

논문 2020-1-9 <http://dx.doi.org/10.29056/jsav.2020.06.09>

콘텐츠 구매이력과 사용시간을 고려한 연관규칙탐색

이봉규*†

Searching association rules based on purchase history and usage-time of an item

Bong-Kyu Lee*†

요 약

디지털 콘텐츠를 사용자 개인별로 차별화하여 서비스하는 방법들이 다양하게 연구되고 있다. 연관규칙 탐색은 디지털 콘텐츠 서비스에서 각 개인의 선호도를 발견하는데 유용한 방법이다. Apriori 알고리즘은 빈발항목집합을 이용한 연관규칙 알고리즘으로 유용하게 사용되고 있다. 그러나 Apriori 알고리즘은 각 콘텐츠의 참조횟수만을 고려하기 때문에 상용 콘텐츠 서비스에서 나타나는 개인의 실제적인 선호도를 반영하기 어렵다. 본 논문에서는 콘텐츠별 구매이력과 사용시간을 기반으로 연관규칙을 탐색하는 Apriori기반의 알고리즘을 제안한다. 제안하는 알고리즘은 빈발항목을 선택할 때 구매여부에 따른 가중치 값을 가지는 사용시간을 활용한다. 이러한 방법을 통하여 실제 사용자의 정확한 선호도를 파악할 수 있다. 제안된 알고리즘을 구현하고, 실제 콘텐츠 서비스 시스템에서 나타난 실제 데이터를 통하여 성능을 검증한다..

Abstract

Various methods of differentiating and servicing digital content for individual users have been studied. Searching for association rules is a very useful way to discover individual preferences in digital content services. The Apriori algorithm is useful as an association rule extractor using frequent itemsets. However, the Apriori algorithm is not suitable for application to an actual content service because it considers only the reference count of each content. In this paper, we propose a new algorithm based on the Apriori that searches association rules by using purchase history and usage-time for each item. The proposed algorithm utilizes the usage time with the weight value according to purchase items. Thus, it is possible to extract the exact preference of the actual user. We implement the proposed algorithm and verify the performance through the actual data presented in the actual content service system.

한글키워드 : 연관규칙, 디지털 콘텐츠, 사용빈도, 구매이력, 알고리즘

keywords : Association rule, Digital content, Frequency of use, Purchase history, Algorithm

1. 서론

디지털 콘텐츠에는 영상/음원, 애니메이션, 전자출판물 등이 포함되어 있으며, 빅데이터 시대에 맞추어 그 생산 및 사용이 급격히 늘어나고 있다. 이런 흐름에 맞추어 디지털 콘텐츠를 유료로 제공하는 서비스 산업이 활성화되었으며, 사

* 제주대학교 전산통계학과

† 교신저자: 이봉규(email: bklee@jejunu.ac.kr)

접수일자: 2020.05.29. 심사완료: 2020.06.08.

게재확정: 2020.06.19.

용자 역시 급격히 확장되고 있는 분야이다. 이러한 디지털 환경에서 유료서비스 사업자들은 특정 개인에 맞춤형으로 콘텐츠에 대한 홍보 및 서비스를 제공하는 ‘개인맞춤형 서비스’의 제공을 위한 다양한 움직임을 보이고 있다. Customer Relationship Management (CRM) 기법은 ‘개인맞춤형 서비스’를 위한 방법으로 소비자 의 사용패턴을 분석하여 개개인에 적용할 정보를 얻는데 활용된다[1].

CRM은 서비스시스템에 저장된 소비자 개인의 서비스 사용정보를 바탕으로 고객의 선호도를 파악하여 효과적이고 효율적인 마케팅 전략을 개발하는데 사용되는 관리체계를 의미한다. 따라서 CRM은 사용자 별 특성을 분석하여 수요자의 행동을 미리 예측하여 다양한 마케팅 방법과 신규 콘텐츠 홍보 등에 연계시키는 것을 가능하게 한다. CRM을 실제 구현하는 방법으로는 아이템 사이의 연관성을 파악하는 연관규칙 탐색방법과 콘텐츠 그룹 사이에 대한 군집분석이 있다. 이중 연관규칙탐색은 콘텐츠에 대한 실제 사용 정보를 이용한 분석방법으로 사용자의 선호도에 관련된 정보를 얻을 수 있는 방법이다[2].

연관규칙(Association rule) 탐색은 분석 대상 집합인 사용자의 트랜잭션에서 미리 결정되어진 최소 지지도를 만족하는 항목인 빈발 항목집합(frequent itemsets)을 이용하여 항목집합간의 연관성을 찾는 방법이다. 연관규칙 탐색에서는 지지도(support), 신뢰도(confidence), 그리고 향상도(lift)에 근거하여 항목집합 사이의 관련성을 찾는다[3]. 대표적인 연관규칙 알고리즘으로는 항목집합들의 발생 빈도를 사용하는 Apriori 알고리즘이 있다[3]. 이 외에도 해쉬(hash)에 기반한 연관규칙 탐색방법[4]이나 샘플링(Sampling) 기반의 연관규칙 탐색 알고리즘[5] 및 제한된 항목집합을 이용하여 성능을 향상시킨 알고리즘[6] 등이 이 제안되었다. 이들 중에서 Apriori 알고리

즘은 다른 방법에 비하여 탐색의 효율성 및 유용성이 뛰어나고, 단순명료하기 때문에 현재 가장 널리 사용되고 있는 방법이다. 이런 이유로 Apriori 알고리즘을 기반으로 한 다양한 연구가 진행되고 있으며 변형된 알고리즘이 제안되었다[7]. 그러나 지금까지의 Apriori 알고리즘에 기반한 방법들은 빈발 항목집합을 선택할 때 대상의 사용 횟수만을 고려하기 때문에 실질적인 콘텐츠 서비스에서의 사용자에 대한 정확한 선호도를 판단하는데 사용이 어렵다. 예를 들면 유료 콘텐츠를 서비스 받으려는 사용자는 콘텐츠를 구매하기에 앞서, 관련되었다고 판단하는 콘텐츠에 대해서 많은 조회를 하는 경향이 있다. 이러한 행위 이후에 실제 구매가 이루어지는 것이 일반적이다. 이런 점을 고려하여 사용빈도와 더불어 콘텐츠 사용시간을 고려하는 방법들이 제안되었다[8][9]. 제안된 방법은 일정 사용시간이 되지 않는 콘텐츠를 빈발항목에서 제외하는 방법을 사용한다. 그러나 이 방법 역시 사용자의 선호도를 정확히 예측하는 것은 어렵다. 작은 사용시간을 가지더라도 실제 구매가 일어날 경우에는 빈발항목집합에 포함되어야하기 때문이다. 서비스에 대한 실제 구매이력은 다른 어떤 척도보다 명확한 사용자 선호도를 나타낼 수 있는 것이기 때문에 구매이력을 고려한 새로운 연관규칙 탐색방법이 필요한 이유이다.

본 논문에서는 사용자의 콘텐츠에 대한 구매이력과 사용시간을 함께 사용하는 효과적인 연관규칙 탐색 알고리즘을 제안한다. 제안된 알고리즘은 사용시간을 기반으로 빈발항목집합을 구성한다. 그러나 만약 구매가 이루어진 경우는 사용시간에 무관하게 빈발 항목집합으로 선정되게 함으로써 구매여부를 함께 고려하여 사용자 선호도를 판단하는 기법이다. 제안된 알고리즘은 사전 시나리오에 따라 사용된 스트리밍 서비스 시스템에서 얻어진 사용자의 사용 정보를 이용하여 실

험한 후, 사전 시나리오와 비교분석하여 성능을 평가한다.

2. 연관규칙 탐색과 Apriori 알고리즘

연관규칙 탐색은 비목적성 기법으로 항목간의 규칙관계를 설명할 수 최소지지도와 최소신뢰도를 만족하는 연관관계를 찾는 방법이다[10]. 연관규칙에서 사용되는 용어들은 다음과 같이 정의한다. 트랜잭션(transaction)은 규칙탐색 대상이 되는 발생한 독립적인 데이터 단위이다. 항목집합(itemset)은 트랜잭션에 포함된 단일 또는 복수의 데이터에 대한 집합을 나타내며 후보 항목집합(candidate itemset)은 개별 항목집합의 결합을 통해 생성된 항목집합이다. 연관규칙에서 가장 중요한 빈발 항목집합(frequent itemset)은 후보 항목집합에서 최소 지지도/신뢰도를 만족하는 항목집합을 표현한다. 이러한 정의를 이용하여 연관규칙은 다음과 같은 의미로 표현된다. 탐색에 필요한 정보를 표현하는 항목 집합 $I = \{I_1, I_2, \dots, I_k\}$ 가 주어질 경우, 개별 트랜잭션 T 는 고유한 아이디를 가지고 중복이 허용되지 않는 I 의 부분집합으로 정의할 수 있다. 만일 특정 트랜잭션이 항목집합 I 의 부분집합인 X 의 모든 항목들을 포함한다면, 이 트랜잭션은 항목집합 X 를 '지지한다(support)'고 할 수 있다. 항목 X 와 Y 가 각각 $X \subseteq T, Y \subseteq T$ 그리고 $X \cap Y = \emptyset$ 을 만족하는 경우 연관규칙 $X \Rightarrow Y$ 는 $X \Rightarrow Y [Spt = s\%, Conf = c\%]$ 로 정의된다. 이 연관규칙은 항목 X 와 Y 를 함께 포함하는 트랜잭션의 지지도가 $s\%$ 이고, X 가 포함된 트랜잭션 중에서 X 와 Y 가 동시에 포함된 트랜잭션의 신뢰도가 $c\%$ 임을 의미한다. 연관규칙은 지지도와 신뢰도를 계산하여, 지정한 최소지지도 임계치(minimum support threshold)와 최소신뢰도 임

계치 (minimum confidence threshold)를 모두 만족하는 항목 간에 연관관계가 있는 것으로 판단한다.

<ul style="list-style-type: none"> • 지지도(support): $Spt(X \Rightarrow Y) = P(X \cup Y)$ • 신뢰도(confidence): $Conf(X \Rightarrow Y) = P(Y X) = \frac{P(X \cap Y)}{P(X)}$
--

연관규칙탐색은 첫 번째 단계에서는 빈발항목 집합을 구성하고 두 번째 단계에서는 첫 번째 단계1에서 구성된 빈발항목집합을 이용하여 연관규칙을 생성하는 2-패스로 구성된다. 규칙 생성 방법은 모든 빈발항목집합에 대해 빈발항목집합의 모든 공집합이 아닌 부분집합을 찾고, 각각의 부분집합에 대하여 최소 신뢰도를 만족하는 항목을 찾는다. 일반적으로 연관규칙 탐색 알고리즘은 빈발항목을 구성하는 단계1에 의해서 성능 및 정확도가 결정된다. 빈발항목이 잘 결정되면 연관규칙의 정확도가 높아진다.

Apriori 알고리즘은 규칙탐색을 위한 항목집합을 사전(priori) 지식을 적용하여 축소하는 방법으로 연관규칙을 탐색하는 방법이다[3]. Apriori 알고리즘은 $k + 1$ 번째 빈발항목집합을 찾기 위해 k 번째 빈발항목집합을 이용하는 레벨단위의 반복 기법을 이용한다. 이런 레벨단위 반복생성에서 탐색공간을 줄이기 위하여 Apriori 알고리즘은 2가지의 비단조적 (Anti monotonicity) 속성을 가진다. 첫 번째는 빈발항목집합에 대한 모든 하위 집합은 반드시 빈번하다는 가정이다. 두 번째는 빈발항목이 아닌 집합의 모든 상위집합은 빈발항목집합이 아니라는 가정이다. 즉 낮은 단계에서의 빈발항목만을 이용하여 상위의 항목을 추출하는 방법으로 탐색공간을 줄이는 방법이다. Apriori 알고리즘에서는 빈발항목 구성을 위한 사전 지식으로 아이템에 대한 사용빈도를 사용한다. 즉, 아이템 사용빈도를 사전지식으로 이용하여 빈발항목을 구성한다. 그러나 실제 유료 콘텐츠 서비스 관련 응용분야에서는 적은 참조가 이

루어짐에도 불구하고 빈발항목으로 간주해야 할 경우가 많다. 이런 경우 Apriori 알고리즘에서는 최소 지지도를 조정하여 문제를 해결할 수 있다. 그러나 최소지지도를 낮출 경우 탐색해야 할 후보항목집합들이 늘어나 탐색시간이 커지는 단점이 있다. 또한 낮은 최소지지도로 인하여 부적절한 빈발항목집합이 생성될 가능성이 크다. 따라서 문제를 해결하기 위해서는 빈발항목집합의 구성을 위한 사전지식으로 사용할 수 있는 새로운 판단기준이 필요하다.

3. 구매이력 가중치를 가지는 Apriori 알고리즘

Apriori 알고리즘은 방문/검색 횟수가 상대적으로 많은 콘텐츠(아이템)를 규칙 탐색을 위한 빈발항목으로 간주하고 방문/검색 횟수가 적을 경우에는 빈발항목에서 제외한다. 그러나 방문/검색 횟수가 적더라도 빈발항목으로 간주할 필요성이 있기 때문에 상대적으로 적게 방문하는 콘텐츠라 하더라도 사용시간이 긴 경우에는 빈발항목으로 간주하는 방법이 제안되었다[8][9]. 그러나 사용시간 역시 실제 사용자의 정확한 관심을 나타내는 것이라 볼 수 없다. 예를 들면 특정콘텐츠는 다른 콘텐츠를 고르기 위해 일정시간 이상 지속적으로 사용될 수 있기 때문이다. 사용시간이 상대적으로 적어도 실제 구매가 이루어진 경우에는 빈발항목집합으로 간주해야한다. 이에 본 논문에서는 사용시간과 구매이력을 함께 고려하여 빈발항목 선택방법을 제안하고 구현한다.

표 1에 구매이력 가중치를 사용한 제안된 알고리즘이 나타나 있다. 알고리즘의 기본 아이디어는 다음과 같다. 그림 1과 같은 계층구조를 가진 서비스의 경우, 사용자가 (E, I) 콘텐츠를 구매하기 위해서는 (A, B, E, B, C, D, I), (A, B, E, B,

A, D, I) 또는 (A, B, E, B, C, A, D, I) 등의 이동 경로가 나타날 수 있다. 일반적으로 콘텐츠 서비스 시스템은 계층 구조를 가지기 때문에 이런 여러 이동 경로가 발생 가능하고 다른 이동 경로를 보인다. (A, B, E, B, C, D, I) 경우에는 A 콘텐츠에서 E콘텐츠를 방문하기 위해서는 B 콘텐츠를 방문해야 한다. 예에서 보듯이 사용자는 E와 I를 구매하기 위해서 B를 2번 참조하지만 사용시간이 짧고 실제 구매가 이루어지지 않은 경우에는 빈발항목으로 적합하지 않다. E의 경우는 방문 횟수는 적지만 사용시간이 상대적으로 크다면 빈발항목일 가능성이 클 수 있다. I콘텐츠는 사용시간이 E콘텐츠보다 적지만 실제 구매가 이루어졌다면 사용자의 선호도를 가장 잘 알 수 있다. 따라서 사용시간과 구매이력을 함께 고려하는 방법이 다른 방법에 비하여 보다 정확한 사용자의 선호도 판단을 가능하게 할 것이다. 이런 아이디어를 실제 구현하는 방법으로 구매여부는 {yes, no}의 형태의 값으로 표준화하고, 사용시간은 표 2 [9]에 나타나는 것과 같이 표준화한 값을 사용한다. 환산범위는 사용되는 콘텐츠 및 데이터의 특성에 따라서 가변적으로 설정이 가능하므로 응용에 따라 달라질 수 있다.

알고리즘은 다음과 같은 방법으로 실행된다. 로그파일에서 수집한 데이터에서 얻어진 사용자 id, 콘텐츠 이름, 사용시간 및 구매여부를 갖는 트랜잭션을 이용하여 수행한다. step1에서는 실제 사용시간을 방문시간단위로 환산한다. 환산된 시간단위를 포함한 데이터들은 step2에서 사용시간이 0인 모든 트랜잭션을 삭제하는 전처리를 한다. 이를 통하여 사용자 선호도를 판단하는 빈발항목후보 수를 줄여서 효율적인 탐색이 가능토록 한다. 3번째 단계에서는 각 사용자 별 사용시간 단위를 더하여 시간단위합을 구하고, 이 값을 이용하여 각 트랜잭션의 콘텐츠에 대한 사용시간을 정규화 한다. 이 때 만약 구매가 이루어졌을 경

우에는 정규화 값을 1.0으로 하여 최소지지도에 상관없이 빈발항목으로 처리되도록 한다. 3번째 단계에서 구해진 값을 이용하여 연관규칙을 step4에서 추출한다. 각 콘텐츠들에 대한 사용자별 가중 값을 모두 더하여 콘텐츠별 가중치 값의 합을 구한다. 구해진 가중치 값의 합이 최소 지지도보다 큰 콘텐츠의 경우에는 1-연관규칙으로 선택한다. 제안한 방법에 의해서 구매가 이루어진 콘텐츠는 설정된 최소지지도에 무관하게 연관규칙에 포함되도록 하는 것이 제안한 방법의 핵심적인 아이디어이다.

2-연관 규칙은 생성된 1-연관규칙에서 얻어진 빈발항목집합을 이용하여 생성할 수 있다. 먼저 1-연관규칙에서 얻어진 빈발항목간의 연관 가중치 합을 계산한다. 이 때 두 항목 중에서 하나가 0인 경우는 가중치 합을 0으로 간주한다. 이러한 과정을 거치면, 두 개의 콘텐츠간의 연관관계를 보여주는 2-연관규칙이 생성된다. 3-연관 규칙을 추출하는 방법은 2-연관규칙에서 빈발항목집합의 가중치 합과 1-연관규칙에서 얻어진 빈발항목집합 가중치 합을 이용한다.

표 1. 제안된 알고리즘
Table 1. The proposed algorithm

purchase weighted Apriori algorithm	
step 1	로그파일에서 사용자별 정보 수집 (사용자 id, 방문경로, (콘텐츠 id, 사용시간, 구매여부 (yes, no)) 방문시간 단위 = 표준화한 단위 (표 2)
step 2	방문시간 0인 콘텐츠 관련 레코드 삭제
step 3	사용자 별 구매 콘텐츠 가중값 계산 가중값 = 콘텐츠 시간단위/사용자별 시간단위합 if (구매이력 = yes) { 콘텐츠 가중값 = 1.0 (절대 빈발항목) }
step 4	콘텐츠에 대한 모든 사용자 가중치의 합 계산 if (가중값의 합 > 1.0) 가중값의 합 = 1.0 (절대 빈발항목) if(가중값의 합 > 최소지지도) 연관규칙생성

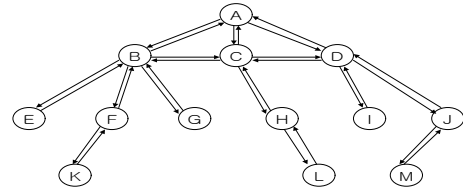


그림 1. 서비스를 위한 방문 순서
Fig. 1. A traversing order for services

표 2. 방문시간단위 환산표
Table 2. The time-translation table

시간단위	유효범위	시간단위	유효범위
1	30초 ~ 1분	6	5분 ~ 6분
2	1분 ~ 2분	7	6분 ~ 7분
3	2분 ~ 3분	8	7분 ~ 8분
4	3분 ~ 4분	9	8분 ~ 9분
5	4분 ~ 5분	10	9분 ~ 10분

4. 실험 및 실험결과

제안된 알고리즘의 유효성을 검증하기 위하여 콘텐츠 서비스 시스템에서 얻은 데이터를 통한 연관규칙 탐색 실험을 한다. 평가를 위한 데이터는 다음과 같은 방법으로 획득한다. 먼저 경제/과학 등 15가지 장르로 구별되는 500개의 오디오북을 스트리밍 서비스하는 시스템을 Ubuntu 16.04.6 LTS 운영체제 기반의 GPU (Graphics Processing Unit)가 설치된 워크스테이션에 구축한다. 구축한 서버를 콘텐츠에 대해서 15명의 실험자들이 미리 계획된 시나리오에 맞추어 콘텐츠를 검색하는 과정을 반복하면서 시나리오에 따라서 특정 콘텐츠에 대해서는 구매의사를 나타내도록 한다. 시나리오는 각 사용자 별로 실제 구매할 아이템과 정해진 선호도에 따른 방문시간 등을 정한 것으로 각 사용자의 선호도에 대한 결과표로써, 실제 탐색되는 연관규칙과 비교하여 알고리즘의 유효성을 판단하는데 사용한다. 스트리

명 서비스 시스템에서는 트랜잭션이 발생할 때마다 사용자 이름 (아이디, ID), 콘텐츠 사용시간, 구매여부와 콘텐츠 이름 (오디오북 이름, 해당 장르)등을 DB에 저장한다. 이런 방법으로 얻어진 15,000개의 트랜잭션 (사용자 당 1000개)을 제안한 알고리즘에 적용하여 연관규칙을 추출한 후 분석한다.

수집된 15,000개의 트랜잭션 데이터를 제안한 알고리즘을 통하여 단일 콘텐츠에 대한 선호도를 알아보는 1-연관규칙을 추출하고, 이를 바탕으로 콘텐츠에 대한 상호연관 선호도를 알아보는 2-연관규칙을 추출하였다. 그리고 1-연관규칙과 2-연관규칙을 이용하여 3개의 콘텐츠의 연관 선호도인 3-연관규칙을 추출하였다. 사용되는 최소 지지도는 0.4이며 추출된 연관 규칙은 표 3에 나타나 있다. 이 추출 결과를 미리 계획한 시나리오에서의 연관패턴과 비교한 결과 사용자 15명에서 대해서 12명이 유사하게 나타났다 (유사도 80%). 이런 결과는 제안 알고리즘이 구매여부를 기준으로 최적의 연관규칙을 탐색하고 있음을 보여준다.

제안 알고리즘을 최소지지도를 변화하면서 추출되는 연관 규칙의 변화를 확인하기 위해 최소 지지도를 변화시키면서 연관규칙을 탐색한다. 사용된 최소 지지도는 0.3과 0.5이다. 0.3의 최소 지지도를 사용한 경우에는 빈발항목이 늘어나기 때문에 알고리즘 수행 시간에 필요한 계산량이 0.4와 비교하여 늘어났다. 또한 0.5를 최소지지도로 사용한 경우는 0.4에서와 비교하여 연관규칙의 수가 적어져 계산량이 줄어드는 결과를 보였다. 그러나 계산량의 증감에도 불구하고 최소 지지도가 변화하더라도 탐색하는 연관규칙의 내용은 유사한 것으로 나타났다. 이러한 결과는 제안 알고리즘이 최소지지도의 변화에도 불구하고 연관규칙을 적절하게 추출할 수 있음을 보여준다.

표 3. 추출된 빈발항목집합
Table 3. The extracted frequent items

규칙 레벨	(L-1) 규칙	추출결과
1-연관	없음	경제, 과학, 소설, 시/에세이, 유아, 인문학, 정치/사회, 제주설화, 클래식, 테스트용 콘텐츠
2-연관	경제	경제-과학, 경제-소설, 경제-시/에세이, 경제-유아, 경제-인문학, 경제-정치/사회, 경제-제주설화, 경제-클래식, 경제-테스트용
	과학	과학-소설, 과학-인문학, 과학-제주설화, 과학-클래식
3-연관	경제-과학	경제-소설-시/에세이, 경제-소설-유아, 경제-소설-인문학, 경제-소설-정치/사회, 경제-소설-제주설화, 경제-소설-클래식, 경제-소설-테스트

비교를 위하여 제안 알고리즘 평가에 사용된 15,000개의 트랜잭션 자료를 기존의 참조횟수 기반의 Apriori 알고리즘에 최소지지도 0.4를 적용한 결과가 표 4에 나타나 있다. 기존의 참조횟수만을 고려하는 알고리즘의 결과를 시나리오와 비교한 결과 실제 시나리오에 의해 기록된 사용자 선호와는 6명이 유사한 것으로 나타났다 (유사도 40%). 이러한 결과는 콘텐츠 사용빈도만을 사전 지식으로 이용하는 기존 Apriori 알고리즘은 사용시간이나 구매이력에 대한 개념을 이용하지 않고 단순히 참조만 빈번하게 발생하는 무의미한 콘텐츠를 빈발항목으로 간주하기 때문에 실질적인 사용자의 선호도를 판단하기 어려운 것이다.

실험을 통하여 사용자의 콘텐츠 구매이력을 고려한 제안 알고리즘이 기존의 단순 참조횟수만을 고려하는 방법보다 우수한 탐색결과를 나타내고 있음을 확인하였다. 이런 결과는 콘텐츠에 대한 사용자 선호를 알 수 있는 가장 중요한 기준인 구매여부를 기반으로 한 기준을 이용함으로써 나타나는 성능 개선으로 판단된다.

표 4. 기존 알고리즘의 결과

Table 4. Results of the original algorithm

규칙레벨	(L-1) 규칙	추출결과
1-연관	없음	소셜, 제주설화, 클래식
2-연관	경제	경제-소셜, 경제-제주설화, 경제-클래식
	소셜	소셜-제주설화, 소셜-클래식
	제주설화	제주설화-클래식
3-연관	소셜-제주설화	경제-제주설화-클래식
	경제-제주설화	소셜-제주설화-클래식

5. 결론

Apriori 기반의 기존 알고리즘은 빈발항목을 선택하는 기준으로 참조 횟수나 사용시간만을 고려하였다. 따라서 적은 참조와 적은 사용시간에도 불구하고 실제 구매가 이루어진 경우와 같이 실제 사용자의 성향을 판단 할 수 있는 요소를 반영하지 못하였다. 그러나 사용자의 선호도를 정확히 예측하고 판단하기 위해서는 실제 구매여부를 반드시 고려해야 한다. 본 논문에서는 유효한 연관규칙을 추출하기 위해 사용자의 실제 구매가 일어난 콘텐츠에 대해 가중치를 부여하는 사전지식을 이용하는 새로운 알고리즘을 통하여 이러한 문제점을 해결하였다. 실험 데이터를 통한 성능평가를 통하여 제안된 방법이 실제 사용자의 선호도를 나타내는 유효한 연관규칙을 추출함을 보였고, 기존 알고리즘과 비교로 성능향상을 확인하였다.

향후 본 연구를 바탕으로 사용자의 선호도를 더 정밀하게 판단할 수 있는 연관규칙 추출을 위한 다양한 방법이 연구될 경우, 디지털 콘텐츠 서비스 분야에 효과적으로 활용이 가능할 것이다.

참고 문헌

- [1] 이상호, 김종배, “캠페인 실행에 영향을 미치는 디지털 마케팅 성과모형 연구”, 한국디지털콘텐츠학회논문지, vol.13, no.2, pp. 205-211, 2012. DOI: <http://dx.doi.org/>
- [2] 황현숙, 박규석, “연관규칙 기반의 상품검색 데이터베이스 최적화 연구”, 한국멀티미디어학회논문지, vol.7, no.2, pp. 145-155, 2004. https://academic.naver.com/article.naver?doc_id=30588026
- [3] R. Agrawal, R. Srikant, “Fast algorithms for mining association rules”, Proceeding of the 20th VLDB Conference, 487-499, Santiago, Chile, 1994. <http://www.vldb.org/conf/1994/P487.PDF>
- [4] J. S. Park, M. S. Chen, S. Y. Philips, “An effective hash-based algorithms for mining association rules”, ACM SIGMOD Record, vol.24, no.2, pp.175-186, 1995. <https://doi.org/10.1145/568271.223813>
- [5] H. Mannila, H. Toivonen, A. I. Verkanmo, “Discovering frequent episodes in sequences”, Data Mining and Knowledge Discovery, vol.1, no.3, pp.259-290, 1997. <http://www.aaai.org/Library/KDD/1995/kdd95-024.php>
- [6] L. Huang, H. Chen, X. Wanf, G. Chen, “A Fast distribution algorithm for mining association rules”, Journal of Computer Science and Technology, vol.15, no.6, pp. 619-624, 2000. https://academic.naver.com/article.naver?doc_id=593065990
- [7] P. T. Kavitha, T. Sasipraba, “Performance Evaluation of Algorithms using a Distributed Data Mining Framework based on Association Rule Mining”, International Journal on Computer Science and Engineering, vol.3, no.12, pp. 3845-3853, 2011. https://academic.naver.com/article.naver?doc_id=81433342
- [8] 김정현, 김재련, “시간을 고려한 연관규칙을 이용한 웹 사용자 접근패턴 분석”, 한국경영

과학회/대한산업공학회 춘계공동학술대회 발표논문집, 852-855, 관동대학교, 양양, 2001. https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE01979990&mark=0&useDate=&bookmarkCnt=0&ipRange=N&language=ko_KR

- [9] 강형창, 양근탁, 김철수, 이윤정, 이봉규, “아이템 사용시간을 고려한 Apriori 알고리즘”, 대한전기학회논문지 D, vol.59, no.7, pp. 1327-1331, 2010. https://www.dbpia.co.kr/pdf/pdfView.do?nodeId=NODE01461451&mark=0&useDate=&bookmarkCnt=1&ipRange=N&language=ko_KR
- [10] 박희창, “연관규칙 마이닝에서 기여 순수 신뢰도의 제안중요지지도를 고려한 연관규칙 탐사 알고리즘”, 한국데이터정보과학회지, vol.22, no.2, pp.235-243, 2011. <http://www.koreascience.kr/journal/JBGHAN/y2003m10a.kr>

저 자 소 개



이봉규 (Bong-Kyu Lee)

1995.2 서울대학교 컴퓨터공학과 박사
1996.3-현재 : 제주대학교 교수
<주관심분야> 인공지능, 패턴인식