전이 확률 행렬에 의한 웹 사용 예측 모델

김영희[O]* 김용모* 정명숙* 강우준**
*성균관대학교 정보통신공학부 컴퓨터공학과, **그리스도신학대학교 경영정보학부
pink77hee@skku.edu[O]*, umkim@yurim.skku.ac.kr*, {chewon*wjkang58**}@hanmail.net

A Web Usage Prediction Model by **Transition Probability Matrix**

Y.H. Kim[O]* U.M. Kim* M.S. Jung* W.J. Kang**

* Dept. of Computer Science & Engineering, SungKyunKwan University
** Dept. of Management information Technology, Korea Christian University

요 약

웹 사용에 대한 다음 요구 사항을 예측하기 위한 마이닝 방법으로 연관규칙이나 순차 패턴 등이 많이 사용되고 있지만, 이러한 방법들은 생성된 규칙들의 지지도(Support)나 신뢰도(Confidence)에 의한 예측만을 고려 하기 때문에 정확한 예측을 하기 어려운 단점을 가지고 있다. 따라서, 본 논문에서는 빈도수에 의한 Markov model을 기반으로 하여 웹 로그 파일에 저장된 사용자들의 행동 패턴에 따라 생성되어지는 여러 형태의 규칙 유형을 찾아내고, 사용 빈도수를 이용한 전이 확률 행렬에 따른 다음 요구사항을 정확하게 예측할 수 있는 모델을 제시하고자 한다. 그 결과 여러 형태의 규칙 유형을 K^{th}-order Markov 과정에서 효율적으로 발견해 낼 수 있다.

1. 서 론

인터넷과 웹의 발전은 여러 가지 웹 마이닝 문제를 새롭게 제시하고 있다. 특히, 웹 사용 정보 마이닝은 일반적인 데이터 마이닝 방법을 웹 도메인에 적용하여 웹 사용 패턴을 찾아내고, 이를 바탕으로 개인화된 서비스를 제공하며, 사용자가 웹 서버와의 상호작용을 하는 동안에 사용자의 행위에 대한 예측과 사용자들의 접근 기록이 저장되어 있는 웹 서버 로그 파일의 분석을 이용하여 웹 페이지의 트래픽에 대한 감시, 웹 사이트 구성의 문제점을 찾아내어 웹으로부터 사용자에게 유용한 정보를 효율적으로 제공한다. 웹 사용에 대한 다음 요구 사항을 예측하기 위한 마이닝 방법으로 연관규칙이나 순차 패턴 등이 많이 사용되고 있지만, 이러한 방법들은 생성된 규칙들의 지지도(Support)나 신뢰도(Confidence)에 의한 예측만을 고려하기 때문에 정확한 예측을 하기 어려운 단점을 가지고 있다. 따라서, 본 논문에서는 Markov model을 기반으로 하여 웹 로그 파일에 저장된 사용자들의 행동 패턴에 따라 생성되어지는 여러 형태의 규칙 유형을 찾아내고, 사용 빈도수를 이용한 전이 확률 행렬에 따른 다음 요구사항을 정확하게 예측할 수 있는 모델을 제시하고자 한다. 이렇게 예측된 결과들은 e-commerce site의 의사결정 과정에 중요한 역할을 함과 동시에 최적의 개인화된 서비스를 웹 사용자들에게 제공할 수 있다. 본 논문의 구성은 다음과 같다. 제2장에서 관련 연구로 웹 사용 마이닝 과정에 대하여 살펴본 후, 웹 로그 데이터를 기반으로 발견되어지는 생성 규칙(rule)을 알아본다. 제3장에서 사용자 행동 패턴 분석을 이용한 웹 사용 예측 모델을 제시하고 마지막으로 제4장에서 결론을 기술한다.

2. 관련 연구

2.1 웹 사용 마이닝

웹 사용자가 웹 사이트를 방문하거나 특정 웹 문서를 클릭했을 때 가장 일반적인 웹 사용 분석(Web Usage Analysis)에는 웹 로그 데이터를 기반으로 한 통계적(statistical) 방법이 주로 사용된다. 예를 들면 가장 많이 방문된 웹 페이지나 방문된 경로 가운데 가장 긴 경로, 평균 트랜잭션 시간 등을 웹 로그 데이터로부터 분석 한다. 이와 같은 분석에 사용되는 툴(tool)로는 Analog[1]와 같은 것이 있다. 웹 사용 정보 마이닝[2]은 그림 1과 같이 3단계로 구성된다. 우선 웹 환경에서 모아진 Site Files, Raw Usage Data, User Profile Data를 입력 자료로 사용한다. 첫 번째 단계는 전처리(Preprocessiong) 단계로 웹 서버에 자동으로 수집된 로그 데이터에서 불필요하고 관련이 없는 자료를 정제하고 cleaning하는 단계이다.

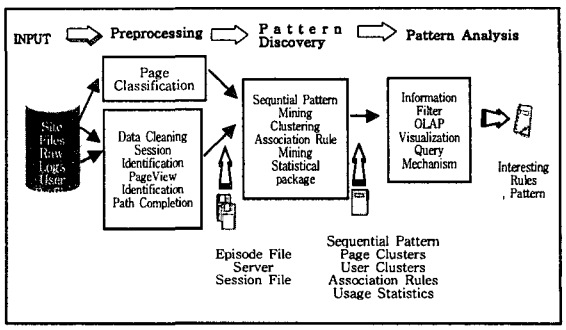

그림 1 웹 사용 마이닝 단계

둘째 단계는 패턴 발견을 위해 정제된 자료 및 정보를 가지고, 통계적인 기법 및 데이터 마이닝, 기계학습, 패턴 인식 등과 같은 알고리즘이나 기술들을 이용하여 특정한 패턴을 찾는 작업을 수행한다. 이때 사용되는 마이닝 기법에는 연관 규칙, 규칙 기반, 순차 패턴, 군집 분석등의 기법들이 사용된다. 마지막 단계는 발견된 패턴들을 이용하여 흥미로운 규칙이나 패턴 발견 단계의 산출물로부터 유용한 패턴을 추출하기 위한 패턴 분석 단계이다. 이렇게 분석된 규칙이나 패턴들은 앞으로 발생되는 다음 사항에 대한 예측을 가능하게 한다.

2.2 규칙 생성

웹 페이지를 방문했을 때 웹 기반 활동의 결과들을 이용하여 사용자의 다음 요구를 예측하는 것은 유용한 정보를 제공함에 있어서 매우 중요한 요소이다. 이러한 문제 해결을 위해 사용되는 패턴 발견 기법으로 CF(Collaborative Filtering), 웹 페이지 또는 세션에 대한 클러스터링(clustering), 연관규칙 생성, 순차패턴 생성과 Markov Model이 주로 사용된다. 생성되는 규칙 유형은 부분집합(subset), 부 시퀀스(subsequence), 최종 부 시퀀스(latest-subsequence), 서브스트링(substring), 최종 서브스트링(latest-substring)의 5가지 형태가 있다. 이러한 규칙은 그림2에서와 같이 이전에 방문된 페이지의 정보(LW)와 다음에 방문되어질 페이지들의 정보(RW)에 대한 두개의 window의 변화를 통해 쉽게 나타낼 수 있다.

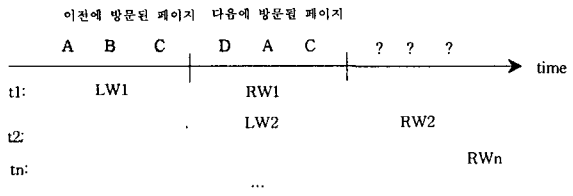

그림 2 시간의 변이에 따른 윈도우(Window) 이동

시퀀스(sequence)A,B,C,→D에서 윈도우의 크기를 4로 했을 때 연속적인 시간에서 윈도우가 변화되면서 예측되는 패턴은 표1과 같이 얻을 수 있고, A,B,C,→D와 B,A,C,→D의 순서 정보에 따른 5가지 규칙 생성은 표2와 같이 발견되어진다.

LW(Previously Visited Pages)			RW(Prediction Page)
L1	L2	L3	P
A	B	C	D
B	C	D	A
C	D	A	C

표 1 윈도우의 변화에 의해 발견된 패턴

Rule Type	Rules (min_sup = 100%)
Subset	{A,B,C}→D, {A,B}→D, {B,C}→D, {A,C}→D, {A}→D, {B}→D, {C}→D
Subsequence	{B,C}→D, {A,C}→D, {A}→D, {B}→D, {C}→D
Latest Subsequence	{B,C}→D, {A,C}→D, {C}→D
Substring Rules	{A}→D, {B}→D, {C}→D
Latest Substring	{C}→D

표 2 규칙 생성 결과

3. 웹 사용 예측 모델

Markov model은 웹 사이트에서 사용자의 브라우징 행동을 예측하기 위해 널리 사용 되어지고 있다. Padbanabham과 Mogul[3]은 웹 캐시(web cache)의 prefetch 개선을 위해 N-hop Markov model을 사용했고, Sarukkai[4]는 Markov model을 이용하여 사용자에 의해 다음에 접근할 페이지를 예측했다. 또한, Cadez et al[5]은 서로 다른 카테고리들로 브라우징 세션들을 분류하는데 Markov model을 사용했다.

3.1 Markov Model

웹 사용 예측을 위해 사용되는 Markov Model은 세 개의 매개변수(parameter) <A, S, T>에 의해 표현된다.

- A : 사용자에 의해 수행될 수 있는 모든 가능한 동작(action)의 집합
- S : Markov model의 모든 가능한 state의 집합
- T : |A| × |S| , Transition Probability Matrix(TPM)
- t_{ij} : 프로세스가 state i 에서 action j 수행 가능성

Markov model은 다음 동작(action)의 예측에서 사용된 이전 동작(action)의 수에 따라 다음과 같이 나타낼 수 있다.

- first-order Markov Model : 마지막 action에서 하나의 state만을 고려한 모델
- second-order Markov Model : 마지막 action에서 두개의 state만을 고려한 모델
- K^{th}-order Markov Model : 마지막 action에서 K의 state를 고려한 모델

Markov model의 설계는 웹 사용자의 동작 시퀀스(action-sequences)들에 대하여 State s_i에서 Action a_i인 사건의 빈도수에 의해 각 t_{ij}를 추정하여 전이 확률 행렬(Transition Probability Matrix)을 구성한다. 웹 사용자의 동작 시퀀스가 표3과 같을 때 위에서 설명한 모델들의 사용 빈도에 따른 전이 확률 행렬(Transition Probability Matrix)은 그림3과 같다.

Sessions	Action Sequences
S1	{ C, B, A}
S2	{ C, E, B, A, D}
S3	{D, E, B, A, E, D}
S4	{C, D, E, B, A}
S5	{A, D, B, D}

표 3 웹 사용자의 동작 시퀀스(action- sequences)

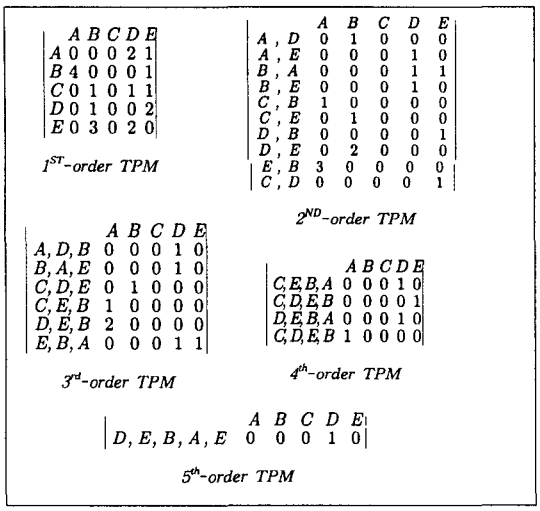

그림3 빈도수에 의한 전이 확률 행렬(Transition Probability Matrix)

3.2 웹 사용 예측 모델

웹 로그 파일에서 기록된 사용자들의 웹 페이지 사용의 정보를 이용하여 시간을 두고 확률적으로 상태가 변하는 과정과 그 결과를 파악하기 위해 시계열을 이산적으로 나타내는 과정을 Markov Process(마코브 과정)라 하고 임의의 상태 Si로부터 다른 임의의 상태 Sj로 변환하는 확률을 $P(Sj|Si) = q_{ij}$의 조건부 확률로 나타낸다. 이때, $t_n \rightarrow t_{n+1}$로의 변환으로 변환 확률 또는 추이 확률을 계산한다. 어떤 시점 t_n에서 일어나는 각 사상의 확률을 벡터 $\overline{P}_n = (P_1^{(n)}\ P_2^{(n)}\ P_3^{(n)})$로 곱해서 다음 시점 t_{n+1}에서의 확률 벡터로 나타내고, 따라서 $\overline{P}_n = \overline{P}_{n+1}$으로의 선형 변환에 의해 다음 상태를 예측할 수 있다. 웹 사용자가 현재 page i 에 있을 때, 이전에 방문되었던 page들의 m 시퀀스(sequence)들을 $\{i_{-m+1}, i_{-m+2}, ..., i_0,\}$이라 가정할 때, 현재 페이지의 벡터 $L_0 = \{l_j\}$는 $l_j = 1$이면 $j=i$ 이고 $l_j = 0$이면 $j \neq i$이다, 반면, 이전 페이지의 벡터 $L_k = \{l_{jk}\}$ $(k=-1,...,-m+1)$, $l_{jk} = 1$이면 $j_k = i_k$ 이고 $l_{jk} = 0$이면 $j_k \neq i_k$ 이다. 따라서, 다음에 방문될 각 웹페이지에 대한 확률 벡터 Pc_1는

$$Pc_1 = a_1 \times L_0 \times Q + a_2 \times L_{-1} \times Q^2 + ... + a_m \times L_{-m+1} \times Q^m$$

과 같다.

위의 확률 벡터 Pc_1에 의한 N^{th}-order에서 다음 n step내에 방문되어질 웹 페이지의 예측을 위한 확률 벡터 Pc_n은 다음과 같이 얻어질 수 있다. 아래 식에서 $a_{1,1}, a_{1,2}, ..., a_{m-1,1}, a_{m-1,2}, ..., a_{m-1,n}$은 히스토리 벡터 $L_0, ..., L_{-m+1}$에 할당된 가중치를 나타내고, Q는 n 계승을 갖는 변환 행렬이다.

$$Pc_n = a_{1,1} \times L_0 \times Q + a_{1,2} \times L_0 \times Q^2 + ... + a_{1,n} \times L_0 \times Q^n +$$
$$a_{2,1} \times L_{-1} \times Q^2 + a_{2,2} \times L_{-1} \times Q^3 + ... + a_{2,n} \times L_{-1} \times Q^{n+1} + ... +$$
$$a_{m-1,1} \times L_{-m+1} \times Q^{m-1} + a_{m-1,2} \times L_{-m+1} \times Q^m + ... + a_{m-1,n} \times L_{-m+1} \times Q^{m+n-1}$$

위 식은 다음에 방문될 웹 페이지의 예측을 위해 모든 히스토리 벡터를 이용하므로 더 정확한 예측을 가져올 수 있다. 뿐만 아니라 session 2의 시퀀스 C,E,B,A,D와 session 3의 시퀀스 D,E,B,A,E,D의 예를 통해 빈도수에 의한 전이 확률 행렬의 차수에 의존하여 위에서 보인 5가지 규칙에 대한 생성을 쉽게 발견 가능함을 알 수 있다.

4. 결 론

본 논문에서는 웹 로그 파일에 저장된 사용자들의 행동 패턴에 따라 생성되어지는 여러 형태의 규칙 유형을 K^{th}-order Markov 과정에서 발견해 내고, 사용 빈도수를 이용한 전이 확률 행렬에 따른 다음 요구사항을 정확하게 예측할 수 있도록 이전에 방문되어진 웹 로그 파일들의 히스토리 벡터를 이용하는 모델을 제시하였다.

향후 과제로 제시된 모델 과정에서 나타나는 규칙 생성시 많은 양의 정보가 있을 때 효율적인 저장을 위한 압축된 정보의 저장 기법과 이전의 정보들을 통한 더욱 더 정확한 예측을 가능하게 할 수 있는 방법들이 지속적으로 연구되어야 할 것이다.

참고문헌

[1] S. Turner. Analog. http://www.analog.cx, 2000.

[2] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastave. Discovery of interesting usage patterns from web data. Technical Report TR 99-022, University of Minnesota, 1999.

[3] V.N. Padmanabham and J.C.Mogul. Using predictive prefetching to improve world wide web latency. Computer Communication Review, 1996.

[4] Ramesh R. Sarukkai. Link prediction and path analysis using markov chains. In Ninth International World Wide Web Conference, 2000.

[5] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Vidualization of navigational patterns on web site using model based clustering. Technical Report MSR-TR-0018, Microsoft Reserch, Microsoft Corporation, 2000.

[6] Mukund Deshpande and George Karypis. Selective Markov Models for Predicting Web-Page Accesses, ACM Transactions on Internet Technology, Volume 4, Issue 2 , May 2004.

[7] Jianhan Zhu, Jun Hong, and John G. Hughes, Using Markov Chains for Link Predicion in Adaptive Web Sites, Springer-Verlag Berlin Heidelberg, 2002