전 방향 참조 경로 탐사 패턴을 이용한 웹 문서 예측

김양규° 손기락
한국외국어대학교 컴퓨터및정보통신공학부
yangkyu0@yahoo.co.kr, ksohn@hufs.ac.kr

Web document prediction using forward reference path traversal patterns

Yang kyu Kim° Kirack Sohn
School of Computer and Information Communication Engineering,
Hankuk University of Foreign Studies

요 약

오늘날 웹을 이용하는 사용자들의 웹 검색 형태를 저장한 웹 로그 데이터들은 데이터 마이닝을 위한 중요한 자료가 되고 있다. 이들 웹 로그들로부터 사용자의 현재 행동을 기반으로 사용자가 다음에 요청할 요구를 예측할 수 있는 예측 모델을 만들 수 있다. 하지만 이들 웹 로그들은 크기가 매우 크고 분석하기가 어렵다. 이런 문제를 해결하기 위해 이미 많은 방법이 제안 되었다. 그 중에서 효과적으로 예측할 수 있도록 제안된 순차적 분류 기반에 연관법칙을 적용한 예측 기법이 있다. 본 논문에서는 전방향 참조 경로 탐사 패턴 알고리즘을 적용하여 연관규칙에 기반한 웹문서 예측 기법을 향상시키는 모델을 제안한다.

1. 서 론

오늘날 실생활에서 빼놓을 수 없을 정도로 급속히 성장한 월드 와이드 웹으로 이용할 수 있는 인터넷 정보의 양은 기하급수적으로 커지고 있다. 또한 많은 웹 사용자들의 웹 이용으로 인해 지난 10년간 네트웍 트래픽 또한 폭주하게 되었다. 이로 인해 웹 사용자들에게 보다 좋은 질의 서비스를 제공할 필요가 증대 되었으며, 그러기 위해 사용자들의 행동을 연구할 필요성 또한 증대되었다. 이러한 연구들을 위한 중요한 데이터로 사용자들의 웹 검색 행동을 기록한 웹 로그 데이터를 이용하게 되었다. 본 논문에서는 웹 로그 데이터를 분석하여 웹을 이용하는 사용자에게 다음 웹페이지 요청을 예측하는 예측 모델을 제안하였다. 정확한 예측의 결과는 소비자에게 상품을 추천하거나, 유용한 링크를 권장하거나, 액세스 지연을 줄이기 위해 웹 페이지의 pre-sending과 pre-fetching, caching에도 이용될 수 있다.

웹 로그에는 웹 사용자들이 이전에 방문한 웹 페이지로 되돌아가기 위해 방문한 웹페이지까지 기록되는데(후방향 참조), 이런 기록들은 웹문서 예측을 위해서는 불필요한 정보이다. 진정한 웹문서 방문 패턴은 전방향 참조(forward reference)이다. 지금까지의 연구에서는 웹 로그 데이터를 전방향 참조와 후방향 참조에 대한 구별없이 하나의 연속된 방문순서열(sequence)로 보고 이를 이용하여 예측하였기 때문에 불필요한 후방향 참조 정보까지 예측에 이용되는 오류를 범할 수 있게 된다. 하지만 경로를 추출하는 과정에 전방향 참조 경로 탐사 패턴 추출[1]을 적용하여 웹 로그 데이터를 하나의 연속된 방문순서열이 아닌 트리 형태로 구성하여 전방향 참조 경로를 추출 할 수 있도록 하며, 결국 더욱 정확한 예측을 할 수 있게 한다.

이 논문은 웹 로그 데이터에서 데이터 마이닝을 통하여 얻게 되는 연관규칙들을 구하기 위하여 전방향 참조 경로 탐사 패턴[1]을 적용시켜 보다 정확한 경로를 구성 할 수 있도록 하였으며, 규칙 구성방법으로는 Latest substring rule[2][3]을 이용하였고, 규칙 선택 방법으로는 pessimistic selection[4] 기법을 적용하였으며, 테스트 데이터에도 전방향 참조 경로 탐사 패턴을 접목 시켜서 결과적으로 보다 정확한 예측을 구하는 것을 목표로 하고 있다.

2. 관련연구

2.1 Moving window

웹 로그 데이터를 하나의 연속된 방문순서열이라고 했을 때, 이 방문순서열을 가지고 예측을 하기 위한 방법으로 moving window[2][3]기법이 제안되었다. 연속된 방문순서열을 일정 크기로 나누어서 규칙을 구성하고 나중에 예측하기 위하여 antecedent window와 consequent window라는 두 개의 window를 정의한다. 본 논문에서는 앞으로 antecedent window는 W1, consequent window는 W2라고 부르겠다. W1에는 사용자가 이미 방문한 웹 페이지들이 나타나 있으며, W2에는 앞으로 방문하게 될 웹 페이지들을 나타낸다. 방문순서열

을 (A, B, C, D, C, B, E)라고 했을 때, W1의 크기를 4, W2의 크기를 1로 정하면 moving window는 다음 표1과 같다.

W1				W2
A	B	C	D	C
B	C	D	C	B
C	D	C	B	E

표1. 방문순서열로부터 추출한 moving window

표1과 같이 window가 하나씩 방문 순서열의 오른쪽으로 옮겨가는 것을 보고 window가 한 칸씩 이동하는 것이기 때문에 moving window라고 부르며, 본 논문에서는 규칙을 구성할 때, 예측 값을 구할 때 모두 W1은 4, W2는 1의 크기로 사용한다. 이 방법에서는 C B와 같이 B로 되돌아가기 위한 의미없는 동작이 moving window의 정보에 나타나는 단점이 있다. 본 논문에서는 경로 패턴을 이용하여 이러한 단점을 보완한다.

3. 알고리즘

순수한 웹 로그 데이터를 이용하여 예측하기까지는 다음의 순서대로 알고리즘을 적용해야한다.

1단계: 웹 로그 데이터에서 필요한 데이터만 추출
2단계: 추출한 데이터를 이용하여 세션별로 분류
3단계: 분류한 데이터를 이용하여 경로 추출(전방향 참조 경로 탐사 패턴 추출[1])
4단계: 추출한 경로를 이용하여 규칙 생성(Latest substring rule[2][3] 적용)
5단계: 테스트 데이터에서 예측하기 위하여 규칙들 추출 (전방향 참조 경로 탐사 패턴 추출[1] 적용)
6단계: 5단계에서 구한 규칙과 4단계에서 구한 규칙과 같은 규칙들 중에 정확한 규칙 선택(pessimistic selection[4] 적용)
7단계: 선택한 규칙을 이용하여 예측하기

3.1 전방향 참조 경로 탐사 패턴 추출 알고리즘

많은 웹 사용자들이 원하지 않는 정보를 가진 페이지로 잘못 이동하여 되돌아가거나, 방금 이용했던 페이지로 다시 돌아가게 되는 경우는 웹 검색에서 매우 빈번하게 발생하게 되는 일들 중 하나이다. 하지만 지금까지는 웹 로그 데이터를 하나의 방문순서열로 보았기 때문에 실제로는 잘못 된 경로를 나타내더라도 해결 할 수 있는 방법이 없었다. 하지만 로그 데이터를 하나의 방문순서열이 아닌 트리의 형태로 보게 되면 잘못 된 경로를 올바르게 표현할 수가 있어진다. 그래서 전방향 참조 경로 탐사 패턴 추출 알고리즘[1]을 적용하여 이런 오류의 범위를 최소화 시켜 보다 정확한 예측 값을 구할 수 있도록 기여한다.

예로 웹 로그에서 세션별로 요구한 웹 페이지들의 순서가 다음과 같을 때: {A, B, C, D, C, B, E, F, G, F, H, A}. 그림 1과 같이 트리를 구성한다.

그림 1에서 구성한 트리에서 추출한 경로는 다음과 같다: {ABCD, ABEFG, ABEFH}. 이를 전방향 참조경로(forward reference path)라 부른다.

그러므로 위의 예에서 후방향 참조경로(backward reference path): {D,C,B}, {G,F}, {H,A}는 새로 만들어진 경로에서 제거되었다.

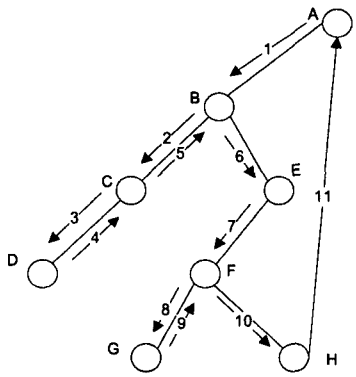

그림 1: 전방향 참조 경로 탐사 패턴을 적용한 트리

3단계에서는 이미 이동한 모든 정보를 가지고 있기 때문에 세션별로 바로 트리를 구성해 경로를 구한다. 하지만 5단계에서 적용할 때는 이미 이동한 모든 정보를 가지고 있지 않고 예측 값을 구해가며 이동 정보를 얻는 경우이기 때문에 moving window의 크기에 맞춰 하나씩 트리에 삽입해 가며 경로를 추출하여 6단계, 7단계에 이용한다.

3.2 Latest Substring Rule

웹 로그 데이터에서 세션 별로 추출한 자료를 가지고 LHS->RHS로 가는 순차적 연관규칙들을 뽑아내게 되는데 이들 방법 중 가장 좋은 효과를 나타내는 Latest substring rule을 적용한다. LHS란 마지막에 요구하게 될 웹 페이지의 전까지 요구했던 웹 페이지의 문자열 목록을 말하며, RHS란 마지막에 요구할 웹 페이지를 말한다.

Latest substring rule이란 결론적으로 RHS를 구하기 전까지 이용했던 웹 페이지를 나타내는 LHS의 suffix들을 말하게 된다. 이 규칙들은 결국 RHS를 구하기전의 가장 중요한 정보는 이미 요구한 웹 페이지들 중에서 가장 최근에 요구한 웹 페이지와 웹 페이지들 요구 순서에 있다고 보는 것이다. 다음 표2는 Latest substring rule의 예를 보여주며 LHS의 마지막 정보인 F는 항상 속하는 것을 볼 수 있다.

LHS	RHS	Latest Substring Rule
B, D, E, F	C	<B, D, E, F>→C, <D, E, F>→C, <E, F>→C, <F>→E

표2. Latest substring rule

3.3 Pessimistic Selection Method

6단계에서 예측을 구하기 위한 테스트 경우의 LHS와 4단계에서 구성해 놓은 규칙들의 LHS가 같은 것들이 하나 이상으로 나타나는 경우 이들 규칙들 중에서 어떤 규칙을 선택해야하는지의 기준이 필요하게 된다. 이 기준이 되는 것을 rule selection 방법이라고 하며, 본 논문에서는 여러 가지 selection 방법 중 가장 효과적인 pessimistic selection 방법을

이용한다.

pessimistic selection 방법[4]은 사람이 인위적으로 최소 지지율을 설정해 줄 필요성이 없으며 지지도와 신뢰도를 결합하여 이용하므로 새로운 selection의 기준을 정할 수 있다.

Pessimistic-Error Estimate는 이미 C4.5[4] 알고리즘에서 최대의 효율을 나타내는 검증된 방법으로, 통계적인 연구의 표본 추출로부터 기인하였다. 이 방법을 차용하여 신뢰값을 이용하는 대신, pessimistic confidence라 부르는 새로운 측정값을 정의하였다. E는 분류된 케이스들 중에서 틀린 개수를 나타내며 모든 분류된 케이스의 개수를 N으로 나타낸다. pessimistic confidence를 다음과 같이 정의한다.

$$conf_p = 1 - \frac{U_\alpha(E,N)}{N}$$

테스트 데이터에서 LHS를 가지고 찾은 규칙들이 다음의 예와 같다면:

규칙1:(C,D)→E, 신뢰도 100%, K=1, E=0

규칙2:(A,B,C,D)→F, 신뢰도 80%, K=100, E=20

여기서 K는 각각의 규칙을 지지하는 개수이고, E는 훈련 데이터에서 틀린 분류의 개수를 나타낸다. 규칙2를 보면 훈련 데이터 집합에서 적용 가능한 100개의 경우가 있고 그중 틀리게 예측한 경우가 20개이다. 신뢰 수준을 75%라고 하면, 실제 에러율을 계산하는 상한값은 $U_{0.25}(0.2, 100)$으로 나타낼 수 있다. 규칙들의 pessimistic confidence를 계산하면 다음과 같이 얻을 수 있다:

규칙1: $1 - U_{0.75}(0,1) = 25\%$

규칙2: $1 - U_{0.75}(0.2,100) = 76.57\%$

위의 두 규칙 중에서 pessimistic selection 방법은 pessimistic confidence 값이 큰 규칙2를 선택하며 더욱 믿을 수 있다고 여긴다. 따라서 예측 값으로 F를 선택하게 된다.

4. 실 험

실험을 위하여 실제 데이터들을 이용하였다. 그래서 Florida에 있는 NASA Kennedy Space Center WWW server로부터 NASA 웹 로그 데이터를 가져왔다. 이 데이터 집합은 Florida에 있는 NASA Kennedy Space Center WWW server의 모든 HTTP 요구들로 한 달간의 자료를 포함하고 있다. 이 웹 로그 데이터는 1995.8.1 00:00:00으로부터 1995.8.31. 23.59:59까지 수집된 것이다. 이 시기에는 총 1,569,898번의 요구가 있었으며, 총 72,151개의 서로 다른 IP가 접속했으며, 총 119,838개의 세션을 가지고 있으며, 총 2,926개의 다른 페이지들이 요구 되었다. 실험을 하기 전에 우리는 CGI script 같이 동적으로 발생되는 내용들은 모두 제거했다. 데이터 집합을 우리는 두 부분의 웹 로그로 분리하여, 규칙을 만들기 위한 훈련 부분, 평가를 위한 테스트 부분으로 나누었다. 훈련 부분은 1-26일까지의 웹 로그를 이용하였으며, 테스트 부분은 27-31까지의 웹 로그를 이용하였다.

훈련 데이터 집합	100,000	150,000	200,000	250,000
테스트 데이터 집합	25,000	37,500	50,000	62,500

표3. 실험에 이용한 훈련, 테스트 데이터 집합의 크기

표3에서 보면 테스트 데이터 집합 크기는 훈련 데이터 집합 크기의 1/4 크기로 정한 뒤 precision을 구한다. 테스트 데이터를 적용해서 구한 예측 값과 실제로 요구하는 웹 페이지가 같은 경우인 정확한 예측의 수를 C, 모든 테스트 케이스의 수를 N으로 나타내며, precision은 다음과 같이 정의한다.

$$precision = \frac{C}{N}$$

다음 그림2는 이 결과를 보여준다.

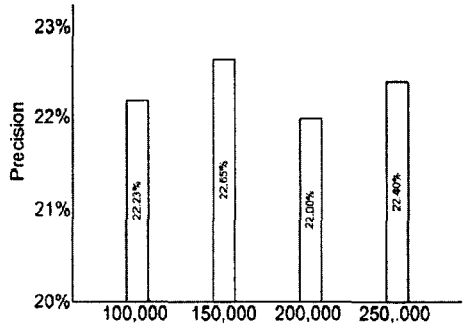

그림2. 훈련 데이터 집합 크기에 다른 precision

5. 결론 및 향후 연구

본 논문에서는 웹 로그 데이터를 하나의 방문순서열이 아닌 트리 구조를 이용하여 예측하는 알고리즘을 제시 하였다. 세션별로 분리해 놓은 데이터에서 경로를 추출하면서 불필요한 경로를 미리 정리했기 때문에 규칙을 구성함에 있어서 정확성을 높여 규칙의 개수를 줄일 수가 있었다. precision은 전방향 참조 경로 탐사 패턴 추출 알고리즘[1]을 적용하지 않았을 때와 비슷한 결과를 나타내었다. 향후에는 전방향 참조 경로 탐사 패턴 추출 알고리즘[1]에 시간과 반복되는 패턴의 횟수에 대한 개념을 도입하여 더욱 정밀하게 필요한 정보와 불필요한 정보를 구분하여 보다 정확한 경로를 추출할 수 있도록 숙고해 보고 싶다.

6. 참고문헌

[1] Ming-Syan Chen, Jong Soo Park and Philip S. Yu IBM Thomas J. Watson Research Ctr., "Data Mining for Path Traversal Patterns in a Web Environment", International Conference on Distributed Computing Systems, pp. 2-4, 1996.

[2] Qiang Yang, Tianyi Li and Ke Wang, "Building Association-Rule Based Sequential Classifiers for Web-document Prediction", Data Mining and Knowledge Discovery, vol. 8, no. 3, pp. 253-273, May 2004.

[3] Tianyi Li, "Web-document Prediction and Presending using Association Rule Sequential Classifiers", Simon Fraser University, July 2001.

[4] J.R.Quinlan, "C4.5:Programs for Machine Learning", Morgan Kaufmann, 1993.