웹 로그 마이닝을 이용한 웹 문서 예측 시스템

이범석° 황병연
가톨릭대학교 컴퓨터공학과
{bslee°, byhwang}@catholic.ac.kr

Web Document Prediction System by using Web Log Mining

Bum-suk Lee° Byung-yeon Hwang
Dept. of Computer Engineering, The Catholic University of Korea

요 약

웹 문서 수의 급격한 증가는 사용자로 하여금 방대한 양의 웹 문서들로부터 필요한 정보를 선별하기 위한 시간과 비용을 낭비하게 만들었다. 따라서 이러한 문제를 해결하기 위한 연구의 필요성이 점차 증가하였는데, 그 중 웹 서버 로그 데이터에 마이닝 기법을 적용하여 사용자들의 사이트 내 문서의 접근 패턴을 분석하고, 그 데이터를 이용하여 동적으로 변화하는 적응형 웹 사이트를 제공하려는 것이 대표적인 연구 사례이다.

본 논문에서는 웹 서버 로그 마이닝을 이용하여 사용자가 필요로 하거나, 관심을 가지고 있는 페이지를 예측하여 추천해 주는 시스템에 대해 소개한다. 이러한 시스템을 구현하기 위해 순차 패턴 마이닝이나 빈발 에피소드 발견 기법 등의 알고리즘을 사용할 수 있다. 제안하는 시스템에서는 사용자 접근 패턴을 분석할 때 순차 패턴 마이닝 기법을 사용하고, 사용자의 이동 패턴을 근거로 웹 문서를 예측하여 추천해줄 때에는 에피소드 발견 기법에서의 window 개념을 이용한다. 본 논문에서 제안한 시스템은 웹 문서를 사용자가 머물었던 시간에 따라 관심 있는 문서와 지나간 문서로 구분하여 관심 있는 문서에 대해서만 마이닝을 수행한다. 또한 일정한 크기를 갖는 History window에 의해 다음 문서를 추천해주기 때문에 사용자의 모든 로그를 저장하지 않으므로 보다 효율적이다.

1. 서 론

월드 와이드 웹(WWW)의 발달은 다양한 분야에 대한 정보량의 증가로 이어졌고, 누구나 손쉽게 원하는 정보를 웹 검색을 통해 찾아낼 수 있게 되었다. 웹은 이러한 장점을 가지고 있지만, 너무 많아진 웹 문서는 사용자로 하여금 검색을 통해 얻어진 방대한 양의 결과물들로부터 필요한 정보를 선별하기 위한 시간과 비용을 낭비하게 만들었다. 따라서 이러한 문제를 해결하기 위한 연구의 필요성이 점차 증가하였는데, 그 중 웹 서버 로그 데이터에 마이닝 기법을 적용하여 사용자들의 사이트 내 문서의 접근 패턴을 분석하고, 그 데이터를 이용하여 동적으로 변화하는 적응형 웹 사이트[1]를 제공하려는 것이 대표적인 연구 사례이다.

적응형 웹 사이트는 웹 마스터의 의도가 반영된 고정된 구조를 가진 초기의 웹 사이트와 달리, 웹 로그 데이터를 마이닝(Mining)하므로써 사용자들의 사이트 내 페이지의 접근 패턴을 분석하여 각각의 사용자에 적합하게 동적으로 변화하는 웹 사이트를 말한다. 최근 이러한 연구는 적응형 웹 사이트에서 더 나아가 연관 규칙 기반으로 구현된, 사용자가 앞으로 요청할 페이지를 예측하고 추천해주는 시스템도 구현되고 있다.

본 논문에서는 대규모의 웹 사이트에서 로그 분석을 통한 보다 효율적인 웹 문서 예측 시스템의 구현에 대해 논의한다. 이러한 시스템을 구현하기 위해서는 웹 서버 로그 데이터의 정제와 분석 과정이 필요하다. 우선 로그 정제 과정에서는 마이닝의 효율성을 높이기 위해 기존의 연구에 고려하지 않았던 웹 페이지의 가중치를 고려하여 사용자가 요청했던 페이지에서 가중치가 낮은 페이지는 제거한다. 로그 분석 과정에서는 순차 패턴 마이닝[2,3]을 사용하여 연관 규칙과 순차 패턴을 생성하여 DB에 저장하고, 빈발 에피소드 발견 기법[4]을 응용하여 새로운 사용자가 특정 페이지 패턴을 요청하게 되면 다음 웹 문서를 예측하여 추천 페이지에 보여주도록 한다.

본 논문의 2장에서는 관련 연구들을 살펴보고, 3장에서는 웹 문서 예측 시스템에 대해 소개한다. 마지막으로 4장에서는 결론 및 향후 연구 계획에 대해 논의한다.

2. 관련연구

데이터 마이닝이란 일정한 기간동안 축적된 다량의 데이터베이스로부터 의미있고 유용한 정보를 찾아내는 과정을 말한다. 데이터 마이닝 기법에 관련된 많은 연구가 이루어졌으나 본 논문에서 사용한 기법은 순차 패턴 마이닝과 빈발 에피소드 발견 기법이다.

첫째로, 순차 패턴 마이닝은 여러 항목들의 집합으로 표현된 트랜잭션에서 각 항목들 사이의 연관성을 반영하는 연관 규칙을 생성한다. X와 Y가 각각 Itemset이라고 할 때, 연관 규칙은 X⇒Y로 표시하며 Itemset X를 포함하고 있는 데이터베이스 내의 트랜잭션이 Y를 포함하고 있을 가능성을 의미한다. 이러한 연관 규칙의 정확성을 측정하기 위해 지지도(Support)와 신뢰도(Confidence)를

이용한다. 지지도와 신뢰도를 구하는 공식은 다음과 같다.

$$(1)\ Support = \frac{X \cap Y}{All} \quad (2)\ Confidence = \frac{X \cap Y}{X}$$

지지도는 DB내의 전체 트랜잭션 중에 X와 Y를 모두 포함한 트랜잭션의 수를 의미하고, 신뢰도는 X를 포함한 트랜잭션 중에 Y를 함께 포함한 트랜잭션의 수를 의미한다. 전체 트랜잭션 데이터베이스로부터 사전에 정의된 최소 지지도를 만족하는 항목들의 집합을 Large Itemset 또는 Frequent Itemset이라고 한다.

순차 패턴은 한 트랜잭션 안에서 발생하는 항목들 사이의 연관 규칙에 시간의 의미를 추가한 것이다. 이를 시퀀스(sequence)라 하며, 순차 패턴의 탐색은 사용자가 정의한 최소지지도를 갖는 시퀀스인 빈도가 많은 시퀀스(large sequence)를 추출하고 이들 가운데 최대시퀀스(maximal sequence)를 찾는 것이다. 순차 패턴 발견에서의 연관 규칙 X⇒Y는 트랜잭션의 Itemset X가 나타나면 일정시간이 경과한 다음에 Itemset Y가 발견된다는 의미로 해석된다. 연관 규칙과 순차 패턴이 다른 점은 연관성 규칙은 X→Y, Y→X가 성립하지만, 순차패턴은 X→Y만 성립한다는 것이다. 다시 말해 연관규칙은 X, Y 중 어느 것이 먼저 일어나도 관계없지만 순차패턴은 반드시 X가 먼저 발생하는 것을 말한다.

이러한 순차 패턴 마이닝 기법 중 의학적 치료, 자연재해와 주식 시장의 예측 등 과학적인 분야와 DNA 서열 및 염색체 구조 분석에서 주로 사용되는 이벤트 서열 내 빈발 에피소드를 발견 기법[4]이 있다. 이 방법은 일정한 크기(width)를 가지는 윈도우를 기준으로 빈발 에피소드를 찾아내고, 에피소드들 사이의 연관 규칙을 찾아낸다. 이것을 응용하면 에피소드가 순차적으로 발생하는 직렬 에피소드(Serial Episode), 순서에 관계없이 함께 나타나는 병렬 에피소드(Parallel Episode), 그리고 둘이 복합적으로 이루어진 비직렬, 비병렬 에피소드(Non-Serial Non-Parallel Episode)의 세 가지 패턴을 찾아낼 수 있다. 그림 1은 이러한 패턴을 보여준다.

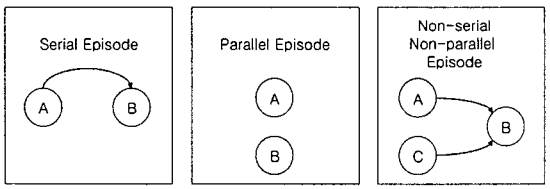


그림 1 에피소드 패턴

3. 웹 문서 예측 시스템

3.1 웹 서버 로그 데이터의 정제

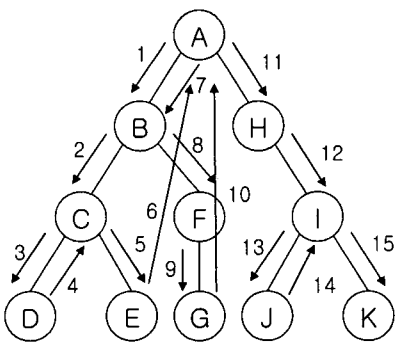
웹 로그 마이닝을 수행하기 위해서는 먼저 웹 로그의 정제 과정이 필요하다. 순수한 웹 로그 데이터는 서버에 접속한 사용자의 IP와 사용자가 요청한 페이지 정보 및 시간 등의 내용을 포함하고 있다. 그림 2는 웹 서버 로그 데이터의 예를 보여준다.

```
203.251.189.47 - - [03/Aug/2000:21:56:55 +0900]
"GET /doc/images/sub.gif HTTP/1.1" 200 6083
203.251.189.47 - - [03/Aug/2000:21:56:55 +0900]
"GET /doc/images/index.gif HTTP/1.1" 200 1540
203.251.189.47 - - [03/Aug/2000:21:57:25 +0900]
"GET /doc/mod/directives.html HTTP/1.1" 200 11339
203.251.189.47 - - [03/Aug/2000:21:57:25 +0900]
"GET /doc/images/home.gif HTTP/1.1" 200 1465
211.52.197.57 - - [03/Aug/2000:21:57:26 +0900]
"POST /cgi-bin/bbs HTTP/1.1" 200 3840
```

그림 2 웹 서버 로그 데이터의 예

이러한 로그 데이터를 정제하는 첫 번째 단계는 사용자가 요구하지 않은 데이터의 삭제이다.. 로그에는 어떤 사용자가 문서를 요청했을 때 해당 문서에 포함 되어있어서 자동으로 전송된 이미지에 대한 정보까지 포함되어 있는데, 이러한 정보는 사용자 패턴 분석에 필요하지 않다. 다음 단계는 사용자 세션을 발견하여 순차적인 데이터로 만드는 작업이다. 예를 들어 웹 로그 데이터가 다음과 같은 서열로 이루어져 있다고 가정하자:{A (by user 1), B (by user 2), C (by user 2), D (by user 3), E (by user 1)}. 이 서열을 IP에 의해 사용자 세션으로 분류한다. 그 결과는 Session 1 (by user 1): (A, E); Session 2 (by user 2): (B, C); Session 3 (by user 3): (D) 의 형태를 가진다. 또한 동일 사용자의 세션이라도 페이지 요청 시간의 간격이 크다면 새로운 세션으로 인식한다. 이때 절대적이지는 않지만 새로운 세션으로 인식하는 시간의 간격은 약 2시간정도가 적당하다[5]. 위의 작업들이 끝나면 세션으로 분류된 사용자의 웹 브라우징 서열을 얻게 된다.

본 논문에서는 각각의 웹 문서에서 사용자들이 머물렀던 시간으로 가중치를 부여하기 위해 웹 문서를 관심 있는 문서(Interesting Document)와 지나간 문서(Passing Over Document)의 두 가지 성질로 분류한다. 우선 사용자가 해당 문서에서 머물었던 시간이 임계값(Threshold) 이상인 것을 관심 있는 문서로 정의한다. 그리고 임계값 미만의 문서들을 지나간 문서로 분류한다. 지나간 문서로 분류된 것들은 사용자 웹 브라우징 서열에서 제거한다. 이 과정이 웹 서버 로그 데이터의 세 번째 정제 단계이다.



{A→B→C→D→C→E→A→B→F→G→A→H→I→J→I→K}

그림 3 사용자 세션의 예

그림 3은 사용자 세션의 예를 그래프 형태로 보여준다. 만약 웹 문서를 분류하기 위한 임계값이 1분, 사용자가 웹 문서 D, G, H, J를 방문했을 때 머문 시간이 각각 45초, 12분, 35초, 15분이라고 가정한다면, G와 J는 사용자가 원했던 문서이고, D와 H는 사용자가 원하는 정보를 찾기 위해 지나간 문서였을 것이라 예측할 수 있는 것이다. 이처럼 머문 시간을 고려하여 사용자 세션을 줄여주는 것은 연관 규칙 생성을 위한 마이닝 작업에 걸리는 시간을 줄여줄 수 있다.

3.2 순차 패턴 마이닝

웹 로그 정제 단계가 끝나면 순차 패턴 마이닝을 이용하여 연관 규칙을 생성한다. AprioriAll 기법은 순차 패턴 마이닝을 위한 다양한 알고리즘들 중 가장 기본적이고, 초기의 알고리즘 이다. 본 논문은 마이닝 알고리즘을 제안한 것이 아니므로, AprioriAll 알고리즘을 사용한다.

정제된 사용자 브라우징 패턴을 서열 형태로 나타내면 <문서 (D1), 문서(D2), ... , 문서(Dn)>으로 표현될 것이다. 서열에 대한 지지도는 전체 사용자 중에서 서열 패턴을 만족하는 사용자의 수로 나타낼 수 있다. 이러한 각각의 최대 서열들을 순차 패턴(Sequential Patterns)이라고 하고 최소 지지도를 만족하는 서열을 빈발 서열(Large Sequence 혹은 Large Itemset)이라고 부른다. 빈발 서열 안의 항목집합들은 반드시 최소 지지도를 만족해야 한다. 따라서 빈발 서열은 항목 집합 목록의 형태로 나타난다.

3.3 웹 문서의 예측

그림 4는 웹 문서 예측에 사용되는 두 개의 윈도우를 보여준다. 순차 패턴 마이닝이 끝나면 연관 규칙을 생성해서 페이지를 예측하고, 사용자에게 추천해줄 수 있게 된다. 마지막 단계로 현재 사용자에게 적합한 웹 문서를 예측하고 추천해주기 위해 두 개의 윈도우를 정의한다. 첫 번째 윈도우는 History window라고 부르고, 현재 사용자가 방문했던 웹 문서들의 정보를 정해진 윈도우 크기(width)만큼 보관한다. 다른 한 개의 윈도우는 Prediction window라고 하며, 사용자가 앞으로 방문할 웹 문서의 경로에 대한 정보를 보관한다. 이 두 개의 윈도우의 내용은 사용자가 요청하는 문서 정보에 의해 실시간으로 변화한다.

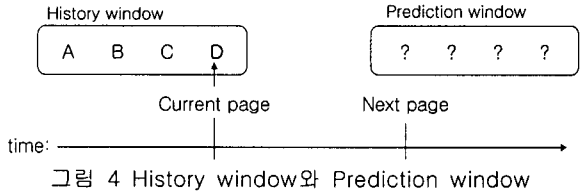


그림 4 History window와 Prediction window

Prediction window의 후보는 선택될 가능성이 높은 순서대로 5개를 생성한다. 시스템에 의해 추천된 5개의 웹 문서 중에서 사용자가 특정 문서를 선택하면 Prediction window의 첫 문서는 History window의 맨 마지막으로 이동한다. 윈도우의 크기가 정해져 있으므로 History window의 내용과 Prediction window는 사용자가 문서를 이동할 때 마다 변경된다.

4. 결론 및 향후 연구 계획

본 논문에서 제안한 웹 문서 예측 및 추천 시스템은 다음의 특징 및 장점을 갖는다.

1) 사용자가 웹 문서에 머물었던 시간을 근거로 문서의 중요도를 고려하여 관심 있는 문서(Interesting Document)와 지나간 문서(Passing Over Document)의 두 가지로 분류하고, 관심 있는 문서에 대해서만 사용자 세션 서열을 생성하고 순차 패턴 마이닝을 수행한다. 이 방법은 순차 패턴 마이닝을 수행할 때 웹 문서의 개수를 줄이고, 의미 없는 문서가 추천되는 것을 제한한다.

2) 사용자에게 추천을 해 줄때 순차 패턴 마이닝을 수행하여 생성한 빈발 서열의 전체를 고려하는 것이 아니라, 일정한 크기를 갖는 History window에 의해 다음 문서를 추천해주기 때문에 사용자의 모든 로그를 저장하지 않으므로 보다 효율적이다. 또한 사용자의 이동 history에 따라서 추천해 줄 문서가 실시간으로 바뀌기 때문에 일반화 시킨 추천 시스템보다 정확하다.

사용자에게 적합한 웹 문서를 예측하여 추천해주는 시스템은 사용자의 입장에서는 원하는 정보를 쉽게 찾을 수 있도록 하고, 웹 사이트 운영자의 입장에서는 불필요한 요청을 줄임으로서 서버의 부하를 줄여줄 수 있다. 이러한 시스템의 웹 문서 추천 방법이나 패턴 마이닝에 관련된 방법의 비교는 가능하지만, 실제 기능이나 성능에 대한 절대적이고 정량적인 비교는 불가능하다. 따라서, 향후 연구로 웹 문서 추천 시스템들의 성능을 정량적으로 비교할 수 있는 방법에 대해 연구해 보고, 순차 패턴 마이닝 알고리즘의 효율성을 비교하는 것도 의미가 있을 것이다.

참고문헌

[1] M. Perkowitz and O. Etzioni, "Adaptive Web Sites: an AI Challenge," In Proc. of the 15th Int'l Joint Conf. AI, 1997.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," In Proc. of the 11th Int'l Conf. on Data Engineering, Taipei, Taiwan, 1995.

[3] R. Srikant and R.Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvement," In 5th Int'l Conf. on Extending DataBase Technology, Avignon, France, 1996.

[4] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences," Data Mining and Knowledge Discovery, 1(3), pp. 259-289, 1997.

[5] Y. Qiang, L. Tianyi and W. Ke, "Building Association-Rule Based Sequential Classifiers for Web-document Prediction," Journal of Data Mining and Knowledge Discovery, 8(3), pp. 253-273, 2004.