웹 로그를 이용한 고객행동모델 분석방법에 관한 연구

Analysis Procedure For Customer Behavior Model Using Web-Log

서 장 훈* 심 상 용* 유 웅 재*
Seo Jang Hoon* Shim Sang Yong* Yoo Woong Jae*

Abstract

In this report, we provide the focus on suggesting a method of estimating and measurement of CBM(Customer Behavior Model). Through the use of internet, a new trend of business for e-CRM on B2C Web Site known as EC has emerged. The purpose of this study is to identify the relationship between the customers of a shopping mall and CBM characteristics. It can be used to gain a better understanding of customers. From this we can determine trends, and so refine business toward customer's needs and target new products to particular customer groups. Result shows that there is a significant relationship between the customers pattern of shopping mall and CBM, CVM(Customer Visit Model).

1. 서론

B2C(Business to Customer) 모델 전자상거래 사이트의 고객들은 세션이라고 하는 1회 방문 시간에 일련의 연속적인 관련된 요청을 통해서 해당 사이트에 접속한다. 세션 내에서 고객들은 로그인(Login), 탐색(Browse), 찾기(Search), 장바구니에 담기(Add to Shopping) 또는 지불(Pay)과 같은 여러 가지 요청을 할 수 있다. 고객마다 전자상거래 사이트 네비게이션 패턴이 다를 수 있기 때문에 전자상거래 사이트가 제공하는 여러 가지 기능들을 서로 다른 방식과 횟수로 실행시킬 수 있다. 자주 구매하는 고객들도 있고, 광범위하게 이것저것 찾아보면서도 구매는 별로 하지 않는 고객들도 있다. 전자상거래 사이트에 접속해 있는 동안 고객이 보여주는 행동은 전자상거래 사이트의 IT 자원과 수입에 지대한 영향을 준다. 따라서 e-CRM을 위한 전자상거래 사이트의

* 썬더 부설 기술연구소

고객들의 행동 특성을 실시간으로 파악하는 것이 매우 중요하다. 특히 웹 로그 분석을 통해 웹고객의 개개인에게 최적화된 서비스를 제공함은 물론, 웹에 제공될 정보들의 효율적인 배치 및 마케팅에 직접적으로 활용할 정보를 분석하는 것이 기업과 고객에 있어서는 쌍방향 커뮤니케이션이 용이하다는 점에서 그 필요성이 대두되고 있다. 그리고 전자상거래를 위한 Web Site 구조가 갖는 e-CRM이 CRM(Customer Relationship Management)에 비해 상대적으로 갖는 특징을 [그림 1]에서 제시하였다.

본 논문에서는 HTTP(HyperText Transfer Protocol) 로그파일(log files)을 이용한 e-Business Site 고객 행동을 포착하는 이유를 논하고, 사례연구로서 제시된 모델을 이용하여 고객 행동 수준을 분석하고, 그 고객과 Site 접속방식에 대한 정량적 통찰력이 제공하는 것이 어떻게 도움이 될 수 있는지에 대한 방법을 제시하고자 한다.[11]

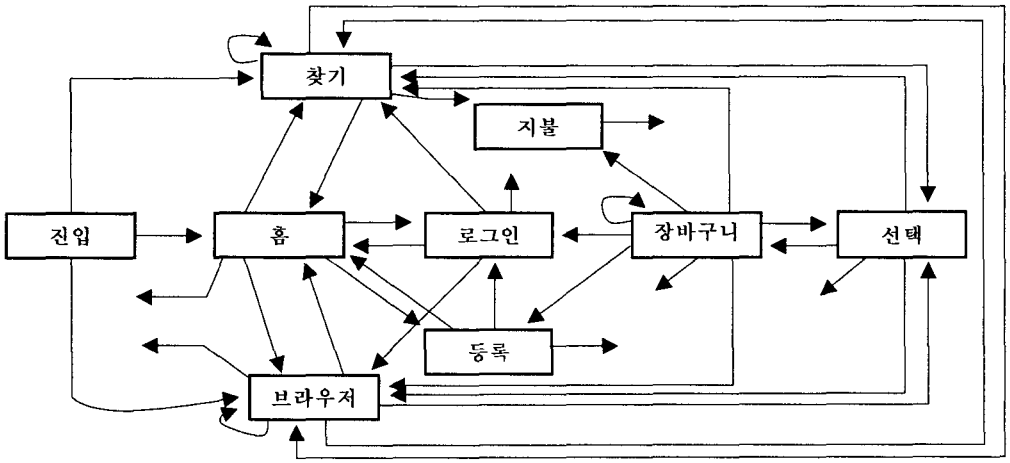
	CRM	eCRM
데이터 수집	영업사원방문, TM(Transaction Model), 데이터 마이닝, 기간계 시스템의 구매데이터 등 복수의 분산된 채널	고객정보, 웹 로그, e-mail 반응, 웹콜센터, 구매 데이터 등 웹 기반의 단일통합 채널
데이터 분석	전통적인 통계기법 데이터 마이닝	실시간 고객성향 분석 고객 행동 분석 등의 웹 마이닝
데이터 활용	마케팅 캠페인 관리, 영업력 강화, 컴퓨터 전화통합, 프로모션 등	원투원 마케팅, 실시간 추천 시스템, 개인화된 컨텐츠 등
비용	높은 인건비, 대규모의 전사적 DW(Data Warehouse) 구축과 시스템 구축으로 상대적으로 높은 비용	초기 IT 비용이 CRM에 비해 저렴하고, 지속적인 관리 비용이 낮음
범위	제한된 시간, 지역적 한계 존재	시간, 지역적 한계 탈피

[그림 1 CRM과 eCRM의 특징 비교]

2. 연구모델

본 논문에서는 사용자 관점에서 가상쇼핑몰 HTTP 로그파일(Log File)을 통하여 e-Business Site 고객 행동을 정량적으로 분석하고, 평가할 수 있는 사용자 네비게이션(Nevigation) 패턴(Pattern) 모델을 제시한다. 이 모델은 네비게이션 패턴, 사용자가 이용하는 전자상거래 기능, 여러 가지 전자상거래 기능들에 대한 접근 빈도, 전자상거래 사이트가 제공하는 여러 서비스에 대한 접근 간격의 측면에서 사용자 행동의 요소들을 파악 할 수 있으며, 사이트 레이아웃(Site Layout) 변경이나 콘텐트 디자인

(Content Design)으로 인한 고객에 대한 영향과 관련하여 여러 가지 문제점을 해결할 수 있는 방안을 제시 할 수 있다. 그리고 이 모델을 이용하여 사용자의 미래 행동을 예측하고, 객체들을 미리 페치(fetch)하여 성능을 향상 시킬 수 있는 이점이 있다.



[그림 1. 가상 쇼핑몰 사이트 CBMG의 상태 전이 모델]

3. Log 파일과 통계 데이터와의 차이점

로그 파일은 웹 서버에 대한 모든 방문객들의 접근을 기록한 데이터로 HTTP Protocol의 일부로 명시된 Common Log Fomat 또는 Extended Log Format을 따라 저장된다. 저장된 로그 데이터는 웹 서버에 접속한 방문객의 IP(internet protocal) 주소, 접근시각, 접근 방법, 대상 URL, 전송 Protocol, 에러 코드, 전송 바이트 수와 같이 방문객을 인식할 수 있는 정보와 웹 페이지에 대한 방문 정보들을 포함고 있으며, 다음 같은 대표적인 특징을 가지고 있다.

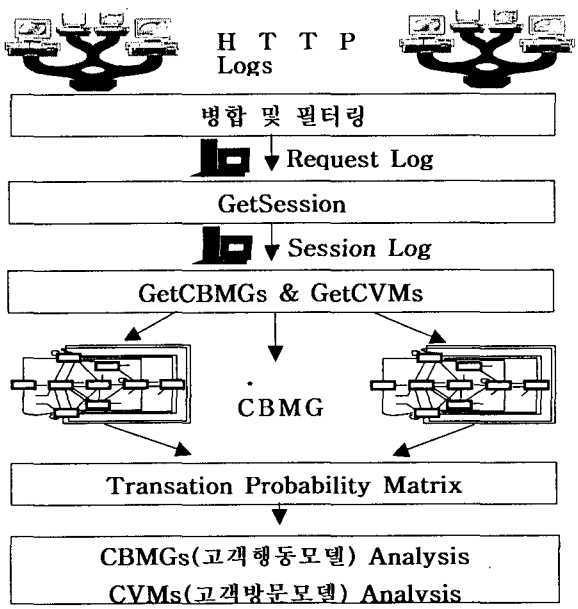
① 데이터를 수집하기 위한 특별한 절차가 필요없다.
② 분석에 불필요한 정보를 많이 포함하고 있다.
③ 분석을 위해서는 추가적인 정보가 필요하다.
④ 일반적으로 대용량이다.

위 로그 데이터의 특징과 같이 현실 세계에서 나타나는 데이터의 대부분은 실험 데이터와는 달리 이와 같은 속성을 반드시 갖지 않으며, 분석에 바로 이용될 수 있는 형태도 아니다. 따라서 이러한 데이터를 분석하기 위해서는 사전처리가 불가피하며 Friedman(1997), Hand(1997) 등에서 지적했듯이 데이터로부터 어떤 패턴을 발견하기 위한 적절한 분석 방법의 개발이 필요하다.[3][4]

4. 연구방법

로그 파일은 사이트를 방문한 사람들의 방문 데이터를 서버가 인식하여 누가 언제 무엇을 요청했고, 가져갔는지, 웹 서버에 얼마나 많은 사람이 왔는지, 어디서 왔으며, 무엇을 제일 좋아하는지, 무엇을 싫어하는지, 가장 오래 또는 가장 많이 보는 페이지는 무엇인지 등등의 방문자 기록을 정리한 하나의 데이터이다. 이러한 로그 파일을 하나하나 분석하여 현 시장 환경 및 고객들의 반응을 예측하며, e-Business site 전략에 활용할 수 있게 하는 것이 e-CRM의 시작이다. 그렇기 때문에, 본 논문에서는 HTTP 로그 파일을 통해서 얻어지는 데이터를 가지고, ID Maching 기법을 이용하여 고객정보를 추출하기 위한 과정을 [그림 2]에서 소개하였다.

로그 파일은 일반적으로 분석에 불필요한 데이터를 많이 포함하고 있으며, 적절한 분석을 위해서는 기본적인 로그 데이터 이외에 추가적인 정보가 요구된다. 또한 데이터의 분석을 위해서는 먼저 웹 서버에 대한 사용자들의 방문정보를 개개인의 그룹으로 구성하는 세션구분(Session Identification)작업이 수행되어져야 한다. 기존 대부분의 연구들에서는 방문객의 웹사이트 체류 시간을 기준으로 세션을 구분하고 있는데, 본 논문에서는 체류시간과 전이횟수를 고려한 GetSession 알고리즘 방법과 이러한 알고리즘을 통해 세션 로그 S가 생성되면 K-평균 알고리즘을 통해서 고객 유형별 세션 클러스터링을 하고, 세션별 고객행동모델 그래프(CBMGs)를 통해 Markov Chain(마코브 사슬) 전이행렬식을 제시하였다.



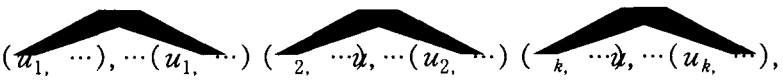
[그림 2. 고객행동모델(CBM) 특성화 방법]

4.1 GetSessions 알고리즘

GetSession 단계에서 세션로그 S를 설명하면, 이 로그의 k번째 항목은 2-튜플(C_k, W_k)로 구성되어 있다. 여기서 $C_k=[c_{i,j}]$는 한 세션에 대한 CBMG의 상태 i와 j간의 전이 횟수의 $n\times n$ 행렬이며, $W_k=[w_{i,j}]$는 한 세션에 대한 CBMG의 상태 i와 j간의 누적 사색 시간의 $n\times n$ 행렬이다. 표기법을 설명하기 위해서, 특정 세션에서 상태 s와 t사이에 3번의 전이가 있었고, 각 전이에 대한 사색시간은 각 20sec, 30sec, 34sec이었다고 가정하자. 그려면, $c_{s,t}=3$이고, $w_{s,t}=20+30+34=84\,\mathrm{sec}$ 이 된다.

1. UserID별로 요청 로그 L을 정렬한 다음 RequesTime별로 요청을 정렬하여 다음 형태와 같이 UserID 당 하나의 부분열로 구성된 정렬 로그 L_s를 생성한다.

u_1의 부분열 u_2의 부분열 u_k의 부분열



2. 각 부분열은 한 개 이상의 세션을 나타낼 수도 있다. 예를 들어, 고객은 연속적인 요청을 생성하고 한 시간 뒤에 사이트에 다시 접속하여 다른 세션을 진행할 수도 있다. 따라서, 부분열은 시간 임계값 T(예를 들어 20분)을 이용하여 세션으로 분할될 필요가 있다.[6] 부분열에 있는 두 개의 연속적 요청 R_1과 R_2 사이의 시간이 T를 초과할 경우, R_1 은 해당 세션의 마지막 요청으로 간주되고 R_2는 다음 세션의 첫 번째 요청으로 간주한다.

3. 이제 부분열은 세션으로 분할되고 세션 내의 요청은 시간순으로 정렬된다. Q를 UserID u에 대한 특정 세션내의 요청 수로 가정하고, $(u, r_1, t_1, x_1), \cdots (u, r_Q, t_Q, x_Q)$를 정렬로그 L_s내에 나타나는 이 세션의 요청들이라고 가정하자. 각 세션에 대해 다음 절차를 반복한다.

$C[i,j]\leftarrow 0$ for *all* $i,\ j=1.\cdots,n$

$W[i,j]\leftarrow 0$ for *all* $i,\ j=1.\cdots,n$

For $k=2$에서 Q까지

Begin $C[r_{k-1}, r_k]\leftarrow C[r_{k-1},\ r_k]+1;$

$W[r_{k-1}, r_k]\leftarrow W[r_{k-1},\ r_k]+(t_k-t_{k-1}-x_{k-1})$

End; $C[r_Q, n]\leftarrow$; 종료 상태로의 전이

[그림 3. GetSessions 알고리즘]

알고리즘 GetSessions는 위 [그림3]에 제시되어 있다. 이것은 세 개의 주요 단계로 구성되어 있다. 첫 번째 단계에서는, 동일 UserID 의 모든 요청이 RequestTime의 순으로 나열되도록 요청 로그를 정렬한다. 이 결과 로그에는 UserID마다 하나의 부분열이 생성된다. 두 번째 단계는, 세션 임계 시간을 이용하여 부분열로부터 세션을 추출한다. 마지막으로, 최종 단계는 각 세션의 요청들을 스캐닝해서 해당 세션에 대한 C와 W 행렬내에 상태들간의 전이 횟수와 사색시간을 누적한다.

HTTP 로그를 이용할 때는 몇 가지 주의할 사항이 있다. 예를 들어, 프로세서와 네트워크의 속도 향상으로 인해 밀리초의 정확도로 요청시간을 기록하는 것은 적절하지 않은 상황이 될 수 있다. 이러한 이유로 용량 설계 연구에서 Apache의 HTTP로그에 더 높은 정밀도의 타임스탬프가 기록되게 되었다.[5]

4.2 GetCBMGs 알고리즘

세션로그가 생성되면, 이에 대한 클러스터링 분석을 실시하여 비교적 적은수의 CBMG들로 구성된 합성 여부를 생성할 필요가 있다. 클러스터의 중심은 CBMG의 특성을 결정한다. 본 논문에서, GetCBMGs 알고리즘을 설명하기 위해서, 클러스터링 알고리즘으로 k-평균 클러스터링 방법을 제시하였다.[7] k-평균 클러스터링 알고리즘은 n개의 속성들을 가지는 각각의 데이터를 n차원의 공간내에 벡터로 표현하면 유사한 특성을 가지는 데이터들끼리 서로 근접하게 위치하게 된다는 가정을 근거로 한다. 다시 말해, 데이터공간내에 랜덤하게 위치하는 k개 클러스터의 중심에서 시작하여 각각의 데이터가 클러스터의 중심을 중심으로 분포되기까지 클러스터의 중심(평균)은 계속 움직여 나가는 과정을 반복함으로서 모든 데이터를 클러스터에 할당 시킨다.

GetCBMGs 알고리즘은 점과 중심사이의 거리 계산에 사용할 거리척도에 대한 정의를 필요로 한다. 세션 로그가 M개의 점 $X_m=(C_m, W_m), m=1,\cdots,M$으로 구성되어 있다고 가장하자. 여기서 C_m과 W_m은 앞에서 정이한 전이 횟수 행렬과 누적 사색시간 행렬이다. 거리에 대한 정의는 전이 횟수 행렬만을 근거로 하는 데, 그 이유는 이 행렬만이 고객과 전자상거래 사이트의 상호작용을 좀더 명확하게 정의한 요소이기 때문이다. 세션 로그내의 두 점 X_a와 X_b 사이의 거리 dX_a, X_b를 유클리드 거리로 정의하기 한다.

$$d_{xa,\ xb}=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(C_a[i,\ j]-C_b[i,\ j])^2}\text{----------(1)}$$

그렇다면, 새로운 점이 클러스터에 추가될 때 새로운 중심의 좌표, 즉 행렬 C와 W의 새로운 값들을 구하는 방법은 다음과 같이 구할 수 있다.

점 $X_m=(C_m,\ W_m)$이 점 $(C,\ W)$로 표현되는 중심 k에 추가된다고 가정하자. 새

로운 중심은 (C', W')로 표현되며, 여기서 매트릭스 C'와 W'의 원소들은 다음과 같이 계산된다.

$$C'[i,j] = \frac{s(k) \times C[i,j] \times C_m[i,j]}{s(k)+1}$$
$$W'[i,j] = \frac{s(k) \times W[i,j] \times W_m[i,j]}{s(k)+1} \quad \text{----------(2)}$$

모든 클러스터가 구해지면, 각 클러스터와 관련된 CBMG를 특성화하는 행렬 P와 Z를 다음과 같이 유도 할 수 있다.

$$p_{i,j} = C[i,j] / \sum_{k=1}^{n} \rightarrow\rightarrow C[i,k]$$
$$z_{i,j} = W[i,j] / C[i,j] \quad \text{----------(3)}$$

클러스터 k의 CBMG에 의해 표현되는 세션의 도착률 λ_k^s는 $\lambda_k^s = s(k)/T$로 구한다. 여기서 T는 요청 로그 L을 구한 시간 간격이다. 각 클러스터에 대한 행렬 P와 Z를 구하면, 세션 유형별로 연구방향에 부합하는 고객특성을 파악할 수 있을 것이다. 그런데, 유의할 사항은 점 X_a와 X_b 사이의 거리 dX_a, X_b를 유클리드 거리를 계산하기전에 부하를 정확히 나타내는 k개의 클러스터수를 결정하는 것이다. 이러한 문제는 두 가지 척도 즉, 클러스터의 점들과 그 중심간의 평균(클러스터내 거리)과 중심들간의 평균거리(클러스터간 거리)의 변이량을 살펴봄으로써 확인할 수 있다. 이 변이량은 변이량 계수(CV:coefficient of variation), 즉 평균과 표준편차 사이의 비율로 정의할 수 있다. 즉, 각각의 k번째 클러스터에 클러스터내 CV를 최소화 하고, 클러스터간 CV를 최대화하는 점에서의 k를 클러스터 수로 결정한다.

4.3 GetCVMs 알고리즘

CBMG 대신 CVM으로 표현된 세션들은 HTTP 로그로부터 [그림 4]의 GetCVMs 알고리즘을 통해 구할 수 있다. 여기서도 CBMG로 표시된 세션의 경우와 마찬가지로 세션을 좀 더 작은 대표적인 그룹들로 묶을 필요가 있다. 클러스터링 기법은 여기서도 또한 사용할 수 있다. 거리 척도는 방문율 벡터 사이의 거리가 된다. 방문율 벡터 $V_a = (V_2^a, \cdots, V_n^a)$와 $V_b = (V_2^b, \cdots, V_n^b)$로 표시되는 세션 a와 b로 계산할 수 있다. 세션 a와 b사이의 거리는 다음과 같다.

$$d_{Va,Vb} = \sqrt{\sum_{i=2}^{n-1} (V_i^a - V_j^b)^2} \quad \text{----------(4)}$$

1. [그림 3]의 GetCBMGs 알고리즘의 단계 1과 2를 실행한다.

2. 이 경우 부분열들은 세션들로 분할되고, 세션 내부에서의 요청은 시간적 순서대로 배열된다. Q를 특정 세션에서 UserID u의 요청횟수라고 하고, $(u, r_1, t_1, x_1), \cdots (u, r_Q, t_Q, x_Q)$를 정렬된 로그 L_s에 다타나는 이 세션의 요청들이라고 하자. 세션별로 다음 절차를 반복한다.

$V_t \leftarrow 0$ for $all\ i = 2, \cdots n-1$

$V_1,\ V_n \leftarrow 1$;

$For\ k = 1$에서 Q까지

$V_{rk} \leftarrow V_{rk+1}$;

[그림 4. GetCVMs 알고리즘]

6. 결 과

본 논문에서는 e-CRM을 위한 데이터마이닝의 한 기법으로 CBMGs와 CVMs 알고리즘을 통하여 고객특성을 파악하는 방법을 제시하였다. 기업은 쌍방향 커뮤니케이션으로 고객에게 제공되는 서비스의 응답성과 활용 가능성 분석을 CBMGs와 CVMs를 통해서 비즈니스나 마케팅 분석을 할 수 있으며, 고객의 Site 탐색 Pattern과 고객행동 관련 유용한 정보들을 효과적으로 추출하고, 분석할 수 있다는 것을 제시하였다. 뿐만 아니라, 앞에서 제시한 연구방법은 e-CRM의 성공과 실패를 구분할 수 있는 첫 걸음이라는 점을 감안할 때, 앞으로 ID matching 기반의 Web 로그분석기법은 [그림 1]에서 각각의 알고리즘을 이용한 CBMG 상태전이와 CVM 사례를 작성하게 될 것이다. 그리고, 클러스터링 기법을 적용하여 구매자와 비구매자의 행동을 분리해서 분석할 수 있다. 이러한 접근법은 보통 세션의 작은 비율만을 차지하는 구매자들을 특별하게 처리할 수 있다는 장점을 가지고 있다. 본 논문에서는 웹로그 파일중에서 접속로그에 대한 분석과정을 보여주고, 접속로그가 누적되면 사용자의 주된 방문시간, 요일, 계절, 검색경로 등을 상세히 알 수 있으므로 사용자의 구매방식과 웹구매 패턴 등을 분석할 수 있게 한다. 따라서 접속로그는 e-CRM의 가장 중요한 도구로서 역할을 담당한다고 볼 수 있다. 다음은 접속로그의 실제의 예이다. 고객 중심의 정보를 수집하고, DB화하는 데 중요한 수단이 될 것이다.

7. 참 고 문 헌

[1] 다니엘 A, 메나세 · 비르질리우 A.F. 알메이다 공저, *전종훈 · 주우석 · 나연묵 공저. 「e-Biz 웹사이트 설계에서 운용까지」*, 2001. p62-71. p357-370

[2] 오라클 연구회 공저, *「e-비즈니스 시스템」*, 2001. p199-206.

[3] Friedman, J.H.(1997). Data Mining and Statistcs : *What's the Connection?, Proceedings of the International Conference on the Interface : Computing Science and Statistics,* http://www.stat.rice.edu/interface97.html

[4] Hand, D.J. (1997). Intelligent Data Analysis : *Issues and Opportunities, Intelligent Data Analysis*, Vol. 2, No. 2, 1-14.

[5] Menasce', D. A. B. Peraino, N. Dinh, and Q. Dinh, "Planning the Capacity of a Web Server; An Experience Report," Proc. 1999 Comp. Measurement Group(CMG) Conf., Reno, NV, Dec. 5-10, 1999.

[6] B. D. Davidson, "Web Traffic Logs: an Imperfect Resoures for Evaluation," Proc. *INET' 99 Conf.* Internet Society, San Jose, CA, June 1999.

[7] D. Ferrari, G. Serazzi, and A. Zeigner, *Measurement and Tuning of Computer Systems*, Upper Saddle River, Prentice Hall, NJ, 1983

[8] http://crmcolumn.sungilsys.co.kr/column_06.html

[9] 정인근, 김운회 공저"The Principles of Digital Economy and e-Business",2002, p233-239