웹로그 데이터에 대한 군집분석 알고리즘에 관한 연구

강현철[1] · 한상태[2] · 선영수[3]

요 약

최근 인터넷은 기업이 고객과 접촉할 수 있는 새로운 수단으로써 기업의 홍보나 서비스를 제공하는 기능을 수행할 뿐만 아니라 사업을 위한 중요한 도구로 여겨지고 있다. 따라서 방문자의 웹사이트 이용형태를 파악하기 위한 다양한 기법들이 제시되고 있으며, 웹로그 데이터에 대한 자료분석 기법들이 여러 학문분야에서 연구되고 있다. 본 연구에서는 웹로그 데이터에 대한 군집분석을 위해 거리측도 및 분석 알고리즘을 제안하였으며, 실제 자료에 이를 적용하여 제안된 알고리즘의 특성을 살펴보았다.

주요용어 : 웹로그, 군집분석, 유사성, 거리측도

1. 서론

인터넷은 기업이 고객과 접촉할 수 있는 새로운 수단으로써 기업의 홍보나 서비스를 제공하는 기능을 수행할 뿐만 아니라 사업을 위한 중요한 도구가 되고 있다. 따라서 인터넷을 통한 고객과의 커뮤니케이션 및 관계유지는 웹사이트를 운영하고 있는 많은 기업들의 주요 관심사항이 되고 있으며, 기존의 CRM(Customer Relationship Management)에 대응되는 eCRM이 새롭게 주목받고 있다.

웹사이트 방문자의 행위들은 각 기업의 웹서버에 웹로그(web log)라는 기록으로 남게 되며, 웹사이트 운영자의 주요 관심은 이러한 방문자의 접속 패턴을 파악하는 것이라고 할 수 있다. 웹마이닝은 크게 웹컨텐츠(web contents) 마이닝과 웹유시지(web usage) 마이닝으로 구분된다(Srivastava et. al., 2000; Cooley et. al., 1999). 특히 웹유시지 마이닝은 데이터마이닝의 한 분야로 웹서버 로그로부터 웹 사용자의 의미 있는 접속패턴을 발견하고 이를 통해 웹사이트의 개선 및 고객에 대한 차별적 서비스 제공 등을 수행하고자 하는 것으로 eCRM의 주된 도구로 사용되고 있으며, 주로 다음과 같은 내용들을 포함한다; (1) 필터링 및 세션 구분 등 웹로그 데이터의 사전처리, (2) 방문회수 등에 대한 기초통계분석, (3) 웹유시지 패턴의 파악 및 분석. 또한 이러한 분석을 통해서 얻어진 결과는 웹사이트 및 컨텐츠의 개선이나 페이지 및 상품의 추천 등에 이용된다.

웹유시지 마이닝에서 방문자의 패턴을 파악하고 분석하기 위해 주로 사용되는 기법으로 연관성규칙발견과 군집분석을 들 수 있다. 웹유시지 마이닝에서의 군집은 유사한 행동패턴을 보이는 사용자의 그룹이며, 이를 통해 시장을 세분화하고 사용자에게 개인화된 웹컨텐츠를 제공하는데 유용하게 이용될 수 있다. 특히 웹로그 데이터에 대한 군집분석에서는 로그 데이터의 특성을 고려한 사전처리나 컨텐츠의 구조를 반영한 거리의 계산 등 일련의 과정이 필요하게 되는데, 본 논문에서는 웹로그 데이터에 대한 군집분석을 위해 웹사이트의 구조를 고려한 거리측도 및 분석 알고리즘을 제안하고자 한다.

1) 호서대학교 자연과학부 정보통계학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1

2) 호서대학교 자연과학부 정보통계학전공 교수, (336-795) 충남 아산시 배방면 세출리 산 29-1

3) 호서대학교 자연과학부 통계전공 석사과정, (336-795) 충남 아산시 배방면 세출리 산 29-1

2. 군집분석을 위한 웹로그 데이터의 사전처리

특정 웹사이트에 방문하는 웹 사용자들이 웹페이지를 클릭하거나 특별한 요청에 대해 웹서버가 응답할 때마다 그 사이트를 관리하고 있는 서버에는 로그라고 불리는 레코드들이 저장된다. 이와 같이 저장된 일련의 레코드들의 집합을 웹로그 데이터라고 한다. 이러한 웹로그 데이터에는 여러 가지 필드들이 저장되는데, 여기에는 Host(방문자의 인터넷 주소, IP 주소), AuthUser (웹서버에 등록된 사용자 이름), Time(접속일자와 시간), Request(GET 및 POST 등의 명령어, 실제 요청대상의 파일 이름, 전송 프로토콜 및 버전), Status(접속상태와 데이터의 이동 현황), Bytes(사용자가 실제로 웹서버에서 가져간 데이터의 양), Referrer(요청이나 링크의 원래 소스), User-Agent(사용자의 요청을 만든 소프트웨어 및 운영체제의 이름과 버전) 등이 포함된다.

이러한 웹로그 레코드를 분석함에 있어 중요하게 고려해야 할 개념 중 하나는 사용자 세션(user session)의 구분이다. 세션(session)이란 사용자가 한 웹사이트를 방문하여 일련의 연속적인 행동을 수행한 후 접속을 중단할 때까지의 과정을 의미한다. 사실 로그 데이터 자체는 여러 사용자의 접속상황이 단지 시간 순서에 의해서 기록된 것이기 때문에 사용자가 언제 새로운 접속을 시도하여 언제 그 접속을 종료하였는지에 대한 정보가 존재하지 않는다. 따라서 웹로그 데이터를 분석하는 경우 세션을 구분하지 않으면 방문 회수나 클릭 빈도가 과장되어 계산될 수 있으며, 연관성규칙발견이나 군집분석을 수행하는 경우에도 의미 있는 결과를 얻기 위해서는 사용자 세션이 기본 분석단위가 되어야 한다.

사용자 세션을 구분하기 위해 몇 가지 방법이 제안되어 있지만, 현재 일반적으로 사용되는 방법은 Time 필드의 시간간격을 이용하는 것이다. 즉, 먼저 사용자 ID(또는 IP 주소)와 Time 필드를 키(key)로 하여 로그 데이터를 정렬하고, 동일 ID 내에서 일정 시간 이상의 시간간격이 발생하면 새로운 세션 ID를 부여하는 것이다(대부분의 상용 소프트웨어에서는 시간간격의 디폴트 설정값으로 30분을 사용하고 있다). 이러한 세션 구분 이외에도 웹로그 데이터를 효율적으로 처리하기 위해서는 불필요한 레코드(예를 들면, Request 필드의 확장자가 gif나 jpg인 이미지 파일, cgi인 스크립트 파일, avi나 mov인 오디오 및 비디오 파일 등인 경우)의 제거 등 여러 단계의 사전처리 과정을 거쳐야 한다(Kang & Jung, 2001).

프로파일 행렬은 동일한 사용자 또는 사용자 세션에서 각 웹페이지가 요청되었는지의 여부(또는 요청된 회수)를 정리한 데이터 행렬이다. 이 때 프로파일 행렬을 작성하기 전에 분석단위를 사용자로 할 것인지 아니면 세션으로 할 것인지를 먼저 결정해야 한다. 사용자를 분류(세분화)하여 마케팅 활동에 활용하기 위해서는 최근 몇 주(또는 개월) 간의 로그 레코드를 이용하여 사용자 단위의 프로파일 행렬을 작성하여 분석하는 것이 일반적이며, 웹사이트의 변환이나 개선을 목적으로 하여 방문정보를 분석하기 위해서는(예를 들어, 동일한 방문 내에서 어떤 컨텐츠 페이지들이 함께 요청되는 경향이 있는지를 파악하기 위해서는) 세션 단위의 프로파일 행렬을 작성하는 것이 좋다(3절 <표 3.3> 참조).

3. 군집분석 알고리즘 및 사례분석

3.1 코사인(유클리드) 거리측도

사용자 또는 사용자 세션 간의 유사성을 측정하기 위해 가장 일반적으로 사용되는 방법은 두 행벡터 간의 정규화된 코사인(normalized cosine)을 이용하는 것이다(Mobasher et al., 2000). 즉, 프로파일 행렬의 i번째 행벡터를 $x_i=(x_{i1}, x_{i2}, \cdots, x_{iN})'$라고 할 때, 두 세션 간의 유사도는 다음과 같이 정의할 수 있다.

$$S_a(x_i, x_j) = \frac{x_i' x_j}{\sqrt{|x_i||x_j|}} \quad (1)$$

이와 같은 유사도 S_a는 통상적인 군집분석에서 흔히 사용되며 계산이 비교적 간단하다는 장점을 가지고 있다.

3.2 차원축소를 이용한 군집분석

프로파일 행렬의 각 원소는 해당 웹페이지에 대한 방문여부 또는 방문회수를 나타내므로 웹페이지의 수가 매우 많은 경우에는 원소 x_{ik}의 대분분이 0의 값을 가지게 된다. 이러한 경우 프로파일 행렬에 기초하여 계산된 유사도는 0에 가까운 값들을 가지기 때문에 군집분석이 적절하게 수행되지 않을 가능성이 많다. 이와 같은 문제점을 해결하기 위한 한 가지 방법으로 프로파일 행렬을 저차원 공간으로 근사시키는 특이값분해(singular value decomposition)를 이용하는 것을 들 수 있다. 즉, $m \times n$ 프로파일 행렬 X에 대해 특이값분해를 적용하여 $m \times k$ 행렬 V를 얻은 후(여기서 k는 n보다 매우 작은 수), 행렬 V에 기초하여 군집분석을 수행하는 것이다.

3.3 웹페이지들 간의 유사성을 이용한 거리측도 및 군집분석 알고리즘

앞에서 설명한 코사인 거리는 웹사이트의 구조 및 웹페이지 간의 연관성을 반영하지 못한다는 단점을 가지고 있다. 특히 웹페이지의 URL이 잘 정리된 계보적 구조를 가지고 있을 때는 유사한 웹페이지를 방문한 두 사용자 세션 간에 더 높은 유사도를 부여하는 것이 바람직한 경우가 있다. 따라서 본 논문에서는 웹페이지들 간의 구조적 관련성을 가중치로 하는 유사성 측도 및 이에 기초한 다음과 같은 군집분석 알고리즘을 제안한다.

단계 1. 각 웹페이지 URL에 대한 $n \times p$ 코드행렬 $A = (a_1, a_2, \cdots, a_p)$를 작성한다(여기서 n은 URL의 수, p는 URL의 최대 길이를 나타낸다). 이는 다음 단계의 과정들을 효율적으로 처리하기 위해 필요하다(표 3.1).

<표 3.1> URL 코드행렬 A의 예

웹페이지 URL	a1	a2	a3	a4	a5	a6	코드
/contents/cont/newmusic/music.html	C2	C5	N3	M8	_	_	C2C5N3M8____
/contents/entertainment.htm	C2	E3	_	_	_	_	C2E3_____
/contents/KMTV/asx/live.asx	C2	K3	K5	L2	_	_	C2K3K5L2____
/contents/sub_html/zlinker_fra.htm	C2	S3	Z2	_	_	_	C2S3Z2_____
/contents/wiselog/broad/index.html	C2	W3	K6	I1	_	_	C2W3K6I1____
/contents/wiselog/frame/contents/cineone.html	C2	W3	K7	C2	C7	_	C2W3K7C2C7__
/contents/wiselog/frame/contents/cool.html	C2	W3	K7	C2	C8	_	C2W3K7C2C8__
/contents/wiselog/frame/contents/happyclinic.html	C2	W3	K7	C2	H4	_	C2W3K7C2H4__
...	...	...	...	...	...	...	...

<표 3.2> URL 연관성행렬 W의 예

	1	2	3	4	5	6	7	8	···
1	1	0.167	0.167	0.167	0.167	0.167	0.167	0.167	···
2	0.167	1	0.167	0.167	0.167	0.167	0.167	0.167	···
3	0.167	0.167	1	0.167	0.167	0.167	0.167	0.167	···
4	0.167	0.167	0.167	1	0.167	0.167	0.167	0.167	···
5	0.167	0.167	0.167	0.167	1	0.333	0.333	0.333	···
6	0.167	0.167	0.167	0.167	0.333	1	0.667	0.667	···
7	0.167	0.167	0.167	0.167	0.333	0.667	1	0.667	···
8	0.167	0.167	0.167	0.167	0.333	0.667	0.667	1	···
···	···	···	···	···	···	···	···	···	···

<표 3.3> 세션 프로파일 행렬 X의 예

User ID	S_ID	C2W3K7O7__	C2W3N2N4__	C2W3S5E7__	C2W3S6___	C2W3W4K8__	...
meba	1	0	0	1	0	0	...
meba	2	0	0	0	0	0	...
meba	3	0	0	0	0	0	...
meba	4	0	1	1	1	0	...
mozala	1	0	0	0	0	0	...
mozala	2	0	0	0	0	0	...
mozala	3	0	0	0	0	0	...
mozala	4	0	1	1	1	0	...
mozala	5	0	0	0	0	0	...
mozala	6	0	1	0	1	0	...
mozala	7	1	1	0	1	1	...
oldtom	1	0	1	0	1	0	...
oldtom	2	0	1	0	1	0	...
...	...	...	...	...	...	...	...

단계 2. URL 코드행렬 A로부터 URL들 간의 연관성행렬 $W(n\times n)$를 작성한다(표 3.2). 여기서 w_{kl}은 두 웹페이지 간의 구조적 연관성을 측정한 것으로 다음과 같이 계산된다. 즉, k번째 웹페이지의 URL 코드를 $a_k=(a_{k1}\ \ a_{k2}\cdots\ a_{kp})$라고 할 때,

$$w_{kl}=\sum_{i}^{p} I(a_{ki}=a_{li}|\ a_{k1}=a_{l1},\cdots,a_{k,i-1}=a_{l,i-1})/p \qquad (2)$$

여기서 $I(a_{ki}=a_{li}|\ \cdot\ )$는 a_{ki}와 a_{li}가 같은 문자코드이면 1 그렇지 않으면 0의 값을 가진다.

단계 3. 방문자 세션 또는 방문자 프로파일 행렬 $X(m\times n)$를 작성한다. <표 3.3>은 세션 프로파일 행렬의 예로써 여기서는 각 방문자 세션에서 해당 URL을 방문하였으면 1 그렇지 않으면 0의 값을 가지도록 작성되었다.

단계 4. 행렬 W와 X를 이용하여 다음 식 (3)과 같이 세션들 간의 유사성행렬 $S(m\times m)$를 작성한다.

<표 3.4> 행렬 D의 예

	JB 12	JUN 1	JUN 2	JUN 3	JUN 4	JUN 5	JUN 6	JUN 7	JUN 8	JUN 9
JB 12	0									
JUN 1	0.786	0								
JUN 2	0.787	0.257	0							
JUN 3	0.804	0.533	0.491	0						
JUN 4	0.722	0.267	0.364	0.457	0					
JUN 5	0.849	0.201	0.282	0.408	0.309	0				
JUN 6	0.881	0.419	0.423	0.457	0.434	0.271	0			
JUN 7	0.931	0.198	0.227	0.435	0.280	0.200	0.311	0		
JUN 8	0.633	0.796	0.824	0.688	0.680	0.797	0.612	0.783	0	
JUN 9	0.633	0.796	0.824	0.896	0.845	0.797	0.744	0.891	0.800	0

$$S_b(x_i, x_j) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} w_{kl} x_{ik} x_{jl}}{\sum_{k=1}^{n} x_{ik} \sum_{k=1}^{n} x_{jk}} \quad (3)$$

단계 5. 식 (3)으로부터 계산된 유사성행렬 S의 대각원소는 일반적으로 1보다 작다. 또한 $s_{ij} \leq Min(s_{ii}, s_{jj})$가 성립하므로, $s_{ij}^{*} = s_{ij}/\sqrt{s_{ii}s_{jj}}$와 같이 보정하여 행렬 S^{*}를 작성한다. 이와 같이 보정된 유사성행렬 S^{*}는 다음 관계를 만족한다. $s_{ij}^{*} \leq 1$, $s_{ii}^{*} = 1$, $i, j = 1, 2, \cdots, m$.

단계 6. 행렬 S^{*}로부터 $d_{ij} = 1 - s_{ij}^{*}$를 계산하여 거리행렬 $D(m \times m)$를 작성한다(표 3.4).

단계 7. 작성된 거리행렬 D를 기초로 통상적인 군집분석을 수행한다.

다음 <표 3.5>는 이와 같은 과정을 수행하여 얻은 군집분석의 결과로써 각 군집의 평균 프로파일(방문비율)의 일부를 제시한 것이다. 이 결과를 살펴보면 군집 1과 2는 많은 웹페이지에 대해 전반적으로 높은 방문비율을 가지고 있고, 반면에 군집 3, 4, 5는 대부분의 웹페이지에 대해 방문비율이 높지 않으나 특정한 웹페이지에 대해 상대적으로 높은 방문비율을 갖는다는 알 수 있다. 따라서 군집 3, 4, 5에 속하는 방문자들은 특정한 소수의 웹페이지에 관심이 많은 사용자들이라는 것을 유추할 수 있다. 본 연구에 사용된 웹로그 데이터는 웹로그 분석 소프트웨어인 wiselog enterprise(www.nethru.co.kr)에 포함되어 있는 것으로, 초기 웹로그 레코드의 개수는 6,012개 이고 이미지나 스크립트 레코드를 제외한 4,253개의 레코드가 분석에 사용되었다.

4. 결론

웹로그 데이터 분석을 이용한 개인화 또는 추천시스템이 현재 많은 관심을 가지고 연구되고 있는데, 이는 방문자의 사용패턴에 근거하여 특정 웹페이지를 사용자마다 다르게 구성해 주거나 특정 페이지를 읽도록 추천하고자 하는 것이다. 또한 쇼핑몰을 운영하고 있는 웹사이트에서는 특별한 사용패턴을 가지는 방문자들이 주로 어떤 속성(예를 들어, 연령 및 성별 등)을 가지고 있으며 어떤 상품을 구매하는 경향이 있는지 등을 파악하여, 실시간 또는 전자메일을 이용한 상품추천에도 응용할 수 있다.

<표 3.5> 군집분석 결과: 평균 프로파일

URL	전체	군집 1	군집 2	군집 3	군집 4	군집 5
C2W3N2N4____	0.361	0.980	0.044	0.000	0.100	0.000
I6_____	0.569	0.940	0.689	0.121	0.000	0.000
I8_____	0.542	0.920	0.711	0.000	0.000	0.000
IS_____	0.556	0.920	0.711	0.061	0.000	0.000
M2_____	0.528	0.900	0.533	0.182	0.100	0.000
I2_____	0.542	0.880	0.711	0.061	0.000	0.000
W1_____	0.507	0.880	0.556	0.121	0.000	0.000
I9_____	0.486	0.800	0.622	0.061	0.000	0.000
I4_____	0.549	0.780	0.867	0.030	0.000	0.000
B7_____	0.368	0.240	0.244	0.848	0.200	0.000
D5_____	0.229	0.140	0.111	0.636	0.000	0.000
G4_____	0.271	0.260	0.133	0.606	0.000	0.000
F8_____	0.236	0.200	0.111	0.576	0.000	0.000
G2_____	0.229	0.180	0.156	0.364	0.500	0.000
G8_____	0.125	0.060	0.022	0.364	0.100	0.167
D8_____	0.201	0.260	0.133	0.212	0.100	0.333
C2W3K7C2P2__	0.063	0.000	0.044	0.121	0.300	0.000
D2_____	0.069	0.060	0.067	0.030	0.300	0.000
F4_____	0.083	0.120	0.067	0.000	0.300	0.000
B3_____	0.042	0.060	0.022	0.000	0.000	0.333
M1M4_____	0.146	0.080	0.222	0.121	0.100	0.333
B6_____	0.014	0.020	0.000	0.000	0.000	0.167
B2_____	0.021	0.040	0.000	0.000	0.000	0.167
...	...	...	...	...	...	...
평균	0.124	0.183	0.124	0.084	0.033	0.013
군집크기	144	50	45	33	10	6

본 연구에서는 웹로그 데이터에 대한 군집분석 절차를 제시하고 웹페이지들 간의 구조적 관련성을 고려한 거리를 제안하였다. 이러한 분석을 통해 얻은 결과는 웹페이지의 개선 및 웹사이트의 구조 변경 등에 이용될 수 있을 뿐만 아니라 웹페이지 및 상품을 추천하기 위한 추천 시스템의 구축에도 응용될 수 있을 것이다.

참고문헌

[1] Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, 1, 5-32.

[2] Kang, H. and Jung B.C. (2001). A Study of Web Usage Mining for eCRM, *The Korean Communications in Statistics*, 8, 831-840.

[3] Mobasher, B., Cooley, R., and Srivastava, J. (2000). Automatic Personalization Based On Web Usage Mining, *Communication of ACM*, 43, 142-151.

[4] Srivastava, J., Cooley, R., Deshpande, M., and Ten, P. N. (2000). Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data, *SIGKDD Explorations*, 1, 12-13.