사용자의 피드백을 통한 퍼지 연관규칙의 웹 사용자 마이닝[1]

장재성[0] 오경환
서강대학교
jsjsang@ailab.sogang.ac.kr, kwoh@ccs.sogang.ac.kr

Web Usage Mining Using Fuzzy Association Rule Considering User Feedback

Jae_Sung Jang[0] Kyung-Whan Oh
Sogang University

요 약

데이터 마이닝은 KDD의 분야로서, 의미 있는 정보와 관심 있는 행동 패턴을 추출해 나가는 과정이다. WWW의 발전으로, 웹 데이터가 거대해지고 있다. 이러한 데이터 마이닝 분야에서도, 웹 사용 마이닝의 목적은 의미 있는 사용자 행동 패턴을 찾아내는 것이다. 특히 현재 전자상거래가 널리 활성화되고 있는 환경에서, 사용자의 특성을 발견해내는 것은 매우 중요한 부분이다. 사용자의 특성에 따라 사용자에게 상품을 추천하거나 메일을 보내는 것이나 사용자에게 적절하게 사이트를 구축하는 것이 가능하다. 전처리 과정을 통해서 추출된 트랜잭션 데이터를 모호한 사용자의 요구를 분석할 수 있는 퍼지 집합으로 변형시켜 Fuzzy Association Rule을 통해 분석한다. 그리고 분석된 결과에 대한 규칙을 사용자의 피드백을 통해서 다시 분석하는 과정을 거치게 된다. 사용자의 요구 사항을 적절히 반영할 수 있다.

1. 서 론

거대하고 복잡한 인터넷 환경에서 사용자가 자신이 원하는 정보를 적절하게 얻는다는 것은 쉬운 일이 아니다. 사용자는 정보 자원을 획득할 수 있는 자동화된 검색 기술을 필요로 하게 된다. 이러한 자동화된 기술은 원하는 정보 자료들을 찾는 것을 가능케 하고, 사용자의 패턴을 추적하여 분석하는 것이 가능하여야 한다[1].

사용자의 성향 분석을 위해서 웹 데이터에 데이터 마이닝(Data Mining) 기술을 도입하게 된 것이 웹 마이닝(Web Mining)이다. 이 웹 마이닝 중에서 사용자가 사이트에 접속한 로그 파일을 가지고 결과를 분석하는 것이 웹 사용 마이닝이다. 웹 사용 마이닝을 이용하여 사용자의 행동 패턴을 분석하는 것이 본 논문의 목적이다.

데이터 마이닝을 위한 기법에는 매우 다양한 방법이 사용될 수 있는데 이러한 다양한 데이터 마이닝 기법 중에서도 연관 규칙을 이용하여 웹 사용 마이닝을 한다.

1 본 논문은 뇌 과학 재단의 지원에 이루어졌음

간결한 사용자 마이닝을 위해서 퍼지 규칙과 연관 규칙을 결합한 형태 퍼지 연관 규칙(Fuzzy Association Rule)을 사용한다. 그리고 추출된 연관 규칙을 사용자의 만족도(Feedback)를 입력받는 부분을 이용하여 웹 사용 마이닝의 성능을 향상시킨다. 웹 사용 마이닝의 성능을 향상시키기 위해서 연관 규칙 즉, 패턴 분석의 결과에 사용자의 만족도를 입력받아서 개선하게 된다. 마이닝 결과에 대해서 사용자가 결과에 대한 만족도를 표시하면 이를 받아 들여서 성능을 개선한다. 이러한 성능 개선을 위해서는 일정 수준의 사용자 만족도를 얻을 수 있는 결과를 가질 때까지 학습을 하게 된다. 학습을 통하여 연관 규칙을 추출을 용이하게 할 수 있다.

2. 연구 배경

웹 사용 마이닝을 위해, 사용 분석(Usage Analysis)하는 것이 필요하다. 이를 위해서는 사용자가 얼마나 많이 특정 페이지에 접근하였는가(pageview)와 웹사이트로의 접근 경로(page path)를 분석하여 웹사이트에 대한 정보를 구축한다. 이를 위해서 웹 서버의 로그 파일에서 사용자 세션 파일(user session file)을 추출해야 한다.[2]

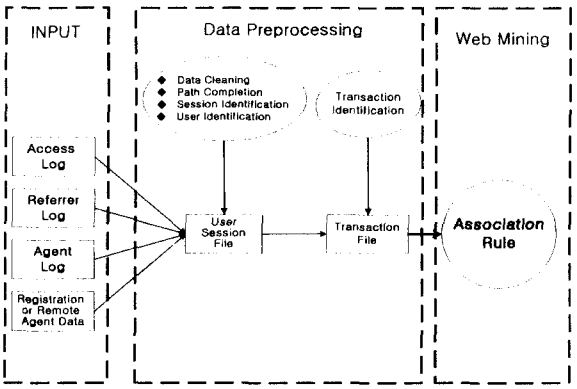

그림1 데이터 전처리

사용자 세션 파일을 추출하기 위해서는 데이터 전처리(Data Preprocessing)가 필요하다. 데이터에 대한 전처리를 하는 방법 중에서 Cooley 등이 사용한 방법으로 사용자 세션 파일을 추출하여 이용한다. 데이터 전처리 과정을 통해 트렌젝션 파일(Transaction File)을 추출한다.

연관 규칙이란 트랜잭션이나 사건에 포함되어 있는 아이템의 경향을 파악해 상호 연관성을 발견하는 것이다. 연관 규칙의 표현은 다음과 같은 식으로 나타내진다. 여기서 X와 Y는 아이템 집합이다.

$$X \Rightarrow Y$$

이러한 규칙은 데이터베이스 내에서 아이템 X를 포함하는 트랜잭션이 아이템 Y 또한 함께 포함하는 경향을 뜻한다. 연관성 규칙을 계산할 수 있는 방법으로는 지지도(Support), 신뢰도(Confidence)가 있다. 신뢰도는 아이템 X를 포함하는 트랜잭션 중에서 아이템 Y가 포함될 확률은 어느 정도인가를 의미한다.

$$C=\frac{P(X\cap Y)}{P(X)}=\frac{\text{항목}X\text{와항목}Y\text{를포함하는거래수}}{\text{항목}X\text{를포함한거래수}}$$

규칙 $X \Rightarrow Y$ 에서 지지도(Support)는 X와 Y를 모두 포함하고 있는 트랜잭션의 비율을 뜻한다.

$$S=P(X\cap Y)=\frac{\text{항목}X\text{와항목}Y\text{를포함하는거래의수}}{\text{전체거래수}N}$$

3. 퍼지 연관 규칙

언어적인 데이터 마이닝을 위해서는 퍼지 집합 이론을 이용하여서 데이터 마이닝을 할 수 있다. 데이터 마이닝 과정에서 모호한 데이터를 처리하는데 있어서 보다 용이하다.

정량 특질(quantitative feature)을 퍼지 집합의 형태로 변환시켜서 연관 규칙을 하는 퍼지 연관 규칙 마이닝(Fuzzy Association Rule Mining)을 생각해 본다.

'If X is A, then Y is B.'[3]

X, Y는 항목(item)이고 A, B는 각각 X, Y를 표현하는 퍼지 집합(Fuzzy Set)이다. 퍼지 집합을 이용하여 표현하면, 이해하기 편한 형태로 나타낼 수 있다.

$I=\{ i_1, i_2, i_3, \ldots, i_n\}$는 아이템의 집합(Itemset)을 의미한다. $T=\{ t_1, t_2, t_3, \ldots, t_n\}$는 데이터베이스를 의미하고, t_i는 T에서의 i번째 튜플(tuple)을 의미하며 정량 어트리뷰트(quantitative attribute)를 나타낸다.

퍼지 집합을 이용한 어트리뷰트 표현을 위해서는 다음과 같은 변수 및 식의 정의가 필요하다.

$F_{i_k}=\{ f^1_{i_k}, f^2_{i_k}, f^3_{i_k}, \ldots, f^l_{i_k}\}$ 은 i_k와 연관된 퍼지 집합을 표현하는 집합 기호이다. 그리고 $f^j_{i_k}$는 F_{i_k}에서의 j번째 퍼지 집합을 의미한다.

트랜잭션 파일을 분석하여, 사용자의 정보 등을 접근 빈도에 따라 위에 제시한 퍼지 집합으로 변환시킨다.

$X=\{ x_1, x_2, \ldots, x_i\}$, $Y=\{ y_1, y_2, y_3, \ldots, y_j\}$의 아이템집합이고 X와 Y는 I의 부분집합이며, 서로 동일한 어트리뷰트를 공유 않는다. $A=\{ f_{x_1}, f_{x_2}, \ldots, f_{x_i}\}$와 $B=\{ f_{y_1}, f_{y_2}, \ldots, f_{y_j}\}$는 각각 X, Y 와 연관된 퍼지 집합을 의미한다.

규칙에서 "X is A"이 만족된다면, "Y is B"도 역시 만족된다. 의미 있는 규칙은 사용자의 최소 요구 신뢰도(MinConf)와 최소 요구 지지도(MinSup)를 만족한다.

퍼지 집합의 중심점을 a_{ij}, $2\le i\le k-1$일 때, 퍼지 집합의 멤버십 함수 다음과 같이 구해진다.

$$\begin{cases} 0 & if\ x \le a_{ij} \\ \dfrac{x-a_{(i-1)j}}{a_{ij}-a_{(i-1)j}} & if\ a_{(i-1)j} < x < a_{ij} \\ 1.0 & if\ x = a_{2j} \\ \dfrac{x-a_{(i+1)j}}{a_{ij}-a_{(i+1)j}} & if\ a_{ij} < x < a_{(i+1)j} \\ 0 & if\ x \ge a(i+1)j \end{cases}$$

위 식으로 추출된 퍼지 집합의 입력 예는 다음과 같다.

Fuzzy Set Label	범위	중심값
Low	0~10	5
Medium	10~100	55
High	100~600	355

표 1 특정 페이지에 머문 시간의 퍼지 집합 예

4. 구현 및 실험

인공지능 연구실 홈페이지에 접속한 다양한 사람의 행동

양식을 분석하고 이에 맞는 명시적인 사용자의 피드백을 요구하여 결과를 분석하고자 한다

명시적으로 피드백을 입력받았을 때의 프로파일 학습을 위한 공식이다.

$$\begin{cases} P = P + f \times U & if\ f \geq \alpha \\ P = 0 & if\ f < \alpha \end{cases}$$

위 식에서 P는 학습된 프로파일, U는 사용자의 관심도, f는 피드백, α는 사용자가 입력한 피드백을 만족하는 임계값(Threshold)이다. 다음이 사용자의 피드백을 고려한 웹 사용자 마이닝의 그림이다.

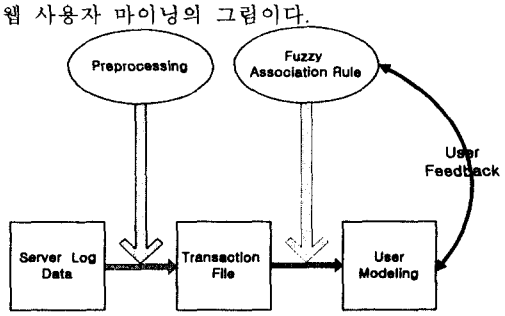

그림 2 사용자 피드백을 고려한 웹 사용 마이닝

2001년 1월 1일부터 3달간 접속한 데이터 로그 파일을 분석하여 구분되는 사용자를 추출하고, 트랜잭션을 클러스터링한다. 사용자 중에서 트랜젝션의 수가 500회를 넘는 사람을 의미 있는 프로파일을 추출할 수 있는 최소의 트랜잭션의 수를 만족한 사람이라 분류하였다.

다음 그림에서 신뢰도를 높게 잡을수록 연관 규칙의 수는 적어진다. 이 과정에서 추출된 규칙이 어느 정도일 때가 가장 적절한가는 정확한 숫자가 정해지는 것은 아니다. 휴리스틱하게 선택이 가능한 것이다.

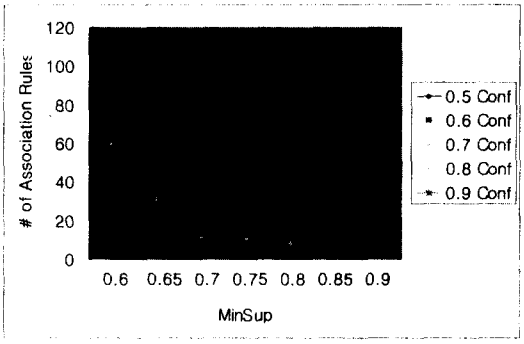

그림 3 MinSup과 추출된 연관 규칙 수의 관계

위의 실험 결과에서 나온 최소 지지도와 최소 신뢰도를 정하여 4월 1일부터 10일간 접속한 사용자를 구분하여 각각의 사용자에게 추출한 연관 규칙의 제시하고 피드백을 입력받았다. 아래 표는 사용자 최소 요구 신뢰도와 사용자 최소 요구 지지도를 각각 0.75%와 80%을 기본으로 해서 추출한 결과이다. 사용자별로 추출된 규칙의 수의 변화가 있다. 기본적으로 규칙의 수는 상당히 줄어들고 있다. 예외적으로 추출된 규칙의 수가 늘어나는 경우는 사용자의 사이트 접속이 늘어나 트랜잭션의 수가 많은 경우이다.

사용자 \ 날짜	4월 1째 주	4월 4째 주	5월 3째 주
A	12 / 14	5 / 6	3 / 3
B	7 / 9	3 / 3	4 / 4
C	4 / 4	2 / 2	8 / 9
D	2 / 3	0 / 0	1 / 1
E	11 / 12	3 / 3	4 / 4
F	2 / 4	3 / 3	3 / 3
G	15 / 16	5 / 5	2 / 2
H	5 / 6	3 / 3	6 / 6

표 2 추출된 규칙에 대한 사용자 만족 규칙의 비율

5. 결론

웹 로그 파일을 전처리하여, 사용자의 트랜잭션을 이진 연관 규칙이나 정량 연관 규칙이 아닌 모호성을 고려한 퍼지 연관 규칙 방법을 사용하여 마이닝하였다. 전처리 과정을 거쳐서 실제적으로 유용한 페이지뷰에 관한 내용을 가중치를 줘서 추출했다. 모호성을 고려하여 퍼지 집합을 이용하였고, 실제적으로 유용한 페이지뷰를 추출하는 과정에서, 휴리스틱한 면이 있기에, 이에 대한 검증을 위해서 사용자의 피드백을 받아 학습하는 웹 사용 마이닝을 제시하였다. 전처리 과정을 거친 트랜잭션 파일을 퍼지 집합으로 변환시켜서, 퍼지 연관 규칙을 적용하였다. 그리고 사용자의 피드백을 입력받는 부분이 있어서, 기존의 방법에 비해서 결과에 대한 분석과 이에 대한 학습의 부분이 첨가되었다. 그 결과 단기적으로는 더 많은 프로세스를 갖는다는 문제점이 있지만, 장기적으로는 더욱 사용자의 의견에 접근할 수 있다는 장점이 있다.

참고문헌

[1] Robert Cooley, Web Mining: Information and Pattern Discovery on the WWW, *ICIA'97* 1997.

[2] Rakesh Agrawal, Mining Generalized Association Rules. *VLDB*, p 407-419, 1995

[10] Kaoru Hirota, Fuzzy Computing for Data Mining, *IEEE, Computer Intelligence* September 1999