웹 로그 분석을 통한 무선인터넷 컨텐츠 추출에 관한 연구

*임영문
*김홍기

Abstract

무선인터넷을 이용한 고객관리는 고객에게 더욱더 세분화된 서비스를 제공할 수 있으며, 고급화된 서비스를 제공함으로써 고객의 만족과 구매욕구를 증진시킬 수 있다. 하지만, 개인화된 서비스를 제공하기 위해서는 고객에 대한 패턴 연구 및 세분화 작업이 먼저 이루어져야 한다. 이러한 작업을 위한 다양한 연구중 한 분야가 웹 로그를 이용한 사용자의 패턴분석일 것이다.
본 연구에서는 웹 로그 분석을 통한 주요 컨텐츠를 추출하는 과정 및 예제시스템의 구현 방향에 대해서 알아보고자 한다.

1. 서론

초기 인터넷 비즈니스 시장은 인프라, 시장규모, 마케팅 테크닉, 비즈니스 인력 등이 거의 부재한 상황에서 진행되었기 때문에 시장진입을 한다는 것 자체만으로 골드러쉬(Gold Rush)에 동참하여 무한히 펼쳐진 미래의 시장기회를 얻을 수 있다고 생각되었다.
하지만, 이러한 단꿈을 꾸는 사이에 비즈니스의 시장에 점차적으로 경쟁자가 늘어나고 우수한 자원과 막대한 자금력으로 무장한 대기업들이 시장에 진입하면서 시장 경쟁은 뛰어난 아이디어 뿐 아니라 경쟁자와 경쟁할 수 있는 다양한 마케팅 테크닉과 우수한 인력들이 필요하게 되었다. 이러한 시장분위기에서 닷컴 기업들이 살아 남기 위하여서는 수익성을 고려하여 고객중심형의 마케팅 전략(Customer Centric Marketing) 전환을 하여야 했다. 고객 중심형의 마케팅 전략 중에서 가장 중요한 것은 현재의 고객과 미래의 잠재고객을 파악하여 고객과 끊임없는 커뮤니케이션(Communication)을 통한 충성고객을 확보하는 것이다. 로그분석을 통한 인터넷 비즈니스 전략의 접근방식은 이러한 고객 중심형의 마케팅 전략을 전개할 수 있는 고객 커뮤니케이션의 방법을 고려하여 단순하게 고객을 확보하는 차원을 넘어서 충성고객데이터를 기반으로 고객학습을 통한 고객 가치를 증대시킬 수 있도록 전개해야 한다[1].

*강릉대학교 산업시스템공학과

또한, 무선 인터넷의 발전과 더불어 고객과 기업이 만나는 채널이 더욱더 다양해지고 있고, 일부 닷컴 기업들은 무선 인터넷을 활용한 마케팅 전략을 구사하고 있으며, 다른 기업들도 무선 인터넷을 활용하려고 하고 있다.
본 연구에서는 기존의 온라인 웹페이지를 가지고 있는 기업들이 무선 인터넷 컨텐츠를 추출하고자 할 때 순회패턴을 이용하여 웹로그 분석 방법을 활용하는 방안을 제시하고자 한다.

2.웹 로그 화일에 대한 패턴 탐사 및 분석

2.1 웹 로그 화일의 특징

웹 액세스 로그 화일 (Web access log file)은 인터넷을 사용하는 사용자가 개설된 웹의 홈페이지를 액세스했을 때부터 찾아보는 모든 페이지 이름과 시간을 기록하는 파일이다. 웹 서버는 모든 개별적인 액세스에 관한 URL과 시간 정보를 기록으로 남긴다[4].
본 논문에서 사용된 데이터는 강릉대학교 치과병원의 웹 액세스 로그 데이터이고, 서버의 OS(Operating System)는 Linux이며, 2001년 8월 12일부터 2001년 9월 13일까지 축적된 것이다. 원형의 데이터 형식은 [그림 1] 과 같다. 전체 트랜잭션의 수는 365,933개 였다. 사용자의 도메인 이름 (또는 IP 주소), 사용한 시간 (일 /월 /년 /시 :분 :초), 사용한 메소드(GET 또는 POST), 그리고 요청한 화일의 이름(URL 주소), HTTP(Hyper Text Transfer Protocol)버전, 사용 상태 코드(성공 또는 에러 코드), 전송된 바이트 수 등의 순서로 기록된다. 즉 [그림 1]의 첫 번째 레코드는 210.95.147.93 이라는 IP 주소를 가진 도메인에서 2001년 9월 9일 07시 16분 00초에 dental.kangnun.ac.kr/title.html 이라는 문서를 요청했으며 그 때의 HTTP버전은 1.1이고 성공적으로 수행되어 8920 bytes 가 전송된 것을 의미한다.

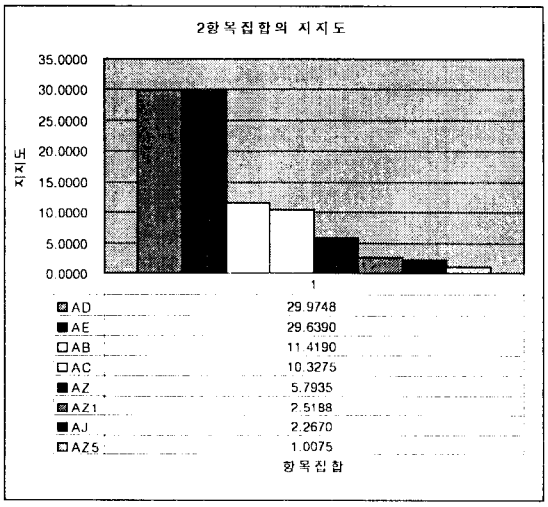
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/title.html	HTTP/1.1"	200	8920
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/main.php3	HTTP/1.1"	200	15525
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/images/intro/top1.jpg	HTTP/1.1"	200	7785
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/images/intro/top2.jpg	HTTP/1.1"	200	16889
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/images/intro/top3.jpg	HTTP/1.1"	200	9869
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/images/bchiju.gif	HTTP/1.1"	200	33994
210.95.147.93	[09/Sep/2001:07:16:00	+0900]	"GET	/images/bbojon.gif	HTTP/1.1"	200	17241

[그림 1] 전처리 전의 원시 웹 엑세스 로그데이터

웹 서버에서 웹 액세스 로그 화일을 저장 관리하기 위해서는 상당히 큰 어려움이 존재한다. 그 이유는 홈페이지를 통해서 액세스하는 모든 IP 주소, 시간, 탐색하는 페이지 등의 기록을 유지하기 위해 필요로 되는 메모리의 대량 사용 때문이다. 그러므로, 대부분의 웹 서버에서 로그 파일의 관리를 위해 일정한 시점이 지나거나 일정한 크기 이상이 되면 그 화일을 삭제하거나, 아주 기초적인 통계 처리를 통하여 기본적인 정보만 추출하고 삭제하는 경우가 보통이다.

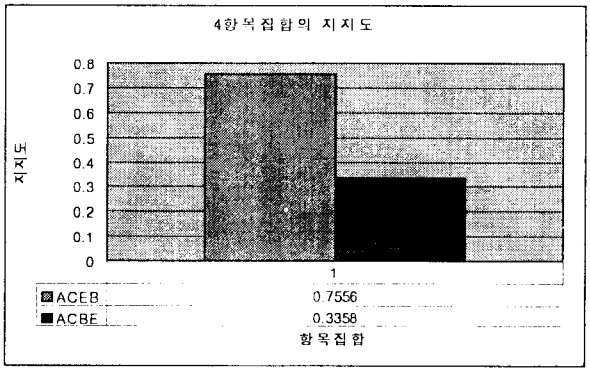
2.2 실험결과

순회패턴 탐사를 실시한 결과(지지도 1%이상) [그림 2],[그림 3]와 같은 결과를 얻었다. [그림 2]에서 보면 AD, AE, AB 순으로 지지도가 높았다. 즉, 메인페이지에서 치학정보로 가는 지지도가 29.97%, 메인페이지에서 자료실로 가는 지지도가 29.63%, 메인페이지에서 사이버진료실로 가는 지지도가 11.41%로 나타났다.



[그림 2] 2항목 집합의 지지도

[그림 3]에서 볼 수 있듯이 최소지지도 1%를 만족하는 4항목 집합은 존재하지 않았지만, 최상위를 나타내는 항목집합은 ACED, ACBE로 나타났다.



[그림 3] 4항목 집합의 지지도

3. 웹로그 파일의 분석을 통한 CRM적 접근

3.1 무선 컨텐츠 추출

2장의 분석결과를 보면 대부분의 사용자들은 2개에서 3개의 페이지를 요청했으며, 주로 치학정보, 사이버진료실, 자료실을 클릭 했다는 것을 알 수 있다. 실제 홈페이지를 살펴보면 사이버진료실에 개시된 글의 건수가 가장 많은 것으로 보이는데, 이는 실제로 병원홈페이지에 접속해서 정보를 얻어 가는 사람보다 아직은 한번쯤 둘러보는 사용자들이 많기 때문에 이러한 결과가 나타나는 것으로 보인다. 특이한 점은 치학정보 사이트를 요구한 사람이 많다는 점이다. 이는 현재 심하게 아프지 않거나 전혀 아프지 않는 사람들이 다양한 정보를 얻고자 하는 것으로 보인다. 따라서, 치과병원에서의 무선 컨텐츠는 치학정보, 사이버진료실, 자료실, 진료안내가 적당할 것으로 보이며, 대부분이 게시판 형식이므로 구현 과정에서 현재 쓰이고 있는 무선 인터넷의 환경을 고려하여야 할 것이다.

3.1.1 회원제 실시 방안

현재 치과대학병원은 회원제를 실시하지 않고 있다. 하지만, 점진적으로 회원제를 추진해야 될 것이다. 회원제를 실시해야만 홈페이지를 접속한 사용자들에 대한 효과적인 관리가 이루어질 수 있기 때문이다. 하지만, 무작정 게시판에 글을 쓰는 사람들에게 사용자 등록을 요구한다면 사용자들은 다른 사이트로 이동을 하게 될 것이다. 초창기에는 회원제를 실시하되 필요로 하는 사람만 실시하면 된다. 즉, 회원에 가입하게 될 경우 지속적으로 이메일이나 핸드폰 문자 메시지로 치학정보를 보내주면 등록된 회원을 유지시킬 수 있을 것이다.

3.1.2 가족관계시스템의 적용

기존의 CRM기법들은 서구적인 기법들로써, 고객을 단순히 마케팅의 대상으로만 여겨왔다. 하지만, 이러한 방법은 다소 유교적이고 친족관계에 집착하는 사람들에게는 마케팅활동 자체를 단순히 물건을 팔기 위한 하나의 수단으로 생각하게 할 것이다. 즉, 고객과의 유대관계와 신뢰감, 공감대 형성이 이루어지지 않고 있는 것이다. 기업이 이러한 고객과의 유대관계와 신뢰감, 공감대를 형성하고 활용 할 수 있다면 보다 효율적인 마케팅을 할 수 있을 것이다.

즉, 현재 치과병원을 내원하는 환자들을 자연스럽게 온라인 사용자들로 유치하면 된다. 실제 진료를 받는 사람이라면, 회원으로 가입하게 될 것이며, 자신이 병원에 가야하는 날짜나 오늘 진료 받은 것에 대한 정보를 이메일이나 핸드폰으로 연락을 받는다면 고객은 신뢰감을 느낄 수 있을 것이다. 이러한 전략을 꾸준히 사용하면서 고객들에게 가족관계를 유도하여 한 집안 전체를 치과병원에서 관리해준다는 느낌을 받았을 때 한번 지나쳐 가는 고객이 아닌 가족전체가 이용하는 병원이 될 것이다.

4. 결론

분석한 로그는 병원이라는 특이성을 가지고 있기 때문에, 병원에 진료차 방문하는 고객들로 하여금 고객의 전화번호나 가족관계를 유도하여 문자 메시지 나 E-Mail을 통하여 지속적으로 구강정보나 진료예약 날짜 등을 제공해 준다면, 병원을 찾는 고객들은 병원과의 유대감을 가지게 될 것이고, 다음에 내원하는 횟수가 증가될 것으로 예상된다. 또한, mCRM(mobile CRM) 적용 시 마케팅에 중점을 두는 것보다는 정보전달에 중점을 둔 mCRM 구현이 병원의 장기적인 수익구조에 도움을 줄 것으로 보인다.

Acknowledgement
본 연구는 2001년도 두뇌한국21 지원사업에 의해 지원되었음.

참 고 문 헌

[1] 김형택, 민옥길, "효과적인 인터넷 마케팅을 위한 웹 로그 분석", 비비컴, 2001.
[2] 류승범, "국내 실정에 맞는 올바른 CRM 접근 방법론", 정보과학회지, vol. 18, no. 11, pp. 46-54, 2000.
[3] 무선인터넷백서편찬위원회, "무선인터넷 백서 2001", (주)소프트뱅크미디어, 2001.
[4] 박종수, "웹 로그 파일에서 빈발 항목집합 탐사", 성신여자대학교 기초과학연구지, pp. 446-449, 1999.
[5] 이경우, 최덕원, "전자상거래에서 상품 추천을 위한 웹 개인화 방안에 관한 연구", 한국경영과학회/대한산업공학회 춘계공동학술대회, 2001.