The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 191-197

A Clustering Algorithm Considering Structural Relationships of Web Contents[1)]

Hyuncheol Kang[2)], Sang-Tae Han[3)], and Young-Su Sun[4)]

Abstract

Application of data mining techniques to the world wide web, referred to as web mining, has been the focus of several recent researches. With the explosive growth of information sources available on the world wide web, it has become increasingly necessary to track and analyze their usage patterns. In this study, we introduce a process of pre-processing and cluster analysis on web log data and suggest a distance measure considering the structural relationships between web contents. Also, we illustrate some real examples of cluster analysis for web log data and look into practical application of web usage mining for eCRM.

Keywords : Web Mining, Web Log, Cluster Analysis, Similarity, Distance Measure

1. 서론

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 연관성 규칙발견과 군집분석을 들 수 있다(Choi, et al., 2001). 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) This research was supported by Hoseo University research grants in 2004.
2) Assistant Professor, Dept. of Informational Statistics, Hoseo University, Asan 336-795, Korea. E-mail: hychkang@office.hoseo.ac.kr
3) Associate Professor, Dept. of Informational Statistics, Hoseo University, Asan 336-795, Korea.
4) Graduate student, Dept. of Statistics, Hoseo University, Asan, 336-795, Korea.

2. 군집분석을 위한 웹로그 데이터의 사전처리

특정 웹사이트에 방문하는 웹 사용자들이 웹페이지를 클릭하거나 특별한 요청에 대해 웹서버가 응답할 때마다 그 사이트를 관리하고 있는 서버에는 로그(log)라고 불리는 레코드들이 저장된다. 이와 같이 저장된 일련의 레코드들의 집합을 웹로그 데이터라고 한다. 웹로그 데이터에는 일반적으로 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 웹로그 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 있다. 또한 연관성규칙발견이나 군집분석을 수행하는 경우에도 의미 있는 결과를 얻기 위해서는 사용자 세션이 기본 분석단위가 되어야 하는 경우가 많다.

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 일반적이다. 반면에 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(3절의 표 3.3 참조).

3. 군집분석 알고리즘

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, 특정 두 세션 간의 유사도를 다음과 같이 정의할 수 있다.

$$S_a(x_i, x_j) = \frac{x_i' x_j}{|x_i||x_j|} \tag{3.1}$$

단, $|x_i| = \sqrt{x_i' x_i}$. 이와 같은 유사도 S_a는 통상적인 군집분석에서 흔히 사용되어지며 계산이 비교적 간단하다는 장점을 가지고 있다.

그러나 위의 유사성 측도는 공통적으로 방문되는 웹페이지(웹컨텐츠)들의 수만을 고려하며, 웹페이지들의 구조적 연관성이 고려되지 않는다는 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 측도를 제안한다. 또한 실제 분석에서 데이터 처리 및 군집분석을 효율적으로 처리할 수 있는 다음과 같은 알고리즘을 제안한다.

단계 1. 각 웹페이지 URL에 대한 $n \times p$ 코드행렬 A를 작성한다(여기서 n은 URL의 수, p는 URL의 최대 길이). 여기서 행렬 A의 k번째 행 $\underline{a_k} = (a_{k1}, a_{k2}, \cdots, a_{kp})'$는 k번째 URL의 코드를 나타낸다(표 3.1). 이는 다음 단계의 과정들을 효율적으로 처리하기 위해서 필요하다.

<표 3.1> URL 코드행렬 A의 예

실제 웹페이지 URL	코드 ($\underline{a_k}$)			
/corporate/backgrounder/intro.htm	$\underline{a_1}$	C2	B3	I8
/corporate/backgrounder/overview.htm	$\underline{a_2}$	C2	B3	O3
/corporate/default.htm	$\underline{a_3}$	C2	D1	_
/corporate/managpc710.htm	$\underline{a_4}$	C2	M1	_
/corporate/profile.htm	$\underline{a_5}$	C2	PF	_
...		...	...	...

단계 2. URL 코드행렬 A로부터 URL들 간의 연관성행렬 $W(n \times n)$를 작성한다(표 3.2). 여기서 행렬 W의 (k, l) 번째 원소 w_{kl}은 두 웹페이지 간의 구조적 연관성을 측정한 것으로 다음과 같이 계산된다(단, $k = 1, \cdots, n;\ l = 1, \cdots, n$).

$$w_{kl} = \frac{1}{p} \sum_{i=1}^{p} I(a_{ki} = a_{li} | a_{k1} = a_{l1}, \cdots, a_{k,i-1} = a_{l,i-1}) \tag{3.2}$$

여기서 $I(a_{ki} = a_{li} | \cdot\ )$는 a_{ki}와 a_{li}가 같은 문자코드이면 1 그렇지 않으면 0의 값을 가진다. 단 $I(a_{k1} = a_{l1} | \cdot\ ) = I(a_{k1} = a_{l1})$이다.

<표 3.2> URL 연관성행렬 W의 예

	a_1	a_2	a_3	a_4	a_5	...
a_1	1	0.667	0.333	0.333	0.333	...
a_2	0.667	1	0.333	0.333	0.333	...
a_3	0.333	0.333	1	0.333	0.333	...
a_4	0.333	0.333	0.333	1	0.333	...
a_5	0.333	0.333	0.333	0.333	1	...
...	...	...	...	...	...	...

단계 3. 방문자 세션 또는 방문자 프로파일행렬 $X(m \times n)$를 작성한다(여기서 m은 세션 또는 방문자의 수를 나타낸다). 표 3.3은 세션 프로파일행렬의 예를 보여주고 있으며, 방문자 세션에서 해당 URL을 방문하였으면 1 그렇지 않으면 0의 값을 가지도록 작성되어 있다.

<표 3.3> 세션 프로파일 행렬 X의 예

User ID	Session ID	a_1	a_2	a_3	a_4	a_5	...
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	...
...	...	...	...	...	...	...	...

단계 4. URL 연관성행렬 W와 프로파일행렬 X를 이용하여 다음 식 (3.3)과 같이 세션들 간의 유사성행렬 $S(m \times m)$를 작성한다. 여기서 x_{ik}는 행렬 X의 (i, k) 번째 원소를 나타낸다.

$$s_{ij} = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} w_{kl} x_{ik} x_{jl}}{\left(\sum_{k=1}^{n} x_{ik}\right)\left(\sum_{k=1}^{n} x_{jk}\right)} \tag{3.3}$$

단계 5. 식 (3.3)으로부터 계산된 유사성행렬 S의 대각원소 s_{ii}는 항상 1보다 작거나 같게 되며, $s_{ij} \leq \sqrt{s_{ii} s_{jj}}$가 성립한다. 따라서 $s^*_{ij} = s_{ij} / \sqrt{s_{ii} s_{jj}}$와 같이 보정하여 행렬 S^*를 작성한다. 이와 같이 보정된 유사성행렬 S^*는 다음 관계를 만족한다; $s^*_{ij} \leq 1$, $s^*_{ii} = 1$, $s^*_{ij} = s^*_{ji} (i, j = 1, 2, \cdots, m)$.

단계 6. 세션들 간의 거리행렬 $D(m \times m)$를 작성한다(표 3.4). 여기서 $d_{ij} = 1 - s^*_{ij}$이다.

<표 3.4> 거리행렬 D의 예

	meba_1	meba_2	meba_3	meba_4	mozala_1
meba_1	0				
meba_2	0.786	0			
meba_3	0.787	0.257	0		
meba_4	0.804	0.533	0.491	0	
mozala_1	0.722	0.267	0.364	0.457	0

단계 7. 거리행렬 D에 기초하여 통상적인 군집분석을 수행한다.

4. 사례분석

본 연구에서 사용된 데이터는 웹로그 분석 소프트웨어인 Webtrends Log Analyzer에 포함되어 있는 것으로(www.webtrends.co.kr), 초기 웹로그 레코드의 개수는 11,370개 이고 이미지 및 스크립트 레코드 등을 제외한 4,522개의 레코드가 분석에 사용되었다. 또한 세션구분 과정을 거쳐 330개의 세션으로 레코드들이 구분되었다.

표 4.1은 앞 절에서 제안된 군집분석 알고리즘을 적용하여 얻은 군집분석의 결과로서 각 군집의 평균 프로파일(방문비율)의 일부를 제시한 것이다(본 연구에서는 계보적 군집분석 방법 중 하나인 Ward 방법을 사용하였다).

이 결과를 살펴보면 먼저 군집 1과 2는 한 번의 세션에서 대략 전체 웹페이지의 4% 정도를 방문(클릭)하며, 반면에 군집 3, 4, 5는 평균 방문비율이 1% 정도로서 소수의 특정한 웹페이지를 방문하는 경향이 있다는 것을 알 수 있다. 또한 군집 1은 P4C8__, P4D1__과 같이 주로 P4로 시작하는 웹페이지들에 대해 방문비율이 상대적으로 높음을 알 수 있고, 군집 2와 군집 3은 R1Z1__, R1C5__과 같이 주로 R1으로 시작하는 웹페이지들에 대해 방문비율이 높음을 알 수 있다. 군집 4는 다른 웹페이지들에 비해 P4D1__과 R1E6__에 대한 방문비율이 매우 높으며, 군집 5는 대부분의 웹페이지에 대해 방문비율이 높지는 않으나 특정한 몇 개의 웹페이지에 대해 상대적으로 높은 방문비율을 갖는다는 것을 알 수 있다. 따라서 군집 4와 5에 속하는 방문자들은 특정한 소수의 웹페이지에 관심이 많은 사용자들이라는 것을 유추할 수 있다.

한편, 표 4.2는 식 (3.1)의 코사인 유사성 측도를 이용하여 얻은 군집분석 결과의 일부이다. 이 결과에서는 군집 1은 한 번의 세션에서 대략 전체 웹페이지의 6% 정도를 방문하며, 반면에 군집 2, 3, 4, 5는 평균 방문비율이 1.2%~1.7% 정도로서 소수의 특정한 웹페이지를 방문하는 경향이 있다는 것을 알 수 있다.

<표 4.1> 제안된 방법을 이용한 군집분석 결과의 일부 (방문비율)

URL	전체	군집 1	군집 2	군집 3	군집 4	군집 5
P4C8__	0.300	0.411	0.458	0.000	0.033	0.050
P4D1__	0.303	0.372	0.323	0.000	0.667	0.050
P4R3__	0.124	0.217	0.135	0.000	0.000	0.000
P4P5D1	0.255	0.318	0.448	0.000	0.000	0.000
R1E6__	0.224	0.194	0.250	0.055	0.733	0.000
R1C5__	0.203	0.155	0.396	0.145	0.000	0.050
R1Y2__	0.115	0.085	0.281	0.000	0.000	0.000
R1Z1__	0.230	0.186	0.458	0.127	0.033	0.000
R1I3__	0.067	0.054	0.125	0.055	0.000	0.000
R1PC__	0.070	0.070	0.115	0.036	0.033	0.000
R1SP__	0.085	0.062	0.156	0.073	0.033	0.000
R1V2__	0.052	0.039	0.104	0.036	0.000	0.000
P1T4__	0.003	0.000	0.000	0.000	0.033	0.000
R1B7__	0.076	0.070	0.125	0.036	0.067	0.000
R1M2__	0.061	0.047	0.104	0.018	0.100	0.000
A1____	0.027	0.016	0.063	0.000	0.000	0.050
C3C7__	0.012	0.008	0.021	0.000	0.000	0.050
O1D1__	0.052	0.078	0.063	0.000	0.000	0.050
Z2____	0.076	0.124	0.083	0.000	0.000	0.050
...	...	...	...	...	...	...
평균	0.029	0.040	0.036	0.009	0.010	0.012
군집크기	330	129	96	55	30	20

<표 4.2> 코사인 유사성 측도를 이용한 군집분석 결과의 일부 (방문비율)

URL	전체	군집 1	군집 2	군집 3	군집 4	군집 5
D1____	0.618	0.771	0.803	0.185	0.242	0.524
R1Z1__	0.230	0.333	0.291	0.074	0.091	0.000
S3D1__	0.255	0.438	0.282	0.000	0.152	0.000
R1C5__	0.203	0.371	0.222	0.019	0.000	0.048
P4C8__	0.300	0.638	0.205	0.037	0.152	0.048
P4D1__	0.303	0.610	0.103	0.019	0.667	0.048
S2____	0.239	0.476	0.051	0.019	0.061	0.952
R1E6__	0.224	0.390	0.026	0.019	0.879	0.000
P2D1__	0.142	0.324	0.060	0.074	0.061	0.000
P4P6D1	0.100	0.295	0.017	0.000	0.000	0.000
P4R3__	0.124	0.219	0.051	0.222	0.000	0.000
R1B7__	0.076	0.171	0.009	0.111	0.000	0.000
R1I3__	0.067	0.143	0.009	0.111	0.000	0.000
R1PC__	0.070	0.124	0.060	0.000	0.091	0.000
D2A0__	0.009	0.019	0.000	0.000	0.030	0.000
C3D1__	0.033	0.057	0.000	0.000	0.030	0.190
D3D1__	0.036	0.086	0.009	0.000	0.000	0.095
A6____	0.003	0.000	0.000	0.000	0.000	0.048
C3C7__	0.012	0.029	0.000	0.000	0.000	0.048
...	...	...	...	...	...	...
평균	0.029	0.059	0.017	0.013	0.015	0.012
군집크기	330	105	117	54	33	21

또한 표 4.2의 결과에서 각 군집에서 방문비율이 높은 웹페이지들을 살펴보면, 표 4.1의 결과에서와는 다르게 여러 주제에 관한 웹페이지들이 서로 섞여 있음을 알 수 있다. 따라서 특정한 공통 주제와 관련된 웹페이지들을 주로 방문하는 방문자 또는 세션을 탐색하고자 하는 경우 본 연구에서 제안된 방법론이 유용하게 사용될 수 있음을 알 수 있다.

5. 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(Sarwar et al., 2000; Schafer et al., 2001; Koh et al., 2002). 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 파악함으로써, 실시간 또는 전자메일을 통한 상품의 추천을 시도하고 있다.

본 연구에서는 웹페이지들 간의 구조적 연관성을 고려한 거리측도를 제안하고 웹로그 데이터에 대한 군집분석 절차를 제시하였다. 이러한 분석을 통해 얻은 결과는 웹페이지의 개선 및 웹사이트의 구조 변경 등에 이용될 수 있을 뿐만 아니라 웹페이지 및 상품을 추천하기 위한 추천시스템의 구축에도 응용될 수 있을 것이다.

참고문헌

[1] Choi, S.B., Kim, K.K., Kang, C.W., Cho, S.K., and Son, J.K. (2001). A study on the comparison of web log analyzers, *Journal of The Korean Data Analysis Society*

[2] Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns, *Journal of Knowledge and Information Systems*

[3] Kang, H. and Jung, B.C. (2001). A study of web usage mining for eCRM, *The Korean Communications in Statistics*, Vol. 8, 831-840.

[4] Koh, B.S., Lee, H.W., and Lee, Y.S. (2002). The empirical study on the issues and improvements about e-commerce Data, *Journal of The Korean Data Analysis Society*, Vol. 4, 217-227.

[5] Mobasher, B., Cooley, R., and Srivastava, J. (2000). Automatic personalization based on web usage mining, *Communication of ACM*, Vol. 43, 142-151.

[6] Sarwar, B.M., Karpis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommender algorithms for e-commerce, *ACM E-Commerce 2000 Conference.*

[7] Schafer, J.B., Konstan, J., and Riedl, J. (2001). Electronic commerce recommender applications, *Journal of Data Mining and Knowledge Discovery*, Vol. 5, 115-152.

[8] 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*,

[2004년 10월 접수, 2005년 2월 채택]