사용자 관심도를 반영한 동적 웹 문서 추천 시스템

김병진[0] 최현우 김용성
전북대학교 컴퓨터과학과
darkh9@orgio.net, hwchoi@cs.chonbuk.ac.kr yskim@moak.chonbuk.ac.kroo

Dynamic Web Documents Recommendation System Using User-Profile

Byoung-Jin Kim[0] Hyun-Woo Choi Yong-Sung Kim
Dept. of Computer Science, Chonbuk National University

요 약

인터넷 이용의 급속한 증가로 웹사이트의 증가뿐만 아니라 웹사이트 내의 웹 문서도 급속한 증가를 보이고 있다. 따라서 이를 효과적으로 사용자들에게 보여주기 위한 동적인 추천 시스템들이 많이 제안되고 있다. 그러나 이러한 추천 시스템들은 전체 사용자들의 브라우징 패턴이나 전체 웹 문서들의 연관성만을 고려하여 서비스를 제공함으로써 개인 사용자들의 관심도를 고려하지 않은 문제점이 있다.

이에 본 논문에서는 웹사이트에 남게되는 로그파일의 분석을 이용한 사용자별 브라우징 패턴과 웹 페이지의 액세스 타임의 측정을 통해, 사용자의 관심도를 측정한다. 그리고 이를 바탕으로 웹 문서들에 대해서 퍼지개념을 적용한 자동분류 알고리즘을 이용하여 사용자의 관심도가 반영된 선별된 웹 문서를 자동분류 및 선별하여 보여줄 수 있는 방안을 제시한다.

1. 서 론

월드 와이드 웹(WWW)의 발달과 빠른 대중화로 인하여 방대한 양의 정보들이 생성되면서 사용자들이 원하는 대부분의 정보들을 컴퓨터 앞에서만 앉아서도 찾아 낼 수 있게 되었다. 하지만, 정보들이 산재하게 되면서 사용자들은 원하는 정보를 찾고자 웹사이트를 찾아다니는 등의 많은 시간과 노동이 필요하게 되었다. 따라서, 많은 웹사이트들 내의 정보들을 효과적으로 검색하고자 검색엔진이 개발되고 있다. 기존의 전문화된 웹사이트들과 통합 서비스를 제공해주는 많은 포털 사이트들은 사용자들의 확보를 위하여 웹사이트 내에 사용자들이 원하는 정보들을 찾아서 서비스를 해주고 있다. 이러한 서비스도 정보의 양이 늘어나면서 사이트의 구조가 복잡해지고, 사이트 내에서도 원하는 정보를 찾아내기는 매우 힘들어 지고 있다. 또한 많은 웹사이트들이 기존의 키워드를 이용한 카테고리 검색과 키워드 검색을 지원하고 있지만, 사용자가 찾았던 정보를 다시 접근하려 하거나 비슷한 정보들을 다시 찾고자 한다면 웹사이트의 카테고리별 분류를 순차적으로 접근하거나 키워드 검색을 통해서 새로운 접근을 되풀이해야하는 어려움을 겪고 있다.

이에 본 논문에서는 사용자가 웹사이트에 접근하여 남긴 로그파일을 분석하여 사용자별 관심도를 측정하고, 측정된 관심도는 사용자별로 관련된 웹 문서의 정보 추출에 이용함으로써 사용자가 로그인할 때 관심 웹 문서 정보에 대해서 빠른 접근 시간을 제공할 수 있도록 한다.

2. 관련연구

인터넷 환경의 빠른 발전과 대중화는 오프라인상의 비즈니스가 온라인서비스로 확대되면서 웹사이트의 객관적인 분석을 위한 자료가 되는 웹 로그파일이 중요시되고 있다. 로그파일의 데이터로부터 직접적으로 얻어 낼 수 있는 접속자나 방문자 수, 시간(월,주,요일,일)별 분석, 요구된 파일별 통계 등의 정보들을 온라인 비즈니스를 위한 마케팅 정보, 시스템 성능의 향상, 효율적인 웹사이트로의 재구성, 트래픽 분석, 사용자들의 행위 패턴 분석을 위한 연구들이 진행 중이다[1][2][3].

또한 정제되지 않은 웹 데이터에는 사용자들의 축적된 경험들을 포함하는 유용한 정보들을 가지고 있기 때문에 추천 시스템들은 이러한 유용한 정보를 마이닝기법이나 다른 측정 방법을 가지고 추출하려는 노력이 시도되고 있다[4][5]. 그러나 이러한 연구들을 웹사이트측면과 효율적인 관리 측면에서만 연구가 이루어질 뿐 접근하는 개인 사용자들을 고려하지 않는 단점이 있다. 또한, 기존의 협력적 추천 시스템들은 사용자에게 평가를 요구하여 축적된 평가 정보를 가지고 추천 집합을 제공하거나, 사용자들로부터 먼저 평가를 받아야 한다는 단점을 가지고 있다. 그리고 사용자들의 브라우징 패턴 정보와 키워드를 가지고 링크된 웹 문서 중에서 사용자의 흥미를 만족할 만한 웹 문서를 추천하는 시스템[6]은 링크되지 않은 연관된 웹 문서를 추천 문서에 포함시키지 않고 있다. 또한 사용자 브라우징 패턴과 전체적인 웹 문서에서 연관 웹 문서를 사용하는 시스템[7]도 있다. 하지만 이런 시스템은 웹 문서의 형식을 고려하지 않고 사용자에게 탐색 페이지와 같은 불필요한 웹 문서까지 추천 문서로 제공하는 문제점을 가지고 있다. 이에 본 논문에서는 로그파일을 이용하여 사용자의 개인별 관심도를 측정하고 사용자에게는 개인별 관심도를 이용하여 관련된 웹 문서를 선별하여 보여주는 시스템을 설계하였다.

3. 시스템 설계

본 논문에서 제안하는 전체 시스템의 구성도는 다음 그림 1 과 같다.

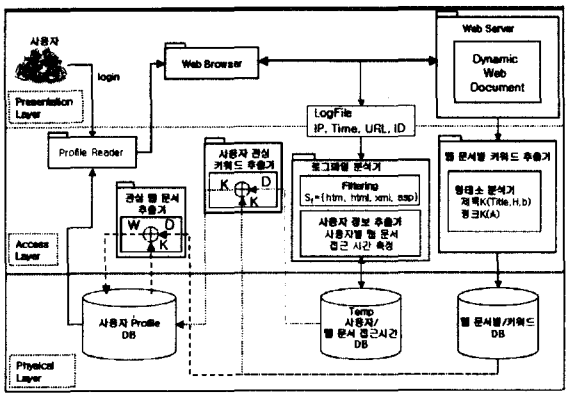

그림 1. 전체 시스템 구성도

시스템은 Presentation Layer, Access Layer, Physical Layer의 3-tier로 구성이 된다. 로그인 과정을 거친 사용자는 자유롭게 웹사이트내의 웹 페이지들을 방문하게 되고, 시스템에서는 로그파일 분석기를 통해서 개인 사용자별 관심도를 분석하게 된다. 로그파일 분석기는 효율을 높이고자 로그파일에서 원하는 정보만을 필터링하고 사용자의 관심도를 측정하게 된다. 이때 측정된 정보는 사용자별로 웹 문서 접근시간 DB에 기록하게 된다. 이에 앞서 전처리 과정으로 웹 문서의 제목과 링크에 대해 키워드를 추출한다. 추출된 키워드들은 웹 문서를 대표하게 된다. 그리고 웹 문서별 키워드 테이블을 생성하여 DB에 저장한다. 이러한 DB는 사이트가 처음에 구축될 당시와 웹 문서를 추가할 때마다 갱신된다. 사용자 관심 키워드 추출기는 이렇게 생성된 DB의 데이터와 사용자별 웹 문서 접근 시간 DB의 데이터를 통하여 사용자 Profile DB에 저장할 사용자별 관심키워드를 추출하게 된다.

사용자별로 관심키워드를 얻게 되면 다시 웹 문서별 키워드 DB를 참조해서 전체 웹문서내에서 사용자의 관심 문서를 추출하게 된다. 이렇게 추출된 웹 문서의 정보는 사용자 Profile DB에 저장되고, 사용자가 다음 로그인 과정을 거치면 사용자 Profile DB에서 Profile Reader를 통해 그 사용자의 관심 문서에 대한 내용을 보여주게 된다.

3.1. 웹 문서별 키워드 테이블 생성

먼저 웹 문서들 사이의 관계성과 사용자별 웹 문서의 관심도를 측정하기 위한 전처리 과정으로, 처음 웹사이트가 구축되어 서비스를 시작할 때와 새로운 웹 문서들이 추가될 때마다 키워드 테이블 생성이 이루어진다. 이때 키워드 추출은 형태소 분석기를 이용한다.

제목(<TITLE>,<H>,<B>)이 되는 키워드와 링크(<A>)가 걸려있는 키워드가 해당 웹 문서를 대표하게 되므로 키워드의 추출은 'TITLE', 'H', 'B' 'A' 태그내의 키워드를 중심으로 한다. 추출 후에는 문서내의 키워드 빈도수를 기록하게 된다.

3.2 사용자별 웹 문서 관심도 측정 알고리즘

사용자가 웹사이트에 접근하게 되면 그 순간부터 서버에는 사용자의 웹사이트 내에서의 모든 행위들을 분석할

표 1. 웹 문서별 키워드 테이블

웹 문서 \ 키워드	k1	k2	k3	k4	k5	…
w1	2	3	2	4	2	
w2	5	1	4	1	0	
w3	1	3	0	4	1	
w4	5	1	0	1	1	
w5	0	5	4	1	0	
⋮						

수 있는 Access log, Error log, Referer log가 존재하게 된다. 이중 사용자의 브라우징 패턴을 분석하는 주요한 웹 로그 분석 행위는 Access log를 대상으로 이루어지게 된다. 본 논문에서 사용자의 관심도 측정을 위하여 분석하고자 하는 Access log에는 요청자의 도메인 이름(IP주소), 요청한 날짜와 시간(일/월/년:시:분:초), 요청한 메소드(GET, POST), 요청한 파일의 이름(URL), HTTP 버전, 요청 상태 코드(성공, 에러), 그리고 전송된 바이트 수의 순서로 기록되고, LogIn과정을 거치면 사용자 ID도 기록을 하게 된다. 사용자가 하나의 웹 문서를 보기 위해 웹 서버로 보내는 Http요구는 웹 문서에 대한 요구 이외에도 웹 문서를 구성하고 있는 이미지 파일이나 동영상 파일들에 대한 기록까지 남아있게 된다. 이러한 파일들의 기록들은 분석 시간을 늘리고 사용자의 행위 패턴이나 관심도를 측정하는데는 필요하지 않으므로 사용자 ID와 'htm', 'html', 'xml', 'asp', … 등의 웹 문서 정보를 나타내는 데이터만 걸러내는 필터링과정을 거친다.

본 논문에서 제안하는 사용자별 웹 문서 관심도 측정 알고리즘은 다음과 같다.

[Step 1] 로그파일 필터링
웹 문서 데이터만 추출
(S_f=htm, html, xml, asp)

[Step 2] LogIn 사용자 추적
1. 로그인한 사용자의 위치 추적
//사용자 ID와 요구한 IP 주소 기록
2. LogIn을 거치지 않은 IP 주소 제거
//등록된 IP와 일치하지 않으면 제거

[Step 3] 접근한 IP주소와 사용자 ID와 매칭 검사
만약 다르면 접근한 웹 문서의 요구 시간만 기록 후 Step 2로 이동

[Step 4] 다음 접근된 웹 문서의 시간에서 앞서 접근한 웹 문서의 시간차(Sec) 측정·기록
Step 3으로 이동

사용자가 원하는 정보가 포함된 웹 문서에 접근하게 되는 경우는 그 문서에 머무는 시간이 길어지게 되므로 다른 웹 문서 요구 시간에 따른 차이를 구하여 사용자 관심도를 나타내는 가중치로 사용한다. 그리고 현재 유동 IP 지원 시스템의 보급으로 IP주소만으로 사용자의 판별이 어려우므로, 다른 사용자 ID에 같은 IP주소가 나타날 경우에는 다른 사용자로 간주하게 된다. 그리고, LogIn을 거치지 않고 접근하여 사용자를 판별할 수 없는 IP주소도 제거한다.

위 알고리즘을 통해서 사용자별 각각의 웹 문서의 접근시간차를 표 2 와 같은 테이블을 얻을 수 있다. 표 2

에서의 수치는 각각의 사용자가 해당 웹 문서의 관심도를 나타내는 기준이 되고, 임계값(30)미만의 웹 문서는 목적 웹 문서를 찾기 위해 거쳐가는 웹 페이지로 가정한다.

표 2. 웹 문서마다 사용자의 접근 시간 테이블

UserID \ 웹 문서	w1	w2	w3	w4	w5	…
User_1	10	15	0	400	0	
User_2	10	9	100	0	900	
User_3	13	5	20	0	324	
User_4	12	20	10	660	10	
⋮						

3.3 웹 문서 자동 선별 알고리즘

우선 사용자가 임계값 이상의 접근한 웹 문서에서 표 1 의 테이블로부터 사용자가 접근한 웹 문서의 키워드들과 값을 추출한다. 문서의 자동 선별을 위해서 다음과 같이 집합을 정의한다.

사용자 집합 U = {u1, u2, u3, u4, …l }

키워드 집합 K = {k1, k2, k3, k4, …n }

문서 집합 W = {w1, w2, w3, w4, …m }

퍼지관계 R은 사용자가 원하는 정보와 웹 문서들의 선별기준이 되는 키워드 사이의 관계정도를 가중치의 값으로 나타낸다..

$R=\{(u_l, w_m, \mu_R \mid (u_l, d_m) \in U \times W, \mu_R = U \times W \rightarrow [0,1]\}$

이때 사용자 관심 키워드는 사용자별로 접근 웹 문서의 전체키워드 중에서 접근 시간이 긴 웹 문서의 키워드들의 값으로 계산되고, 높은 값을 가진 키워드 순으로 n 개의 키워드를 가지게 된다.

본 논문에서의 문서 선별 알고리즘은 다음과 같다.

[Step 1] 사용자가 접근한 웹 문서와 키워드 테이블을 통해서 사용자별 프로파일 테이블 생성

표 3. 사용자 프로파일 테이블

User ID	관 심 키 워 드						카테고리 정보(URL)
User_1	k5	0.9	k1	0.7	…	…	Null
User_2	k4	0.9	k2	0.8	…	…	Null
User_3	k2	0.9	k3	0.8	…	…	Null
User_4	k1	0.9	null	0	…	…	Null
⋮							⋮

[Step 2] 사용자 프로파일을 이용하여 사용자와 웹 문서간의 관계도 측정(α)

표 4. 사용자별 α-cut

UserID \ 웹 문서	w1	w2	w3	w4	w5	…	α
User_1	0.4	0	0	0	0.9		0.5
User_2	0.2	0.7	0.1	0.8	0.2		0.5
User_3	0.3	0.9	0.8	0.2	0.1		0.6
User_4	0.9	0.3	0.3	0.3	0.3		0.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

[Step 3] α-cut을 적용한 사용자별 적합 문서 추출

표 5. 사용자별 적합 문서 추출 테이블

UserID \ 웹 문서	w1	w2	w3	w4	w5	…
User_1	0	0	0	0	1	
User_2	0	1	0	1	0	
User_3	0	1	1	0	0	
User_4	1	0	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮

[Step 4] 사용자 별로 추출된 적합 문서들의 URL을 사용자별 프로파일 테이블의 카테고리 정보에 기록

표 6. 마지막으로 얻어진 프로파일 테이블

User ID	관 심 키 워 드						카테고리 정보(URL)
User_1	k5	0.9	k1	0.7	…	…	w5, …
User_2	k4	0.9	k2	0.8	…	…	w2,w4,…
User_3	k2	0.9	k3	0.8	…	…	w2,w3,…
User_4	k1	0.9	null	0	…	…	W1,…
⋮							⋮

4. 결론 및 향후 연구 과제

본 논문에서는 기존의 정적인 웹 페이지 구성에서 벗어나 사용자의 웹사이트 이용 패턴을 분석하여 개별 사용자의 의도나 성향, 취향에 따라 개인화된 동적인 웹 문서의 정보를 제공하는 방법을 제안하였다. 키워드 검색이 아닌 사용자들이 웹 문서들 사이를 브라우징하는 것만으로도 관심도를 측정할 수 있게 하기 위하여 로그파일 분석을 통한 웹 문서 관심도 측정 알고리즘을 제시하였고, 사용자별 측정된 관심도를 사이트 내의 모든 웹 문서를 대상으로 다시 선별 작업을 하여 사용자가 관심을 가지지만 브라우징시 지나칠 수 있는 사용자와 관련이 있는 웹 문서의 URL 정보도 같이 제공하는 자동 문서 선별 알고리즘을 제안하여 보다 사용자에게 편리성을 제공할 수 있다. 현재 로그파일 분석기와 웹 문서별 키워드 추출기는 구현하였고 사용자 관심 키워드 추출기와 관심 웹 문서 추출기를 구현 중에 있다.

향후 연구 과제로는 급속도로 방대해지는 웹 사이트내에서 문서 선별의 정확성을 높이고 보다 신속한 처리로 서비스를 향상시키는 것이다.

참고 문헌

[1] 아이비즈 넷, http://www.i-biznet.co.kr/log/default.asp

[2] Osmar R Zaiane, Man Xin Jiawei Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," In Proc. Advances in Digital Libraries ADL'98, pp.19-29

[3] A.G. Buchner, S.S. Anand, M.D. Mulvenna and J.G.Hughes, "Discovering Internet Marketing Intelligence through Web Log Mining," Proc. Unicom99 Data Mining & Datawarehousing: Realising the full Value of Business Data, pp.127~138, 1999

[4] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," Proc. of the 20th VLDB Conference, pp. 487-499, 1994.

[5] R. Agrawal and R. Srikant, "Mining Sequential Pattern," Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.

[6] 박영규, 김진수, 김태용, 이정현, "연관 웹 문서 분류와 사용자 브라우징 패턴을 이용한 동적 링킹 시스템," 한국정보처리학회 추계학술발표 논문집, pp. 305-308, 2000.

[7] 최봉진, "Fuzzy Logic을 기반으로 한 SDI 서비스 설계" 한국정보과학회 가을술발표논문집(I) Vol.25,No,2 pp.333-305, 1998.