사용자 활동과 페이지 이용 시간을 이용한 웹 페이지 평가 기법

Evaluation of Web Pages using User's Activities in a Page and Page Visiting Duration Time

이동훈[1], 윤태복[2], 김건수[3], 이지형[4]

경기도 수원시 성균관대학교 전자 전기 컴퓨터공학과
E-mail: {idoun[1], tbyoon[2], kkundi[3]}@skku.edu,
jhlee@ece.skku.ac.kr[4]

요 약

웹 사용 마이닝은 사용자의 웹 이용 패턴에 대해 분석하여 정보를 찾아내는 분야이다. 사용자에 대한 분석은 웹을 통한 비즈니스의 근간이 되고 있다. 때문에 웹 마이닝 분야에서 주목받고 중요시 되는 기술이 되었다. 그러나 최근에는 공개된 기술의 취약점을 이용해 악의적으로 정보를 교란하는 일이 발생되고 있어 사회적으로 이슈가 되고 있다. 이러한 문제는 특히 단순한 페이지 뷰 횟수에 기반을 둔 정보 추출 방식에 주로 발생하고 있다. 따라서 본 논문에서는 이러한 추출 방식의 단순함을 줄이고 사용자의 정보를 더 반영하기 위하여 페이지 이용 시간과 페이지 내의 행동을 분석하여 콘텐츠의 질을 평가하는 방안을 제시한다. 구현 부분에는 사용자의 개인정보 침해 없이 사용자의 행동을 수집하기 위하여 최근 인기를 얻고 있는 Ajax 기술을 사용하였다. 그리고 실시간으로 웹 페이지에 대한 평가를 수행하기 위해 서버에 로그 필터 모듈을 추가하는 수집 기법을 제안하였다.

Key Words : Information Retrieval, Ajax, Web Usage Mining, Activities in a Page, Page Visiting Duration Time

1. 서 론

현대의 일반적인 검색엔진들은 사용자가 입력한 질의어에 따라 가장 관련성이 높은 웹 페이지를 찾아주는 형태를 취하고 있고 이를 위해 각자의 알고리즘을 가지고 웹 페이지를 평가하고 있다[1]. 이러한 평가 순위의 상위를 차지하기 위해 전략적인 방안들이 널리 알려져 왔으며 활용되고 있다.

그러나 이러한 방법을 이용하여 상업적으로 혹은 악의적으로 조작된 검색 결과를 만드는 문제를 낳게 되었다[1]. 페이지 점수법 알고리즘의 약점을 이용한 구글 폭탄(Google bombing)은 이미 세계적인 문제로 인식되고 있다. 국내에서도 공격적인 단어들이나 사회적 이슈, 자극적인 미리보기 그림 등을 이용하여 콘텐츠의 조회 수를 높이는 일은 흔해졌고, 포털 사이트에서 제공하는 '추천 검색어 순위' 혹은 '실시간 인기 검색어'를 공격하여 변경 가능하다는 것은 이미 널리 알려져 있다. 이는 정보화 사회의 부정적인 측면으로 작용하여 검색 결과와 콘텐츠에 대한 불신으로 이어질 가능성이 크며, 정보의 공유라는 웹의 기본적인 이념을 무시한 '순위' 자체에 집착하는 부작용을 낳게 되었다.

네이버[2]가 공개하고 있는 '실시간 인기 검색어'의 산정 방법에 의하면 검색어 입력 횟수의 증가 폭이 큰 것을 순서대로 보여주며, 일정 단위시간에 하나의 IP당 하나의 키워드만 인정하고, 검색창을 통해 입력된 키워드만 대상으로 하고 있다[3]. 이러한 측정 방식은 콘텐츠의 내용과 품질에 대한 평가가 배제되어 '추천'이라는 본래의 목적을 만족시키기에 부족한 방법이다. 웹의 구조상 링크의 깊이가 깊을수록 적은 방문 횟수가 나올 수밖에 없고, 그다지 중요하지 않은 중간단계나 낮은 깊이의 페이지가 더 높은 점수를 받을 수 있다. 또한 페이지의 이해력 정도나 페이지 액세스 능력, 네트워크 환경의 차이점 등과 같은 요소들을 완전히 배제하고 있다[4].

본 논문에서는 페이지의 내용에 대한 새로운 평가 방법을 제안하고자 한다. 우선 사용자 관

심도를 반영하는 좋은 측정치중의 하나인 '페이지 이용 시간'을 평가 기준에 도입한다[5]. 페이지 이용 시간을 페이지 평가 기준에 추가하여 사용자가 페이지를 이용하는데 소모한 시간을 반영한다. 그리고 이것을 강화하는 측면에서 페이지에 대한 사용자의 활동이라는 개념을 추가한다. R. Badi[6]의 연구에 의하면 사용자가 흥미를 보이고 문서의 가치가 높을 경우 사용자는 단순히 페이지를 보는 것과 함께 페이지에 대하여 특정 활동(Activity)을 행한다고 말한다.

먼저 웹 페이지에서 사용자의 활동과 페이지 이용시간을 수집하기 위해 최근 주목받고 있는 Ajax 기술을 이용한다. 사용자 정보를 얻기 위하여 사용자의 쿠키정보를 수집하거나 로컬 컴퓨터에 로그 수집 에이전트를 설치하는 방식을 사용해 왔으나 보안적인 문제의 부각 때문에 현재로서는 적용하기 어려운 방법이다[7]. 따라서 브라우저 내부에서 동작하는 자바스크립트를 이용한 Ajax를 이용하여 개인정보 침해우려가 없는 로깅 방법을 제안하고, 시간과 페이지 이용 시간을 이용한 페이지 평가 시스템을 연구하였다.

2. 사용자 활동/시간 기반 웹 로깅 시스템

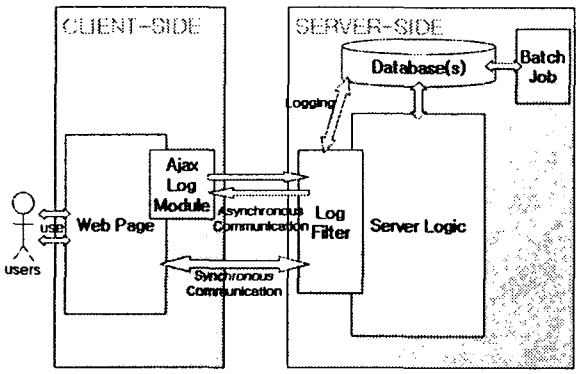

그림 1. Web Logging System Architecture

웹 사용 마이닝(Web usage mining)은 웹 로그 파일(Web log file)로부터 정보를 추출하는 것이 보통이다[8]. 그러나 실시간으로 처리하기에는 I/O에 따른 시스템적인 부하가 너무 많아 서비스에 영향이 적은 심야 시간대에 일괄작업으로 처리하는 것이 일반적이다. 그러나 사용자가 행하는 활동을 즉시 처리해야 하는 요구사항에 부합하지 못하게 되므로 실시간으로 정보를 수집하여 처리할 수 있는 시스템이 필요하게 되었다.

따라서, 제안하고자 하는 로깅 시스템에서는 서버의 웹 어플리케이션의 맨 첫 진입점에 로그 필터 모듈을 두었다. 로그 필터 모듈에서는 로그의 저장과 평가를 컨트롤하도록 하였다. 그리고 로그의 저장 대상이 로그 파일이 아닌 데이터베이스에 저장하는 방식으로 수행되게 된다. 그림 1은 실시간 로그 처리를 위해 구성한 시스템의 전체적인 아키텍처이다.

2.1 로그 필터(Log Filter)

로그 필터는 로그를 기록하기 위한 관문이다. 또한 웹 어플리케이션 서버로의 요청(Request)을 가로채 데이터베이스에 로그를 기입하고 페이지에 대한 평가를 수행하게 되는 핵심적인 부분이다. 우선, 정보 수집 대상을 확인한다. 웹에 존재하는 페이지들의 가치는 제각각 다르고 의미 없는 정보 수집을 피할 필요가 있다. 정보 수집의 대상은 모든 페이지가 될 수도 있고, 페이지 단위로도 지정가능하며, 특정 URL 패턴과 일치하는 경우로도 지정할 수 있도록 하였다. 다음으로는 로그 처리에 대한 컨트롤 역할을 한다. 그림 1에서 동기식(Synchronous Communication)으로 표현된, 통상적인 웹 브라우징 방식으로 접근하는 경우에는 해당 페이지에 대한 신규 접속으로 간주한다. 이 경우에는 데이터베이스에 새로운 접속 로그로 처리되도록 하였다. Ajax에 의한 비동기식 통신으로 사용자의 활동 정보 수집이 이루어질 경우에는 페이지에 대한 사용자의 활동 정보가 포함된 경우로 인식하고 로깅과 함께 페이지 평가 프로세스가 수행되도록 하였다. 그림2는 위에서 말한 로그 필터 처리 프로세스를 도식화 한 것으로 URL필터링 부분부터 로그필터 모듈이 시작한다고 볼 수 있다. 로그 필터 처리가 끝난 후 어플리케이션 서버에서 비즈니스 로직을 수행하게 된다.

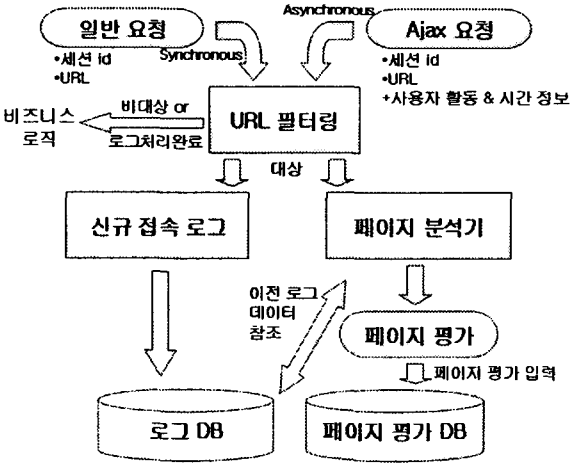

그림 2. 로그 필터 처리 프로세스

2.2 Ajax 로그 모듈(Ajax Log Module)

Ajax 로그 모듈은 웹 페이지 내부에서 일종의 정보 수집 에이전트 역할을 하게 된다. 비동기적으로 통신할 수 있는 기술인 Ajax를 이용하여 사용자가 일반적인 웹 서핑을 하는 동안에도 웹 브라우저로부터 사용자 활동 정보를 수집하도록 할 수 있다. 브라우저 내부의 정보만을 대상으로 하고 있어 개인정보 침해의 우려가 없고 속도 면에서도 빠르다[9].

우선 사용자가 수행하는 많은 행위들을 자바스크립트의 '이벤트 리스너'를 통해 모니터링한다. 이것은 원시 데이터를 생성하는 직접적인 역할을 한다. 이렇게 수집된 사용자의 활동 데이터를 일정 시간 혹은 사용자가 의미 있는 행위를 충분히 했을 때 서버에 전송해주는 역할을 하게 된다.

2.3 데이터베이스와 일괄처리(Batch Job)

로그로서 기록하기 위해 필요로 하는 값들은 아래와 같다.

- ◇ 사용자가 본 페이지 로그를 나타내는 시퀀스(PK)
- ◇ 사용자 구분을 위한 세션ID
- ◇ 대상 URL
- ◇ 페이지 흐름을 파악할 수 있도록 하는 이전 페이지의 시퀀스
- ◆ 페이지 최초 접속 시각
- ◆ 페이지의 사용자 활동정보가 조사된 마지막 시각
- ◆ 사용자의 활동 정보 셋(Set)

사용자의 구분은 웹 어플리케이션 내부에서 생성해주는 세션ID를 사용하였다. 이 세션ID와 시퀀스를 이용하여 사용자가 어떤 페이지를 어떻게 이용하는 중인지 관찰이 가능하게 된다. 여기에, 이전 페이지에서 받은 시퀀스를 함께 기록하여 어디서 유입된 것인지도 파악 가능하게 설계하였다.

일괄처리 부분에 위 정보를 활용하여 사용자의 성향, 웹 액세스 패턴 분석 등 저장된 로그로부터 추가적인 정보를 찾아내는 작업을 간단히 추가할 수 있다. 게다가 계속 증가하는 로그를 정리하기 위해 일괄처리 부분은 반드시 필요하다. 일정 기간이 흐르면 로그들을 백업받는다던가, 의미가 불분명하거나 더 이상 필요치 않는 로그는 지워나가야 한다. 웹 서비스를 하는 DB와 같이 사용할 경우에는 쌓이는 로그의 양 때문에 서비스에도 문제를 발생시킬 수 있다. 따라서 로그만을 기록하는 로그 DB를 독립적으로 구성하여 로그 저장에 대한 부분을 일임할 수 있다.

3. 웹 페이지 유용성 평가

3.1 흥미와 활동 간의 관계

사람이 어떤 웹 문서에 대해 흥미를 가지고 볼 때에는 여러 가지 활동이 수반된다는 연구결과가 있다. 예를 들어, 즐겨찾기에 추가를 하거나, 페이지를 여러 번 훑어보느라 스크롤이 자주 발생하게 된다는 것이다. 페이지를 보는데 시간을 소모하는 것도 같은 관점으로 볼 수 있다.[5][6] 따라서 이러한 반응이 많은 문서일수록 사용자의 흥미를 끌고 유용한 정보를 지니고 있다고 판단할 수 있는 것이다.

활동(Activity)	가중치
마우스 휠의 움직임	5
마우스의 왼쪽 버튼 클릭	2.5
마우스 포인터의 움직이는 정도	2.5
문서를 블록으로 선택하는 것	7
개체 끌기(Drag)	2.5
키 누름	5

표 1. 활동에 대한 가중치의 예

본 연구에서는 표 1과 같은 사용자의 특정 활동에 초점을 맞추어 문서를 평가하고자 한다. 그리고 각각의 활동에 R. Badi[6]가 제안한 가중치 방식을 적용하였다.

3.2 흥미와 시간과의 관계

문서의 가치를 인식하는 요소 중 시간은 다른 요소들에 비해 그 중요성이 높다[5][6]. 따라서 표 1에서 나열된 활동들보다 더 높은 가중치를 부여하여야 할 필요가 있다. 그러나 인터넷 환경이라는 측면을 고려할 때 사용자가 브라우저에 페이지를 열어둔 후 거의 활동 없이 일정 시간이 경과한 경우 등의 흔히 발생할 수 있는 예외상황들을 고려해야 한다. 그림 3의 그래프는 이러한 예외상황을 반영해 단순화한 그래프이다. 자신에게 맞지 않거나 잘못된

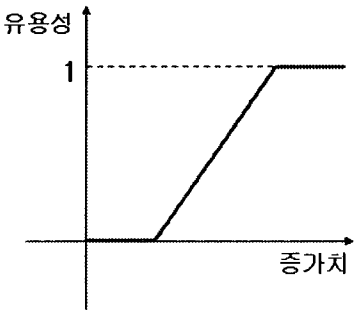

그림 3. 증가치에 대한 유용성 평가 그래프

페이지의 경우 사용자들은 미련 없이 브라우저를 닫아버린다는 점에서 초반의 유용성 점수를 낮게 평가하였다.

이러한 특징은 사용자의 활동에 대한 평가에도 동일하게 적용하였다.

3.3 웹 페이지 유용성 평가

페이지 m에 대한 수집된 활동들의 값 집합을 A라 하고,

$$Set\,A_m = \{\alpha | \alpha \in A_m\}$$

로 정의하였다. 그리고 그림3의 함수를 $f(\alpha)$라 하고 각각의 α_i에 대응하는 가중치를 ω_i라 하면, 시간을 제외한 다른 활동들의 평가 값 Θ_m는

$$\Theta_m = \frac{\sum \omega_i f_i(\alpha_{mi})}{\sum \omega_i}$$

와 같이 나타낼 수 있다. 이 값을 3.2에서 언급한 시간과 활동과의 관계로 사용한다. 이것을 이용하여 페이지 m의 평가 식을 다음과 같이 표현할 수 있다.

$$R(m) = \frac{\sum \omega_i f_i(\alpha_{mi}) + \Theta_m \omega_t f_t(\beta_m)}{\sum \omega_i + \omega_t}$$

4. 결론 및 향후과제

본 논문에서는 사회적 이슈가 되고 있는 악의적인 정보 조작에 대응하기 위한 콘텐츠의 유용성을 평가하는 방법을 제안하였다. 이를 위해 실시간으로 사용자의 활동을 수집하는 시스템을 설계하고 최근 인기를 얻고 있는 Ajax 기술을 도입하여 에이전트를 설치하지 않고서도 사용자의 활동을 관찰할 수 있는 방법을 제시하였다. 보안 문제가 시간이 중요시 되는 현 시점에서 간단한 정보 수집에는 유용한 대안이 될 것이다. 그러나 자바스크립트라는 기술에 종속적이기 때문에 한계가 존재할 수밖에 없다. 자바스크립트는 다른 서버에서 동작하는 페이지에는 접근이 불가능하여 정보 수집이 불가능하다. 따라서 세션이 공유되는 하나의 시스템에는 유용하지만 범용 검색엔진이나 다양한 CP(Contents Provider)를 보유한 포털의 경우에는 사용이 한정적일 수밖에 없다.

또한 실시간 분석이라는 목표를 위해 로그 파일을 분석하는 방법이 아니라 데이터베이스를 이용한 로그 시스템을 구축하는 방안을 도입하였다. 이러한 방법은 데이터베이스 혹은 웹 어플리케이션 서버에 부하를 발생시킬 것이라는 편견으로 꺼려져 왔던 것이 사실이다. 그러나 실제 전자상거래 프로젝트에 도입하고 이용해 본 결과[10] 일괄처리를 통해 데이터에 지속적인 관리를 해 줄 경우 큰 문제없이 사용 가능하였다. 로그 파일을 따로 처리해 줄 필요가 없고, 로그 파일로부터 추출하는 것보다 훨씬 더 정확하며 다양한 데이터를 추출할 수 있어 활용가치 측면에서도 높은 점수를 줄 수 있다.

참 고 문 헌

[1] 이우기, 신광섭, 강석호, "링크내역을 이용한 페이지 점수법 알고리즘", 정보과학회논문지:데이타베이스, 제 33권, 제 7호 pp.708-714, 2006.

[2] http://www.naver.com/

[3] http://story.nhncorp.com/

[4] 성현정, 용환승, "페이지 소요 시간을 고려한 웹 액세스 패턴 마이닝", 한국정보과학회 학술발표논문집, 제 28권, 제 2호, pp.55-57, 2001.

[5] M. Claypool, P. Le, M. Wased, D. Brown, "Implicit Interest Indicators", *Proc. of the 6th international conference on IUI,* pp.33-40, 2001.

[6] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, F. Shipman, C. C. Marshall, "Recognizing User Interest and Document Value from Reading and Organizing Activities in Document Triage", *Proc. of the 11th international conference on IUI,* pp.218-225, 2006.

[7] 최영환, 이상용, "웹 이용 마이닝을 위한 데이터 전처리에서 사용자 구분에 관한 연구", 한국정보과학회 학술발표논문집, 제 28권, 제 2호, pp.118-120, 2001.

[8] 이시헌, 이지형, "웹 사용 마이닝을 위한 퍼지 카테고리 기반의 트랜잭션 분석 기법", 성균관대학교 석사학위논문, 2004.

[9] http://www.adaptivepath.com/publications/essays/archives/000385.php

[10] http://partyngift.nate.com/