웹 사용 마이닝의 정확도 향상을 위한 인기도 기반 전진 참조 기법

조현웅[0] 김유성

인하대학교 전자계산공학과

g2001416@inhavision.inha.ac.kr yskim@inha.ac.kr

Popularity-weighted Forward Reference Scheme for High Accuracy in Web Usage Mining

Hyun-Woong Cho[0] Yoo-Sung Kim

Dept. of Computer Science & Engineering, Inha University

요 약

웹 사용 마이닝의 단계중 패턴 발견을 위해 초기 데이타를 정제하는 전처리 과정은 매우 중요한 작업이다. 전처리 과정의 결과가 높은 정확도를 가지고 있다면 마이닝의 결과 역시 보다 정확한 결과를 생성한다는 것은 여러 연구를 통해 널리 알려진 사실이다. 본 논문에서는 전처리 과정중 내용 페이지를 구분하기 위해 자주 이용되는 기법중 하나인 최대 전진 참조(M.F.R : Maximal Forward Reference) 기법을 개선한 인기도 기반 전진 참조(P.F.R : Popularity-weighted Forward Reference) 기법을 제안하고 예제를 통해 두 기법의 결과를 비교하였다. 그 결과 최대 전진 참조 기법에서 발생할 수 있는 오류를 극복한 인기도 기반 기법이 좀더 정확한 내용 페이지 구분이 가능하여 웹 사용 마이닝 단계에서 유용하게 활용할 수 있음을 보였다.

1. 서 론

웹 사용 마이닝은 보통 전처리 과정(preprocessing), 패턴 발견(pattern discovery), 패턴 분석(pattern analysis)의 세 부분으로 나누어 생각할 수 있다. 인터넷의 특성상 방대하고 부정확한 초기 데이타로부터 패턴 발견을 위해서 필요한 데이타를 추출하고 변환, 정제하는 작업을 전처리 과정이라고 하는데 이는 보다 정확한 결과를 얻기 위해서 매우 중요한 작업이다[1,2].

웹 사용 마이닝에서는 어떠한 마이닝 작업이 실행되기 이전에 페이지 참조들의 시퀀스는 트랜잭션 혹은 사용자 세션으로 대표되는 논리적인 단위로 나뉘어 그룹 지워진다. 사용자 세션은 어떤 사이트에 방문한 한 사용자에 의해 만들어진 모든 페이지 참조들이다. 트랜잭션은 사용자 세션과 달리 구분 기준에 따라 하나의 페이지 참조로 이루어진 것부터 모든 페이지 참조로 이루어진 것까지 다양하다. 트랜잭션은 세션에 의미를 부여한 것으로 데이타 마이닝을 위한 기본 단위이다. 이는 패턴 발견 과정에서 사용될 연관 규칙(association rules), 웹 방문 패턴(web navigation patterns), 떠남 패턴(exit patterns)을 발견하는데 사용되는 기본적인 단위이다[1].

전처리 과정중 세션에서 트랜잭션을 구분하는 기법중 하나인 최대 전진 참조 기법(M.F.R : Maximal Forward Reference)은 내용 페이지만을 고려한 트랜잭션을 정의하기 위하여 사용된다. 이 기법은 세션중 후진 참조가 발생하기 이전의 참조를 최대 전진 참조라 하고 이 참조를 사용자가 관심을 가지고 찾고자 했던, 즉 내용 페이지로 간주한다[3]. 그러나, 이러한 최대 전진 참조가 반드시 내용 페이지라는 보장이 없기 때문에 만일 사용자의 실수 등으로 인하여 보조 페이지가 최대 전진 참조였다면 잘못된 내용 페이지 구분 결과를 생성하게 된다.

본 논문에서는 보다 정확한 전처리 과정을 통해 좀더 향상된 데이타 정제 효과를 얻고자 기존의 최대 전진 참조 기법을 개선하여 각 페이지의 참조 횟수를 고려한 인기도 기반 전진 참조 기법(P.F.R : Popularity-weighted Forward Reference)을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 간단하게 소개한다. 3장에서는 인기도 기반 전진 참조 기법의 알고리즘을 설명하고 예제를 통하여 정확도를 비교한다. 4장에서는 결론 및 향후 연구방향으로 본 논문을 맺는다.

2. 관련 연구

웹 사용 마이닝 분야와 관련된 연구는 매우 다양하지만 본 논문에서는 전처리 과정 단계의 트랜잭션 구분과 관련된 연구중 최대 전진 참조 기법에 대해서만 언급하기로 한다.

최대 전진 참조 기법은 전체 참조 집합을 부분 참조 집합들로 구분하기 위한 기법이다. 이 기법을 요약하면 전체 참조 집합을 순차적으로 검색하며 전진하다가 후진방향으로 참조가 발생할 경우, 후진 참조가 발생하기 이전의 참조들을 하나의 부분 집합으로 판단한다. 예를 들어 {ABFOFBG, AD, ABACJ, LR}의 참조 집합이 있으면 이는 {ABFO, ABG, AD, AB, ACJ, LR}의 부분 집합들로 구분될 수 있다. 각 부분 집합들이 가지는 의미는 마지막의 참조가 사용자가 찾고자 하였던 참조이며 다른 참조들은 원하는 참조를 찾기 위한 보조적인 참조들이라는 것이다. 결과적으로 각 부분 집합의 마지막 참조들의 집합인 {O-G, D, B-J, R}이 내용 데이타 페이지로서 이용된다[3,4,5].

그러나, 이 결과를 [그림 1]에 적용하여 보면 보조 페이지로 정의된 O, D, B 페이지들이 내용 데이타 페이지로 선택된 것을 알 수 있다. 이는 최대 전진 참조 기법이 내용 페이지를 판단하는데 있어 다른 정보는 이용하지 않고 오직 참조 경로 정보만을 이용함으로써 발생하는 오류이다.

3. 전처리 과정과 인기도 기반 전진 참조 기법

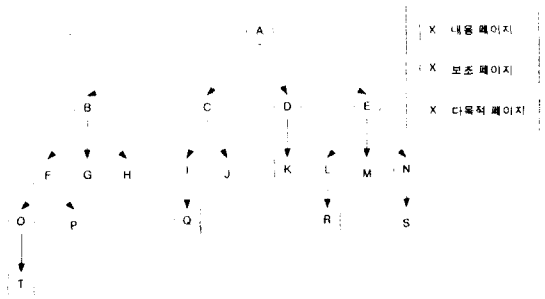
본 논문에서는 전처리 과정에서 많이 사용되고 있는 최대 전진 참조 기법을 개선하여 정확도를 향상시킨 인기도 기반 전진 참조 기법

을 제안하고 그 정확도를 비교하여 본다.

3.1 마이닝의 대상이 되는 데이타

웹 사용 마이닝시 사용할 수 있는 데이타로 가장 대표적인 데이타는 웹 서버가 사용자의 요청을 처리할때 그 요청을 기록한 웹 로그(web log)이다. 대부분의 웹 서버가 제공하는 로그의 형식은 CERN과 NCSA에 의해 HTTP 프로토콜(protocol)의 일부로 명시된 "Common Log Format"을 따른다[4]. 보다 정확한 사용자 정보를 얻고자 할 때에는 쿠키(cookie)를 이용하거나 혹은 웹 서버가 지원하는 부가적인 데이타를 이용할 수 있으나, 이러한 데이타들은 표준이라고 볼 수 없으므로 본 논문에서는 고려하지 않는다.

그 외 본 논문에서 사용되는 데이타로는 웹 사이트의 논리적 구조와 각 페이지의 인기도 정보가 있다. 이러한 정보들은 웹 로그를 분석함으로써 간접적으로 얻을 수 있다[4,5]. [그림 1]은 예제로 사용된 웹 사이트의 구조를 나타낸 그림이다. "내용 페이지"는 사용자에게 제공할 정보를 포함하고 있는 페이지를 의미하며 "보조 페이지"는 사용자가 "내용 페이지"를 찾아 가는데 있어 편리함을 제공하기 위한 페이지를 의미한다. 다목적 페이지는 위의 두 가지 목적을 모두 가지고 있는 페이지를 의미한다.



[그림 1] 전처리 과정에서 이용되는 사이트 구조도 예제

3.2 데이타 정제 및 세션 구분

표준 형식의 웹 로그는 사용자가 특정 페이지를 서버에 요청하였을 때 발생하는 요청을 비롯한 수많은 이미지 파일들과 다른 추가적인 데이타(아래 한글 파일, 자바 스크립트 파일, 플래시 파일등)에 대한 요청도 모두 기록한다. 일반적인 웹 사용 마이닝시에는 오직 웹 페이지에 대한 요청만이 관심 대상이므로 이러한 정보들은 제거되어야 한다. [그림 2]는 표준을 따르는 웹 로그의 일부를 나타내며 필요 없는 정보는 제거된 것이다[4,5]. IP Address열은 사이트에 접속한 사용자의 IP를 의미한다. User ID열은 인증이 있을 경우 사용자의 ID를 기록한다. 인증이 없을 경우에는 "-"로 대체된다. Time열과 Method/URL/Protocol열은 각각 사용자가 요청이 발생한 시각과 그 요청을 나타내며, Agent열은 사용자가 사용한 브라우저의 정보를 기록한다.

세션을 구분하기 전에 먼저 사용자를 구분하는 작업이 선행되어야 한다. 웹 로그에서 사용자를 구분하는 것은 매우 어려운 작업이다. HTTP 프로토콜이 연결 지향(connect-oriented) 방식이 아니기 때문에 사용자를 구분하기 위해서는 사용자의 IP를 이용해야 하지만, 인터넷 서비스 제공자가 그들의 고객에게 동적인 IP를 할당하거나 프록시(proxy) 서버를 이용한다면 단지 IP만으로는 사용자를 구분할 수가 없게 된다. 그러므로 사용자의 브라우저 소프트웨어의 정보를 나타내는 Agent 정보를 이용함으로써 사용자를 구분 할 수 있다. 이러한 방법으로 [그림 2]에서는 세 명의 사용자를 구분할 수가 있고, 그 사용자들의 방문 경로는 각각 A-B-F-O-G-A-D, A-B-C-J, L-R 이다[4,5].

#	IP Address	User ID	Time	Method/URL/Protocol	Status	Size	Referer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.04(Win95;I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.04(Win95;I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	GET L.html HTTP/1.0	200	4130	-	Mozilla/3.04(Win95;I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	GET F.html HTTP/1.0	200	5096	B.html	Mozilla/3.04(Win95;I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.01(X11;I;IRIX6.2-IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	GET B.html HTTP/1.0	200	2050	A.html	Mozilla/3.01(X11;I;IRIX6.2-IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	GET R.html HTTP/1.0	200	8140	L.html	Mozilla/3.04(Win95;I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	GET C.html HTTP/1.0	200	1920	A.html	Mozilla/3.01(X11;I;IRIX6.2-IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	GET O.html HTTP/1.0	200	2270	F.html	Mozilla/3.04(Win95;I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	GET J.html HTTP/1.0	200	9430	C.html	Mozilla/3.01(X11;I;SunOS 5.5.1 sun4u)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	GET Q.html HTTP/1.0	200	7220	R.html	Mozilla/3.04(Win95;I)
12	123.456.78.9	-	[25/Apr/1998:05:05:22 -0500]	GET A.html HTTP/1.0	200	3290	-	Mozilla/3.04(Win95;I)
13	123.456.78.9	-	[25/Apr/1998:05:06:03 -0500]	GET D.html HTTP/1.0	200	1680	A.html	Mozilla/3.04(Win95;I)

[그림 2] 표준을 따르는 웹 로그의 일부

세션은 한 사용자가 어떠한 웹 사이트에 요청을 처음으로 시도해서 사이트를 떠날 때까지 발생한 일련의 과정이다. 그렇지만, 웹 로그에는 사용자가 떠나는 시점에 대한 기록이 없으므로 보통 일정 시간을 기준으로 세션을 구별하게 된다. 대부분 경험적인 연구 결과로 25.5~30분을 기준으로 한다[4,5]. 30분을 기준으로 사용자들의 방문 경로를 세션으로 구분한 결과는 A-B-F-O-G, A-D, A-B-C-J, L-R 이다.

위의 작업을 거쳐 사용자 세션을 구하여 보면 실제로 웹 사이트의 구조상 생성될 수 없는 세션이 발생하는데 이는 주로 로컬 캐쉬(cache) 때문이다. 그러므로 구하여진 세션은 웹 사이트의 구조를 참고하여 부족한 부분을 보충하여야 한다. 보정 작업을 거쳐 최종적으로 생성된 세션은 다음과 같다. A-B-F-O-F-B-G, A-D, A-B-A-C-J, L-R[4,5].

세션은 웹 사용 마이닝의 실제의 입력 데이타로서 사용되는 매우 중요한 정보이다. 그러나, 세션은 사용자가 자신의 목적에 부합하는 페이지를 찾기까지 중간에 발생한 과정도 포함하고 있다. 만일 이러한 세션 정보를 전부 입력 데이타로 사용하게 되면 매우 많은 무의미한 패턴과 오버헤드가 발생하게 되므로 각 세션마다 의미 있는 페이지만을 고려한 세션이 유용하다[4,5].

3.3 인기도 기반 전진 참조 기법에 의한 트랜잭션 구분

일반적인 최대 전진 참조 기법에서는 세션의 각 참조들을 순차적으로 검색을 진행하다가 후진 참조가 발생하는 경우 후진 참조가 발생하기 이전의 참조를 최대 전진 참조라 하여 사용자가 관심을 가지고 방문한 참조로 판단한다.

인기도 기반 전진 참조 기법에서는 그 페이지의 인기도를 고려하여 일정 이상의 인기도를 가진 페이지일 경우에만 내용 페이지로 판단한다. 만일 해당 참조가 보조 페이지일 경우에는 무시하고 분기점(breakpoint)까지 역순으로 검색한다. 만일 분기점까지 모두 보조 페이지 뿐이라면 이 트랜잭션은 사용자의 "실수"로 생긴 트랜잭션이며 무의미한 트랜잭션으로 간주하여 삭제한다. 인기도는 웹 사이트 설계자가 직접 지정할 수도 있지만 사용자의 관점에서의 인기도가 보

다 의미 있으므로 일정 기간이상의 로그를 분석하여 얻는다. 인기도 정의는 다음과 같다.

$$\text{인기도} = \frac{in-link}{in-link+out-link}$$

위의 수식에서 *in-link*는 해당 페이지로의 참조가 발생한 횟수이며, *out-link*는 해당 페이지에서 다른 페이지로의 참조가 발생한 횟수를 의미한다. 즉, 다른 페이지들로부터 어떤 페이지로의 참조 발생 비율이 높다면 그 페이지는 높은 인기도를 갖게 되며 내용 페이지로 정의할 수 있다.

[표 1] 인기도 계산 결과

페이지	in-link	out-link	인기도	페이지	in-link	out-link	인기도
A	931	1337	0.41	K	280	0	1.00
B	686	826	0.45	L	98	119	0.45
C	210	245	0.46	M	182	0	1.00
D	154	280	0.35	N	105	154	0.41
E	287	385	0.43	O	119	189	0.39
F	539	196	0.73	P	77	0	1.00
G	147	0	1.00	Q	147	0	1.00
H	140	0	1.00	R	119	0	1.00
I	133	147	0.48	S	154	0	1.00
J	112	0	1.00	T	189	0	1.00

[그림 1]을 참고로 예제 사이트를 구축하고 1주일 동안의 로그를 수집, 분석하여 [표 1]에 결과를 나타내었다. 0.5 이상의 인기도를 기준으로 인기도 기반 기법을 적용하여 트랜잭션을 구분한 경우 내용 페이지로 간주된 페이지는 F-G, R, J 이다.

3.4 정확도 비교

어떠한 패턴을 찾기 위해서는 세션을 의미 있는 트랜잭션으로 구분하는 것이 필요하다. 예를 들어 연관 법칙을 찾기 위한 경우에는 사용자가 참조한 내용 페이지가 관심 대상이다. 다른 페이지들은 사용자가 정보를 찾는데 있어 편리함을 제공해 주는 것이며 이런 참조들은 보조 페이지의 특성을 갖는다. 어떠한 사용자에게 보조 페이지인 것이 다른 사용자에게 내용 페이지가 되는 경우는 드물다. 내용 페이지만으로 구성된 트랜잭션을 이용하면 두 내용 페이지들간의 관계를 발견할 수 있다. 중간 경로에 대한 어떠한 정보도 필요 없이 말이다. 내용 페이지들 사이의 참조를 발견하고자 할 때에는 내용 페이지들만을 고려한 트랜잭션들 만이 유용하다. A→B라는 연관 법칙은 A가 집합에 있을 때, B도 그 집합에 존재한다는 것을 의미한다. 만일 이 결과가 내용 페이지만을 고려한 트랜잭션의 결과로 생성되었다면 이는 좀더 특별한 의미를 갖는다. 즉, A와 B 모두가 내용 페이지로서 사용되었을 때에만 A→B가 성립한다는 의미를 갖는다. 이러한 성질은 데이타 마이닝시 로그의 모든 페이지를 포함한 트랜잭션을 이용하면 놓칠 수 있는 법칙을 발견할 수 있도록 한다. 만일 A 페이지를 보조 페이지로 이용한 사용자는 B 페이지로 가지 않을 것이므로 A 페이지와 B 페이지가 같은 집합에 존재하지 않는 경우가 많이 발생할 것이다. 그러므로 보조 페이지를 포함한 트랜잭션을 이용한 마이닝 결과는 A→B 법칙의 신뢰도를 떨어뜨린다. 이는 패턴 분석의 목적에 따라 장단점이 될 수 있지만 중요한 것은 보조 페이지를 포함한 트랜잭션을 사용할 경우는 제한되어 있다는 것이다[4].

일반적인 최대 전진 참조 기법을 이용한 경우와 인기도 기반 기법을 사용한 경우의 결과를 정리하여 보면 [표 2]와 같다.

[표 2] 트랜잭션 구분 결과

원래 세션	A-B-F-O-F-B-G, A-D, A-B-A-C-J, L-R
M.F.R	O-G, R, B-J, D
P.F.R	F-G, R, J

[표 2]의 결과를 비교하여 보면 일반적인 최대 전진 참조 기법에서는 보조 페이지인 O, B, D 페이지가 선택되었으나, 인기도 기반 기법을 이용하면 보조 페이지는 모두 누락되고 대신 내용 페이지인 F가 추가로 선택되었다. 이러한 결과는 내용 페이지만을 고려한 트랜잭션을 생성하고자 하는 목적에 좀 더 잘 부합됨을 알 수 있다.

4. 결론 및 향후 연구계획

데이타 마이닝의 단계중 패턴 발견을 위해 초기 데이타를 정제하는 전처리 과정은 매우 중요한 작업이다. 전처리 과정의 결과가 높은 정확도를 가지고 있다면 마이닝의 결과 역시 보다 정확한 결과를 생성한다는 것은 여러 연구를 통해 널리 알려진 사실이다. 본 논문에서는 전처리 과정에서 자주 이용되는 일반적인 최대 전진 참조 기법을 개선한 인기도 기반 전진 참조 기법을 제안하고 그 결과가 더 정확한 결과를 산출함을 보였다.

향후 연구과제로는 단순하게 *in-link*와 *out-link*의 횟수만으로 계산하는 현재의 인기도 계산 방법의 개선이 주요한 과제로 남아있다. 각 페이지들의 인기도를 사용자 관점에서 좀 더 정확하게 계산하고 기준치를 정의할 수 있는 방법의 연구와 실험이 필요하다. 또한 이 기법을 이용하여 실험적으로 구축된 사이트가 아닌 실제 웹 사이트를 분석하는 실험을 수행하여 제안된 기법의 유용성을 검증할 것이다.

참고문헌

[1] R. Cooley, M. Bobasher, J. Srivastava, *Web Mining : Information and Pattern Discovery on the World Wid Web*, IEEE, 1997.

[2] J. Srivastava, R. Cooley, M. Deshpande, Pang-Ning Tan, *Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data*, SIGKDD, Jan 2000.

[3] Ming-Syan Chen, Jong Soo Park., Philip S. Yu, *Efficient Data Mining for Path Traversal Patterns*, International Conference on Distributed Computing Systems, May 1996.

[4] R. Cooley, M. Bobasher, J. Srivastava, *Data Preparation for Mining World Wide Web Browsing Patterns*, Knowledge and Information Systems, Feb 1999.

[5] 김종달, *웹 로그에서 웹 방문 패턴을 이용한 사용자 웹 방문 패스 클러스터링*, 포항공대 석사학위 논문, 2000년 12월.