사용자 행동 패턴 분석을 이용한 규칙 기반의 컨텐츠 사이트 관리 모델

김정민*, 김영지, 옥수호+, 문현정, 우용태
창원대학교 컴퓨터공학과,+고신대학교 컴퓨터과학부
{jmkim, yjkim}@ce.changwon.ac.kr, {mun, ytwoo}@sarim.changwon.ac.kr, shok@kosin.ac.kr

A Content Site Management Model by Analyzing User Behavior Patterns

Jeong-Min Lee*, Young-Ji Kim, Soo-Ho OK+, Hyeon-Jeong Mun, Yong-Tae Woo
Dept. of Computer Engineering, Changwon National University, +Kosin University

요 약

본 논문에서는 컨텐츠 사이트에서 디지털 컨텐츠를 보호하기 위하여 사용자 행동 패턴을 분석을 이용해 특이한 성향을 보이는 사용자를 탐지하기 위한 모델을 제시하였다. 사용자의 행동 패턴을 분석하기 위한 탐지 규칙(detection rule)으로 Syntactic Rule과 Semantic Rule을 정의하였다. 사용자 로그 분석 결과 탐지 규칙에 대한 위반 정도가 일정 범위를 벗어나는 사용자를 비정상적인 사용자로 추정하였다. 또한 제안 모델은 eCRM 시스템에서 이탈 가능성이 있는 고객 집단을 사전에 탐지하여 고객으로 유지하기 위한 promotion 전략 수립에 응용될 수 있다.

1. 서 론

최근에 인터넷 비즈니스 분야에서 수익성 향상을 위하여 컨텐츠 유료화를 비즈니스 모델로 채택하는 사이트가 늘고 있다. 하지만 디지털 컨텐츠는 복제가 용이하고, 복사본의 품질이 원본과 동일하게 유지되는 특징으로 인해 저작권 보호 문제가 야기되고 있다. 특히 유료 컨텐츠 사이트에서 컨텐츠 불법 복제나 아이디 도용은 사이트의 수익성을 저하시키고 정상적인 사용자에게도 피해를 입힐 수 있는 문제가 발생한다.

이에 따라 디지털 컨텐츠를 보호하기 위한 다양한 연구가 진행되고 있다. 이러한 기술은 크게 시스템 차원에서 컨텐츠를 보호하기 위한 방법과 컨텐츠 자체를 보호하기 위한 방법으로 구분될 수 있다. 시스템 차원의 보호 기술은 네트워크를 통한 불법적인 침입 탐지 기법[1], 데이타베이스 자원에 대한 접근 패턴 분석 기법[2] 등이 연구되고 있다. 그리고 컨텐츠 자체를 보호하기 위한 기술은 암호화, 워터마킹(watermarking)[3], 디지털 저작권 관리(DRM)[4], 디지털 컨텐츠 식별 시스템(DOI)[5] 등과 같은 다양한 기술이 연구되고 있다.

또한 컨텐츠 사이트에서는 고객들과 친밀한 관계를 유지하고 우수 고객에 대한 차별화 된 서비스 제공이나 이탈 가능한 고객을 사전에 탐지하기 위하여 eCRM 시스템을 활발하게 도입하고 있다[6]. 이러한 eCRM 시스템에서는 고객을 체계적으로 관리하기 위하여 고객을 적절하게 분류하기 위한 방법은 RFM 모델[7] E-metrics[8], 의사 결정 트리(decision tree)[9], 인구통계학적 정보와 클릭스트림을 이용한 방법[10] 등이 있다.

기존의 컨텐츠 보호 방식은 주로 네트워크나 데이타베이스를 통한 불법 침입이나 권한 오용을 탐지하거나 컨텐츠 자체를 보호하기 위한 연구가 대부분이다. 하지만 컨텐츠 사이트에서 정상적인 고객이 컨텐츠를 불법적으로 복사하는 행위와 같은 비정상적인 접속 패턴을 탐지하기 위한 연구는 거의 이루지지 않았다. 또한 eCRM 시스템에서 고객들의 접속 패턴을 이용하여 고객을 분류하기 위한 연구가 부족하다.

본 논문에서는 컨텐츠 사이트에서 디지털 컨텐츠를 보호하기 위하여 로그 기록에 남겨진 고객들의 행동 패턴을 분석하여 특이한 성향을 보이는 사용자를 탐지하기 위한 모델을 제시하였다. 일반적으로 컨텐츠 사이트를 이용하는 정상적인 고객들은 인구통계학적인 특징이나 관심도에 따라 일정한 형태의 접속 패턴을 가지게 된다. 하지만 컨텐츠를 불법적으로 복제하려는 의도를 가지고 접속하는 사용자는 정상적인 사용자와 다른 특이한 접속 패턴을 보일 수 있다.

또한 본 모델에서는 eCRM 시스템에서 고객들의 행동 패턴을 분석하여 고객을 분류하기 위한 새로운 방법을 제시하였다. 컨텐츠 사이트의 고객은 자주 접속하면서 수익을 많이 올려주는 우수 고객, 평균적인 고객, 신규 고객 그리고 이탈 가능한 고객 등과 같이 다양한 형태로 분류할 수 있다.

본 논문에서 사용자의 행동 패턴 유형을 분류하기 위해 제안한 탐지 규칙(detection rule)은 컨텐츠의 종류나 응용 분야에 무관하게 적용할 수 있는 Syntactic Rule과 특정 분야에 종속적으로 적용 가능한 Semantic Rule으로 구분된다.

제안 모델의 효율성을 입증하기 위해 채용 정보 사이트에서 로그 기록을 분석하여 컨텐츠 접속 패턴을 수집하였다. 그리고 접속 패턴에 대한 통계적인 분석을 기반으로 Syntactic Rule과 Semantic Rule에 대한 위반 범위를 설정하였다. 이것을 기반으로 특이한 성향을 보이는 사용자를 추출하였다.

2. 기존 연구의 문제점

그 동안 인증되지 않은 사용자가 자원을 불법적으로 사용하는 행위를 탐지하거나 디지털 컨텐츠의 저작권을 보호하기 위해 네트워크 차원에서 침입 탐지 및 데이타베이스 자원의 오용에 대한 탐지 등과 같은 여러 가지 연구들이 진행되고 있다[1-2]. 최근에는 컨텐츠 자체에 보안 정보를 포함하는 연구도 활발하게 진행되고 있다[3-5]. 하지만 기존의 연구는 컨텐츠 사이트에서 정상적으로 가입한 사용자가 컨텐츠를 불법으로 복사하거나 다른 사람의 아이디를 도용하여 사용하는 등의 비정상적인 접속 패턴을 탐지하여 컨텐츠 사이트를 보호하는 방법으로 적용하기 어렵다. 또한 eCRM 시스템에서 고객들의 접속 패턴을 이용하여 고객을 분류하는 기법에 대한 연구도 부족하다. 따라서 컨텐츠 사이트에서 고객을 체계적으로 관리하기 위하여 사용자의 행동 패턴을 이용하는 연구가 필요하다.

3. 사용자 행동패턴을 이용한 컨텐츠사이트 관리모델

본 논문에서는 디지털 컨텐츠 사이트에서 사용자의 행동 패턴을 분석하여 정상적인 행동 범위를 벗어나는 특이 사용자 그룹을 추출하고, 모니터링 할 수 있는 컨텐츠 사이트 관리 모델을 제안하였다. 제안 모델에서는 사용자의 행동 패턴을 분석하기 위한 탐지 규칙(detection rule)을 Syntactic Rule과 Semantic Rule에 의해 정의하였다. 먼저, Syntactic Rule은 일정 기간 동안 사이트에 접속한 사용자의 로그 기록에서 접속시간, 접속빈도 등을 통계적으로 분석하여 특이한 형태의 접속 패턴을 탐지하기 위한 규칙이다. 이러한 Syntactic Rule은 컨텐츠의 종류나 응용 분야에 무관하게 적용할 수 있다. 그리고 Semantic Rule은 응용 분야나 사이트 성격에 따라 종속적으로 적용할 수 있는 탐지 규칙이다. Semantic Rule은 사용자의 거주 지역, 학력, 관심분야 등 인구통계학적 프로파일 정보나 사이트에서 제공하는 컨텐츠의 종류에 따라 휴리스틱(heuristic)하게 정의할 수 있다. 그리고 컨텐츠를 사용한 접속 로그를 분석하여 탐지 규칙에 대한 위반 정도가 일정 범위를 벗어나는 사용자를 특이 사용자로 추출하였다.

제안 시스템의 구성은 Detection Rule Engine, User Behavior Pattern Engine, 그리고 User Behavior Monitoring Engine으로 나누어진다. 다음 그림 1은 본 논문에서 제안한 사용자 행동 패턴 분석을 이용한 규칙 기반의 컨텐츠 사이트 관리 모델의 전체적인 구조이다.

3.1 Detection Rule Engine

컨텐츠 사이트를 이용하는 사용자들은 인구통계학적 특성이나 관심 분야에 따라 컨텐츠의 이용 형태가 다르게 나타난다.

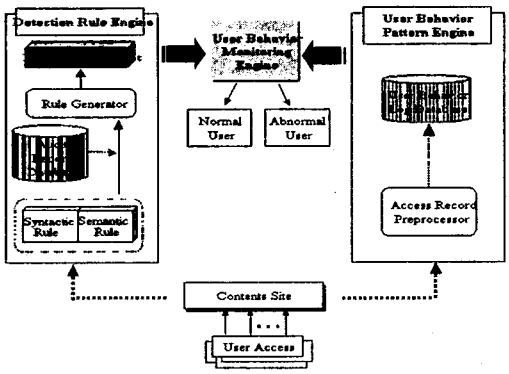

그림 1. 사용자 행동 패턴 분석을 이용한 규칙 기반의 컨텐츠 사이트 관리 모델

즉, 사용자의 관심도가 높은 종류의 컨텐츠는 조회 수가 높거나 접속 시간이 긴 반면, 관심도가 낮은 분류의 컨텐츠는 접속 시간이 짧고 이용 횟수가 적다. 하지만 컨텐츠를 불법적으로 복사하거나, 다른 사람의 아이디를 도용하여 사이트를 이용하는 사용자는 일반 사용자들과 다른 행동 패턴을 보일 가능성이 있다. 제안된 모델에서는 컨텐츠 사이트에 접속하는 사용자의 행동 패턴을 분석하기 위해 정의한 탐지 규칙은 다음과 같다.

3.1.1 Syntactic Rule

·Reading Time Rule : 사용자가 하나의 컨텐츠를 읽는 시간이 일반적인 사용자에 비해 지나치게 길거나 짧은 경우를 탐지하기 위한 규칙이다. 사용자가 컨텐츠를 읽는 시간은 해당 컨텐츠에 대한 선호도로 볼 수 있다. 이 시간이 지나치게 짧으면 선호도가 낮음을 의미할 수 있고, 반대로 지나치게 길면 로그아웃을 하지 않고 다른 작업을 하고 있을 가능성이 있다.

.Visit Frequency Rule : 사용자가 사이트를 방문하는 횟수가 일반적인 사용자에 비해 지나치게 많거나 적은 경우를 탐지하기 위한 규칙이다. 사이트 방문 횟수는 컨텐츠 사이트에 대한 사용자의 관심도를 반영하므로 빈번히 접속하는 사용자와 지나치게 접속 횟수가 적은 사용자를 분석한다.

.Session Time Rule : 사용자의 세션 시간이 일반 사용자보다 지나치게 짧거나 긴 경우를 탐지하기 위한 규칙이다. 세션 시간이 지나치게 긴 사용자는 컨텐츠를 많이 이용하거나 하나의 컨텐츠를 오래 보는 경우이다. 그리고 지나치게 짧은 사용자는 사이트에 대한 흥미를 잃어간다는 의미로 해석할 수 있다.

.Content Usage Rule : 컨텐츠의 이용 정도가 일반적인 사용자의 평균 컨텐츠 이용량 보다 지나치게 많거나 적은 경우를 탐지하기 위한 규칙이다. 일반적인 사용자보다 지나치게 많은 컨텐츠의 이용은 비정상적인 행동으로 의심해 볼 수 있다. 또한 사용량이 지나치게 적으면 이탈 가능성이 있는 사용자로 간주할 수 있다.

.Menu Usage Rule : 사이트의 사용자가 평균적으로 이용하는 메뉴의 수보다 지나치게 많거나 혹은 지나치게 적은 경우를 탐지하기 위한 규칙이다. 일반적으로 인터넷 사이트는 여러 메뉴로 구성되어 있다. 따라서 사이트의 모든 메뉴를 이용하거나 특정 메뉴만 집중적으로 사용하는 경우는 특이한 행동으로 간주할 수 있다.

3.1.2 Semantic Rule

·User Profile Mismatch Rule : 일반적인 사용자는 자신의 성별, 연령, 학력 같은 인구통계학적 프로파일 정보와 유사한 컨텐츠를 이용하는 경향이 있다. 하지만 비정상적인 의도를 가진 사용자는 자신의 프로파일 정보와 무관한 컨텐츠를 빈번히 열람하는 행동 패턴을 보일 수 있다. 예를 들면, 채용 정보 사이트에서 학사 학위 소지자가 박사 자격을 요구하는 채용 정보 컨텐츠를 자주 이용하는 경우이다.

.Content Depth Access Rule : 계층 구조를 가지는 컨텐츠 사이트에서는 하위 단계로 내려갈수록 상세한 내용을 포함한다. 또한 사용자는 관심도에 비례하여 하위 단계의 상세한 내용을 열람하게 된다. 따라서 관심도와 무관하게 모든 컨텐츠에 대해 상세 정보 단계까지 이용하는 사용자는 사이트에서 제공하는 여러 분야의 컨텐츠에 관심이 많거나 컨텐츠 도용과 같은 비정상적인 목적을 가진 사용자로 추정할 수 있다.

.Content Dependent Mismatch Rule : 서로 종속적인 관계나 순서관계가 있는 컨텐츠를 순서 없이 자주 이용하는 경우이다.

3.1.3 Detection Rule Set 생성

Syntactic Rule이나 Semantic Rule 중에서 하나의 규칙에 대해 정상적인 범위를 벗어나는 경우를 특이한 사용자 혹은 비정상적인 사용자로 판단하기 어렵다. Detection Rule Engine에서는 두 개 이상의 규칙을 결합하여 일반적인 사용자 그룹과 특이 사용자 그룹으로 사용자를 분류한다. 둘 이상의 규칙을 결합할 때 사이트의 성격 및 비즈니스 룰에 따라 특정 규칙에 대해 가중치를 부여할 수 있다.

3.2 User Behavior Pattern Engine

User Behavior Pattern Engine은 사용자가 인터넷 사이트를 이용하는 패턴을 수집하는 기능을 담당하는 모듈이다. 사용자의 행동 패턴을 수집하기 위하여 인터넷 사이트에 접속하는 시점과 로그인하여 컨텐츠를 이용하는 동안 컨텐츠에 대해 발생하는 클릭스트림 정보를 수집하여 데이타베이스에 저장한다.

3.3 User Behavior Monitoring Engine

User Behavior Monitoring Engine은 User Behavior Pattern Engine에서 수집한 데이타를 분석하여 사용자를 그룹화 하는 기능을 담당한다. 로그 데이타에 대한 분석 목적은 사용자의 행동 패턴, 인구통계학적인 프로파일 정보, 그리고 컨텐츠의 속성 정보의 일치 여부를 판단하기 위한 것이다.

4. 실험 결과 및 고찰

사용자 행동 패턴을 분석하기 위한 실험 데이타는 채용 정보 사이트인 하이브레인넷(http://www.hibrain.net)을 통해 수집하였다. 사용자 16,414명이 1개월 동안 1,679건의 채용 정보를 사용한 로그 기록을 통계적으로 분석하였다.

다음 표 1은 Audit Record에서 Syntactic Rule의 평균과 사용자별 측정값에 대해 오름차순으로 정렬하여 상·하위 10%에 해당하는 값을 나타낸 것이다.

표 1. Audit Record에서 Syntactic Rule에 의한 사용자 행동 분석 결과

Syntactic Rule(Unit)	Avg.	Lower 10%	Upper 10%
Content Usage(num.)	11.35	1≤ cu ≤2	15≤ cu ≤152
Visit Frequency(times)	3.84	vf=1	17≤ vf ≤147
Session Time(min.)	16.69	0≤ st ≤1	18≤ st ≤20143
Reading Time(sec.)	36.4	1≤ rt ≤24	81≤ rt ≤20008

Semantic Rule에 대해 Audit Record를 분석한 결과 이용자들이 전공과 불일치한 채용 정보를 평균 11.70% 정도 이용하였고, 자신의 학력과 무관한 채용 정보를 사용한 것은 13.50%의 불일치를 보였다. 전공 및 학력 두 요소 모두 불일치한 비율은 2.48%였다. 사용한 컨텐츠의 요약 레벨에 대한 상세 레벨의 비율을 분석한 결과 전체 사용자가 평균 19.22%의 채용 정보를 상세 레벨까지 이용하였다.

이와 같은 16,414명의 1개월간 로그 기록을 이용하여 탐지 규칙 집합을 생성하였다. 그리고 이 규칙에 의해 12,682명의 2주간 로그 기록을 이용하여 사용자 행동 패턴을 분석을 하였다. 실험 결과 390명을 특이 사용자로 추정할 수 있었다.

아래 그림 2는 특이한 사용자로 추출된 사용자들과 전체 사용자들의 Syntactic Rule과 Semantic Rule로 분석한 평균값을 비교한 것이다. 그림에서처럼 Reading Time을 제외하고 특이 사용자 그룹과 일반 사용자들의 컨텐츠를 사용하는 형태가 다르게 나타났다.

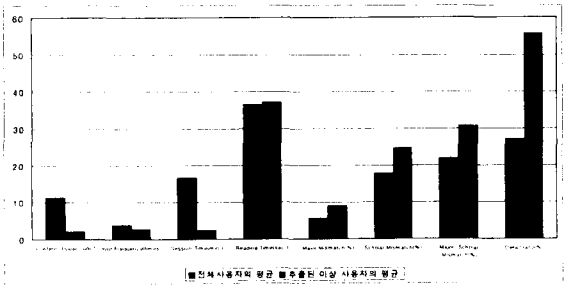

그림 2. 전체 사용자와 특이 사용자로 추출된 그룹의 비교

실제 이들 중 일부는 최근 접속일이 오래 전이고, 회원 가입시 등록한 사용자 프로파일과 모순되는 컨텐츠를 반복적으로 열람하거나 회원 가입 후 접속 횟수나 컨텐츠 이용 비율이 현저히 떨어지는 행동 패턴을 보였다.

5. 결 론

본 논문에서는 디지털 컨텐츠 사이트에서 사용자의 인구통계학적인 프로파일 정보와 컨텐츠를 이용하는 행동 패턴을 비교 분석하여 특이한 성향을 보이는 사용자를 탐지하기 위한 모델을 제안하였다. 사용자의 행동 패턴을 분석하기 위한 탐지 규칙으로 Syntactic Rule과 Semantic Rule을 정의하고 탐지 규칙에 대한 위반 정도가 일정 범위 이상 벗어나는 사용자를 특이 사용자로 추정하였다.

실험 결과 특이 사용자로 추정된 사용자들은 회원 가입시 등록한 사용자 프로파일과 모순되는 컨텐츠를 반복적으로 열람하거나 회원 가입 후 접속 횟수나 컨텐츠 이용 비율이 현저히 떨어지는 행동 패턴을 보였다. 따라서, 제안한 탐지 규칙이 디지털 컨텐츠 사이트 관리에 효과적으로 응용될 수 있음을 보였다.

또한 본 논문에서는 디지털 컨텐츠 관리를 위한 새로운 기법과 eCRM 시스템에서 고객을 분류하는 새로운 방법을 제시하였다. 제안 모델의 사용자 분류 기법에 의해 특이 사용자 그룹만 집중적으로 모니터링하여 사용자를 효율적으로 관리할 수 있다. 그리고 제안된 기법을 이용하여 유료 컨텐츠 사이트에서 컨텐츠 및 사용자를 관리하는 시스템의 개발이 가능하다.

참 고 문 헌

[1] K. Ilgun, R. Kemmerer and P. Porras, "State Transition Analysis: A Rule-Based Intrusion Detection Approach," IEEE Transaction on Software Engineering, Vol.21, No.3, pp.181-199, 1995

[2] 박정호, "데이터베이스 보안을 위한 사용자 정상행위 패턴 탐사에 관한 연구," 연세대학교 대학원, 2000

[3] N. Memon and P. W. Wong, "Protecting Digital Media Content," Communication of the ACM, Vol.41, No.7, pp.35-43, 1998

[4] 이용효, 황대준, "에이전트 기반의 동적 디지털저작권관리 시스템 설계 및 구현," 한국정보처리학회 논문지D, Vol.8, No.5, pp.613-622, 2001

[5] "디지털 유통체계," http://www.com-world.co.kr/html/200105/interent_intra/e-book-3.htm, 2001

[6] A. Berson, S. Stephen and T. Kurt, "Building Data Mining Applications for CRM," McGraw-Hill, pp.155-164, 1999

[7] Mark Sakalosky, "우수 고객을 논리적으로 선별하는 방법, R F M ," http://korea.internet.com/channel/content.asp?kid=10&nid=19890&cid=71, 2002

[8] J. Sterne and M. Cutler, "E-Metrics: Business Metrics For The New Economy," White Paper, NetGenesis Co., 2001

[9] 최중훈, 이은, 공은배, "Decision Tree를 이용한 고객 취향 관리 시스템," 한국정보과학회 추계 학술발표논문집, Vol.27, No.2, pp.60-62, 2000

[10] H. D. Rozanski, G. Bollman and M. Lipman, "Seize the Occasion: Usage-Based Segmentation for Internet Marketers," White Paper, BoozAllen & Hamilton Inc., 2001