사이버쇼핑에서 페이지 체류시간을 고려한 연관규칙 마이닝

조영상*, 김원영*, 김응모*
*성균관대학교 정보통신공학부
e-mail: youngsang.jo@gmail.com

A association data mining on the Cyber-shopping using page duration time

Youngsang Jo*, Won-Young Kim*, Ung-Mo Kim*
*Dept of Information and Communication Engineering,
Sungkyunkwan University

요 약

인터넷의 보급으로 크게 성장한 전자상거래 중에서 사이버쇼핑은 고객과 직접 거래가 가능하다는 특징이 있다. 여기에 연관규칙 데이터마이닝을 적용하여 고객이 필요로 할 물건들을 사전에 보여 줌으로써 고객에게는 쇼핑의 편의성의 높이고 기업에는 판매량 증진을 도모할 수 있다. 또한 고객의 본 물건을 알 수 있는 사이버쇼핑의 특성을 이용하여 페이지체류시간을 고려함으로서 구매에 이르지 않았더라도 흥미를 가지는 패턴을 찾아내 좀 더 많은 마이닝 데이터를 얻을 수 있다.

1. 서론

현대 사회는 인터넷이 보급됨에 따라 엄청난 변화를 겪었고, 또 겪고 있다. 흔히 정보화시대라고 불리는 사회로 진입 하였으며, 인터넷과 연관된 다양한 형태의 서비스들이 나타나고 있다. 전자상거래는 인터넷이 보급되기 이전에도 홈쇼핑 등의 형태로 존재하였다. 그러나 전자상거래는 인터넷이 보급됨에 따라 그 수요가 폭발적으로 증가하여 현재 전자상거래라 함은 인터넷과 연관된 상거래를 나타내고 있다. 국내 전자상거래는 올해 2/4분기 총 거래액이 167조에 이를 정도로 큰 규모를 이루었다. 그 중에 B2B와 B2G를 제외한 사이버쇼핑(B2C) 총 거래액은 4조 8,430억 원으로 전년 동 분기에 비해 11%가 증가 하였고, 소매판매액에서 차지하는 비율은 점차 증가하는 추세이다.[1] 이러한 사이버쇼핑에 연관규칙 마이닝기법을 적용하여 고객이 흥미를 가질만한 상품을 연관규칙 마이닝으로 추출하여, 고객에게 제공함으로써 고객에게는 쇼핑의 편리성 향상과 판매자의 소득 증진이 가능하다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 연관규칙마이닝과 Apriori 알고리즘의 관련연구에 대해 기술한다. 3장에서는 페이지 체류시간을 연관규칙에 사용하는 방법을 제시한 후 마지막으로 4장에서 결론을 맺고자 한다.

2. 관련연구

2.1. 연관규칙 마이닝(Association rule mining)

연관규칙[2]이란 어떠한 사건이 일어나면 다른 사건이 일어나는 관련성을 의미한다. 이는 어떤 데이터 집합에서 빈발패턴(frequent pattern)을 찾는 것을 말한다. 예를 들자면 자동차와 자동차 액세서리를 구입한 고객의 98%가 자동차 서비스를 받았다면 자동차와 자동차 액세서리를 구입한 고객에게 자동차 서비스 광고를 노출시키는 것은 판매 증진에 크게 도움이 된다. 여기서 자동차와 자동차 액세서리를 구입한 고객이 자동차 서비스를 받는 다는 규칙을 발견해 내는 것이 연관규칙 마이닝이다.[3]

$I=\{i_1,i_2,i_3...il\}$ 인 I는 l개의 항목을 가지는 집합이고, 트랜잭션 T는 $T \subseteq I$ 인 항목들의 집합이다. D는 트랜잭션 T들로 이루어진 데이터베이스이다. 각 트랜잭션은 고유한 트랜잭션 번호(TID)를 갖는다. A를 항목들의 집합이라 할 때, 트랜잭션 T가 필요충분조건으로 $A \subseteq T$를 만족하는 경우에 트랜잭션 T가 항목 A를 포함한다고 한다. 여기서 $A \subset I$, $B \subset I$, $A \cap B = \emptyset$을 만족할 때 연관규칙은 $A \Rightarrow B$로 나타낸다. 규칙 $A \Rightarrow B$의 지지도(support) s는 트랜잭션 집합 D에서 집합 A와 B를 동시에 포함하는 트랜잭션의 백분율이고, 신뢰도(confidence) c는 집합 A를 포함하는 트랜잭션 중에서 집합 B도 포함하고 있는 트랜잭션의 백분율을 나타낸다. 즉

$$support(A \Rightarrow B) = P(A \cup B)$$
$$confidence(A \Rightarrow B) = P(A|B)$$

예를 들면, 연관규칙 $A \Rightarrow B$가 20%의 지지도와 60%의 신뢰도를 가지고 있다면 전체 트랜잭션의 20%의 트랜잭션이 A와 B를 포함하고 있고, A를 포함한 트랜잭션의 60%가 B를 포함하고 있다는 말이다.

보통 지지도와 신뢰도가 사용자가 정한 최소지지도(minimum support)와 최소신뢰도(minimum confidence)보다 클 때 빈발항목집합이라고 한다.

2.2 Apriori Algorithm

Apriori 알고리즘[2,3]은 연관규칙 마이닝 알고리즘의 하나로 기본적이며 유용하다. Apriori는 k번째 항목집합이 (k+1)번째 항목집합을 발견하기 위해 사용되는 반복적 접근방법을 사용한다. 더 이상의 (k+1)번째 빈발항목집합이 없을 때까지 반복되는데, 각 단계바다 한 번의 데이터베이스 스캔이 필요하다. 크게 k+1 번째 빈발항목집합의 후보 항목집합을 구하는 결합단계와 그 후보항목집합에서 빈발하지 않은 항목집합을 삭제하는 가지치기 단계로 이루어져있다.

3. 체류시간을 이용한 연관규칙 마이닝

사이버쇼핑에서 어떠한 항목을 고객이 본다는 것은 그 물품에 관심이 있다는 이야기이다. 물품의 구매까지 이르지 않더라도 충분한 시간을 들여 상품에 대한 설명을 읽는 다는 것은 그러한 물품에 관심이 높다고 할 수 있다. 이러한 물품의 설명을 읽는 시간, 즉 페이지 체류시간을 고려하여 연관규칙을 찾아내 관련이 있는 다른 물품을 제시하는 것은 고객의 쇼핑에도 도움이 되고 판매증진에도 도움이 될 것이다. 사이버쇼핑에 있어서 고객이 살펴본 상품과 그 상품이 나와 있는 페이지에 머문 시간을 아는 것은 페이지 진입시간과 페이지 이탈시간을 이용하여 계산할 수 있으므로 용이하다. 한 트랜잭션을 한명의 고객이 살펴본 물품들과 그 페이지에 머문 시간이라고 한다면 <표1> 과 같은 트랜잭션 데이터를 얻을 수 있다.

TID	항목
T100	(I1,66), (I2,12), (I3,120), (I5,87)
T200	(I2,55), (I4,60)
T300	(I1,12), (I2,72), (I3,20), (I4,77), (I5,8)
T400	(I1,57), (I2,67), (I3,B), (I5,54)

<표 1> 체류시간을 포함한 트랜잭션 데이터

각 페이지마다 정보의 양에 따라서 최소체류시간(min_time)을 설정하여 항목에서의 체류시간이 최소체류시간을 만족하지 못한다면 그 항목은 버린다. 또한 너무 많은 시간을 그 페이지에서 머문다는 것은 페이지를 열어두고 다른 용무를 볼 가능성이 높기 때문에 최대체류시간(max_time)을 정해 그 이상의 시간을 가진 항목을 버린다. 고객이 그 항목을 구매한 경우에는 머문 시간과 관계없이 그 물품에 충분한 관심이 있는 것으로 볼 수 있으므로 페이지체류시간을 B로 표기하고 버리지 않는다. <표 1>에서 각각의 최소체류시간과 최대체류시간을 50과 100이라고 할 때 유효한 항목들의 트랜잭션 데이터는 다음과 같다.

TID	항목
T100	I1, I5
T200	I2, I4
T300	I2, I4
T400	I1, I2, I3, I5

<표 2> 체류시간을 고려한 유효 트랜잭션 데이터

이렇게 생성된 유효 트랜잭션 데이터를 가지고 Apriori 알고리즘을 통해 연관규칙을 마이닝 할 수 있다.

4. 결론

본 논문에서는 연관관계 마이닝의 기초가 되는 Apriori 알고리즘에 체류시간을 통한 유효판정을 넣어 마이닝을 하였다. Apriori 알고리즘은 각 단계의 후보 집단을 생성시 데이터베이스를 전체 스캔해야 한다는 단점이 있다. 그러므로 이를 개선하기 위해 FP-tree 알고리즘[4]을 이용하거나 해쉬 트리를 이용한 알고리즘[5]을 사용하는 것이 가능하다.

또한 본 논문에서는 각 페이지마다 최소체류시간과 최대시간이 정해져있어 각각의 페이지마다 최소체류시간과 최대체류시간이 상이할 경우 항목의 유효 여부를 판단하기 위해 페이지 수만큼의 많은 저장 공간을 필요로 한다. 그러므로 향후에 페이지에 들어있는 정보량을 기준으로 최소체류시간과 최대체류시간을 구하는 공식을 발견해 내는 연구가 필요하다.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771).

참고문헌

[1] 통계청 “2009년 2/4분기 및 상반기 전자상거래 및 사이버쇼핑 동향” 통계청

[2] Jiawei Han, Micheline Kamber "Data mining : Concepts and Techniques" 2nd Ed. Morgan Kaufmann 2006

[3] Rakesh Agrawal, Ramakrishnan Srikant "Fast Algorithm for Mining Association Rules" Proc. of the 20th Int'l Conference on Very Large Databases, 1994

[4] Jiawei Han, Jian Pei, Yiwen Yin " Mining Frequent Patterns without Candidate Generation" Proceedings of 2000 ACM SIGMOD Int. Conf. Management of Data(SIGMOD'00)

[5] 이재문, 박종수 “복합 해쉬트리를 이용한 효율적인 연관 규칙 탐사 알고리즘”, 정보과학회 논문지(B) 제26권 제3호