2005 Proceedings of the Spring Conference, Korean Statistical Society

웹 페이지 방문 시간을 고려한 연관 규칙 탐색

강 형 창[1], 김 익 찬, 김 철 수[2]

요 약

웹 사이트를 이용하는 사용자들은 정보를 편리하게 얻고자 한다. 웹 사이트 운영자들은 웹 사이트를 이용하는 사용자들에게 차별화된 서비스를 제공하기 위해 사용자에 따른 패턴 분석을 해야 한다.

연관 규칙은 패턴 발견을 위해 데이터 마이닝 기법중의 하나이다. 사용자에 따른 패턴을 찾아내면, 사용자에 따른 차별화된 서비스를 제공할 수 있다. 사용자에 따른 패턴은 연관 규칙 탐색으로 알 수 있고, 웹 페이지 방문 시간을 고려한 연관 규칙 탐색 결과는 차별화된 웹 구조 서비스 및 추천 서비스가 가능하다.

주요용어 : 연관 규칙, 패턴 발견, 웹 페이지 방문 시간

1. 서론

인터넷 기술의 발달과 사용 환경의 편리함으로 많은 사람들이 인터넷을 이용하고 있다. 웹 사이트와 사용자와의 효과적인 커뮤니케이션이 이루어지지 않으면 사용자들의 호응을 얻기 힘들 것이다. 그러므로 사용자의 흥미에 맞는 정보를 제공한다면 좋은 호응을 얻을 수 있을 것이다. 대부분의 웹 사이트는 사용자의 단계적인 클릭행위를 통해 접근하는 방식으로 운영되고 있기 때문에 고객별로 개인화되지 않은 일률적 정보를 제공함으로써 사용자에게 필요로 하는 정보 탐색에 있어 많은 시간과 노력을 허비하게 된다. 이는 결국 웹 사이트를 이용하는 사용자의 전반적인 만족도 하락을 초래한다.

웹 사이트에서 사용자에게 알맞은 정보를 제공하는 전략을 세우기 위해서는 사용자 개개인의 행동 패턴에 대한 정보가 필요하다. 이와 같은 정보를 기반으로 사용자 개개인의 특성에 맞은 동적인 웹 페이지 구성이나 링크정보를 제공할 수 있다. 사용자의 정보와 행동 패턴을 분석하고 이를 활용하여 사용자의 다음 행동이나 상품 추천 등에 대한 정보를 제공하는 것을 개인화라 한다.

인터넷 사용자가 웹 사이트를 방문하면 웹 서버에는 사용자가 요청한 서비스나 방문 페이지, 방문 시간 등에 대한 정보를 파일 형태로 저장하게 되는데 이를 웹 로그 파일이라 한다. 웹 사이트 운영자는 웹 로그 파일을 이용하여 웹 사이트 사용자들의 행동 패턴을 분석하여 유용한 정보를 얻어내기 위해 데이터 마이닝을 웹에 적용하게 되었고, 이를 웹 마이닝이라 한다.

웹 마이닝은 웹 사이트 개선, 광고 효과 측정 또는 상품 추천 등에 응용할 수 있다. 웹 로그 파일은 파일 구조로 인해 전처리 과정이 필요하고, 이를 통해 연관 규칙, 순차 패턴, 군집화, 분류 등의 데이터 마이닝 기법을 이용하여 사용자들의 행동 패턴을 발견하고, 패턴 발견 과정에서 유용하고 의미 있는 규칙과 패턴을 찾아낸 후 패턴을 분석하는 과정으로 구성된다.

마이닝 기법 중 연관 규칙은 데이터들 사이에 숨겨져 있는 패턴을 탐색하는 기법이다. 연관 규칙은 빈발 항목집합을 찾아내고, 이들로부터 연관 규칙을 생성한다. 빈발 항목집합을 발견하

1) 제주대학교 전산통계학과 대학원
2) 제주대학교 전산통계학과 교수

는 대표적인 알고리즘으로 알려져 있는 Apriori 알고리즘은 데이터베이스에 저장되어 있는 트랜잭션을 단계마다 계속적으로 스캔해야 하는 문제를 가지고 있다. 따라서 효율적인 빈발 항목집합을 찾기 위한 다양한 알고리즘들에 대한 연구가 진행되고 있다.

본 논문에서는 효율적인 빈발 항목집합 탐색과 개인화를 위해 사용자가 웹 사이트에 접속하여 웹 페이지를 방문한 시간을 고려하여 사용자 접근 패턴을 분석하고자 한다.

2. 관련연구

2.1. 웹 로그 파일과 웹 마이닝

웹 로그 파일은 사용자가 웹 사이트 접속할 때마다 요청 시간, 요청 페이지 등과 같은 정보를 웹 서버에 파일 형태로 저장되는 정보이다. 웹 로그는 웹 마이닝 분야에서 가장 많이 사용되는 정보로 사용자들의 행동 패턴을 분석하기 위한 정보들이 포함되어 있다. 하지만 실제 웹 로그 파일에는 이미지 정보나 스크립트 정보와 같은 불필요한 정보가 많이 포함되어 이를 제거하기 위한 전처리 과정이 필요하다.

다음 [그림 1]은 CLF(Common Log Format)의 구조이다. [그림 1]에서처럼 특정 IP를 가진 사용자가 페이지끼리 이동한 경로를 추출하여 웹 사이트의 구조적인 정보를 유추할 수 있으며, 사용자의 행동 패턴 분석이 가능하다. 이러한 웹 로그 파일은 웹 로그를 수집하는 수준에 따라 클라이언트 수준, 프록시 수준, 서버 수준으로 구분할 수 있는데 분석을 위하여 사용되고 있는 대부분의 웹 로그 데이터는 웹 서버 수준에서 얻어지는 것이다.

[그림 1] CLF(Common Log Format) 구조

```
203.253.221.147 - - [21/Aug/2003:18:52:59 +0900] "GET /index.html HTTP/1.1" 200
1180
203.253.221.136 - - [21/Aug/2003:18:56:32 +0900] "GET /index.html HTTP/1.1" 200
1180
203.253.221.147 - - [21/Aug/2003:18:57:29 +0900] "GET /spss/SPSS_05.hwp
HTTP/1.1" 200 499073
203.253.221.136 - - [21/Aug/2003:18:57:41 +0900] "GET /spss/07.zip HTTP/1.1"
200 1175
```

웹 마이닝은 분석 대상의 유형에 따라 웹 구조 마이닝(web structure mining), 웹 내용 마이닝(web content mining), 웹 사용 마이닝(web usage mining)으로 구분할 수 있다. 웹 구조 마이닝은 웹 내용을 기술하는데 사용하는 구조화된 정보를 분석하는 과정으로서 하이퍼텍스트로 구성된 문서들의 구조에 대해 마이닝하는 것이다. 웹 내용 마이닝은 웹 사이트를 구성하는 페이지 내용 중에서 텍스트, 이미지, 오디오, 동영상 등과 같은 다양한 데이터로부터 얻어지는 내용에 대한 분석을 통하여 웹 사이트에 대한 유용한 정보를 찾아내는 과정이다. 마지막으로 웹 사용 마이닝은 웹 서버에 저장된 웹 로그 파일을 이용하여 사용자들의 행동 패턴에 대한 정보를 분석하는 과정이다. 웹 사용 마이닝은 웹 사이트에서 사용자들의 웹 페이지 사용 패턴을 분석하고, 사용자가 웹 서핑하면서 발생되는 로그 데이터와 사용자가 직접 작성한 등록정보 등에 의해 얻어지는 데이터를 사용하여 수행한다. 웹 사용 마이닝은 접속 패턴을 찾는 작업과 개별 사용자의 사용 패턴을 분석하여 차별화된 서비스를 제공하기 위해 사용된다.

2.2. 적응형 웹 기술

적응형 웹 기술은 사용자의 특성에 맞게 변경하는 웹 사이트를 변경하는 기술로 개인화와 고객화로 나눌 수 있다. 개인화는 사용자가 원하는 형태로 웹 사이트 구성을 선택하는 것을 의미

한다. 현재 대부분의 인터넷 포털 사이트에서 My Page, My Menu 등을 사용하여 개인화 서비스를 제공하고 있음을 강조하고 있다.

고객화는 사용자들의 웹 로그 파일을 분석하여 차별화된 서비스를 제공하는 것을 의미한다. 사용자들의 행동 패턴 분석을 통하여 관심을 보이는 상품 또는 웹 페이지를 제공하여 차별화된 서비스를 제공할 수 있다.

2.3. 연관 규칙 탐색

웹 마이닝 분석 목적 중에서 가장 중요한 것은 사용자들이 웹 사이트에서 어떤 패턴(성향)을 가지는가를 알아내는 것이기 때문에 '패턴발견'은 매우 중요하다. '패턴발견'을 위해 적용 가능한 분석기법에는 연관 규칙(association rule), 시차 연관 규칙(sequential association rule), 군집화(clustering), 분류(classification) 등이 있다. 이러한 패턴 발견의 여러 기법을 적용하여 웹 사이트에서 사용자들의 패턴을 분석함으로써 사용자의 성향을 파악하고 행동을 예측할 수 있다.

Apriori 알고리즘은 연관 규칙을 생성하기 위해 사전 지식(priori knowledge)을 이용하여 빈발 항목집합을 생성한다. Apriori는 k번째 항목집합이 k+1번째 항목집합을 발견하기 위해 레벨 단위로 진행하여 반복 접근한다. 빈발 항목집합을 생성하기 위한 순서는 다음과 같다. 첫째 1-빈발 항목집합을 찾는다. 이 집합을 L_1으로 나타내면, L_1은 2-빈발 항목집합인 L_2를 찾는데 사용되며, 이것은 다시 3- 빈발 항목집합 L_3를 찾는데 이용되는 식으로 계속되어 더 이상의 빈발 k-항목집합이 없을 때까지 진행된다. 각 k- 빈발 항목집합 L_k를 찾기 위해서는 전체 데이터베이스에 대한 스캔이 요구된다.

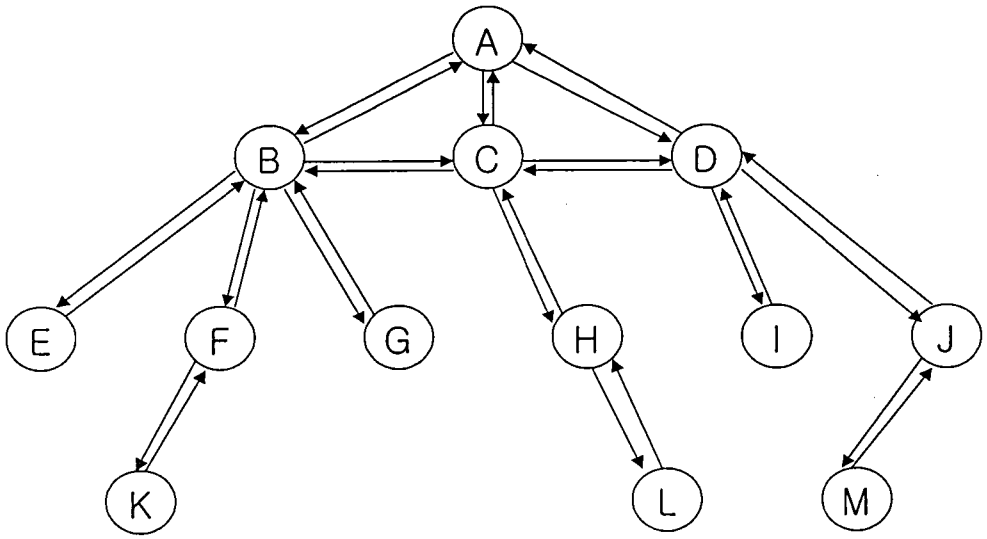
빈발 항목집합을 레벨단위로 생성하는 것을 효과적으로 개선하기 위해서 Apriori 알고리즘 특성인 '모든 공집합이 아닌 빈발 항목집합의 부분집합은 반드시 빈번하다는 특성'을 이용하여 탐색 공간을 감소시키는데 사용할 수 있다. Apriori 알고리즘은 결합(join)과 가지치기(prune)의 두 과정으로 이루어진다.

3. 웹 페이지 방문 시간을 고려한 연관 규칙 탐색

사용자가 웹 사이트에 접속하여 웹 페이지를 방문하는 시간과 방문하는 순서는 다르다. 다음 <그림 2>에서 어떤 한 사용자가 (A, E, I) 페이지를 방문하는 것은 (A, B, E, B, C, D, I), (A, B, E, B, A, D, I) 또는 (A, B, E, B, C, A, D, I)의 페이지 이동 경로가 나타날 수 있다. 사용자들의 웹 페이지 방문은 여러 이동 경로가 발생할 수 있으며 웹 페이지 방문에 따른 이동 경로는 다르다. (A, B, E, B, C, D, I) 페이지 이동 경로는 A 웹 페이지에서 E 웹 페이지를 방문하기 위해 B 웹 페이지를 방문해야 한다. 그러나 B 웹 페이지는 E 웹 페이지를 방문하기 위한 경로에 불과하다. 이런 경우 방문하였다 하더라도 방문 시간이 상대적으로 적을 수밖에 없다. 이러한 웹 페이지의 경우 빈발 웹 페이지로 볼 수 없다. 즉 사용자들이 웹 사이트에 접근한 경로 및 방문 순서, 방문 페이지, 방문 시간 등의 정보가 다르기 때문에 웹 로그 파일에 대한 연관 규칙 탐색을 하는 경우 사용자에 따른 웹 페이지 방문 시간을 고려하여야 한다.

본 논문에서 사용자들의 웹 페이지 방문 사이에 '패턴발견'을 위해 페이지 방문 시간을 고려한 연관 규칙을 다루고, Apriori 알고리즘에 의한 연관 규칙 탐색과 비교한다.

<그림 2> 웹 페이지 구조



3.1 웹 로그 파일을 이용한 웹 페이지 연관 규칙 탐색

<표 1> 웹 로그 데이터

User ID	(방문 페이지): (이동 및 방문 페이지, 방문 시간)
ID001	(A, E, I) : (A, 3), (B, 1), (E, 5), (B, 0), (A, 0), (D, 1), (I, 2)
ID002	(A, G, H, M) : (A, 0), (B, 1), (G, 3), (B, 0), (A, 0), (C, 1), (H, 2), (C, 0), (A, 0), (D, 1), (J, 1), (M, 9)
ID003	(A, K, L) : (A, 2), (B, 3), (F, 1), (K, 6), (F, 0), (B, 0), (A, 0), (C, 1), (H, 2), (L, 5)
ID004	(A, K, M) : (A, 1), (B, 1), (F, 2), (K, 4), (F, 0), (B, 0), (A, 0), (D, 2), (J, 2), (M, 10)
ID005	(A, I, L) : (A, 0), (D, 1), (I, 4), (D, 0), (A, 1), (C, 3), (H, 7), (L, 5)

<표 1>은 웹 페이지를 방문한 이동 페이지 및 방문 페이지와 그 시간을 의미하고, 방문 시간이 0시간은 다른 페이지로 이동하기 위해 거쳐야하는 페이지를 의미한다. 다음 <표 2>는 웹 로그 파일을 정제한 데이터이다.

<표 2> 웹 페이지 방문 시간을 고려한 웹 로그 데이터

User ID	(방문 페이지): (이동 및 방문 페이지, 방문 시간)
ID001	(A, E, I) : (A, 3), (B, 1), (D, 1), (E, 5), (I, 2)
ID002	(A, G, H, M) : (A, 0), (B, 1), (C, 1), (D, 1), (G, 3), (H, 2), (J, 1), (M, 9)
ID003	(A, K, L) : (A, 2), (B, 3), (C, 1), (F, 1), (H, 2), (K, 6), (L, 5)
ID004	(A, K, M) : (A, 1), (B, 1), (D, 2), (F, 2), (J, 2), (K, 4), (M, 10)
ID005	(A, I, L) : (A, 1), (C, 3), (D, 1), (H, 7), (I, 4), (L, 5)

<표 2>는 각 사용자에 따라 웹 페이지를 방문했을 때 시간 단위를 주어 해당 웹 페이지의 방문 시간이 최대인 값으로 할당한 것이다. 예를 들어 (A, E, I) 페이지를 방문하는 경우 이동

경로는 (A, 3), (B, 1), (E, 5), (B, 0), (A, 0), (D, 1), (I, 2)이고, A 페이지는 3시간 단위와 0시간 단위 2개중 최대값인 3시간 방문을 의미한다. 단위당 시간 가중치는 다음 <표 3>과 같다.

<표 3> 방문 시간에 따른 방문 시간 단위 가중값

방문 시간 단위	방문 시간	방문 시간 단위	방문 시간
1	10초 ~ 1분	6	5분 ~ 3분
2	1분 ~ 2분	7	6분 ~ 3분
3	2분 ~ 3분	8	7분 ~ 3분
4	3분 ~ 3분	9	8분 ~ 3분
5	4분 ~ 3분	10	9분 ~ 10분

방문 시간이 10초 미만은 방문 시간 단위를 0으로 처리하고, 방문 시간이 10분 이상은 방문 시간 단위를 10으로 처리하였다. 그리고 연관 규칙을 생성하기 위한 각 웹 페이지 방문 시간 단위에 따른 가중값은 사용자가 웹 페이지를 방문한 총 시간 단위 합을 계산한 후 각 페이지 방문 시간 단위로 나누어 계산하였다.

웹 페이지 당 가중값=웹 페이지 방문 시간 단위/사용자의 전체 웹 페이지 방문 시간 합

<표 4> 방문 페이지에 따른 가중값

User ID	(이동 및 방문 페이지, 방문 시간) : 웹 페이지 방문 시간 합
ID001	(A, 3), (B, 1), (D, 1), (E, 5), (I, 2) : 12
ID002	(A, 0), (B, 1), (C, 1), (D, 1), (G, 3), (H, 2), (J, 1), (M, 9): 18
ID003	(A, 2), (B, 3), (C, 1), (F, 1), (H, 2), (K, 6), (L, 5) : 20
ID004	(A, 1), (B, 1), (D, 2), (F, 2), (J, 2), (K, 4), (M, 10) : 22
ID005	(A, 1), (C, 3), (D, 1), (H, 7), (I, 4), (L, 5) : 21

<표 5> <표 4>에 대한 실험 결과

	Apriori 알고리즘 min_sup=0.4(min_count=4)	웹 페이지 방문 시간을 고려한 탐색 알고리즘 min_weight=0.4
1-빈발 항목집합	A, B, C, D, F, H, I, J, K, L, M	A, B, E, H, I, K, L, M
2-빈발 항목집합	AB, AC, AD, AF, AH, AI, AK, AL, BC, BD, BF, BH, BJ, BK, BM CD, CH, CL, DH, DI, DJ, DM, FK, HL, JM	AB, AE, AH, AI, AK, AL, AM, BE, BH, BK, BL, BM, EI, HI, HK, HL, HM, KL, KM
3-빈발 항목집합	ABD, ABF, ABK, ACH, ACL, ADI, AFK, BCH, BDJ, BDM, BFK, BJM, CDH, CHL, DJM	ABE, ABH, ABI, ABK, ABL, ABM, AEI, AHI, AHK, AHL, AIK, AIL, AKL, AKM, BEI, BHK, BHL, BHM, BKL, BKM, HIL, HKL
4-빈발 항목집합	ABFK, ACHL, BDJM	ABEI, ABHK, ABHL, ABKL, ABKM, AHIL, AHKL, BHKL

웹 페이지 방문 시간에 따른 연관 규칙 탐색을 위한 빈발 항목집합을 찾기 위해 Apriori 알고리즘에는 min_sup=0.4(min_count=2)로 하고, 시간을 고려한 연관 규칙 탐색 알고리즘에는 min_weight=0.4로 하였다.

<표 5>에서 Apriori 알고리즘 빈발 항목집합 생성 결과와 웹 페이지 방문 시간을 고려한 알고리즘 빈발 항목집합 생성 결과는 다르다. 그 이유를 예를 들면 ID001 사용자는 E 웹 페이지를 방문하기 위해 B 웹 페이지를 거쳤으며, 가장 많은 시간 단위동안 머물렀다. 이를 Apriori 알고리즘을 사용하여 빈발 항목 집합을 생성하는 경우 E 웹 페이지는 빈발 항목집합으로 선택되지 않지만 시간을 고려한 경우는 빈발 항목집합으로 선택된다. 다른 사용자들에 대한 웹 페이지에 대한 빈발 항목집합도 같은 방법으로 해석된다. 이는 사용자 마다 관심을 가지는 웹 페이지는 다르며 방문하는 시간도 같지 않게 된다. 또한 원하는 페이지로 이동하기 위해 거쳐야 하는 웹 페이지 방문 시간도 고려해야 한다.

4. 결론 및 향후과제

사용자에 따라 웹 서비스를 개인화 또는 고객화 하기 위해서는 사용자에 따른 웹 페이지 연관 관계를 파악하여 구조를 변경하거나 추천할 수 있다. 웹 구조는 일반적으로 hierarchy한 구조를 가지고 있기 때문에 사용자가 원하는 페이지를 방문하기 위해서는 다른 페이지를 거쳐야 한다. 이런 경우 페이지를 그냥 지나칠 수도 있지만 잠시 방문했다가 이동할 수도 있기 때문에 전혀 무시할 수 없다. 따라서 사용자에 따른 전체 웹 페이지 방문 시간을 이용하여 방문 페이지 시간 가중값을 계산할 수 있으며, 이를 바탕으로 연관 규칙을 생성하여 차별화된 서비스를 제공할 수 있게 된다.

앞으로의 연구 방향은 사용자에 따라 특정 웹 페이지 방문 횟수와 방문 시간간의 관계 파악 및 시간 가중치를 계산함으로써 발생되는 문제와 사용자 패턴 분석을 위한 시간 가중치 관계를 좀더 정확하게 파악할 것이다.

참고문헌

김정현, 김재련 (2001), 시간을 고려한 연관규칙을 이용한 웹 사용자 접근패턴 분석, 한국경영과학회/대한산업공학회 춘계공동학술대회

R. Agrawal and R. Srikant (1994), Fast algorithms for mining association rules, Proceeding of the 20th VLDB Conference, Santiago, Chile

Jiawei Han, Miceline Kamber, Data Mining: Concepts and Techniques. Simon Fraser University

R. Kosala, and H. Blockeel (2000), Web mining research : A survey, SIGKDD Explorations - Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, Vol2. No.1. pp. 1-15

An Efficient Algorithm for Updating Discovered Sequential Patterns in Data Mining

H. Mannila, H. Toivonen, and A. I. Verkanmo(1995), Discovering frequent episodes in sequences, In Proc. of the First Int'1 Conference on Knowledge Discovery and Data Mining, pages 210-215

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava(1999), Data preparation for mining world wide web browsing patterns, Knowledge and Information Systems, 1

Abstract

Users who use Web site wish to get information conveniently. To users who web site operators use Web site differentiation to provide done service pattern analysis by user do must.

Association rule is one of data Mining techniques for pattern discovery. If search for pattern by user, differentiation by user done service offer can. Association rule search result that pattern by user can know, and considers web page visiting time for association rule search differentiation done web structure service and recommendation service possible.