사용자 군집을 이용한 개인화 된 웹 페이지 추천

이은경*,이기현* 조근식**
인하대학교 전자계산공학과
silver13@eslab.inha.ac.kr gsjo@inha.ac.kr

The personalized web page using the Users clustering method

Eun Kyung Lee*,Gi-Hyung Lee*,Geun Sik Jo**
Dept. of Computer Science & Engineering, Inha University

요 약

기존의 웹 로그를 이용한 추천 system에서의 추천 문서 집합은 웹 페이지의 연관성과 웹 문서 사이의 거리를 이용하여 사용자들에게 추천 문서 집합을 제공해 주는 방식을 사용하였다. 이 방법에 의하면 추천 페이지로 제공되는 페이지는 사용자별 연관성이 고려되지 않으므로 모든 사용자들이 웹 페이지의 연관성만을 이용한 페이지를 추천 받는다. 따라서 처음 웹사이트를 방문한 새로운 사용자들에게는 추천해주는 페이지는 사용자가 보고 있는 웹 페이지의 연관성에 의한 웹 페이지만을 추천 받게 되므로 생각하지 못했던 페이지나 비슷한 취향을 가진 사용자들이 방문을 했던 페이지에 대해서는 추천 받지 못한다는 문제점을 가지고 있다. 따라서 본 논문에서는 동일한 페이지를 방문한 사용자별로 클러스터링 하여 같은 그룹에 속한 사용자들의 브라우징 패턴 정보를 발견, 분석화 하여 DB에 저장하였으며, 새로운 사용자에 대해서 웹 페이지 추천 집합을 제공하였다.

1. 서 론

웹사이트 중 Yahoo에 가보면 Yahoo의 기본 화면을 사용자의 구미에 맞게 편집하여 볼 수 있는 기능을 비롯해 사용자의 스타일에 맞는 정보를 선별하여 볼 수 있게 해준다. 또한 Amazon.com이나 Cdnow.com를 비롯한 전자 상거래 업체들도 사용자의 개인적 취향에 따라 자신의 페이지를 구성하고 사용자의 구매기록, 취향에 맞는 제품을 추천 받을 수 있는 기능들을 제공한다. 이 모든 것들이 개인화(Personalization)라고 불려진다. 그러나 대부분의 경우 개인화(Personalization)라고 하면 웹사이트와 연관지어서 웹사이트 개인화라는 용어를 사용하는 것이 보통이다. 또한 아직까지 개인화 방법에 대한 분류 기준조차도 사람에 따라 다르게 표현된다[1].

그러나 기존의 웹 로그를 이용한 추천 system에서의 추천 문서 집합은 웹 페이지의 연관성과 웹 문서 사이의 거리를 이용하여 사용자들에게 추천 문서 집합을 제공해 주는 방식을 사용하였다. 이 방법에 의하면 추천 페이지로 제공되는 페이지는 사용자별 연관성이 고려되지 않으므로 모든 사용자들이 웹 페이지의 연관성만을 이용한 페이지를 추천 받는다[2]. 따라서 처음 웹사이트를 방문한 새로운 사용자들에게는 추천해주는 페이지는 사용자가 보고 있는 웹 페이지의 연관성에 의한 웹 페이지만을 추천 받게 되므로 생각하지 못했던 페이지나 동일한 취향을 가진 사용자들이 방문을 했던 페이지에 대해서는 추천 받지 못한다는 문제점을 가지고 있다. 따라서 본 논문에서는 동일한 페이지를 방문한 사용자별로 클러스터링 하여 같은 그룹에 속한 사용자들의 브라우징 패턴 정보를 발견, 분석화 하여 DB에 저장하였으며, 새로운 사용자에 대해서 웹 페이지 추천 집합을 제공하는 것을 목적으로 한다. 웹 사용 마이닝(Web usage mining)을 이용하여 사용자 패턴 발견 과정 및 사용자별 그룹화 하는 과정은 은 다음과 같은 2가지 단계로 구성되었다. 첫 번째 단계에서는 웹 페이지들과 사용자와의 연관성을 분석하기 위하여 클러스터링(Clustering) 방법을 사용하였으며, 이 방법을 이용하여 유사한 웹 페이지를 본 사용자들을 그룹화(Group)할 수 있었다. 그룹화 된 사용자에 의한 추천 페이지 구성 방법은 현재 접근중인 웹 문서 V_i와 클러스터 내에 포함된 사용자들의 웹 문서 방문 페이지 사이 V_j의 $Sim(V_i,V_j)$를 측정하여 상위 N개의 추천 집합을 생성하였다.

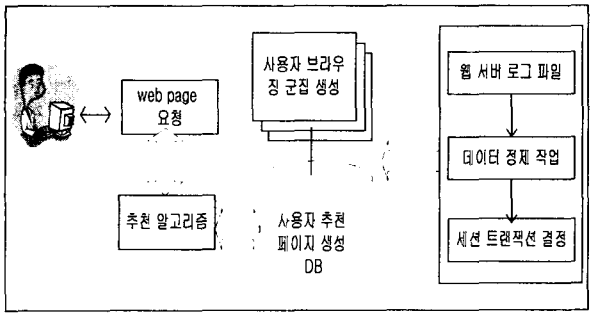
본 연구의 구성은 다음과 같다. 2장은 웹 마이닝에 대해서 살펴보고 3장에서는 본 연구에서 제안하는 웹 사용 마이닝을 이용한 웹 페이지 추천 페이지에 대하여 살펴보고 4장은 제시한 방법으로 만들어진 그룹과 각 그룹별로 나타난 웹 추천 페이지에 대해서 실험하였다. 마지막 5장은 결론 및 향후 연구과제에 대해서 논의한다.

2. 이론적 배경

사용자의 성향 분석을 위해서 웹 데이터에 데이터 마이닝(Data Mining) 기술을 도입하게 된 것이 웹 마이닝(Web Mining)이다. 또한 이 웹 마이닝 중에서 사용자가 사이트에 접속한 로그 파일을 가지고 결과를 분석하는 것이 웹 사용 마이닝이다. 일반적으로 웹 마이닝은 대상이 되는 웹 데이터에 따라 구조(web Structure mining), 내용(web content mining), 사용(web Usage mining)의 세 가지 분야로 나눌 수 있다[2]. 본 연구에서 사용한 웹 사용 마이닝은 초기 데이터를 마이닝 알고리즘의 입력 형태로 적절하게 바꾸어 주는 전처리(preprocessing) 과정, 전처리 과정에서 얻은 데이터에서 유용

한 정보를 얻기 위한 패턴 발견(pattern discovery) 과정, 마지막으로 생성된 규칙과 패턴을 분석(pattern analysis) 하는 과정으로 나눌 수 있다[3].

3. 개인 웹 페이지 추천 시스템



[그림 1] 전체 구성도

본 논문에서 설계한 시스템은 사용자 그룹화를 위해 웹 서버에 저장되어 있는 웹 로그 파일로부터 데이터 정제 작업 후, 세션 트랜잭션을 생성하고 사용자별 그룹을 생성한다. 이 과정에서 추출된 사용자별로 그룹화 된 웹 페이지들 간의 연관 분석을 이용하여 새로운 사용자가 웹 문서에 접근하였을 때 추천 집합으로 각 그룹에서 사용자의 요청 문서와 연관도가 높은 문서를 제공 한다.

3.1 전처리 과정

웹 서버에 기록된 원시 로그 파일(raw access log file)은 사용자가 필요한 웹 페이지를 요구하면 웹 서버는 요청한 문서에 존재하는 이미지 등의 파일까지도 로그 파일에 기록하므로 양이 방대하다. 따라서 마이닝 작업을 하기 전에 분석에 필요한 내용을 필터링 해야 한다[4]. 사용자 별로 방문을 결정하고, 한번의 방문이동 경로를 세션으로 결정한다. 이때 사용자는 IP 주소로 구별되며 동일 IP라도 마지막 접속 후 30분 이후의 접속은 새로운 세션이라고 가정한다[5]. 본 연구에서는 Access log file에 있는 여러 필드 중 에러가 나지 않고, 접속 페이지의 확장자가 *.htm, *.html인 화일을 우선 필터링 한 후 사용자IP, 접속시간, 접속페이지 필드를 추출하였다.

필드 추출 다음 과정으로 엑세스 로그 파일에 전처리 작업을 수행한 최종 결과는 사용자 트랜잭션(User-Transaction)파일로써 다음과 같이 구성되어 있다.

T_i=<ID_i, {R_1=(R_1.url, R_1.time), ..., R_{n-k}=(R_{n-k}.url, R_{n-k}.time)}>

ID_i :세션(session)의 IP_i에 의해 생성된 사용자 고유번호

R_i: 사용자 요구(request)페이지인 R_i.url

사용자가 머문 시간인 R_i.time

T_i: 사용자에 의해서 생긴 트랜잭션

3.2 웹 페이지와 사용자간의 연관성 (EM 알고리즘에 의한 사용자 군집)

사용자별 군집화 과정은 웹 페이지들과 사용자 사이간의 연관성 분석이다. 클러스터링은 전체 웹 페이지들과 사용자들이 비슷하게 보는 페이지에 의해서 사용자들을 그룹화 하여 동일한 페이지를 방문하게 될 사용자들에게 추천 페이지를 제공할 것이다.

클러스터링 (Clustering)이란 주어진 데이터 셋을 서로 유사성을 가지는 몇 개의 클러스터로 분할해 내가는 과정으로, 하나의 클러스터에 속하는 데이터 점들 간에는 서로 다른 클러스터 내의 점들과는 구분되는 유사성을 갖게 된다[4]. 클러스터링을 수행하기 위해서는 전처리 된 데이터를 그냥 이용할 수 없기 때문에 클러스터링을 수행하기 전에 클러스터링 데이터를 만드는 작업이 필요하다. 이 작업은 전처리 과정과 연계하여 사용자가 열(row)에 위치하며, 행(column)은 사용자에 따라서 그 페이지를 방문하였으면 "0", 방문하지 않았으면 "1"를 대입한 자료이다. 이렇게 만들어진 자료는 Clustering을 할 때, 사용자들이 같은 페이지를 방문한 사용자를 기준으로 사용자 그룹(User-Group)으로 나누어지게 되므로 방문 페이지에 의한 사용자 그룹 분석을 할 때 사용한다. 데이터 행렬은 다음 [표1]과 같다.

[표 1] 사용자 그룹화를 위한 전처리 단계

user page	1	2	3	4
a28	0	0	0	1
a29	0	1	1	0
a30	1	0	0	1
a31	1	0	0	0

3.3 새로운 사용자 추천 페이지 생성 과정

새로운 사용자의 추천 페이지를 구성하기 위한 과정은 사용자별로 그룹화 된 웹 페이지와 사용자가 요청한 문서의 상관 관계에 의해서 제공된다. 다음은 페이지 추천 생성 과정에 대한 단계이다.

첫째, 사용자가 요청한 page(V_i)를 벡터로 표현한다.

둘째, Group안에 있는 page(V_j)을 벡터로 표현한다.

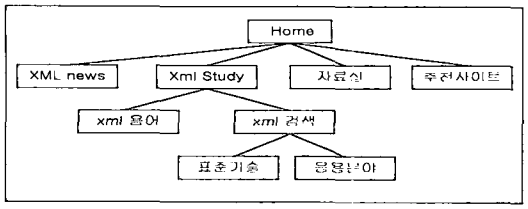
셋째, 벡터로 표현된 각 페이지 V_i와 V_j 간의 Cosine Similarityf를 [식 1]에 의해서 계산한다.

$$Sim(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i} \cdot \vec{v_j}}{|\vec{v_i}| \times |\vec{v_j}|} = \frac{\sum_{j=1} v_{i,j}^2 \times v_{i,q}}{\sqrt{\sum_{j=1} v_{i,j}^2 \times \sum_{j=1} v_{i,q}^2}} \quad [식1]$$

넷째, 계산된 유사도에 따라서 추천 집합을 정렬하여 Top-N개를 추천 집합으로 보여준다.

4. 실험

본 연구의 실험은 www.xmlgo.net [6]라는 xml에 관한 학습 사이트의 web access log file을 중에서 1월 2일부터 1월 12일까지 일주인간의 데이터를 사용하였다. 본 사이트는 평균 200명 정도의 사용자들이 하루동안 방문하며, 총 12개의 카테고리 아래 2-3개의 세부적인 문서가 존재한다. [6]



[그림 2] 웹 사이트 구성도

4.1 전처리 과정 (문서 필터링과 사용자 필터링)

해당하는 사이트는 Webalizer라는 tool을 이용하여 자체적인 로그 분석을 한다. tool에 의하여 제공하는 내용 중 방문자 hit 률을 살펴보면 /bbs1/list.html와 /top.html 같은 페이지가 최고의 hit를 기록한다고 나타내고 있다. 그러나 top.html 같은 페이지는 내용초기 화면으로 실제적으로 사용자에게 유용한 정보를 주는 페이지라고 볼 수 없다. 따라서 본 연구에서는 사용자의 의미 있는 행위 패턴을 발견하기 위하여 위 사이트에 존재하는 문서에 중 top.html과 같은 단순한 링크만을 제공하는 페이지들을 제거 한 결과 총 126490개의 자료 중 51904개의 자료로 필터링을 할 수 있었다. 패턴 발견을 위한 attribute 값으로 사용된 웹 페이지는 사용자가 적어도 3번 이상 본 페이지 대상으로 모두 27페이지의 유용한 페이지를 찾을 수 있었다. 또한 사용자 중 웹 페이지를 3 페이지 이상 보지 않고 종료된 사용자를 필터링 하여, 단순한 방문 보다는 해당 사이트에서 의미 있는 정보를 찾고자 하는 사용자를 구별 하였다. 필터링 한 결과 해당 사이트는 평균 200명 정도의 사용자들이 방문하나, 필터링 한 후에는 약 181명의 의미 있는 사용자를 발견하였다.

4.2 웹 페이지와 사용자 연관성

[표3]은 클러스터링을 수행한 결과를 나타낸 것이다. 181명의 사용자를 대상으로 실시한 군집은 총3개로 나타났으며, 웹 페이지를 방문했던 사용자의 그룹은 각각 25, 113, 43으로 구별되었다.

[표 2] 각 그룹에 할당된 사용자 수

cluster	1	2	3
사용자 수	25	113	43

또한 요청된 페이지(V_i)와 그룹에 있는 페이지(V_j)와 추천해줄 페이지의 연관 페이지, $P=\{p1,p2,... \mid Sim(V_i,V_j)\}$는 다음 [표 4]처럼 나타났다.

[표 3] 그룹2에서의 연관성 분석

pages of Group2
1. /book_m.html=1 ==> /document/lecture/xml_base/index.html=1
2. /book_m.html=1 & /test/memo5/naksu_kernel.html=1
==> /class/resource.html=1
3 /book_m.html=1& /document/cool.htm=1
==> /howow.html=1

4.3 새로운 사용자를 위한 추천 페이지

다음 [표6]는 사용자가 요청한 페이지에 따라서 다른 추천을 보여준다. 각 페이지는 서로 다른 군집에 속한 사용자들에 의해서 유사도의 계산에 따른 추천집합이 구성되므로 사용자가 어떠한 페이지를 요청하였는가에 따라서 동적인 추천 페이지를 볼 수 있다.

[표 4] 그룹1과 그룹 2에서 Top-3 추천 페이지 비교

그룹 2의 추천 문서	book_m.html	/document/lecture/xml_base/index.html
		/class/resource.html
		/howow.html
기존 추천 시스템	book_m.html	/document/lecture/xml_base/index.html
		/howow.html
		/document/cool.htm

5. 결론

본 연구에서는 개인화 된 웹 시스템을 위한 추천 집합으로 웹 사용 마이닝의 기술 중 웹 페이지와 사용자간의 연관성을 살펴보기 위하여 클러스터링 방법을 통하여서 비슷한 패턴을 보인 사용자끼리 그룹으로 묶을 수 있었다. 사용자별 군집화를 이용하여 각 그룹별 특성이 나타나는 웹 페이지를 저장하였고, 새로운 사용자가 나타났을 때 사용자가 요청한 페이지와 그룹안에 존재하는 페이지 사이의 유사도를 계산하여 추천 집합으로 웹 페이지를 정렬하여 보여 주었다. 그러나 같은 집합 안에 존재하는 페이지에 대해서는 어떠한 우선 순위가 존재하지 않으므로 동일한 페이지의 추천도 이루어지는 것을 확인 할 수 있었다. 따라서 앞으로의 과제는 발견된 패턴을 기반으로 다양한 그룹에 별로 개인화 된 웹 페이지를 서비스하는 시스템 구축과 웹 사용 마이닝(web usage mining)과 웹 내용 마이닝(web content mining)을 결합하여 웹 사용 마이닝에서 찾을 수 없었던 웹 페이지의 내용까지도 분석을 하여 좀더 구체적인 개인화 된 웹 페이지를 서비스 할 수 있어야 하겠다.

6. 참고문헌

[1] www.personalization.co.kr

[2] J.Srivastava,R.Cooley, M.Deshpande, and P.N.Tan, "Web-Usage Mining:Discovery and Application of Usage Patterns for Web data", In ACM SIGKDD Exploration, Vol.1(2),pp12-23,Jan 2000

[3]김종달, 김성민, 남도원, 이동하, 이전영, "웹 마이닝 시스템 설계 및 유용한 접근패턴 정의", 한국지능정보시스템학회 춘계 학술발표, pp.283-291, 6월 2000

[4]Ian H.Witten and Eibe Frank, "Data mining: Practical machine learning tools and techniques with Java Implementations", Morgan Kanufmann publishers, 1999

[5]Bamshed Mobasher, R.Cooley and J.Srivastava, "Automatic Personalization Based on Web Usage Mining," Communications of the Association of Computing Machinery (CACM), pp. 142-151,August 2000

[6] http://www.xmlgo.net