웹 로그 화일에서 순회 패턴 탐사를 위한 시스템†

박 종수[0] 윤 지영
성신여자대학교 컴퓨터정보학부
{jpark, jyyoon}@cs.sungshin.ac.kr

A System for Mining Traversal Patterns from Web Log Files

Jong Soo Park[0] and Ji Young Yoon
School of Computer Science and Engineering, Sungshin Women's University

Abstract

In this paper, we designed a system that can mine user's traversal patterns from web log files. The system cleans an input data, transactions of a web log file, and finds traversal patterns from the transactions, each of which consists of one user's access pages. The resulting traversal patterns are shown on a web browser, which can be used to analyze the patterns in visual form by a system manager or data miner. We have implemented the system in an IBM personal computer running on Windows 2000 in MS visual C++, and used the MS SQL Server 2000 to store the intermediate files and the traversal patterns which can be easily applied to a system for knowledge discovery in databases.

1. 서 론

최근 웹(WWW) 비즈니스가 급증하면서 웹을 이용하는 사용자의 사용 패턴에 대한 관심이 고조되고 있다. 그것은 인터넷이 고객과 서비스 제공자간에 밀접한 상호작용이 이루어지는 매체이기 때문이다. 이미 발생된 사건에 따라 저장되거나 처리된 데이터에서 유용한 정보를 추출해내는 것을 데이터 마이닝이라 하고[1,2,3], 이것의 한 분야로 웹을 사용하는 환경에서 중요한 패턴 정보를 찾아내는 것을 웹 마이닝이라 부른다[4, 5, 6, 7, 8, 9, 10, 11].

웹 마이닝의 대상이 되는 웹 데이터로는 여러 가지가 있지만, 그 중에서도 사용자가 웹 페이지를 접근할 때마다 그에 관련된 정보가 기록되는 웹 로그 화일이 많이 사용되고 있다. 이것은 웹을 이용하는 사용자들에 대한 의미있는 행동 패턴을 찾아내는데 있어 매우 중요한 자료가 된다. 사용자가 웹 페이지를 탐방하는 일정 순서를 순회라 하고, 순회 패턴(Traversal Pattern)이란 일정 수 이상의 사용자들이 공통적으로 순회하는 웹 페이지의 순서를 일컫는다[4]. 웹 로그 화일에서 빠르게 패턴을 찾아내는 알고리즘의 연구도 중요하고[4,6,10], 입력 데이터에서 결과 패턴을 보여주는 전체적인 시스템의 연구도 중요하다[9,11]. 이 논문에서는 연구실에서 개발된 패턴 탐사 알고리즘을 기본으로 하여 전체적인 순회 패턴 탐사 시스템에 대하여 기술하고 있다.

이 논문의 구성은 다음과 같다. 제 2장에서는 순회 패턴 문제를 살펴보고, 제 3장에서는 제안한 순회 패턴 탐사 시스템의 설계 및 구현에 대하여 설명한다. 제 4장에서는 실제 웹 로그 화일에서의 순회 패턴 탐사 실험과 그 결과에 대하여 밝히고, 마지막으로 제 5장에서는 결론 및 향후 과제로 끝을 맺는다.

2. 순회 패턴 문제

웹(WWW) 환경과 같이 문서 또는 객체들이 서로 연결되어있는 정보 제공 환경 하에서는 사용자들이 제공된 링크(Links)와 아이콘(Icons)에 따라 앞뒤로 객체를 순회하는 연속적인 행동을 하게 된다[4]. 순회 패턴 탐사에서 문제 정의는 다음과 같다. 웹 환경 하에서 사용자의 일반적인 접근 패턴을 보면 형제 노드를 방문하기 위해서 새로운 링크를 클릭하는 대신 역방향(Backward) 아이콘을 누르고 링크를 통해 다시 순방향(Forward)으로 형제 노드를 선택한다[4]. 이러한 관점에서 이 논문의 순회 패턴 탐사에서 역방향 참조는 주로 내용을 가져오기보다는 순회의 편리를 위해 행한다고 가정하고 역방향 참조가 일어나면 순방향 참조 경로는 종결된다고 정의한다. 그리고 이렇게 얻어지는 사용자의 순방향 참조 경로를 최대 순방향 참조(Maximal Forward Reference)라 칭한다[4]. 순회 패턴 탐사는 일반적으로 다음과 같은 3단계를 가진다[4,10,11]:

단계 1: 원시 데이터베이스에서 최대 순방향 참조들을 구한다.
단계 2: 최대 순방향 참조 집합에서 빈발 참조 시퀀스를 구한다.
단계 3: 빈발 참조 시퀀스에서 최대 참조 시퀀스를 구한다.

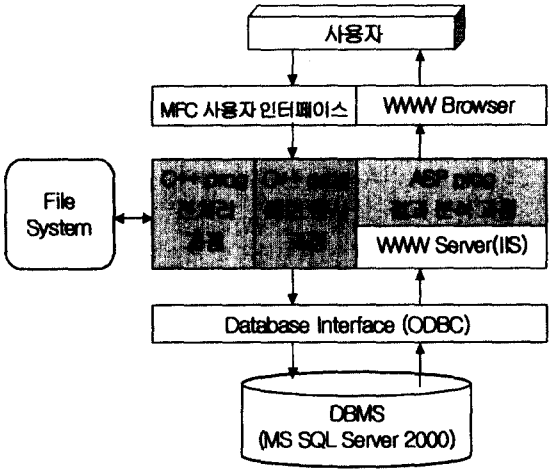
위의 3단계를 설명하면 다음과 같다. 만약 원시 데이터베이스로부터 한 사용자가 순회한 시퀀스 $<s_1, ..., s_n>$이 주어지면 이것을 각각 최대 순방향 참조를 나타내는 여러 부분 시퀀스들로 나눌 수 있다. 최대 순방향 참조를 구하는 자세한 내용은 [4,10]을 참조한다.

모든 사용자에 대한 최대 순방향 참조들을 저장한 데이터베이스 D_F가 만들어지면 D_F안에서 빈발하게 발생되는 빈발 참조 시퀀스들을 찾고, 그 빈발 참조 시퀀스들 가운데서 다른 빈발 참조 시퀀스 어느 하나에도 포함되지 않는 최대 참조 시퀀스를 찾아냄으로써 최종적인 순회 패턴을 구한다. 빈발 참조 시퀀스에 대한 정의는 다음과 같다. $1 < j < k$ 에 대해 $s_{i+j} = r_j$ 인 i 가 존재한다면 시퀀스 $<s_1, ..., s_n>$은 연속된 부분 시퀀스로서 $<r_1, ..., r_k>$를 포함한다고 말한다. 예를 들어 시퀀스 BAHPM은 시퀀스 AHP를 포함한다. 연속된 부분 시퀀스인 $<r_1, ..., r_k>$를 포함하는 최대 순방향 참조들을 가진 사용자의 수가 최소 지지도 이상이면 이러한 k-참조 시퀀스 $<r_1, ..., r_k>$를 빈발 k-참조 시퀀스라 부른다.

† 이 논문은 과학기술부 지원 연구 과제에 의한 연구 결과의 일부임(연구과제번호: 00-B-WB-05-A-02).

3. 순회 패턴 탐사 시스템 설계 및 구현

실험에 사용된 PC는 Intel P-Ⅲ 600 MHz Dual CPU, 512MB 주 메모리, Windows 2000 서버 운영체제로 구성되어 있고, 제안된 순회 패턴 탐사 시스템은 MS Visual C++ 6.0을 사용하여 구현하였다. 또한 사용자 인터페이스를 위해 MS Visual C++ 6.0의 각종 MFC를 이용하였고, 결과 패턴들은 ASP 프로그램을 이용하여 사용자가 웹 브라우저를 통하여 분석할 수 있게 설계하였다. 각 단계별 결과를 저장할 DBMS로는 MS SQL Server 2000을 사용하였다. [그림 1]은 연구된 시스템의 개발 환경을 나타내고 있다.



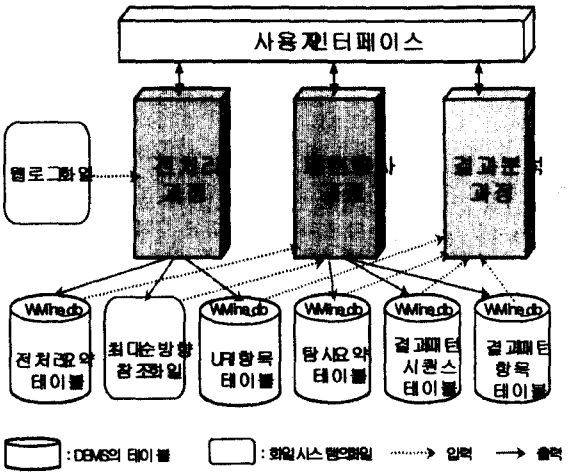
[그림 1] 시스템 개발 환경

이 논문에서 제안한 순회 패턴 탐사 시스템은 그 처리 과정을 기능에 따라 크게 전처리 과정, 패턴 탐사 과정, 그리고 결과 분석 과정으로 구분하여 각각 하나의 시스템 안에서 독립적으로 수행되도록 모듈별로 설계하였다. 각 과정의 역할과 주요 기능은 [표 1]과 같다. 최대 순방향 참조는 웹 로그 화일의 종류에 따라 달라지므로 전처리 과정에 포함시킨다.

전처리 과정	웹 로그 화일을 패턴 탐사에 적합한 형태로 변환하는 과정	- 데이터 정제 - 정렬 및 사용자 트랜잭션 식별자 부여 - URI 항목 식별자와 사용자 식별자 부여 후 웹 트랜잭션 로그 화일에 대한 데이터 인코딩 - 최대 순방향 참조 생성
패턴 탐사 과정	순회 패턴 탐사 과정	- 빈발 참조 시퀀스 생성 (CHT_FS 또는 TPADE 알고리즘 이용) - 최대 참조 시퀀스 생성
결과 분석 과정	탐사된 패턴 분석 과정	- 최대 참조 시퀀스의 각 항목 식별자에 대한 데이터 디코딩 - 탐사 결과의 시각화

[표 1] 순회 패턴 탐사 시스템의 과정별 역할과 주요 기능

[그림 2]는 구현된 시스템의 연계도를 데이터베이스 시스템의 테이블과 화일 시스템과의 데이터 흐름을 나타내고 있다. 이 시스템에서는 관련 데이터의 저장 및 관리를 관계형 데이터베이스로 사용함으로서 데이터 관리의 안정성을 기할 수 있고, 그리고 이런 데이터를 표준 데이터 변환을 통해 다른 시스템의 입력 데이터로도 사용하여 지식 탐사에 응용할 수 있다.



[그림 2] 순회 패턴 탐사 시스템의 각 과정간 연계도

이 시스템의 패턴 탐사 과정에서는 전처리 과정에서 생성된 최대 순방향 참조 화일을 읽어 빈발 참조 시퀀스를 찾은 후 최대 참조 시퀀스를 찾는 과정을 수행하게 된다. 이 과정에서 사용하는 탐사 알고리즘은 연구실에서 개발된 CHT_FS 알고리즘과 TPADE 알고리즘이다[10]. 한 알고리즘을 지정한 뒤 전처리 과정에서 생성시킨 전처리 이름을 선택하여 탐사할 최대 순방향 참조 화일을 결정하고 패턴 탐사 결과 이름을 입력하여 WMine_db의 탐사 요약 테이블에서 기본 키로 사용한다. 마지막으로 최소 지지도를 입력한 뒤 실행 버튼을 눌러 패턴 탐사를 수행하고, 탐사 작업이 완료되면 결과 보기 버튼을 눌러 웹 환경에서 직접 패턴 탐사 결과를 분석할 수 있게 설계되었다. 지정된 알고리즘으로 패턴 탐사가 끝나면 생성된 빈발 참조 시퀀스들 중에서 최대 참조 시퀀스를 찾아냄으로써 최종적인 순회 탐사 패턴을 얻게 된다.

패턴 탐사가 끝난 후 결과보기 버튼을 누르면 웹 브라우저가 호출되어 웹 환경에서 결과를 분석할 수 있게 하였다. 사용자가 패턴 탐사 결과를 분석하고자 하면 패턴 탐사 과정에서 사용자가 입력한 패턴 탐사 결과 이름을 선택한다. 그러면 해당 순회 패턴 시퀀스들이 시퀀스 길이별, 지지도 순으로 URI 항목 식별자 형태로 화면에 나타나고, 맨 하단에 패턴 탐사 결과 이름, 최소 지지도, 최대 참조 시퀀스의 총 개수가 출력된다. 사용자가 각 결과 시퀀스 내의 URI 항목 식별자를 실제 URI 이름으로 보고자 하면 URI 이름보기 버튼을 누르면 된다. 그러면 시스템은 ASP 프로그램을 통해 SQL 질의를 수행하여 데이터를 디코딩한 후 실제 URI 이름으로 보여준다.

4. 실험 및 결과 분석

이 논문에서는 구현한 시스템의 성능을 실험하고자 '(주)물가와 실적산 정보'(www.pricedw.com) 웹 서버의 웹 로그 화일 가운데 2000년 5월 23일부터 2000년 11월 22일 사이에 생성된 웹 로그 화일의 로그 데이터를 실험 데이터로 사용하였다. 이 웹 로그 화일은 IIS 웹 서버의 W3C 확장 로그 화일 형식으로 저장된 것으로서 일일단위로 생성된 183개의 개별 화일이며 그

크기는 462MB이고 약 8,000,000개의 웹 로그 레코드로 구성되어 있다.

4.1 전처리 과정의 결과 분석

[표 2]에서와 같이 전처리 과정을 통해 생성된 인코딩된 웹 로그 트랜잭션 화일의 로그 레코드 수는 1,225,632개로서 원본 웹 로그 화일의 약 15%정도로 감소한 것을 알 수 있었다. 이는 원본 로그 화일에 기록된 약 8,000,000번의 액세스 로그 중 85%는 실제 페이지 뷰와 상관없는 정보라는 것을 의미한다고 할 수 있다. 또한 전체 화일 크기는 462MB에서 22MB로 줄어 처음의 약 5%로 감소한 것을 확인했는데 이는 정제된 데이터를 다시 정수형 식별자로 인코딩한 결과라고 할 수 있다.

따라서 실험 데이터에 대한 전처리를 통하여 실제 패턴 탐사에 앞서 데이터 처리비용을 95% 정도 감소시킬 수 있었음을 확인할 수 있었다.

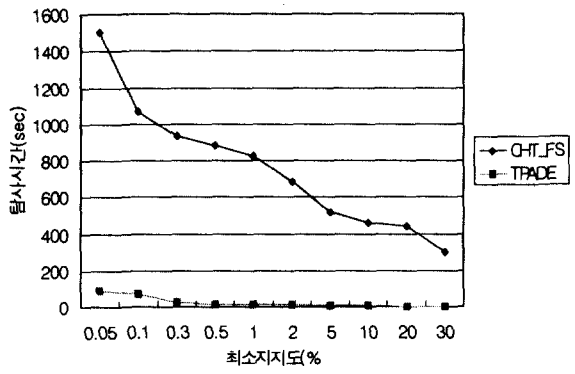
	원본 웹 로그 화일	인코딩된 웹 로그 화일	비고
로그 레코드 수	약8,000,000개	약1,200,000개	약 15%로 감소
전체 화일 크기	462MB	22MB	약 5%로 감소

[표 2] 원본 웹 로그 화일과 인코딩된 웹 로그 트랜잭션 화일의 비교

4.2 패턴 탐사 과정의 결과 분석

전처리 과정을 통하여 생성된 최대 순방향 참조 화일은 13MB로서 사용자 수(=사용자 트랜잭션 개수)는 81,458명, URI 항목 개수(=사용자들이 접근한 웹 페이지들의 개수)는 13,367개, 트랜잭션 수(=최대 순방향 참조 개수)는 259,100개 였다. 이 최대 순방향 참조 화일에 대하여 CHT_FS 알고리즘과 TPADE 알고리즘을 각각 이용하여 패턴 탐사를 한 결과 TPADE는 최소 지지도 0.03%까지, CHT_FS는 0.05%까지 실험할 수 있었다.

[그림 3]은 CHT_FS와 TPADE의 최소 지지도별 패턴 탐사 실험을 통해 탐사 시간을 측정한 그래프이다. 이 실험을 통하여 이 시스템은 최소 지지도를 상당히 작게 주었을 때에도 안정적인 패턴 탐사를 할 수 있음을 알 수 있었다.



[그림 3] CHT_FS와 TPADE의 최소 지지도별 패턴 탐사 실험

5. 결론

이 논문에서는 웹 로그 화일로부터 웹 사용자 트랜잭션 데이터베이스를 추출하여, 사용자들이 순방향으로 접근하는 일정한 패턴을 찾아내는 순회 패턴 탐사 시스템을 설계하고 구현하였다. 이 순회 패턴 탐사 시스템은 크게 전처리 과정, 패턴 탐사 과정, 그리고 결과 분석 과정의 세 모듈로 설계되었다.

전처리 과정에서는 웹 로그 데이터를 정제하고 인코딩하는 작업을 통하여 순회 패턴 탐사에 대한 입력 데이터를 얻고, 패턴 탐사 과정에서는 순회 패턴 탐사 알고리즘인 CHT_FS와 TPADE 알고리즘을 적용하여 패턴 탐사를 한다. 그리고 결과 분석 과정에서는 디코딩 과정을 거쳐서 찾아낸 패턴을 분석할 수 있도록 순회 패턴을 이루는 해당 페이지들을 웹 브라우져를 통하여 볼 수 있는 기능을 제공한다. 구현한 시스템을 실험하고자 웹 데이터베이스 서비스를 제공하고 있는 회사의 웹 서버에 생성된 웹 로그 화일을 구하여 실험하였다. 패턴 탐사 과정의 실험을 통하여 구현된 시스템이 최소 지지도를 아주 작게 주었을 경우에도 안정적인 패턴 탐사를 할 수 있는 시스템임을 확인할 수 있었다.

향후 과제는 다음과 같다. 먼저 다양한 웹 로그 화일의 특징을 보다 자세히 연구하여 더 유용한 데이터로 가공할 수 있도록 전처리 과정을 강화하는 것과 대용량의 웹 로그 화일에 대한 실험을 통해 시스템의 견고성을 높이는 작업이 필요하다. 또한 찾아진 결과 패턴을 효율적으로 분석할 수 있도록 사용자 인터페이스를 보다 효과적으로 개발하여야 한다. 나아가 지식 탐사 시스템 및 웹 데이터 웨어하우징 시스템과 연계할 수 있는 확장된 웹 마이닝 시스템에 대한 연구가 필요하다.

참고문헌

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[2] J. S. Park, M.-S. Chen and P. S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules", In Proceedings of ACM SIGMOD, pages 175-186, 1995.

[3] Mohammed J. Zaki, "Efficient Enumeration of Frequent Sequences", In Proceedings of the 7th ACM CIKM, pp. 68-75, Nov. 1998.

[4] M.-S. Chen, J. S. Park and P. S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE TKDE, Vol. 10, No. 2, pp. 209-221, Mar. 1998.

[5] A.G. Bchner and M.D. Mulvenna "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", ACM SIGMOD Record, 27(4):54-61, 1998.

[6] J. Pei, J. Han, B. Mortazavi-Asl and H. Zhu: "Mining Access Patterns Efficiently from Web Logs", in Proceedings of PAKDD'00, 2000.

[7] J. Srivastava, R. Cooley, M. Deshpande and P-T. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, (1) 2 : 12-23, 2000.

[8] 박종수, "웹 로그 화일에서 빈발 항목집합 탐사", 성신여자대학교 기초과학연구지, 17권, pp. 73-88, Feb. 1999.

[9] 김지현, 최영란, 박종수, 지원철, 최지윤, "웹 로그 화일에서의 순차 패턴 탐사", '99 한국데이터베이스 학술대회 논문집, pp.29-35, Feb. 1999.

[10] 하미라, "웹 로그 화일에서 순회 패턴 탐사 알고리즘", 성신여자대학교 석사 학위 논문, Feb. 2000.

[11] 윤지영, "웹 로그 화일에서 순회 패턴 탐사 시스템 설계 및 구현", 성신여자대학교 석사 학위 논문, Feb. 2001.