웹 사용 마이닝을 위한 퍼지 카테고리 기반의 트랜잭션 분석 기법

Fuzzy category based transaction analysis for web usage mining

이시헌, 이지형
성균관대학교 정보통신공학부 컴퓨터 공학과

Si-Heon Lee, Jee-Hyong Lee
School of ICE, Dept of ECE
Sungkyunkwan University
E-mail : clsoju@skku.edu

요 약

웹 사용 마이닝(Web usage mining)은 웹 로그 파일(web log file)이나 웹 사용 데이타(Web usage data)에서 의미 있는 정보를 찾아내는 연구 분야이다. 웹 사용 마이닝에서 일반적으로 많이 사용하는 웹 로그 파일은 사용자들이 참조한 페이지의 단순한 리스트들이다. 따라서 단순히 웹 로그 파일만을 이용하는 방법만으로는 사용자가 참조했던 페이지의 내용을 반영하여 분석하는데에는 한계가 있다. 이러한 점을 개선하고자 본 논문에서는 페이지 위주가 아닌 웹 페이지가 포함하고 있는 내용(아이템)을 고려하는 새로운 퍼지 카테고리 기반의 웹 사용 마이닝 기법을 제시한다. 또한 사용자를 잘 파악하기 위해서 시간에 따라 관심의 변화를 파악하는 방법을 제시한다.

1. 서론

데이터 마이닝은 대용량의 데이터베이스로부터 기존에 알려지지 않은 즉 단순한 질의어로 추출할 수 없는 형태의 '유용한' 정보를 찾아내고 이를 바탕으로 데이터에 대한 통찰을 얻는 것으로 정의할 수 있다[4]. 웹 마이닝은 웹에서 발생하거나 웹 사이트에 저장한 데이터를 대상으로 유용한 패턴을 찾아내는 것이다. 일반적으로 웹 마이닝은 대상이 되는 웹 데이터에 따라 웹 구조 마이닝, 웹 내용 마이닝, 웹 사용 마이닝으로 나눌 수 있다[5].

이중에서 웹 사용 마이닝(Web usage mining)은 웹 로그 파일(web log file)이나 웹 사용 데이타(Web usage data)에서 의미 있는 정보를 찾아내는 연구 분야이다. 웹 사용 마이닝에서 일반적으로 많이 사용하는 웹 로그 파일은 사용자들이 참조한 페이지의 단순한 리스트들이다. 따라서 단순히 웹 로그 파일만을 이용하는 방법만으로는 사용자가 참조했던 페이지의 내용을 반영하여 분석하는데에는 한계가 있다.

웹 페이지의 내용을 기술하는 가장 좋은 방법은 페이지가 포함하고 있는 아이템 그 자체를 다루는 것이다. 그러나 웹 사이트 마다 다루는 개별 아이템의 개수는 상당히 많을 수 있으므로, 개별 아이템을 다루는 것은 효율적이지 못할 수 있다. 이러한 이유로 본 논문에서는 웹 페이지에 기술된 아이템을 카테고리 별로 구분하여 분석하였다.

또한 사용자를 잘 파악하기 위해서는 시간에 따라 관심의 변화를 잘 파악할 수 있어야 한다. 그러나 기존의 마이닝 방법은 대부분 시간을 제대로 반영하지는 못했다.[1]

본 연구에서는 하나의 웹 페이지 안에 아이템

이 두 개 이상의 카테고리에 포함 될 수 있기 때문에 이를 분석하기 위해서는 퍼지이론을 도입하여 새로운 웹 마이닝 기법을 제시한다. 특히 사용자들이 참조한 시간의 요소를 반영한 웹 마이닝 기법을 제시하였다.

2. 개선된 웹 사용 마이닝

웹 사용 마이닝은 웹 로그에서 불필요한 정보를 제거하고 필요한 정보만을 추출하는 전처리(Preprocessing)와 다양한 로깅 정보에서 한 사용자의 트랜잭션을 규정하는 트랜잭션 검증(Transaction Identification)을 수행한다. 이 과정이 끝나면, 로그 데이터 상에 나타나는 사용자의 구매 패턴을 발견하는 패턴 발견(Pattern Discovery)을 수행하게 되면, 발견된 패턴을 사용 목적에 맞게 분석하는 패턴 분석(Pattern Analysis)을 수행하다[5].

2.1 퍼지 카테고리(Fuzzy Category)

본 연구에서는 페이지 단위의 분석이 아닌 아이템 위주로 분석하기 위해서 퍼지 이론을 도입하였다.

이미 언급하였듯이 각 페이지의 내용을 분석하기 위하여 페이지의 아이템을 카테고리로 구분하였는데, 일반적으로 하나의 아이템은 두 개 이상의 카테고리에 포함 될 수 있기 때문에 퍼지 개념의 카테고리 분석이 필요하다. 우리는 이것을 퍼지 카테고리라고 정의했다. 예를 들어 운동화 아이템은 스포츠 카테고리에 0.8정도 속하고 신발 카테고리에 0.2만큼 속한다고 할 수 있다. [표는 1]은 각 페이지의 퍼지 카테고리 소속도를 나타낸 표이다.

	Category1	Category2	Category3	Category4	Category5
A	0.1	0	0	0.3	0.6
B	0.4	0	0	0.1	0.5
C	0.1	0	0.3	0.1	0.5
D	0	0	0.4	0.6	0
E	0	0	1	0	0
F	0.4	0.6	0	0	0
G	0	0	0	0.2	0.8
H	0	0	0	0.9	0.1
I	0.2	0.8	0	0	0
J	0.8	0.2	0	0	0
K	0	0	0	0.1	0.9

[표 1] 각 페이지의 카테고리 소속도

본 장에서는, 어느 웹 사이트에 아래와 같이 A-K까지 11개의 페이지가 있고, 각 페이지에 포함된 아이템은 크게 5개의 카테고리로 나눌 수 있다고 하고, 각 페이지의 내용이 각 카테고리에 속하는 정도는 [표 1]과 같다고 가정한다. 예를 들면, 페이지 A에 속하는 아이템들이 각 카테고리에 속하는 정도는 카테고리1에는 0.1, 카테고리4에는 0.3, 카테고리5에는 0.6이다.

2.2 퍼지 지지도(Fuzzy Support)

사용자를 파악하기 위해서는 사용자가 어떤 카테고리 아이템에 흥미가 있는지를 파악하여야 한다. 사용자가 자주 보는 페이지에 있는 아이템에 사용자는 흥미를 갖고 있다고 판단할 수 있다. 따라서 사용자를 파악하기 위해서는 사용자가 어느 카테고리 아이템을 많이 보았는지를 파악하여야 한다. 사용자가 어느 카테고리 아이템을 어느 정도로 자주 보았는가를 측정한 값을 지지도(Support)라고 하는데, 지지도는 웹 마이닝에서 가장 기본적으로 사용되는 것이다.

기존의 웹 마이닝 방법에서의 지지도를 구할 때 웹 페이지 방문 시각은 고려하지 않고 단지 몇 번 방문하였는가를 위주로 구하였다. 그러나 사용자의 관심은 시간에 따라 변하므로 사용자의 흥미를 잘 반영하기 위해서는 방문한 시각도 고려하여야 한다. 본 연구에서는 이를 보완하기 위해서 사용자들이 참조한 시간의 요소를 반영하여 지지도를 구한다.

카테고리의 카운터와 지지도를 구하는 수식을 [수식 1]과 [수식 2]와 [수식 3]을 정의하였다.

$$Count(C) = \sum_{t=1}^{T} \mu_{T(t)}(C) \quad \cdots \text{ [수식1]}$$

$$Support(C) = \frac{Count(C)}{\sum_{t=1}^{T} t - \sum_{t=1}^{T} t * \mu_{T(t)}(C) + T} \quad \cdots \text{ [수식2]}$$

$$\mu_{T(t)}(C) = \frac{\sum_{P \in T(t)} \mu_C(P)}{T(t)에 포함된 페이지} \quad \cdots \text{ [수식3]}$$

이 수식은 트랜잭션의 순서에 따라 가장 최근에 참조한 페이지에 더 높은 가중치를 둔 식이다. 여기서 T는 트랜잭션의 개수이고, 는 트랜잭션 t에 포함된 페이지의 카테고리의 소속도를 나타낸다. 트랜잭션 t에 대한 카테고리 소속도를 구할 때 페이지 개수로 나눈 이유는 소속도 합을 1로 정규화 시키기 위해서이다. 따라서 $\mu_{T(t)}(C)$는 트랜잭션에 포함된 페이지들의 카테고리 소속도의 합을 트랜잭션에 있는 총 페이지를 나눈 값이 된다.

예를 들어 어떤 사용자가 웹 사이트에 5번 방문하여 {A,G,K}, {B,C,G}, {C,F,G,H}, {D,G,H} 트랜잭션을 만들었다고 하자. 즉, 첫 방문에서는 페

이지 A, G, K를 방문하였고, 두 번째는 B, C, G를 방문하였다는 것을 기록한 것이다. 이 때 [표 1]의 각 페이지의 카테고리 소속도를 사용하여 5개의 카테고리의 카운터와 지지도를 구하면 [표 2]와 같다.

Category	Count	Support
1	0.33	2.46%
2	0.15	1.11%
3	0.30	2.36%
4	1.20	11.58%
5	2.02	20.61%

[표 2] 카테고리의 카운터와 지지도

그리고 최소 카운트(Minimal Count)와 최소 지지도(Minimal support)의 문턱값(Threshold)으로 정의하여, 최소 카운트 값 보다 크고, 최소 지지도 보다 큰 카테고리에 대해서 사용자가 관심을 갖고 있다고 가정하고 최소 카운트와 최소 지지도를 만족하지 않는 카테고리에 대해서는 사용자가 관심을 가지지 않는다고 가정한다.

2.3 퍼지 지지도의 속성

정의된 [수식 1]은 다음과 같은 3가지 속성을 가진다.

첫째, 모든 트랜잭션에 카테고리가 속한 페이지가 없으면 지지도는 0이고, 모든 트랜잭션에 카테고리 소속도가 1이면 지지도는 1이다.

속성1. $\forall$Tn 에 대해서 $\mu_{T(n)}(C)=0$이면
Support(C)=0 이다.
$\forall$Tn 에 대해서 $\mu_{T(n)}(C)=1$이면
Support(C)=1 이다.

둘째, 예전에 참조한 페이지의 카테고리 보다 최근에 참조한 페이지의 카테고리의 지지도가 더 높다.

속성2. T1={C1}, T2={C2}이고,
$\mu_{T1}(C_1)=\mu_{T2}(C_2)$ 이면
Support(C_1) < Support(C_2) 이다.

증명1. Count(C_1) = Count(C_2) 이고,
$\sum^{T} t*\mu_{T(n)}(C_2) > \sum^{T} t*\mu_{T(n)}(C_1)$ 이기 때문에
Support(C_2) - Support(C_1) > 0 이다.
따라서 Support(C_1) < Support(C_2)이다.

셋째, 많이 참조된 카테고리일수록 사용자의 관심이 많은 것이므로, 많이 참조된 카테고리가 시간의 영향을 적게 받는다. 즉,

속성3.

Transaction group1	Transaction group2
$T_1 = \{C_i\}$	$T_1 = \{C_j\}$
$T_2 = \{C_i\}$	$T_2 = \{C_j\}$
⋮	⋮
$T_m = \{C_i\}$	$T_n = \{C_j\}$

m<n 이면 Support(C_i)<Support(C_j)이다.

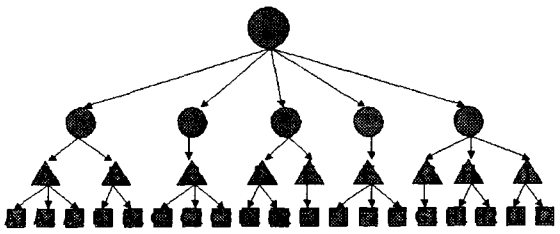
증명2. m<n 이고, $\mu_T(C_j)=\mu_T(C_i)$이므로

$$(\sum_{t=1}^{n} t-\sum_{t=1}^{n} t*\mu_{T(n)}(C_j)+n)-(\sum_{t=1}^{m} t-\sum_{t=1}^{m} t*\mu_{T(n)}(C_i)+m)>0$$

이고, Count(C_i) < Count(C_j) 이므로
Support(C_i) < Support(C_j)이다.

이러한 분석을 통하여 본 논문에서 제시한 방법을 통한 분석이 시간을 적절히 반영하고 있는 것을 알 수 있다.

3. 실험



[그림 1] Sample Web page

[그림 1]은 실험에서 사용 할 간단한 웹 페이지의 구조를 나타낸 그림이다. 퍼지 지지도의 시간 요소에 대한 영향과 기존 방법과의 비교를 하기 위해서 카테고리 9개를 정하여 각 웹 페이지마다 퍼지 카테고리 소속도 [표 3]과 같이 정의하여 실험을 하였다. 각 서브페이지(A1,A2, ... , I2)는 상위 페이지(A,B, ... ,I)와 같게 설정하였다.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A	0.7	0.3	0	0	0	0	0	0	0
B	0.3	0.7	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	0.8	0.2	0	0	0	0
E	0	0	0	0.2	0.8	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	0.6	0.3	0.1
H	0	0	0	0	0	0	0.2	0.7	0.1
I	0	0	0	0	0	0	0.1	0.1	0.8

[표 3] 각 페이지의 카테고리 소속도

우선 사용자1, 2, 3이 [표 4]와 같이 각각 9개의 트랜잭션을 만들었다고 하자. 사용자 1과 사용자 2는 각각의 카테고리마다 카운트는 같지만 트랜잭션 순서를 반대로 한 트랜잭션이고 [그림 2]와 [그림 3]은 9개의 카테고리의 지지도(Support)를 나타낸 것이다. [그림 2]와 [그림 3]

에서 나타나듯이 시간에 따라 가장 최근에 참조한 페이지의 카테고리가 높게 나타났다. 또한 사용자가 참조한 페이지가 각각 1번씩 밖에 없으므로 지지도의 차이는 많이 나지 않는다는 것을 알 수 있다.

사용자 3은 처음에는 카테고리 3에 관심이 있었지만, 최근에는 카테고리 6에 더 관심이 있는 사용자이다. [그림 4]는 기존의 방법으로 구한 각 카테고리의 지지도를 나타내고, [그림 5]는 제안한 퍼지 카테고리 기반의 퍼지 지지도를 나타낸 그림이다.

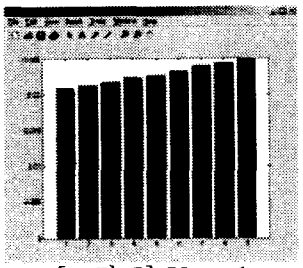
기존의 지지도를 구하는 방법은 다음과 같다.

$$Support = \frac{X\text{를포함하고있는트랜잭션수}}{\text{전체트랜잭션수}}$$

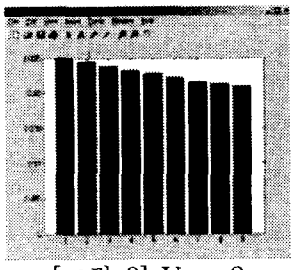
기존에 방법에서는 단지 페이지의 카운트만 반영하기 때문에 카테고리 3이 더 높게 나타났지만, 제한한 방법으로는 카테고리 6이 더 높게 나타났다. 사용자 3은 현재 카테고리 3에 더 많은 관심이 있는 사용자이므로 우리가 제한한 방법이 사용자를 더 잘 반영한다는 것을 실험을 통해서 알 수 있었다.

	User 1	User 2	User 3
T1	{A,A1,A2,A3}	{I,I1,I2}	{C,C1,C2}
T2	{B,B1,B2}	{H,H1,H2}	{C,C2,C1}
T3	{C,C1,C2,C3}	{G,G1}	{C,C1,C3}
T4	{D,D1,D2}	{F,F1,F2,F3}	{C,C2,C3}
T5	{E,E1}	{E,E1}	{C,C3,C2}
T6	{F,F1,F2,F3}	{D,D1,D2}	{F,F1,F2}
T7	{G,G1}	{C,C1,C2,C3}	{F,F1,F3}
T8	{H,H1,H2}	{B,B1,B2}	{F,F2,F3}
T9	{I,I1,I2}	{A,A1,A2,A3}	{F,F3,F2}

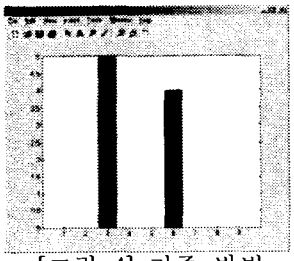
[표 4] 사용자 3명의 트랜잭션



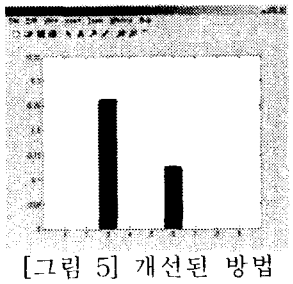
[그림 2] User 1



[그림 3] User 2



[그림 4] 기존 방법



[그림 5] 개선된 방법

4. 결론 및 향후과제

본 연구는 웹 마이닝 기법에서 퍼지 카테고리 중심의 새로운 트랜잭션 분석 방법을 제시하였다. 이를 위해서 퍼지 카테고리를 정의하고, 트랜잭션의 시간의 요소를 반영한 퍼지 지지도와 속성을 정의하였으며, 이를 증명하였다. 또한 실험을 통하여 퍼지 지지도의 시간의 요소에 대한 영향과 기존 방법과의 비교를 하였다. 따라서 우리가 제시한 방법은 페이지의 내용과 사용자들이 페이지를 참조한 시간을 고려할 수 있었다. 본 연구에서는 사용자가 관심이 있는 카테고리까지만 분석하였지만, 사용자의 관심을 더 자세히 파악하기 위해서는 카테고리 사이의 연관성을 파악하기 위해 2개 이상의 카테고리도 분석할 필요가 있다. 그리고 분석 결과를 토대로 실질적으로 사용자 모델링(User Modeling)에 적용시킬 수 있는 차후 연구가 필요하다.

5. 참고문헌

[1] Yi-Hung, Yong-Chuan Chen, Arbee L.P.Chen "Enabling Personalized Recommendation on the Web Based on User Interests and Behaviors", Proc. of 11th International Workshop , IEEE, 2001.

[2] A Gyenesei, "A Fuzzy Approach for Mining Quantitative Association Rules", TUCS Technical Reports no, 336, 2000.

[3] Jae-Sung Jang, Sung-Hae Jun , Kyung-Whan Oh, "Fuzzy Web Usage Mining for User Modeling" International Journal of Fuzzy Logic and Intelligent Systems, vol. 2, no. 3, pp204-209, 2002.

[4] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information System, vol. 1, no. 1, pp8-19, 1999

[5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, "Web mining : Information and Pattern Discovery on the World Wide Web", proc of the 9th IEEE International Conf. on Tools with Artificial Intelligence, pp61-62, 1997.