제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

클릭스트림 분석을 이용한 사용자별 메뉴 생성 시스템

박종배, 유남현, 김원중
순천대학교 컴퓨터과학과
e-mail:kwj@sunchon.ac.kr

Personalized Menu Creation System On Clickstream Analyzing

Jong-Bae Park, Nam-Hyun Yoo, Won-Jung Kim
Dept. of Computer Science, Sunchon National University

요 약

인터넷 사용 환경이 편리해지고 일상생활화되면서 웹 사이트 운영자들은 다양한 컨텐츠 제공과 차별화된 웹 서비스의 제공을 위해 노력하고 있다. 본 논문에서는 웹 사이트에서 발생하는 사용자의 클릭스트림 정보를 사용자별로 구분하여 분석하고, 분석된 클릭스트림 정보를 이용해 사용자에게 가장 관심있는 컨텐츠를 효과적으로 전할할 수 있는 사용자별 메뉴를 생성하는 시스템을 구현하였다. 웹 사이트에서 사용자가 자주 이용하는 컨텐츠 메뉴를 별도로 구성하여 제공함으로써, 사용자가 필요로하는 컨텐츠를 접근하기 위해 소요되는 시간을 절약할 수 있을 뿐만 아니라, 연관성있는 컨텐츠 메뉴를 함께 제공함으로서 사용자 중심의 차별화된 서비스를 제공할 수 있다.

1. 서 론

인터넷을 이용한 정보의 공유, 활용 및 전자상거래가 활성화되면서 많은 양의 웹 컨텐츠가 웹 사이트에 서비스되고 있으며 지금도 계속 증가하고 있다. 사용자가 많은 양의 웹 컨텐츠 중에서 자신에게 필요한 정보를 빠른 시간 안에 정확하게 찾는 것은 쉽지 않은 일이다. 또한 웹 사이트가 많은 양의 컨텐츠를 보유하고 있다고 해도 다른 경쟁사이트와의 차별화가 힘들어지게 되었다. 웹 사이트에서 서비스되고 있는 정보에 비해 사용자가 필요한 자료는 극히 적기 때문이다. 따라서 관심도가 높은 정보제공이나 컨텐츠에 접근을 용이하게 하기 위한 최적화된 관련 메뉴제공 등 웹 서비스의 차별화가 요구되고 있다.

다른 웹 사이트와 차별화 시키는 방법 중의 하나는 각각의 사용자별 선호도에 따라 서비스를 선택할 수 있도록 하여 사용자들이 필요한 정보를 보다 쉽게 찾을 수 있게 하는 것이다. 이런 개인화(Personalization)된 웹 사이트는 사용자에게 무엇을 원하는지 직접 물어보는 것이 아니라, 사용자들의 특성을 통계학 및 인공지능 기법을 활용하는 데이터 마이닝을 통해 분류 또는 세그먼트화하여 개인이 원하는 것을 예측하여 서비스하는 것이다[1][3].

웹 사이트의 운영자들은 개인화된 웹 사이트를 구축하여 차별화된 서비스를 제공하기 위해 웹 서버의 로그파일과 사용자별 클릭스트림을 수집해서 분석한다. 클릭스트림 데이터를 분석하면 사용자가 무엇을 좋아하고 무엇을 싫어하는지를 비롯해 사용자의 행동에 대해 많은 것을 빠르게 파악할 수 있으며, 사용자가 필요로 하는 정보 제공 및 사용자 성향에 따른 메뉴제공 서비스를 하여 기존 사용자 유지 및 새로운 사용자 확보 등 웹 투자의 효율성을 크게 높일 수 있다[4][5].

본 논문에서는 거대화된 웹 사이트에서 익명의 사용자를 구별하고 사용자별 클릭스트림 정보를 분석하여 많은 양의 컨텐츠 중에서 개인화된 사용자별 메뉴를 생성하는 시스템을 구현하였다.

2. 관련연구

2.1 웹 로그

웹 서버는 사용자의 웹 서비스에 대한 요청과 제공하는 정보를 로그 파일에 저장하게 된다. CLF(Common Logfile Format)는 NCSA (National Center for Supercomputing Applications) 계열의 웹 서버에서 사용하는 표준 로그 파일 형식은 Access Logfile 또는 Transfer Logfile이라고 불리며 Host, AuthUser, Time, Request, Status, Volume과 같은 필드들로 구성되어 기록된다[2].

웹 사용자들의 클릭스트림 정보를 분석하기 위해 Access Logfile을 이용하여 데이터 수집이 가능하지만 Batch로 작업이 이루어져야 하기 때문에 실시간 분석이 불가능하다는 단점이 있다. 본 논문에서는 이러한 단점을 해결하기 위해 웹 페이지에 로그 추출을 위한 Event Extractor 코드를 삽입하여 클릭스트림 정보를 데이터베이스에 저장하고 실시간으로 분석하였다.

2.2 웹 마이닝

웹 마이닝은 데이터 마이닝의 한 분야로서 사용자들의 클릭스트림의 집합체인 로그파일을 이용하여 사용자들의 패턴을 분석하는 기술이다. 웹 마이닝은 크게 Web Content Mining, Web Usage Mining, Web Structure Mining의 세 가지로 분류 된다[3].

2.2.1 Web Content Mining

Web Content Mining은 웹 사이트의 컨텐츠, 자료, 문서로부터 유용한 정보를 찾아내는 과정을 말한다. 정보 검색 부문에서 자연어 처리 시스템이나 개인화된 웹 에이전트에 이용될 수 있고, 데이터베이스 부문에서는 데이터베이스에 축적된 데이터 중에서 원하는 자료를 쉽게 찾을 수 있게 하는 구조화된 질의 언어를 개발하는데 이용된다.

2.2.2 Web Usage Mining

Web Usage Mining은 웹상에서 사용자의 패턴을 발견하고 분석하는 과정이며 전처리(Preprocessing), 패턴 발견(Patten Discovery), 패턴 분석(Pattern Analysis) 과정으로 나눌 수 있다. 사용되는 데이터는 웹 로그, 쿠키정보, 컨텐츠 데이터, 사용자의 마우스 클릭을 포함하고 있다. 이러한 데이터를 이용해서 사용자의 접속 패턴이나 관심부분에 따라서 개인화 맞춤 서비스, 시스템 개선과 웹 마케팅 등에 적용할 수 있다.

2.2.3 Web Structure Mining

Web Structure Mining은 웹 사이트와 웹 페이지의 하이퍼 링크 정보를 데이터 마이닝 과정을 통해 구조화시키고, 표준화 시키는 프로세스를 말한다. 링크된 하이퍼 링크에 근거하거나 관리자의 컨텐츠 생성단계에서의 분류로 범주화(Categorization)와 군집화(Clustering) 작업을 수행함으로서 유사한 웹 페이지들을 분류한다. 본 논문에서 메뉴 생성 시스템을 적용한 웹 사이트는 CMS (Contents Management System)에 의해 개발되고 관리되고 있기 때문에 대부분 컨텐츠 생성단계에서 그룹핑 작업을 거쳐 유사 컨텐츠를 분류하였다.

2.3 클릭스트림

클릭스트림(Clickstream)은 웹 사이트에서 사용자의 이동 경로를 분석할 수 있는 모든 행위 패턴 정보를 말한다. 사용자가 웹 사이트에 접속한 순간부터 웹 페이지를 이동하기 위해 마우스를 이용하여 클릭 이벤트를 발생시키는 일련된 흐름의 정보들이다.

3.메뉴 생성 시스템

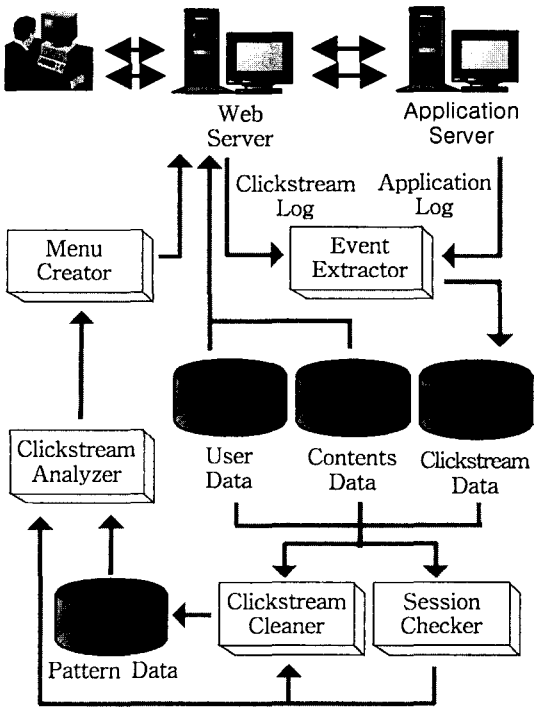
본 논문에서는 웹 사이트에서 발생하는 웹 서버 로그, 웹 어플리케이션 로그와 Event Extractor에서 추출된 사용자별 클릭스트림을 분석하는 시스템을 구축하고 사용자에게 가장 적합한 정보를 효과적으로 제공하는 메뉴 생성 시스템을 구현하였다. 먼저 필요한 모든 정보를 웹 로그에 저장하고 있다고 해도 전처리 과정을 거쳐야 한다[4]. 전처리 과정에서는 세션을 이용한 사용자 식별(User Identification)과 사용자별 클릭스트림 데이터중에서 분석에 필요하지 않은 데이터를 제외시키는 데이터 정제(Data Cleaning) 과정을 거쳐야 한다.

3.1 사용자 식별

사용자별 클릭스트림을 저장하기 위해서는 사용자 식별이 이루어져야 한다. 하지만 대부분의 웹 사용자들은 개인정보 유출 등의 문제로 인해 자신들의 익명성을 원하기 때문에 인증이 꼭 필요한 곳에서만 인증절차를 거치게 된다. 그리고 표준 웹 로그만으로는 사용자를 완전히 구분할 수 없으므로 정확한 사용자 구분을 위해서 사용자가 웹 사이트에 접속하

면 고유한 세션 ID를 부여하고 Event Extractor에 의해 추출된 클릭스트림 정보를 Session Checker가 분석하여 Pattern Data에 저장한다.

세션은 사용자가 접속을 시작한 시점부터 접속을 종료한 시점까지를 말한다. 세션이 종료되기 전까지는 동일한 사용자에 의한 레코드로 처리하지만 사용자가 일정한 시간동안 아무런 응답이 없을 경우 새로운 세션 ID를 부여하고 다른 사용자로 구분하여 처리한다. 사용자가 필요에의해 인증절차를 거친 후에는 사용자 ID로 새로운 세션을 생성하고 인증 이전에 저장된 클릭스트림 정보 중에서 세션 ID를 사용자 ID로 업데이트 시켜 동일한 사용자로 처리하게 된다.



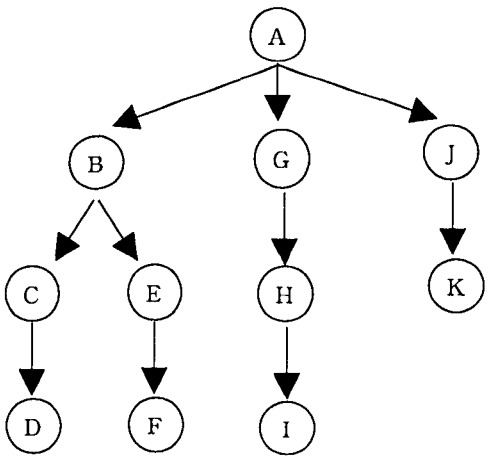
[그림 1] 사용자별 메뉴 생성 시스템

3.2 데이터의 정제와 클릭스트림 분석

Access Logfile에서 Request는 사용자가 요청한 실제 컨텐츠 파일 이름이 저장된다. 하지만 웹 페이지에 포함되어 있는 부수적인 gif 또는 jpg 등의 그림파일까지도 Logfile에 저장된다. 본 논문에서 구현한 메뉴 생성 시스템에서는 Event Extractor가 웹 페이지 URL 정보만을 추출하여 데이터베이스에 저장하였기 때문에 부수적인 파일에 대한 정제과정은 생략할 수 있다. 그러나 저장된 모든 웹 페이지 URL 정보를 사용자별 메뉴 생성 시스템의 Menu Creator 입력 데이터로 사용하지 않는다. Clickstream Analyzer는 사용자가 찾고자하는 컨텐츠 URL 정보만을 분석해서 일정시간 이상의 체류시간을 가진 웹 페이지 URL 정보만을 Menu Creator에게 입력 데이터로 전달한다. 체류시간은 사용자가 브라우저를 통해 특정 웹 페이지에 실제로 머문 시간을 말한다. 체류시간이 매우 짧은 웹 페이지는 사용자가 이 페이지를 실수로 열었거나 실제 찾고자하는 웹 페이지까지의 이동 경로 역할을 하는 경우가 대부분이다[5].

본 논문에서는 Menu Creator를 위한 입력 데이터 정제를 위해 Clickstream Cleaner에 의해 분류된 사용자별 패턴 데이터를 Clickstream Analyzer가 체류시간의 평균을 구하고 웹 페이지의 데이터 크기와 비교하여 분석하였다.

[그림 2]는 Clickstream Analyzer가 분석한 사용자 패턴을 트리형식으로 나타낸것이다.



[그림 2] 분석된 사용자 패턴 트리

[표 2]는 사용자의 패턴 트리중 하나를 추출하여 패턴 데이터의 평균 체류시간과 데이터의 크기를 나타낸 것이다. 사용자는 A⇒B⇒C⇒D 경로로 이동하였다.

[표 2] 웹 페이지별 체류시간과 데이터 크기

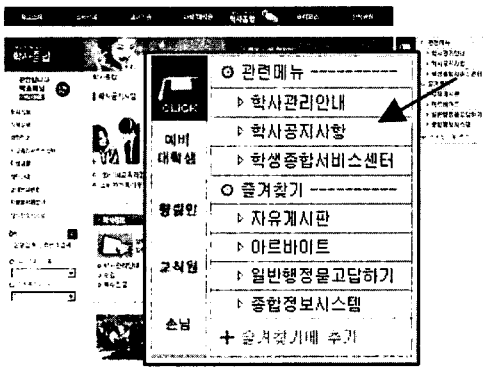
웹 페이지	체류시간	데이터 크기
A	3초	48464 Byte
B	3초	14913 Byte
C	7초	71488 Byte
D	27초	77244 Byte

A 페이지에서의 체류 시간을 계산하기 위해서는

B 페이지에 접속한 시간에서 A 페이지에 접속한 시간을 빼면 된다. 각각의 웹 페이지 체류시간 중 데이터 크기가 비슷한 C와 D페이지의 경우 D 페이지는 평균 체류시간이 27초이지만 C 페이지에서의 평균 체류시간은 7초로 많은 차이가 있다. 따라서 A ⇒B⇒C 까지의 정보는 단지 D 페이지를 찾아오기 위한 경로로만 사용된 것을 알수 있기 때문에 D 페이지만 입력데이터로 사용된다. 그리고 사용자가 D 페이지에서 더 이상의 클릭스트림을 발생시키지 않았을 경우 세션이 Time Out되고 체류시간을 구할 수 없기 때문에 메뉴 생성 시스템의 입력 데이터로 사용하지 않고 정제하게 된다. D 페이지가 메뉴 생성 시스템의 입력 데이터로 사용되어 사용자에게 제공 되었을 경우 A⇒B⇒C⇒D 경로를 통해 D 페이지에 접근하는데 소요되는 평균 시간 13초 중 B와 C 페이지를 거치지 않아도 된다. 이 사용자의 경우 사용자별 메뉴를 제공하였을때 D 페이지에 접근하는데 평균 10초 이상의 시간이 줄었다. 또 다른 패턴 데이터에서도 사용자별 메뉴를 제공하였을 경우 경로 역할을 하는 페이지에서 소요되는 시간이 줄어드는 결과를 얻을 수 있었다.

3.3 사용자별 메뉴 생성

사용자별 메뉴 생성은 데이터베이스에 저장된 클릭스트림 데이터 중 입력 데이터로 사용 가능한 정보를 사용자별로 분류하고 Clickstream Analyzer에 의해 실시간으로 사용자별 메뉴를 생성한다.



[그림 3] 사용자별 관련 메뉴 및 즐겨찾기

관련메뉴는 현재 사용자가 브라우저를 통해 보고 있는 웹 페이지와 연관성이 있는 컨텐츠에 대한 메뉴를 출력하고, 즐겨찾기 메뉴는 사용자가 방문한 각각의 웹 페이지별 체류시간이 길고 클릭스트림 데이터중에서 방문 회수가 많은 순서로 출력하였다. 또한 새로운 웹 페이지가 추가되었을때 사용자의 방문 회수가 적어 즐겨찾기에서 제외되는 경우를 해결하기 위해 사용자가 직접 즐겨찾기에 추가할 수 있는 메뉴를 제공하였다.

4. 결 론

본 논문에서는 세션을 이용하여 사용자 식별을 하고 웹 페이지에서 발생하는 클릭스트림 정보를 데이터베이스에 저장하였다. 그리고 저장된 클릭스트림 정보를 기반으로 평균 체류시간을 계산, 정제 하고 분석하여 사용자별 메뉴를 생성하는 시스템을 구현하였다. 사용자별 메뉴 생성 시스템은 서비스되는 많은 양의 컨텐츠 중에서 사용자가 자주 찾는 웹 페이지와 연관성있는 웹 페이지에 대한 메뉴를 분류하여 제공함으로써 사용자 중심의 웹 서비스가 가능하게 되었다. 부가적으로 분석된 클릭스트림을 이용해 웹 사이트의 우수 또는 취약한 컨텐츠를 선별하여 효율적으로 관리할 수 있게 하였다.

보다 정밀한 체류시간 및 클릭스트림 정보를 얻기 위해 세션이 종료되기 전까지의 사용자의 행동을 파악할 수 있는 연구가 이루어져야 할 것이다.

참고문헌

[1] Personalization Consortium http://www.personalization.org

[2] http://www.w3.org/Daemon/User/Config/Logging.html

[3] Mobasher, B., Cooley, R., Srivastava, J., "Web Mining: Information and Pattern Discovery on the World Wide Web," In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97), November, 1997.

[4] Srivastava, J. Cooley, R., Deshpande, M. & Tan P.N. Web Usage Mining: Discovery and Application of Usage Patterns from Web Data. SIGKDD Explanations, 1. 2000.

[5] Ralph Kimball, Richard Merz, "The Data Webhouse Toolkit", 8, 2000.