데이터 마이닝 기술을 이용한 웹 분석 시스템의 개발

전재범, 양성모, 윤석호, 김상욱
한양대학교 정보통신대학
e-mail : powerory@hanmail.net

Development of a Web Analyzing System based on Data Mining Techniques

Jae-Bum Jun, Sung-Mo Yang, Seok-Ho Yoon, Sang-Wook Kim
College of Information and Communications
Hanyang University

요 약

최근 웹 분석 시스템은 단순히 통계 정보를 이용한 분석을 넘어서서 웹 마이닝 기술을 이용한 웹 분석 시스템의 형태로 변화하고 있다. 이는 기존의 단순 통계 분석으로는 점차 거대하고 복잡해져가는 현재의 웹 사이트를 분석하는 데 한계가 있기 때문이다. 따라서 앞으로 웹 분석 시스템은 웹 마이닝 기술을 활용한 다양한 측면의 연구와 구현이 이루어 질 것으로 보인다. 본 연구에서는 기존의 웹 마이닝 기술을 이용한 웹 마이닝 분석 시스템을 구현하여 웹 마이닝 기술에 대한 분석과 응용을 고찰한다. 또한, 실제로 한양대학교 웹사이트를 대상으로 웹 분석 시스템을 설계 구현함으로써 웹 마이닝 기술을 이용한 웹 분석 시스템의 가능성을 타진한다.

1. 서론

컴퓨터 기술과 인터넷의 발달로 인해서 웹사이트의 수는 기하급수적으로 증가하고 있다. 웹사이트 수의 급격한 증가는 새로운 비즈니스 모델을 형성시키고 있으며, 웹사이트를 이용한 수익의 극대화를 위해서 웹사이트의 분석이 시도되고 있다[3,4,5,7].

웹사이트의 로그에서 통계적 방법을 이용해 유용한 정보를 추출하는 것은 이러한 시도들 중에 하나이다. 통계적 방법은 웹사이트를 분석하는데 유용한 정보를 주었지만, 점차 거대하고 복잡해져가는 웹사이트를 분석하기 위해서는 좀 더 깊이 있는 분석이 요구된다[7].

이러한 요구로 인해 데이터 마이닝 기술을 기반으로 한 웹 마이닝 분석 시스템이 등장하고 있다. 웹 마이닝 분석 시스템은 단순 통계정보가 주는 지식의 한계를 넘어서 웹사이트 방문자들의 행동 패턴 등과 같은 고급 정보를 제공하며, 웹사이트의 구조 정보와 함께 분석되어서 웹사이트의 구조적 문제를 해결하는 데에도 사용된다[3,5]. 본 연구에서는 한양대학교 웹사이트(www.hanyang.ac.kr)를 대상으로 웹 마이닝 분석 시스템을 구현해보고, 앞으로의 웹사이트 분석을 위한 웹 마이닝 시스템의 방향을 타진하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 구현하고자 하는 웹 분석 시스템의 기반이 되는 웹 마이닝에 대해서 알아본다. 제3장에서는 분석 대상이 되는 한양대학교 웹사이트를 소개하고 제공하고자 하는 지식 및 요구사항을 정의한다. 제4장에서는 웹사이트 분석을 위한 본 연구의 접근 방안에 대해서 논의한다. 제5장에서는 제안된 접근 방안의 요구 사항 만족 여부와 응용에 대해서 알아본다. 그리고 마지막으로 제6장에서는 웹 분석 시스템의 방향을 제시하며 결론을 내린다.

2. 웹 마이닝

2.1 웹 마이닝의 정의와 과정

웹 마이닝은 데이타 마이닝 기술을 이용해서 자동적으로 웹사이트로부터 유용한 지식을 추출하는 방법이다. 웹 마이닝은 웹사이트의 증가와 전자상거래의 활성화로 더욱 중요시 되고 있는 기술이다[8].

웹 마이닝은 총 네 단계로 진행된다. 먼저, 전처리(preprocessing)는 로그 정보를 분석이 용이한 상태

로 정제, 가공하는 단계며, 다시 데이터 정제(data cleaning), 방문자 식별(user identification), 세션 식별(session identification), 경로 완성(path completion)의 단계로 구성된다. 데이터 정제는 gif 같은 그림 파일 및 cgi 등의 불필요 파일을 제거하는 과정이다. 방문자 식별은 각 방문자별로 로그에 있는 정보를 구별해 주는 단계로 ip 주소와 브라우저 정보를 통해 방문자별로 로그 정보를 구분해 주는 과정이다. 세션 식별은 각 페이지 간의 시간의 차이를 이용해서 각 방문자들의 로그 정보를 논리적인 작업단위인 세션으로 나누는 과정이다. 경로 완성은 캐시로 인하여 로그에 남지 않는 페이지를 유추해서 방문자가 이동한 페이지들의 경로를 완성하는 과정이다.

둘째, 트랜잭션 식별(transaction identification)은 데이터 마이닝을 하기 위해 페이지들을 그룹화 시켜서 분석 목적에 의미 있는 단위인 트랜잭션 데이터로 세션을 나누거나 병합하는 단계이다. 트랜잭션 식별에는 참조 길이(reference length), 최대 전진 길이(maximal forward length), 시간 간격(time window)과 같은 요소를 사용한다.

셋째, 규칙 마이닝(rule mining)은 데이타 마이닝 기술을 이용하여 의미 있는 패턴을 뽑아내는 단계이다. 규칙 마이닝에는 연관 규칙 마이닝(association rule mining)과 순차 패턴 마이닝(sequential pattern mining) 등이 사용되는데 이 기술에 대해서는 2.2절에서 자세하게 알아본다.

끝으로, 패턴 분석(pattern analysis) 단계는 규칙 마이닝 단계를 통해서 얻게 된 여러 가지 패턴들을 평가하고 얻은 지식이 유용한지를 판별하는 작업이다[2,6].

2.2 대표적 규칙 마이닝 기술들

연관 규칙 마이닝은 페이지간의 상관관계를 분석하기 위하여 사용된다. 두 페이지 A, B가 있을 때, 두 페이지간의 상관관계 여부를 알기 위해서는 페이지 A, B가 동시에 포함되는 트랜잭션이 전체 트랜잭션의 몇 퍼센트인지(지지도), 그리고 페이지 A를 포함하는 트랜잭션 중 페이지 B가 포함될 확률은 몇 퍼센트인지(신뢰도) 계산한다. 이 지지도와 신뢰도가 미리 정해진 값보다 크다면 그 두 페이지 A, B는 상관관계가 있다고 판정한다. 이때 페이지 A, B의 상관관계 여부를 알아내는 것이 목적이기 때문에 페이지 A, B의 순서는 상관이 없다. 따라서 연관 규칙 마이닝은 트랜잭션 안에 있는 페이지를 정렬 한 후에 분석을 수행한다[2,6].

순차 패턴 마이닝은 연관 규칙 마이닝과 매우 유사하나, 페이지 액세스 순서를 고려한다. 즉, A와 B, B와 A는 다른 액세스 패턴으로 간주한다. 또한, 순차 패턴 마이닝은 신뢰도 없이 지지도만을 계산해서 그 패턴의 유용성을 판정한다. [1,2,6].

3. 요구사항 분석

3.1. 분석 대상 사이트

본 연구에서는 한양대학교 웹사이트를 분석 대상으로 한다. 한양대학교 웹사이트는 2001년 이후부터 현재까지 약 62,800,000회의 충분히 많은 방문이 이루어진 웹사이트이다. 많은 수의 카테고리와 페이지가 있어서 방문자들의 사용 패턴과 웹사이트의 구조적 문제를 분석 할 수 있는 좋은 대상이다[9].

3.2 분석 목표

한양대학교 웹사이트에서 분석하고자 하는 목표는 방문자의 편의성을 위한 분석과 웹사이트의 구조적인 문제 분석 두 가지로 크게 구분될 수 있다.

방문자의 편의성을 위한 분석은 컨텐츠 페이지들간의 연관 분석과 이탈페이지에서 이탈하지 않는 방문자들의 네비게이션 패턴 분석으로 구분된다. 컨텐츠 페이지 분석은 방문자들이 필요로 하는 정보와 관심사를 파악하기 위한 분석으로 많은 방문자들이 주로 보는 페이지들의 상관관계를 분석하는 것이다. 이탈페이지에서 이탈하지 않는 방문자들의 네비게이션 패턴 분석은 방문자들이 주로 어떤 페이지에서 웹사이트를 이탈하는지 조사하고 이탈하지 않는 방문자들의 네비게이션 패턴을 분석한다.

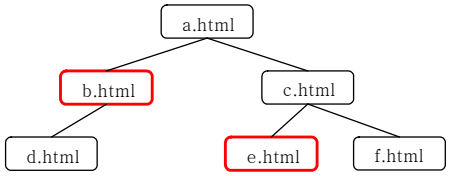
웹 사이트 구조 분석은 컨텐츠 페이지까지의 경로 분석과 연관된 컨텐츠 페이지간의 위치 분석으로 구분된다. 컨텐츠 페이지까지의 경로 분석은 방문자들의 네비게이션 패턴을 분석해서 웹 사이트의 네비게이션 구조가 바르게 구성되어 있는지 분석한다. 연관된 컨텐츠 페이지간의 위치 분석은 연관성이 높은 페이지들을 분석하고 웹사이트 구조와 비교 분석함으로써 웹 사이트의 컨텐츠의 분류가 잘 구성되어 있는지 분석한다.

4. 접근 방안

4.1 연관 규칙 마이닝을 통해 웹사이트 분석

4.1.1 연관된 컨텐츠 페이지간의 연관 분석

연관 규칙 마이닝을 통해서 연관된 컨텐츠 페이지들을 분석한다. 연관된 컨텐츠 페이지들은 방문자들이 일반적으로 한 세션에서 함께 보는 페이지들이 집합이다. 따라서 내용의 상관성이 높다. 그림 1는 연관 규칙 마이닝을 통해서 찾은 연관된 페이지들을 표시하고 있다.

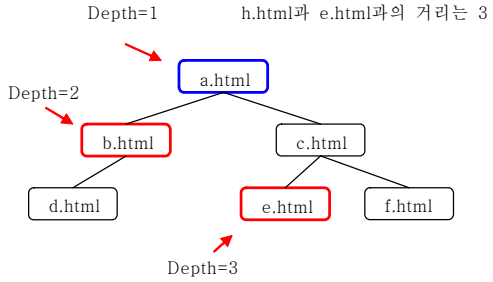


(그림 1) 연관된 페이지들.

4.1.2 연관된 컨텐츠 페이지간의 위치 분석

웹사이트는 일반적으로 연관성이 높은 페이지들을 카테고리별로 구분해 놓는다. 따라서 연관 규칙 마이닝을 통해서 얻은 연관된 컨텐츠 페이지들은 이러한 웹사이트의 일반적이 규칙에 맞게 구분되어 있어야 한다. 하지만 웹사이트의 페이지들에 대한 분류가 잘못되어 있는 경우에는 연관된 컨텐츠 페이지들이 다른 카테고리에 포함되어 질 수 있다. 따라서 연관 규칙 마이닝을 통해 얻은 연관된 켄텐츠 페이지들과 웹사이트 구조정보를 비교해서 연관된 컨텐츠 페이지들이 다른 카테고리에 포함되어 있는 문제를 찾을 수 있다.

실제적인 구현은 연관된 페이지들 간의 거리를 측정함으로 계산되어진다. 두 페이지간의 거리는 각 페이지의 깊이를 더한 값에서 두 페이지를 모두 포함하는 페이지들 중에서 가장 깊이가 깊은 페이지의 깊이를 빼준 값이다. 예를 들면, 그림 2에서 b.html과 e.html은 각각 깊이가 2와 3이다. 두 페이지를 포함하는 깊이가 가장 깊은 페이지가 a.html이므로 1×2을 빼준 값, 즉 3이 b.html과 e.html의 거리이다. 두 페이지간의 거리가 컨텐츠 페이지간의 평균거리보다 긴 경우에 두 연관된 컨텐츠 페이지들의 위치에 문제가 있다고 판단한다.



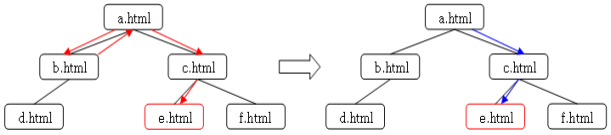
(그림 2) 연관된 컨텐츠들의 거리 측정.

4.2 순차 패턴 마이닝을 통해 얻을 수 있는 지식

4.2.1 컨텐츠 페이지까지의 경로 분석

순차 패턴 마이닝을 통해서 방문자들의 일반적인 웹사이트 네비게이션 패턴을 찾는다. 네비게이션 패턴과 웹사이트 구조정보로 알 수 있는 컨텐츠 페이지까지의 최단 경로 정보를 서로 비교한다. 이러한 비교 분석은 방문자가 컨텐츠 페이지를 방문하기 위한 네비게이션 패턴과 웹사이트 구조상의 최단 경로와 유사한지를 분석함으로써 웹사이트의 구조적 문제를 발견한다. 단, 방문자들이 여러 컨텐츠 페이지들을 동시에 방문할 수 있으므로 4.1절에서 찾은 연관된 컨텐츠 페이지와 함께 분석한다.

예를 들어, 그림 3에서 왼쪽에 있는 그림은 실제 방문자의 네비게이션 패턴이고 오른쪽은 웹 사이트 구조로 본 최단 경로이다. 만약, e.html이 컨텐츠 페이지고 다른 페이지들은 보조 페이지라면 방문자들이 a.html에서 b.html으로 이동 후 다시 a.html에서 c.html로 이동하는 것은 구조적 문제일 가능이 있다. 하지만 만약 b.html이 컨텐츠 페이지이고 e.html과 연관된 페이지라고 한다면 웹사이트의 구조적 문제는 없다고 판단한다.

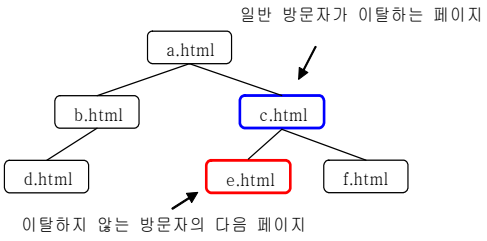


(그림 3) 최단 경로와 실제 방문자 네비게이션 패턴의 비교.

4.2.2 이탈페이지에서 이탈하지 않는 방문자들의 네비게이션 패턴 분석

먼저, 각 컨텐츠 페이지에서 웹사이트를 이탈하는 방문자들의 수를 계산하여 페이지들의 이탈율을 구한다. 이탈율이 높은 페이지가 포함되어 있는 순차 패턴을 찾는다. 단, 이때 이탈율이 높은 페이지가 순차 패턴의 가장 끝에 나타나는 경우는 방문자들의 이탈 패턴이므로 분석 대상에서 제외시킨다. 이렇게 구해진 패턴은 이탈율이 높은 페이지에서 방문자들이 이탈하지 않고 계속적으로 웹 사이트를 이용하는 패턴이 된다.

예를 들어, 그림 4에서 방문자들의 일반적으로 이탈페이지가 c.html이라고 할때, c.html이 포함된 순차 패턴이 a.html -> c.html ->e.html이라고 가정하자. 일반적인 방문자는 c.html을 방문한 후 웹사이트를 빠져나가지만 일부 방문자는 c.html을 방문한 후 e.html을 방문한다. 이런 경우는 일부 방문자가 어떤 이유인지는 몰라도 c.html과 e.html이 연관이 있다고 생각하는 경우이다.



(그림 4) 이탈 페이지에서 계속적으로 이용하는 방문자의 다음 페이지.

5. 응용

제4장에서 본 연구에서 목표로 했던 네 가지 요구사항을 만족하기 위한 접근 방안을 알아보았다. 컨텐츠 페이지 분석은 방문자들이 함께 보는 페이지들을 파악해서 방문자들이 연관된 페이지들을 쉽고 빠르게 볼 수 있도록 관리자에게 링크를 추천하고 방문자들이 좀 더 쉽게 연관된 페이지를 볼 수 있도록 목록을 보여 줄 수 있다. 이는 웹사이트 방문자들의 편의성을 향상시킬 수 있다.

방문자의 이탈 페이지를 분석하고 이탈 페이지에서 이탈하지 않는 방문자들의 네비게이션 패턴을 분석했다. 실제적으로 웹사이트 관리자가 방문자들의 정보 이용률을 높이기 위해서 이탈하지 않는 방문자들이 이탈 페이지 이후에 보는 페이지들을 추천함으로써 이루어 질 수 있다. 또한, 이탈 페이지 이후에 보는 페이지는 웹사이트 관리자가 이탈페이지의 문제점을 보완하기 위한 참조 페이지로써의 역할도 할 수 있다.

컨텐츠 페이지까지의 경로 분석은 방문자들이 컨텐츠 페이지까지 네비게이션 패턴을 분석하여 웹 사이트가 가지고 있는 구조적 문제를 관리자에게 알려주며, 방문자가 쉽게 볼 수 없는 링크를 찾아 낼 수 있다. 방문자들이 원하는 컨텐츠를 보다 쉽고 빠르게 찾게 함으로써 웹사이트의 이용률을 증가시킬 수 있다.

연관된 컨텐츠 페이지간의 위치 분석을 통해서 관리자는 컨텐츠가 카테고리에 맞게 구성되어 있는지를 확인할 수 있으며, 방문자들이 연관된 컨텐츠 페이지들를 중심으로 새로운 카테고리를 구성할 수 있게 도와준다. 또한, 새로운 컨텐츠 페이지의 추가 시 이를 통해서 새로운 컨텐츠 페이지의 위치를 결정할 수 있게 돕는다.

6. 결론

본 연구에서는 웹 마이닝 기술을 이용하여 웹 분석 시스템을 구현함으로써 기존의 통계 위주의 분석 시스템에서 찾을 수 없었던 유용한 정보를 제공하는 방안에 관하여 논의하였다. 특히, 제안된 방안은 방문자에 대한 사이트 이용의 편의성 제공 및 잘못된 사이트 구조 개선에 큰 도움이 된다. 웹 분석 시스템에서 사용 가능한 웹 마이닝 기법에는 여러 가지가 있으며 이를 실제 이용하는 웹 분석 시스템이 점차 증가하고 있다. 그러나 아직 이렇게 발견된 유용한 정보를 관리자가 쉽게 이해하고 적용시키기에는 부족한 면이 많다. 따라서 차후의 연구에서는 추출한 유용한 패턴을 효과적으로 관리자에게 전달 할 수 있는 기법에 대해서 연구하고자 한다.

감사의 글

본 연구는 정보통신부 및 정보통신진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음
(IITA-2005-C1090-0502-0009)

참 고 문 헌

[1] 김지현, 최영란, 박종수, "웹 로그 화일에서의 순차 패턴 탐사", 『한국정보과학회 데이타베이스 연구회 동계 논문 발표집』, 1999, pp. 29-35.

[2] B. Mobasher, N. Jain, E. H. Han, and J. Srivastava, “Web Mining: Pattern Discovery from World Wide Web Transactions,” Technical Report TR-96050, Department of Computer Science, University of Minnesota, Minneapolis, 1996.

[3] C. C. Aggarwal and P. S. Yu, "An Automated System for Web Portal Personalization," In Proceedings of the 28th VLDB Conference, 2002. pp. 1031-1040.

[4] F. Liu, C. Yu and W. Meng, "Personalized Web Search by Mapping User Queries to Categories," In Proceedings of the 11th International Conference on Information and Knowledge Management, 2002. pp. 558-565.

[5] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, Vol. 3, No. 1, 2003. pp. 1-27.

[6] R. Cooley, B. Mobasher, and J. Srivastava, “Data Preparation for Mining World Wide Web Browsing Patterns,” Knowledge and Information Systems, Vol. 1, No. 1, 1999. pp. 5-32.

[7] R. Kohavi and R, Parekh, "Ten Supplementary Analyses to Improve E-commerce Web Sites," Proceedings of the Fifth WEBKDD Workshop: Webmining as a Premise to Effective and Intelligent Web Applications, 2003.

[8] U. Fayyad, S. Djorgovski, and N. Weir, "Automating the Analysis and Cataloging of Sky Surveys," In Advances in Knowldge Discovery and Data Mining, 1996. pp. 1-34.

[9] www.hanyang.ac.kr, 한양대학교 홈페이지.