에 기록된 고객들의 구매 성향을 분석하여, 상품 만족도가 높은 항목들에 대한 순위를 부여한 후, 관련된 다른 상품에 대한 구매 가능성을 예측하여, 평균적인 고객의 구매 결정력을 모델링 하였다. 하지만 실제 구매율이 희박한 여행상품에 대하여 분석하기에는 상품의 만족도가 높다고 여러 개의 여행 상품을 구매하지는 않기 때문에 희박한 구매율을 보이는 여행상품 행동평가 및 구매율 측정에는 한계가 있다.

온라인 쇼핑 고객의 장바구니 품목을 분석하여, 해당 장바구니 품목의 빈도를 활용한 고객 행동 평가 및 상품추천 시스템[16]은 고객의 구매 의도와는 별개로 추가된 상품들로 인해 해당 고객의 정확한 행동을 추적하기 어렵고, 고객의 잠재적인 구매 예정 상품에 대한 예측이 어렵다는 점이 존재 하며, 단순히 온라인 고객의 행동에 대한 일반적인 판단만 할 수 있어서, 본 연구에는 적합하지 않다.

웹 로그 분석을 이용한 실시간 온라인 마케팅 시스템에 관한 연구[6]는 웹 사이트 방문자의 행동을 수집하기 위해 웹 로그 분석을 진행 하였고, 고객의 행동을 분석하여 고객의 성별, 나이, 연령, 직업 등을 고려한 고객 프로파일 기술을 제안하였다. 그러나 전체 고객에 대한 분석을 통하여, 방문 고객을 프로파일 하면서, 실제 구매력이 있는 고객과 없는 고객의 구분이 없고, 과거 방문자의 기준을 최근 24시간 이내로 한정 지어, 구매 빈도가 낮은 온라인 여행상품에 대한 고객 행동 평가 시스템에는 적합하지 않다.

소셜 미디어 활동 기록을 통해 제품의 브랜드별/카테고리별 선호도를 측정한 연구[9][10]도 진행되었지만, 실시간 스트림 분석과 즉시적인 대응이 어렵다는 한계점이 존재하였다. 순차패턴 마이닝을 이용한 연구[15]도 디스크 기반의 데이터베이스를 반복적으로 스캔해야 하는 문제점으로 인해 실시간 대응이 어렵다는 한계점이 존재한다.

III. 실시간 고객 행동평가 시스템

3.1 고객 행동평가 시스템 설계

본 논문에서는 웹 사이트에서 발생되는 클릭 로그 스트림 데이터를 분석하여 고객별 집중도 분석을 통해 분류한 관심고객을 실시간으로 모니터링하고, 분석가가 사전에 정의한 분석 프로파일을 토대로 각 프로파일별 관심고객을 모니터링 하는 시스템을 개발한다. 본 시스템은 클릭 이벤트가 발생될 때마다 추가되는 URL 클릭 로그 스트림과 상품 조회 및 구매 내역에 대한 스트림을 입력하는 단계, 입력된 스트림을 분석하여 각 프로파일별 관심고객을 스코어링 하는 단계, 관심고객을 확인하는 단계로 구성되어 있고 본 연구에서 제안하는 시스템의 구성도는 그림 1과 같다.

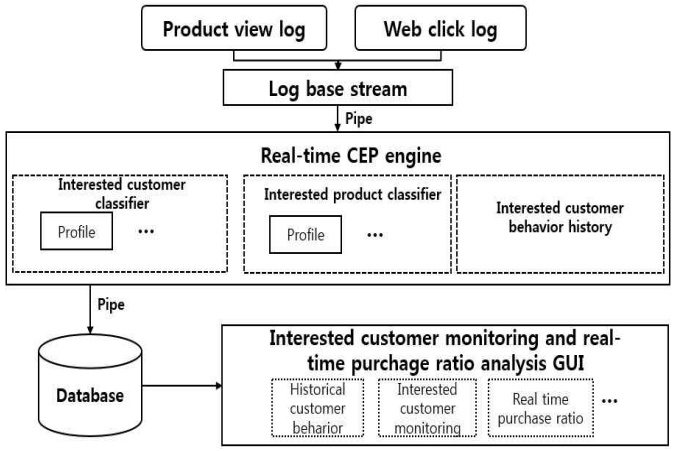

그림 1. 고객 행동평가 시스템 흐름도
Fig. 1. Customer behavior evaluation system

첫번째 단계에서는 웹사이트에서 발생된 URL 클릭 로그 와 상품조회 및 구매내역이 존재하는 스트림을 수집하고, 분석가가 정의한 프로파일에 적절하게 속성 선택과 프로파일별 고객 행동 평가에 맞는 데이터 스키마 형태인 로그 베이스 스트림으로 변경한다. 고객별 집중도 분석 및 관심고객 분류 단계에서는 입력된 로그 베이스 스트림에서 각 프로파일에 맞게 속성들이 선택되고, 스코어링을 진행한다. 관심고객으로 분류된 고객들은 마리아 데이터베이스에 적재되며 이를 활용하여 해당 결과를 실시간 시각화 단계를 수행하여 분석 결과를 확인한다. HTML 파서의 기능은 대상 웹 페이지에 존재하는 상품 페이지 등의 전체 웹 페이지 구조를 크롤링하여, 모든 URL을 반복적으로 추출하여 해당 웹 페이지의 구조를 추출하는 모듈로 추출된 데이터는 향후 GUI에서 확인하여 프로파일 생성 시 필요한 정보 제공에 사용된다. 실시간 CEP(Complex Event Processing) 엔진은 실시간으로 생성되는 로그 베이스 스트림을 분석가가 설정한 프로파일에 맞게 속성을 선택하여 집계를 통해 스코어링 단계를 수행하는 엔진이다.

3.2 고객 행동 웹 로그 데이터

본 연구에서 사용되는 웹 로그 스트림의 종류는 표 1과 같이 URL 클릭 로그 스트림, 상품 조회 로그 스트림, 로그 베이스 스트림으로 구분할 수 있다. 첫번째로 URL 클릭 로그 스트림은 모든 고객이 사이트에 접속하여 발생되는 모든 URL 클릭 로그에 대한 스트림으로 Timestamp, 클릭 시각, 접속IP, 클릭URL 구조로 구성 되어있다. 여기에는 상품을 조회한 내역, 상품을 구매한 내역을 포함하여, 사이트 내에 방문한 모든 기록이 남아있다. 두번째로 상품 조회 로그 스트림은 고객이 조회한 여행 상품에 대한 상세 정보가 존재한다. 상품 조회 로그 스트림에는 Timestamp, 접속IP, 클릭 시각, 여행사코드, 여행상품코드, 출발일, 대륙, 국가, 도시, 항공사, 테마, 가격, 여행일에 대한 정보가 존재한다.

본 연구에 사용된 데이터는 사전에 사용자의 개인정보 보호를 위해, 접속 IP 주소에 대한 기록은 사전에 자동 증가 식별자 매핑 테이블을 이용하여 비식별 처리된 데이터를 수집하였다.

표 1 웹 로그 데이터
Table 1. Web log data

Data	Description
URL click log stream	It generates whenever users click web pages, which includes timestamp, click time, IP address and URL.
Product view log stream	Detail information about travel product. It has user's IP address, click time, product information, price and so on.
Log base stream	The stream data preprocessed by data analyst based on URL click log stream and product view log stream.

그림[2]과 같이 클릭 로그가 새롭게 생성될 때 매핑 테이블 내에 존재하는 IP정보와 비교하여 신규방문자인 경우 자동 증가 식별자 컬럼을 새롭게 만들어 변환된 스트림을 출력한다.

로그 베이스 스트림은 고객 행동평가 시스템의 각 프로파일별 관심고객 분류를 위해 로그 베이스 스트림 생성단계를 수행한다. 실시간으로 발생되는 URL 클릭 로그와 상품조회 로그를 매핑하여 여행상품의 세부정보와 해당 상품의 구매 여부를 확인하는 데이터로 사용되며, 분석가가 사전에 생성한 규칙으로 로그가 분석되기 전 원본 데이터를 의미한다. 로그 베이스 스트림의 상품정보 속성에는 여행사정보, 여행상품코드, 여행 시작일 정보, 대륙, 국가, 도시, 항공사, 테마, 가격, 여행기간 등의 정보가 포함 되어있다.

3.3 실시간 관심고객 및 관심상품 분류

각각의 고객별 전체 행동 로그 조회 수와 각각의 상품별 조회 수를 이용하여, 임계치 이상 조회기록이 있는 고객을 관심고객으로 분류하고, 관심고객으로 분류된 고객은 1개 이상의 관심상품을 가지고 있어야 한다. 각 상품별 행동 집중도 집계 흐름도는 그림 2와 같다.

초기 원본 로그 베이스 스트림을 이용하여, 고객 식별자를 통해 고객들을 분류한 이후, 해당 고객의 전체 상품 조회수와 상품별 조회수를 이용하여 관심고객 집계 분류단계로 수행된다. 행동 집중도 집계결과 X_i에서 i 값은 각각의 고객들을 식별하는 식별자이다. 행동 집중도를 통해 고객이 특정 상품에 대하여 행동 집중도를 계산할 수 있고, 이를 통해 특정상품에 해당 고객이 관심을 보이는지를 측정할 수 있는 척도로 사용될 수 있다.

추후 고객의 상품별 관심도 점수에 임계치를 적용하여, 임계치 이상 특정상품에 관심을 보이는 고객들을 관심고객으로 분류한다. 관심도 최저 임계치는 0.0 ~ 1.0 값을 가지며, 임계치가 0.2인 경우를 예를 들어, i고객의 전체 상품 조회수(Total Product Count(i))가 100일 때, 관련 상품 A, B, C 가 각각 50, 40, 10 인 경우 i고객의 관심상품은 A, B가 선출 되며, 각각의 상품 집중도는 0.5, 0.4 값이 스코어링 된다.

이후 관심고객으로 분류된 고객의 실제 구매 내역 관심상품 리스트를 비교하여, 특정 상품에 관심을 보이는 고객이 해당 상품을 구매하는지 여부를 측정하는 단계를 수행하여 본 연구에서 제안하는 시스템의 구매 정확도를 측정한다.

3.3.1 사전정의

로그 스트림을 분석하여 관심고객을 분류하는 단계에서는 다음과 같은 사전정의가 필요하다. 관심고객 분류를 위한 집계 윈도우 단위는 아래 정의 1과 같으며, 관심고객 분류를 위한 집계범위 임계치 정의와 관심상품에 대한 임계치의 정의는 각각 정의 2와 정의 3과 같다.

정의 1. 관심고객 분류를 위한 집계 윈도우 단위 ITE_{thres}: Inter 임계치(ITE_{thres})는 각 고객별 행동로그가 발생하는 시점에서 활동하지 않는 기간을 정의하기 위해 사용된다. Inter 임계치의 단위는 분 단위 또는 초 단위의 시계열 임계치로 Inter 임계치 이상 활동하지 않았을 때 고객의 활동이 정지된 시점으로 인식하여 해당 고객의 다음 활동이 발생할 때까지 해당 고객의 상품의 집중도 집계 점수 내역을 유지한다. 그림 3과 같이 Inter 임계치 이하인 구간을 고객활성구간, 임계치 이상인 구간, 즉 고객이 행동을 중단한 구간을 고객 비활성 구간이라고 한다.

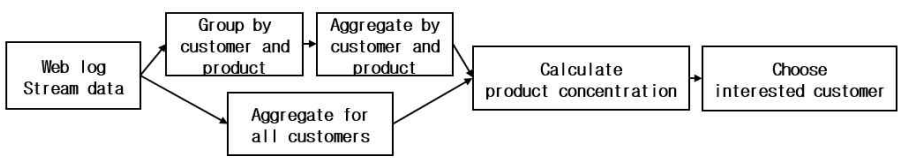

그림 2 행동 집중도기반 관심고객 선정 흐름도
Fig. 2. Flow of interested customer selection based on behavior concentration

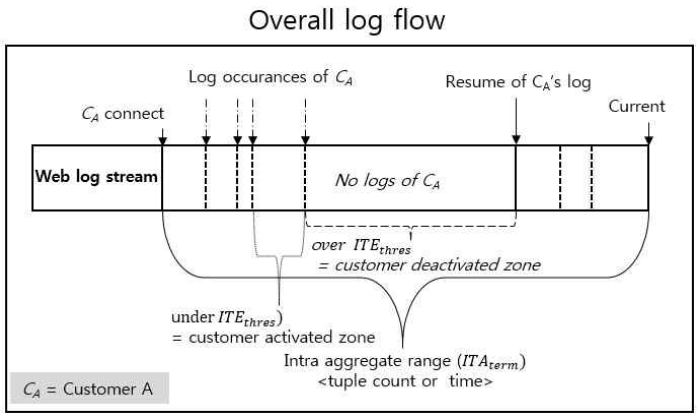

그림 3 웹 로그 스트림 임계치 정의
Fig. 3. Threshold of web log stream

ITA_{term}: Intra 집계범위(ITA_{term})는 일정 기간 이상 전체 고객 대상 집중도 집계 점수 내역을 사용하여 산출한 평균집중도와 누적집중도를 조회함에 있어 그 집계 범위를 나타낼 때 사용된다. 이는 해당 고객이 Inter 임계치 내에 활동한 횟수를 집계하며, Intra 집계범위는 과거의 특정시점 이후에 집계된 집중도의 평균, 누적스코어를 집계할 때 범위를 나타낸다.

정의 2. 관심고객 분류를 위한 집중도 임계치

ITE_{thres} 내에 상품을 클릭한 기록인 고객별 활동 로그를 집계하여 상품의 집중도를 산출하고, 최소 상품 집중도(SP_{thres}) 이상의 상품을 관심 상품으로 분류하고, 1개 이상의 관심 상품을 가진 고객을 관심 고객으로 분류한다.

3.3.2 관심상품 및 관심고객 분류

관심상품과 관심고객을 분류하기 전 불용로그 필터링 단계가 요구된다. 온라인 웹에서 발생되는 로그 스트림에는 본 연구에서 사용될 상품 클릭 로그를 제외한 로그는 사전에 불용로그로 정의하여 그림 4와 같이 2단계에 걸쳐 불용로그를 제거 한다. 첫 번째 단계는 클릭로그의 속성 중 특정 상품 검색 및 구매 기록이 아닌 모든 데이터를 제거하여, 시스템 처리 효율을 증가시킨다.

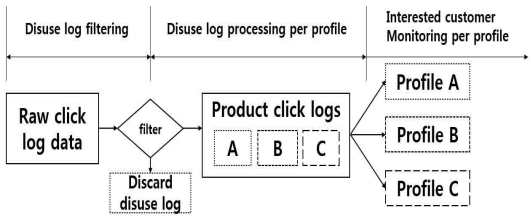

그림 4 단계별 불용로그 처리 흐름도
Fig. 4. Process flow of disuse logs

두 번째 단계는 관심고객 모니터링을 위한 프로파일 설정 조건에 부합하지 않는 상품 정보 조회 및 구매 로그데이터를 의미한다. 불용로그 2단계 전지작업을 통해, 실시간성 확보 및 분석결과 정확도를 향상 시킨다

불용로그가 제거된 로그 베이스 스트림을 생성하여 해당 로그를 활용하여 관심상품과 관심고객으로 분류하기 위해서는 집계 범위내의 로그 스트림에서 관심고객정보와 관심상품을 나누어 각각 분석한다.

상품 관심고객은 SP_{thres} 이상 관심을 보이는 고객을 의미하며 ITE_{thres} 이내의 관심도 점수는 다음과 같이 산출 한다.

$$SP_{i,p,seq} = \frac{PC_p}{TPC_i} \quad (1)$$

i와 p는 각각 고객 식별자와 특정상품을 의미하며 seq는 p 고객의 seq번째 고객 활성구간을 의미한다. PC_p 은 p 상품을 클릭한 로그 스트림의

횟수, TPC_i는 Inter 임계치 이내에서 상품을 클릭한 전체 로그 스트림의 횟수를 의미한다.

집계된 관심도 점수내역은 Inter 집중도 단위별로 데이터베이스에 적재하여 추후 3.3.3장에서 진행하는 누적 집중도 및 평균 집중도를 집계할 때 활용된다. 관심상품 대한 모니터링 흐름도는 로그 베이스 스트림에서 각각의 고객별로 그루핑을 진행한 이후, 각 고객의 상품집중도를 집계 하여 모니터링 GUI 단계에서 각각 관심상품을 확인한다.

그림 5는 세부 상품과 관심 고객을 집계하는 예제이며, 최소 상품 집중도는 0.4으로 설정되다고 가정한다. 그림 5에서 보는 바와 같이, A, B, C 고객의 여행 상품 조회 정보가 존재하며, 고객별로 상품 조회 로그를 분류한 후, 상품집중도를 집계한다. A 고객이 조회한 상품은 아시아/일본과 아시아/태국 여행 상품이며, 각 상품의 상품집중도는 0.5이다.

두 상품 모두 최소 상품 집중도(0.4)보다 크므로, A고객의 관심 상품은 아시아/일본, 아시아/태국이 되며, A고객은 관심 고객으로 분류된다. B 고객은 아시아/일본과 유럽/프랑스 상품을 조회하였지만, 아시아/일본 상품의 상품집중도는 0.33이므로, 유럽/프랑스(0.67)만 관심 상품으로 분류된다. C 고객의 경우에는 3개의 상품을 조회하였지만, 모두 상품집중도가 0.33이므로, 모두 관심상품이 아니며, 그 결과 C고객은 관심고객으로 분류되지 않는다.

이러한 방식을 통해 각 상품별 관심고객을 분류하여, 해당 고객이 어떤 상품에 관심을 보이는지 확인할 수 있고, 추후 실시간 구매율 분석과 매칭하여, 전체 고객대비 관심고객으로 분류된 고객의 구매율을 비교 분석하여 본 시스템의 구매 예측 정확도를 측정한다.

3.3.3 누적 관심도 및 평균 관심도 집계

평균 집중도와 누적 집중도를 집계하는 것은 기존 고객의 경향을 분석하기 위해서 반드시 필요한 요건이다. 집계 단위를 설정 할 때는 Intra 집계범위 파라미터(ITA_{term})를 사용하며, Intra 집계범위 내의 고객 활성구간의 개수 또는 시계열 데이터 기준으로 집계 한다. 기존 각 고객 활성구간별 집중도 $SP_{i,p,seq}$의 전체 합과 평균을 통해 각 i 고객별 누적집중도와 평균 집중도를 아래와 같이 구할 수 있다.

정의 3. 관심상품 누적 집중도 집계

$$i\text{고객 상품 누적 집중도} = \sum_{seq=1}^{ps} SP_{i,p,seq} \quad (2)$$

$(ps = \text{고객 활성구간 개수})$

여기에서 $SP_{i,p,seq}$란 seq번째 고객활성구간에서 i고객의 p 상품의 상품집중도를 의미한다.

정의 4. 관심상품 평균 집중도 집계

평균 집중도는 Intra 집계범위 내 고객 활성구간별 집중도들의 평균을 나타내며, 아래와 같이 산출한다.

$$i\text{고객 평균 집중도} = \frac{\sum_{seq=1}^{ps} SP_{i,p,seq}}{ps} \quad (3)$$

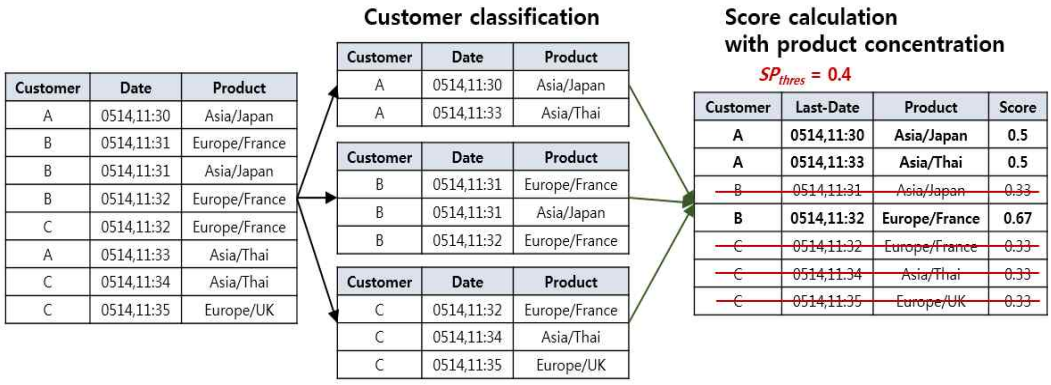

그림 5. 상품 집중도 집계 예제
Fig. 5. Example of product concentration calculation

3.3.4 실시간 구매율 및 상품 집중도 분석

고객별 상품집중도 집계 단계와 관심고객 분류 단계 이후 탐사 된 관심고객목록을 바탕으로 실시간 구매율을 산출 할 수 있다. 비교 분석을 위해 전체 고객에 대한 구매율과 관심고객으로 분류된 고객에 대한 구매율 분석결과를 비교할 수 있다.

본 논문에서는 F-Measure를 통하여 본 연구에서 제안하는 시스템의 정확도를 예측한. F-Measure란 Precision과 Recall의 개념을 통합하여 정확성을 측정하는 지표로 가중치를 가진 조화평균을 구할 수 있다. F-Measure 수식에서 가중치를 Precision과 Recall에 각각 0.5씩 균등하게 반영한 F1-score를 측정한다. Precision은 본 시스템에서 관심고객 중 실제 상품을 구매한 고객의 비율이 대응되고, Recall값은 실제 상품을 구매한 고객 중 관심고객의 비율을 나타낸다.

$$f1-score = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

Ⅳ. 성능 평가

4.1 실험 환경

본 장에서는 실험을 통하여 본 연구에서 제안한 웹 로그 스트림을 이용한 실시간 고객 행동평가 및 상품 추천 시스템의 성능을 확인한다. 실험 환경은 16GB의 메모리를 가진 Intel Core i7-4790 3.40GHz 서버에 리눅스 Centos 7환경에서 진행되었다. 실시간 CEP엔진은 java 환경에서 구현 되어있다.

4.2 실험 데이터셋

실험에 사용된 데이터셋의 정보는 다음 표 2과 같다. 데이터셋 D1은 온라인 여행상품 웹 사이트에서 사용자에 의해 발생된 클릭 이벤트의 URL 주소에 대한 로그 데이터이다. 데이터셋 D2는 온라인 여행상품 웹 사이트에서 사용자가 조회한 여행 상품과 실제 구매한 여행상품에 대한 정보가 있는 데이터이다.

표 2 실험 데이터 집합 정보
Table 2. Experimental data sets

Data	Description	Size	# of tuples
D1	URL click log	65.6 MB	631,791
D2	Product view log	15.3 MB	374,274
D3	Product purchase log	1.5 MB	1,930

데이터셋 D1은 시계열 유닉스 타임스탬프, 클릭시간, 접속 아이피, 클릭한 URL등으로 구성되어있다. 실험 데이터를 통해 구매로그, 상품조회로그, 접속자수 등의 정보를 확인할 수 있으며, 전체 데이터 집합의 접속자 및 상품조회 정보는 표 3과 같다.

실험에 사용된 전체 데이터셋 접속자 행동 통계는 월 평균접속자수 3천1백여명, 실제 상품을 구매한 내역은 월평균 1백6십여명의 데이터 집합을 사용하여 실험을 수행 하였다.

표 3 데이터셋 통계
Table 3. Experimental data statistics

Period	Purchase log	View log	# of users
2016.02~2016.05	551	67,441	4,589
2016.06~2016.09	624	258,964	26,975
2016.10~2017.01	755	52,539	6,196
Total	1,930	374,274	37,760

4.3 성능 평가

상품 집중도 집계 실험에 필요한 실험 변수 조건은 다음과 같다.

표 4 실험 변수
Table 4. Experimental parameters

Variables	Description	Range
SP_{thres}	Product concentration threshold	0.1 ~ 0.5
ITE_{thres}	Inter threshold	300 s
ITA_{term}	Intra-aggregate range	Monthly
$Window\ Size$	window size for aggregate	1,000

상품에 대한 집중도 최소 임계치와 Inter 임계치 및 집계범위와 집계단위를 설정할 수 있으며, 실험 변수 값의 정보는 표 4와 같이, 상품집중도 임계치 값과 고객의 활동기간 단위를 구분할 수 있는 Inter

임계치 단위, 매월 말일 시점에서의 최근 1달간의 집계 범위를 설정하고, 실시간 스트림 환경을 고려하여 1,000개 로그가 입력될 때마다 집중도를 집계하도록 윈도우 크기를 1,000으로 기본 설정하여 실험을 진행 하였다.

그림 6은 월별 집계된 전체 접속고객에 대한 데이터셋 통계정보이다. 8월의 경우, 19,185명이 접속(166,762건 조회)하여 가장 많은 접속량을 기록했고, 12월의 경우 1,124명(10,552건 조회)이 접속하여 가장 적게 접속한 것으로 나타나 있다.

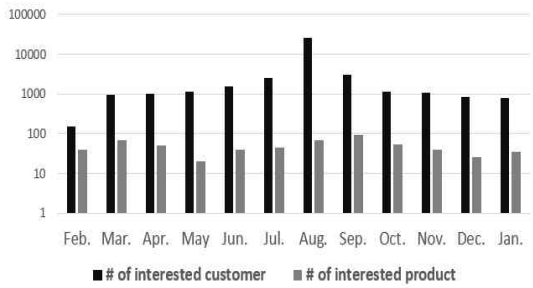

그림 6. 월별 관심상품/고객 결과
Fig. 6. Monthly interested product/customer statistics

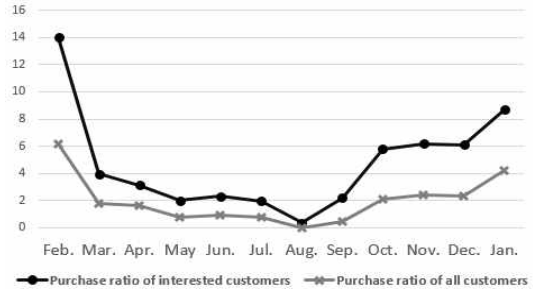

그림 7. 관심고객과 전체고객의 구매율 비교
Fig. 7. Comparison of purchase ratio of interested customers and total customers

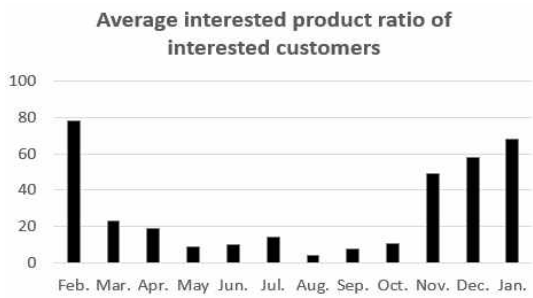

그림 8. 평균 관심고객의 관심상품 비율
Fig. 8. Interested product ratio of interested customers

그림 7은 월별 관심고객의 구매율과 전체 고객의 구매율을 비교한 실험이다. 월별로 구매율의 차이는 존재하지만, 관심고객의 전체 구매율은 5.12%을 나타내었고, 전체 고객의 전체 구매율은 2.06%로, 관심고객의 구매 확률이 약 2.48배 높았다. 마케팅적 관점에서 해당 관심고객들만 집중적인 상품 프로모션과 같은 행위를 수행하여 보다 효율적인 마케팅 지표로 활용이 가능하다.

그림 7에서 보듯이, 월별로 구매율이 매우 상이함을 알 수 있는데, 이를 확인하기 위하여 관심상품 비율 실험을 진행하였으며, 결과는 그림 8과 같다.

11월~2월 집계된 데이터의 경우 관심상품비율이 매우 높게 측정되었는데, 이는 관심고객들이 특정 상품에 지속적인 관심을 보인 것으로 의미하며, 다른 집계 기간에 비해, 적은 종류의 상품을 집중적으로 조회한 것으로 판단할 수 있다. 반면, 그 외 기간에 관심상품비율이 낮은 이유는 고객들의 구매성향이 특정 상품들에 집중하기보다는 여러 상품의 정보를 비교하여 구매하기 때문이며, 이런 경우에는 특정 상품에 관심을 보이는 시기보다 구매율이 떨어짐을 그림 7을 통해 알 수 있다. 이후 실험을 통해 평균적인 관심상품의 비율이 낮은 3~10월 사이와 상대적으로 높게 기록된 11월~2월의 기록에 대한 관심고객 분류의 정확도와 재현율을 통해 본 시스템의 정확성을 측정한다.

그림 9와 그림 10은 각각 3월~10월과 11월~2월의 정밀도, 재현율, F1-스코어를 측정한 결과이다.

관심상품의 비율이 상대적으로 높은 기간(11월~2월)의 정밀도(평균 0.0876)가 비율이 낮은 기간(3월~10월)의 정밀도(평균 0.0294)에 비해 약 3배 높다.

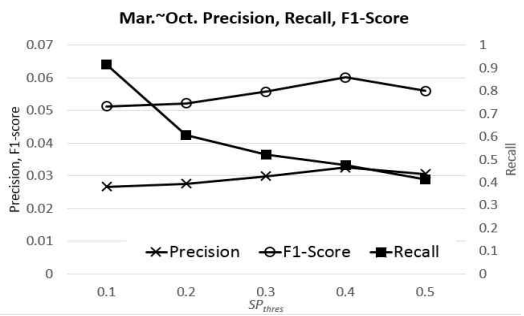

그림 9. 3~10월 정밀도, 재현율, f1스코어
Fig. 9. Precision, recall, f1-score (March~October)

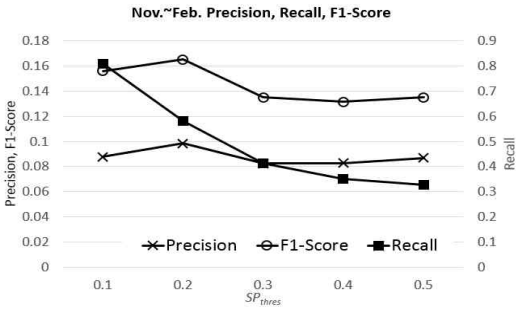

그림 10. 11~2월 정밀도, 재현율, f1스코어
Fig. 10. Precision, recall, f1-score (November~Febuary)

정밀도란 관심고객 중 실제 구매 고객의 비율로써, 정밀도가 높다는 뜻은 구매의사가 있는 고객을 관심고객으로 잘 분류했다는 의미이며, 관심상품 비율이 높은 기간에는 특정 상품에 지속적으로 관심을 가진 고객이 많았기 때문에 보다 정확하게 관심고객을 분류할 수 있었다고 해석할 수 있다.

반면, 상품집중도 임계치(SP_{thres}) 값이 커질수록 관심고객의 수가 줄어들기 때문에, 구매고객 중 관심고객의 비율을 나타내는 재현율은 낮아지는 것을 볼 수 있다. 재현율이 낮다는 의미는 실제 구매의사가 있는 고객임에도 불구하고, 관심고객으로 분류하지 못했음을 의미하므로, SP_{thres}를 너무 높게 설정하면 집중 광고 효과가 떨어질 수 있다.

V. 결론 및 향후 과제

본 연구에서 제안하는 웹 로그 스트림을 이용한 실시간 고객 행동 평가 시스템 구현하였다. 본 시스템은 고객의 여행상품 조회 기록과 여행상품 구매 기록과 같은 웹 로그 데이터를 분석하여, 고객별 상품 집중도 임계치 이상의 관심을 보인 고객을 관심고객으로 분류하였다. 분류된 관심고객과 구매내역 데이터 집합을 활용하여, 월별 관심고객의 여행상품 구매율을 측정하였고, 전체 고객의 구매율 대비 약 2.5배 정도 관심고객의 구매율이 높게 측정되었으며, 재현율 측정을 통해 구매자들 중 관심고객의 비율(약 88%)이 높은 것을 확인하였다. 이를 통해 구매율/구매임박도가 높은 고객을 효과적으로 관심고객으로 분류한 것을 정량적으로 확인하였다. 그러므로 전체고객에게 마케팅을 하는 것이 아니라, 관심고객 대상으로 집중적인 상품 추천을 함으로써 추천 비용을 감소시키고 효과는 극대화하여 마케팅 효율성을 증대시킬 수 있을 것으로 기대된다. 또한 관심상품의 비율이 높은 겨울(11월~2월)에는 관심고객을 효과적으로 분류할 수 있지만, 고객이 특정 상품을 집중적으로 조회하지 않고, 여러 상품들을 산발적으로 조회하여 관심상품 비율이 낮은 시기(3월~10월)에는 상대적으로 관심고객을 효과적으로 분류할 수 없는 한계가 존재한다. 이러한 한계를 극복하기 위하여 사전 데이터셋 정제 작업 연구와 다양한 계층정보와 여러 개의 차원을 활용하여 보다 정확한 상품 구매 예측은 미래 연구로 남아있다.

References

[1] UNWTO World Tourism Barometer Vol. 15, http://www.travelbizmonitor.com/ [accessed: Jan. 05, 2017]

[2] G. C. Lee, S. J. Kwon, and J. K. Kim, "Analysis of User`s Purchasing Pattern on the Web Shopping Mall by Using Web Log Analysis and Fuzzy Cognitive Map Approach", Korea Society of Business Administration Vol. 32, No. 2, pp 567-595, Apr. 2003.

[3] D. H. Lee, S. M. Kim, J. H. Oh, D. R. Seo, and K. G. Im, "User behavior analysis in e-shopping mall with web log mining", Korea Intelligent information system, pp. 305-312, Nov. 2004.

[4] G. Zhu, J. Cao, C. Li, and Z. Wu, "A recommendation engine for travel products based on topic sequential patterns", Multimedia Tools and Applications, Vol. 76, No. 16, pp. 17595-17612, Aug. 2017.

[5] O. Besbes, Y. Gur, and A. Zeevi, "Optimization in Online Content Recommendation Services: Beyond Click-Through Rates", Manufacturing & Service Operations Management, Vol. 18, No. 1, pp. 15-33, Sep. 2015.

[6] J. H. Oh, J. H. Kim, and J. W. Kim, "A Study on the Development of Realtime Online Marketing

System Using Web Log Analytics", Society for e-Business Studies, Vol. 16, No. 3, pp. 249-261, Aug. 2011.

[7] D. Poel and W. Buckinx, "Predicting online purchasing behaviour", European Journal of Operational Research, Vol. 166, No. 2, pp. 557-575, Oct. 2005.

[8] A. K. Kau, Y. E. Tang, and S. Ghose, "Typology of online shoppers", Journal of Consumer Marketing, Vol. 20, No. 2, pp. 139-156, 2003.

[9] Y. Zhang and M. Pennacchiotti, "Recommending branded products from social media", Proceedings of the 7th ACM conference on Recommender systems, 2013 Article, pp. 77-84, Oct. 2013.

[10] Y. Zhang and M. Pennacchiotti, "Predicting purchase behaviors from social media", WWW '13 Proceedings of the 22nd international conference on World Wide Web, pp. 1521-1532, May 2013.

[11] X. Wu and A. Bolivar, "Predicting the conversion probability for items on C2C ecommerce sites", CIKM '09 Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1377-1386, Nov. 2009.

[12] E. Manohar and D. S. Punithavathani, "Hybrid Data Aggregation Technique to Categorize the Web Users to Discover Knowledge About the Web Users", Wireless Personal Communiacations, Vol. 97, No. 4, pp. 5289-5303, Aug. 2017.

[13] R. Orit, G. Anat, and F. Lior, "Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach", Electronic commerce research and applications, Vol. 26, pp 1-12, Sep. 2017.

[14] Google Analytics, https://analytics.google.com [accessed: Feb. 15, 2018]

[15] S. P. Mary and E. Baburaj, "A novel framework for an efficient online recommendation system using constraint based web usage mining techniques", Biomedical Research 2016; Special Issue: S92-S98. Jan. 2016

[16] A. G. Closea and M. Kukar-Kinney, "Beyond buying: Motivations behind consumers' online shopping cart use", Journal of Business Research, Vol. 63, No. 9-10, pp. 986-992, Sep. 2009.

저자소개

이 한 주 (Hanjoo Lee)

2016년 : 가톨릭대학교 컴퓨터공학 학사
2018년 : 연세대학교 컴퓨터과학 석사
2018년 ~ 현재 : (주)레이언스 연구소 연구원
관심분야 : 데이터스트림 마이닝, 사용자패턴분석

박 홍 규 (Hongkyu Park)

2004년 : 연세대학교 정보산업공학과 학사
2007년 : 연세대학교 컴퓨터공학 석사
2012년 : 연세대학교 컴퓨터공학 박사
2012년 ~ 2015년 : 삼성전자 책임연구원
2015년 ~ 2017년 : SK텔레콤 종합기술원
2017년 ~ 현재 : 동양미래대학교 조교수
관심 분야 : 실시간 빅데이터 처리 및 분석, NoSQL

이 원 석 (Wonsuk Lee)

1985년 : Boston University 컴퓨터과학 학사
1987년 : Purdue Uiversity 컴퓨터과학 석사
1990년 : Purdue Uiversity 컴퓨터과학 박사
2004년 ~ 현재 : 연세대학교 컴퓨터과학과 정교수
관심분야 : Sensor Data Stream Processing, Data Stream Mining, OLAP/Data Warehouse