무선 인터넷 웹 로그 파일을 이용한 사용자 클러스터링

한상훈[0] 신주리 이건명
충북대학교 컴퓨터과학과 / 첨단정보기술 연구센터
likelamb@aicore.chungbuk.ac.kr

User Clustering from Wireless Internet Web Logs

Sang-Hoon Han[0] Ju-Ri Shin Keon-Myung Lee
Dept. of Computer Science, Chungbuk University and AITrc

요 약

무선 인터넷 보급이 확산되고 그 활용범위가 날로 넓어짐에 따라 무선 인터넷 기술에 데이터 마이닝을 접목시키고자 하는 노력은 당연하면서도 필요한 것이다. 이 논문에서는 무선 인터넷에서 사용자를 대표할 수 있는 정보와 무선 인터넷 웹 서버의 로그 파일에 저장된 정보를 접목시킴으로 일정한 액세스 패턴을 가지고 있는 사용자의 클래스를 추출하는 방법을 제시한다. 일정한 액세스 패턴을 가지고 있는 사용자들의 클래스를 추출함으로써 무선 인터넷 서비스를 사용하는 사용자에 대한 서비스의 질을 향상시키는데 기여할 수 있을 것이다.

1. 서론

유선 인터넷이 발달하여 유선 인터넷에서 얻을 수 있는 정보와 그 활용 범위가 넓어짐에 따라 유선 인터넷을 통해 얻을 수 있는 서비스에 대한 사용자의 요구사항도 다양해지고 있다. 이러한 사용자의 욕구가 증가됨에 따라 사용자를 만족시켜야 하는 서비스를 제공하는 개인이나 기업들의 의해 서비스의 질적 향상을 위한 여러 가지 시도들이 시행되어지고 있다. 무엇보다도 서비스의 향상은 다양하면서도 그 서비스의 폭이 넓어지며 보다 개인에 가까운 서비스를 제공하고자 하는 시도들로 집중되어져 가고 있다. 그러나 유선 인터넷에서는 서비스를 제공하는 개인이나 기업에서 사용자에 대한 개별적이면서도 구체적인 서비스를 제공하기 위해 개인 정보를 사용자로 하여금 입력하게 하는 것이 필수적이다. 이것은 유선 인터넷 웹 서버에 접근하는 사용자에 대한 개별적인 정보를 추출할 수 없다는 데도 크게 원인이 있다.

유선 인터넷에서 서비스를 제공하는 웹 서버에 접근하는 사용자들에 대한 로그 파일에 남는 정보 중 개인적인 정보에 가장 가까운 정보라고 한다면 IP 주소를 말할 수 있는데, 이 IP 주소는 컴퓨터에 부여되어 있는 주소이므로 해당 컴퓨터를 사용하는 사용자는 언제든지 바뀔 수 있다는데 그 문제가 있다. 그러므로 IP 주소 정보를 사용자 정보라고 생각하고 이를 이용하여 사용자에 대한 여러 가지 정보들을 추출한다는 것은 의미가 없게 되는 것이다.

무선 인터넷의 경우에는 사정이 다르다. 무선 인터넷의 경우에는 무선 인터넷 서비스를 제공하는 웹 서버에 접근하기 위해 사용하는 클라이언트가 컴퓨터가 아닌 무선 인터넷 단말기(예: 핸드폰)이기 때문에 무선 인터넷 웹 서버에 전달하는 헤더정보에는 무선 인터넷 단말기에 대한 정보가 담겨있다. 이 헤더 정보는 유선 인터넷에서 컴퓨터에 할당되어지는 IP 주소처럼 유일하면서도 컴퓨터가 아닌 사용자를 대표할 수 있는 무선 인터넷 단말기 정보가 포함되므로 사용자를 구별하기 위한 별도의 작업 없이도 개별적인 사용자 정보를 추출할 수 있을 뿐 아니라, 그 정보를 이용하여 사용자에 대한 가치 있는 정보들을 추출해낼 수 있다.

무선 인터넷 웹 서버의 로그 파일에는 유선 인터넷 웹 로그 파일과 같이 다음과 같은 데이터들이 남게 된다.

- 웹 서버에 액세스한 시간
- 요구된 페이지의 URL
- 사용자가 요구한 페이지의 전송 여부(성공, 실패, 에러 발생 등)
- 사용자에게 보내진 데이터의 크기
- 사용자가 사용하는 사용자 에이전트

유선 인터넷 웹 로그 파일과 무선 인터넷 웹 로그 파일만을 보면 두 인터넷 웹 로그 파일은 별다른 차이점이 없

* 이 논문은 첨단정보기술 연구센터(AITrc)를 통해서 과학재단의 지원을 받고 연구된 것임.

을뿐 아니라, 서비스를 이용하는 브라우저의 제약으로 오히려 무선 인터넷 웹 로그 파일의 데이터가 더 부족한 것이 사실이다. 이와 같은 사정으로 지금까지 유선 웹 로그 파일만을 이용하여 사용자 액세스 패턴을 찾아내는 등의 유선 인터넷 웹 로그 파일에 데이터 마이닝 기술을 적용하여 여러 가지 정보를 추출하고자 하는 연구[1,2,3]들은 많이 이루어져 왔으나, 무선 인터넷의 특성을 이용한 데이터 마이닝 기술의 적용은 시도되지 않았다.

이 논문에서는 무선 인터넷의 헤더정보를 통해 얻을 수 있는 사용자 정보를 가지고 사용자를 구분하며 구분된 사용자가 일정한 시간동안 사용자의 액세스 패턴을 추출하여 비슷한 패턴을 가지고 있는 사용자들의 클래스를 찾아내는 시스템을 구현하고 시스템 구현에 사용된 사용자 클러스터링 방법을 제시한다.

2. 사용자 클러스터링

이 논문에서 제시하는 사용자 클러스터링 과정은 크게 세가지 단계로 나누어질 수 있다.

1. 사용자를 구별할 수 있는 개인 정보와 서비스의 액세스 정보를 각각 헤더정보와 로그 파일에서 추출하여 하나의 데이터 베이스를 구축한다.
2. 구축된 데이터 베이스에서 각각의 사용자별로 액세스 패턴을 찾아낸다.
3. 클러스터링 알고리즘을 적용하여 사용자의 액세스 패턴을 기준으로 사용자들의 클래스를 찾아낸다.

위의 과정에서 사용자를 구별할 수 있는 무선 인터넷 단말기 정보를 추출하기 위해 모든 서버를 접속하는 사용자들이 처음 액세스하는 페이지에 사용자의 무선 인터넷 단말기 정보를 추출하는 과정을 삽입하고, 첫번째 단계인 사용자 클러스터링을 위한 데이터 베이스 구축은 사용자의 액세스 이벤트가 발생할 때마다 새로운 데이터 정보를 데이터 베이스에 추가하게 된다.

3. 데이터 베이스 구축

3.1 사용자 정보 추출

앞에서도 언급한 바와 같이, 유선 인터넷 웹 로그 파일과 같이 무선 인터넷 웹 로그 파일에 남게 되는 정보로는 사용자를 구별할 수 없다. 무선 인터넷 웹 로그 파일에 사용자를 구별할 수 있는 유일한 정보가 나타나야 하는데, 그 정보는 일반적으로 로그 파일에 남지 않기 때문이다. 그러므로 사용자를 구별할 수 있는 정보를 추출하여 로그 파일에 있는 정보와 연결하는 특별한 작업이 필요하게 된다. 무선 인터넷의 경우에는 사용자를 유일하게 구별할 수 있는 정보가 헤더에 포함된다. 즉, 무선 인터넷 웹 서버로 보내지는 헤더정보에 사용자의 무선 인터넷 단말기 번호나 통신 회사에서 부여된 특별한 아이디가 나타나는데, 이 정보는 유일한 사용자를 대표할 수 있는 정보라고 할 수 있다.

이 헤더 정보에서 사용자 정보와 함께 무선 인터넷 로그 파일에도 남아있는 데이터를 추출한다. 헤더 정보와 로그 파일에 동시에 남는 데이터를 선택하는 것은 무선 인터넷 단말기와 사업자, 그리고 무선 인터넷 서비스에 사용하는 인터넷 서버에 따라 약간의 차이가 있을 수 있다.

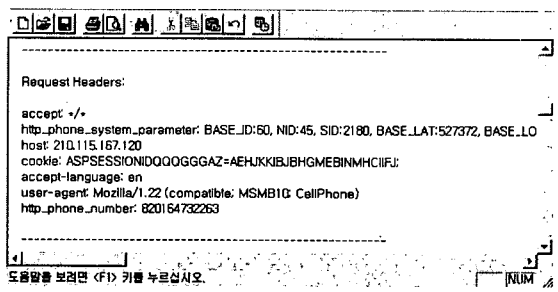

그림 1. 헤더 정보

3.2 로그 파일에서의 액세스 정보 추출

무선 인터넷에서 제공되는 서비스들은 유선 인터넷에서 제공되는 서비스와는 달리 서비스 페이지와 서비스를 찾아가는 링크 페이지가 확실히 분리된다. 즉, 무선 인터넷 서비스는 일정한 트리 구조를 가지고 있어 그 트리의 단말 페이지에서 서비스가 제공된다. 그러므로 사용자가 액세스한 서비스의 패턴을 찾아내려면 해당 서비스로 가기 위해 액세스하게 되는 페이지들은 제거하고 서비스 페이지에 대한 로그 파일의 기록만을 추출한다.

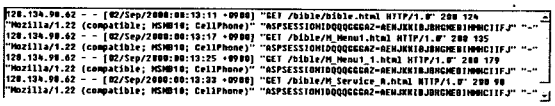

그림 2. 무선 인터넷 로그 파일

3.3 데이터 베이스 구축

추출된 사용자 정보와 액세스 정보를 연결하여 데이터 베이스에 추가한다. 사용자 정보와 액세스 정보의 연결은 사용자 정보와 로그 파일에서 모두 추출할 수 있는 세션 정보나 쿠키 등을 이용하여 데이터 베이스의 조인 오퍼레이션을 통해 하나의 데이터 베이스로 구축한다.

Time	Page	status	Phone_num
02/Sep/2000:00:09:57 +0900	A	200	0114005000
02/Sep/2000:00:10:06 +0900	B	200	0195006000
02/Sep/2000:00:10:11 +0900	C	200	0114005000
02/Sep/2000:00:13:05 +0900	F	200	0164560000
02/Sep/2000:00:13:07 +0900	A	200	0164560000

그림 3. 구축된 데이터 베이스

4. 사용자의 액세스 패턴 추출

사용자의 액세스 패턴은 각 서비스 페이지에 대한 액세스 수의 합으로 나타낸다. 즉, 서비스 페이지들의 집합 $S=\{s_1,s_2,s_3,s_4,s_5\}$ 일 때, 사용자 패턴 U_{userID} 는 다음과 같이 나타낸다.

$$U_{userID}=s_1^a s_2^b s_3^c s_4^d s_5^e \ (0\le a,b,c,d,e)$$

User	Access pattern
0114005000	$A^2B^3C^0D^2E^4F^7$
0195006000	$A^3B^5C^1D^2E^5F^0$
0114005000	$A^5B^5C^6D^3E^4F^2$
0164560000	$A^4B^3C^2D^2E^3F^3$
0164560000	$A^3B^4C^1D^3E^5F^3$

그림 4. 추출된 사용자 액세스 패턴

5. 클러스터링

사용자 별로 나타난 사용자의 패턴 U_{userID} 은 사용자를 대표하는 데이터 값이 된다. 해당 웹 서버에서 제공하는 서비스 $S=\{s_1,s_2,\cdots,s_N\}$ (N : 서비스 페이지 수)에서 다음의 클러스터링 알고리즘을 적용한다.

1. 모든 사용자의 패턴 U_{userID} 을 비교하여 두 사용자의 패턴 d_{ij} 의 값이 0 인 것을 모아 초기 클래스로 사용한다.

   사용자 패턴 U_i 와 다른 사용자 U_j 의 거리

   $$d_{ij}=\sqrt{\sum\{U_i(s_k)-U_j(s_k)-\Delta\}^2}/N$$
   $$(i,j\subset userID,\ 1\le k\le N)$$

   단,

   $U_i(s_k)$: 사용자 i가 k번째 서비스 페이지를 액세스한 횟수.

   $$\Delta=\sum\{U_i(s_k)-U_j(s_k)\}/N$$

2. 각 클래스의 대표값과 모든 사용자의 거리를 구한 뒤, 거리가 가장 가까운 클래스에 포함시킨다.
3. 클래스의 대표값을 계산한다. 클래스에 포함되어 있는 사용자의 모든 패턴들의 평균값으로 결정한다.

   클래스의 대표값

   $$C_{ClassID}=s_1^{n_1}s_2^{n_2}\cdots s_N^{n_n}$$

   단,

   $$n_n=\sum_{userID\subset C_{ClassID}}U_{userID}(s_n)/|C_{ClassID}|$$

4. 클래스의 대표값들과 각 클래스에 포함된 사용자의 거리를 계산하여 가까운 클래스에 포함시킨다.
5. 모든 사용자의 소속 클래스가 더 이상 바뀌지 않을 때까지 2 번 과정으로 되돌아간다.
6. 모든 과정이 끝나면 사용자 클러스링의 결과는 일정한 액세스 패턴으로 가지고 있는 클래스 정보와 각 클래스의 대표값으로 나타난다.

6. 결론 및 향후과제

이 논문은 무선 인터넷 환경에서 사용자를 대표할 수 있는 정보를 추출하여 무선 인터넷 로그 파일과 연결함으로써 유사한 액세스 패턴을 가지고 있는 사용자들의 클래스를 추출할 수 있는 방법을 제시하였다.

무선 인터넷 환경에서 얻을 수 있는 로그 데이터들에 대한 데이터 마이닝을 통해, 많은 유용한 정보를 추출할 수 있으며 그 정보를 이용하여 무선 인터넷에서 제공하는 서비스의 질을 향상시키고 사용자에 특성화된 서비스를 제공하는데 도움을 줄 수 있다. 한편, 무선 인터넷 사용자의 개인정보(연령, 직업, 서비스 지역 등)을 함께 이용하면 보다 유용한 정보를 추출할 수 있지만 사생활 침해의 여지 등의 문제를 일으킬 소지도 생길 수 있다. 무선 인터넷 로그 데이터에 대한 데이터 마이닝을 아직 충분히 수행하지 않은 상태로, 사용자 클러스터링 이외에 데이터 마이닝을 적용하여 추출할 수 있는 정보의 종류와 그 적용 범위를 넓히는데 대한 계속적인 연구가 필요하다.

7. 참고 문헌

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *Proc. 16th Int'l Conf. Distributed Computing Systems*, pages 385-392, May 1996.

[2] O. Zaiane, M. Xin, and J. Han. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In *Proc. Advances in Digital Libraries Conf. (ADL'98)*, Melbourne, Australia, pages 144-158, April 1998.

[3] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu. Mining Access Pattern efficiently from Web logs. In *Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00)*, Kyoto, Japan, April 2000.

[4] Van Wijk, J.J.; Van Selow, E.R. Cluster and calendar based visualization of time series data. In *Proc. 1999 IEEE Symposium on, Information Visualization, 1999. (Info Vis '99).*

[5] 남기범, 이건명. 무선 웹 기술과 전망. 정보 과학 회지 제18권 제6호 2000년, pages 32-38

[6] T. Zhang, R. Ramakrishnan, M. Livny. BIRCH: An efficient data clusering method for very large databases. In *Proc. ACM SIGMOD Int. Conf.* On Management of Data, pages 103-114, 1996.