효율적인 트리구조 기반의 웹 사용 마이닝 기법

김상영, 박병준
광운대학교 컴퓨터과학과
{ ssang1024, bjpark }@kw.ac.kr

A Web Usage Mining Technique based on an Efficient Tree Structure

Sang Young Kim, Byung Joon Park
Department of Computer Science, Kwangwoon University

요 약

웹 사용 마이닝은 웹에서 발생하는 모든 데이터를 분석 대상으로 삼는다. 그 중 사용자들이 어떠한 경로를 통해 웹 사이트를 빈발하게 참조하며 이동했는지에 대한 패턴을 찾아 분석하는 것은 중요한 문제이다. 그래서 웹 사용 마이닝에서는 사용자들의 웹 로그 파일을 이용하여 사용자들의 패턴을 트리구조로 만들어 효율적으로 관리하고 마이닝 알고리즘을 통해 빈발 패턴을 분석한다.

본 논문에서는 이러한 트리로 구조화된 사용자들의 패턴 구조를 이용하여 보다 효율적이고 빠른 패턴 발견 마이닝 알고리즘을 제안한다. 효과적으로 사용자들의 패턴을 트리구조로 구조화한 CATS Tree를 이용하여 사용자들의 빈발 패턴을 추출하는 FELINE 알고리즘을 개선한 효율적인 빈발 패턴 추출 알고리즘인 FPM 알고리즘을 제안하고, 개선된 성능을 보여주는 실험 결과를 제시한다.

1. 서 론

World Wide Web 환경에서의 대용량의 데이터로부터 필요한 관련 정보를 탐색하고, 다양한 형태의 정보로부터 지식을 창출하는 일은 매우 어려운 일이다. 많은 데이터를 반복적으로 검사하여 가능한 많은 연관성들 중에서 일정 수준 이상의 신뢰도와 지지도를 가지는 규칙들만을 뽑아내는 일은 많은 메모리와 많은 디스크 엑세스 그리고 많은 시간을 필요로 하는 일이다. 데이터 마이닝은 이러한 어려운 정보 추출을 하는 방법으로 최근 정보에 대한 가치 인식이 높아지면서 많은 연구가 진행되어 왔다.

이러한 대용량의 데이터로부터 잠재적인 정보를 추출하는 방법으로 데이터 마이닝 기법이 있다. 그 중 웹 환경에서의 사용자들의 패턴을 발견하는 방법이 웹 마이닝이다. 데이터로부터 정보를 추출하는 마이닝 기법으로는 연관규칙(Association Rule), 순차 패턴(Sequential Patterns), 분류(Clustering), 구분(Classification) 등으로 분류 되며 연관규칙은 사용자의 여러 패턴들 사이에서의 연관성을 발견하는 기법이고, 순차패턴은 시간 순서적으로 발생한 패턴들 사이의 연속성을 발견하는 기법이고, 분류는 데이터의 여러 속성들을 비교하여 서로 유사한 항목들 간의 집합을 의미 하며, 구분은 새로운 데이터 발생 시 기존의 구성된 집합 유형 중 어디에 속할 것인가를 예측하는 기술이다.

본 논문에서는 이러한 마이닝 기법 중 하나인 연관 규칙을 이용하여 사용자들 사이의 패턴들의 연관성을 가지는 빈발 패턴을 추출하는 알고리즘을 제시하고 연관규칙 알고리즘 중 웹 데이터에 대한 효율적인 저장구조를 가지고 있는 CATS Tree[4]를 기반으로 하여 빈발 패턴을 추출하는 FELINE 마이닝 알고리즘[4]을 개선하여, 마이닝에 정보를 빠르게 분석, 유지 할 수 있도록 한다.

본 논문의 총 5장으로 구성되어 있다. 2장에서는 관련 연구에 대해 소개하고, 3장에서는 본 논문에서 제안하는 FPM 알고리즘에 대해 소개를 하며 FELINE 알고리즘과 비교한다. 4장에서는 두 알고리즘 간의 성능비교 및 실험을 통한 결과를 도출한다. 마지막으로 5장에서는 결론과 향후 연구 과제에 대해서 논한다.

2. 관련 연구

웹 마이닝[1]은 웹 컨텐츠 마이닝, 웹 구조 마이닝, 웹 사용 마이닝의 3가지 영역으로 분류 될 수 있다. 웹 사용 마이닝은 웹 서버 로그로부터 사용자들의 접속 유형을 자동으로 찾아주어 사용자들이 자주 방문한 웹 페이지와 평균 접속 시간, 자주 발생되어지는 정보, 웹 트래픽과 같은 정보 제공과 사용자들의 웹 페이지 방문 데이터로 웹 페이지 간의 연관성과 빈발하게 발생되어지는 패턴을 제공하므로 웹을 사용 할 때 그 안에 잠재된 유용한 정보를 효율적으로 제공 할 수 있다.

웹 사용 마이닝은 그림 1과 같이 웹 서버의 로우데이터를 전처리 단계(Preprocessing), 패턴 발견 (Pattern Discovery), 패턴 분석(Pattern Analysis)의 3단계를 거쳐 진행 될 수 있다.