효율적인 웹 로그 분석을 위한 웹 정보구조 설계에 관한 연구

박홍규*, 강환창*, 최진영*
*고려대학교 컴퓨터공학과
e-mail:hkpark@korea.ac.kr

Information Architecture Design of Web Site for Effective Web Log Data Analysis

Hong-gyu Park*, Hwan-chang Kang*, Jin-young Choi*,
*Dept. of Computer Science and Engineering, Korea University

요 약

인터넷 이용자의 급증에 따라 보다 효과적인 웹 사이트의 구현을 위한 과학적이고 체계적인 분석의 필요성은 더욱 절실해지고 있다. 웹 로그 파일은 사용자가 웹 사이트를 이용하면 서버에 남는 기록으로 이 데이터를 기반으로 다양한 정보를 추출해낼 수 있다. 본 논문은 실제 인터넷 사이트를 운영하는 기관의 로그를 상용화된 로그분석 툴을 이용하여 사이트 정보구조의 개편 전과 후를 비교 분석하고, 로그분석의 정확성을 높이기 위한 웹 사이트내 정보구조설계 방안을 제시 한다.

1.서론

디지털 정보기술 및 네트워크 기술의 발전이라는 최근의 세계적인 환경변화에 따라 삶의 방식이나 사람들의 교류형태가 변화하고 있다. 특히 인터넷 관련 정보기술의 급속한 발전으로 많은 기업들이 웹 사이트를 단순한 정보검색이 아니라 기업을 운영하는데 있어서 주요자원의 하나로서 인식하고 있다[1].

이로 인해 웹 사이트를 방문하는 이용자들의 접속패턴을 파악하고, 이를 기업의 비즈니스 전략에 활용함에 따라 사이트를 운영하는 기업들은 사용자가 서버에 남긴 로그를 분석하는데 관심을 기울이게 되었다. 로그분석이란 사이트 이용자가 웹 사이트를 이용하면 서버에 남는 기록으로 이 데이터를 기반으로 다양한 정보를 추출해내는 것을 말하며[2], 효과적인 웹 사이트 구축을 위한 자료로서도 그 가치가 있다고 할 수 있다.

최근 성능이 향상된 많은 상용화된 로그 분석툴들이 출시됨에 따라 방대한 량의 로그를 보다 손쉽게 분석하고, 이를 사이트 운영을 위한 소중한 자료로 이용하고 있지만, 웹 사이트를 구성하고 있는 디렉토리와 파일이름을 기반으로 처리하는 분석 툴의 한계[3][4]로 인하여 웹 사이트의 구조에 따라서 분석데이터의 값이 상당히 유동적으로 나타나게 된다.

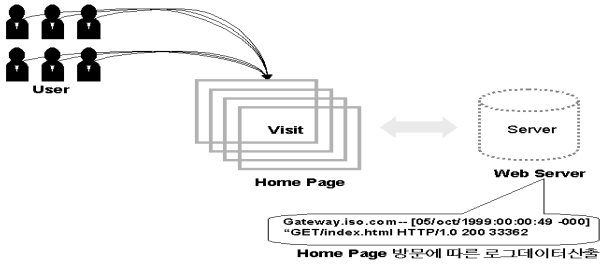
본 논문에서는 웹 사이트의 정보구조를 개편한 a기관의 사례를 통해 효율적 정보구조설계가 보다 명확한 로그분석 데이터 값을 추출할 수 있음을 실험을 통해 보여준다. 본 논문의 구성은 2장에서 웹 로그의 이해와 관련연구에 대해 살펴보고, 3, 4장에서는 로그분석과 정보구조와의 관계 및 사이트 개편전후 로그분석을 비교하였으며, 5장에서는 실험을 토대로 효율적 로그분석을 위한 웹 사이트 정보구조의 설계방안과 향후 연구과제에 대해서 기술한다.

2. 관련연구

2.1 로그파일의 정의

일반적으로 사용자가 웹 사이트에 접속하기 위해 웹 브라우저에서 웹 사이트의 주소(URL)을 입력하는데, 이를 웹 사이트에 보내기 위한 요청(Request)라 하고, 이러한 요청이 인터넷을 통해 해당 사이트를 호스팅하는 웹 서버로 보내진다. 특정사이트에 대한 모든 요청들은 해당 웹 서버에 '로그파일'이라고 불

리는 파일로 저장되어진다[1].



[그림 1] 로그파일 생성

2.2 로그파일의 종류

웹 서버에 따라 한 개가 아닌 여러개의 로그파일을 만들 수도 있는데, 엑세스(access), 로그, 에러(error)로그, 리퍼럴(referral)로그 및 에이전트 로그 등 크게 4가지로 로그파일로 분류된다[5][6].

2.3 로그파일의 분석방법

로그파일의 분석방법에는 Tag, TCP/IP Packet Sniffing, Server-Add In, 웹서버 로그분석 방식 등으로 분류할 수 있다[5][6].

2.4 로그분석의 한계

일반적으로 로그분석 툴은 페이지뷰나 방문자 수의 카운트 통계정보를 기반으로 하기 때문에 사이트현황 정보를 파악해주는 기능들로 구성되어있다. 로그분석을 하는 담당자의 능력이나 로그분석 툴의 기능, 로그데이터의 조건지정 등에 따라 결과가 다르게 나타나는 한계가 있다.[3][6][8]

첫째, 사용자에 대한 구분이 접속 IP주소 단위로 이루어져 부정확하다. 이는 사용자의 인구통계 특성을 감안한 트래픽 측정이 중요한 경우에는 치명적인 결함이 될 수 있다.

둘째, 웹 서버에 요청된 트래픽만 측정할 뿐 중간의 프록시(proxy)서버나 웹 브라우저의 캐쉬에 의해 보여지는 페이지에 대해서는 측정되지 않는다. 이로인한 트래픽 측정 손실은 사이트마다 다르므로 일률적으로 말하기는 어려우나 일반적으로 페이지 뷰 기준으로 20~30% 정도로 추정하고 있다.

셋째, 로그파일에 의한 분석은 로그파일을 소유한 주체가 측정의 대상이 되는 사이트 자신이므로 측정의 객관성에 문제가 있을 수 있다.

로그분석을 활용한 웹 사이트의 구조 설계에 관련된 연구들로 웹 사이트를 분석하고 사이트 구조와 링크 관계를 통해 사이트 이용자들이 보다 편리하게 정보를 탐색할 수 있는 웹 사이트 디자인 방안[9], 웹 서버 관리 및 사이트 디자인 등의 전략을 수립하는 방안[10], 웹 사이트의 재구성을 위한 방향 제시[11] 등의 있다.

그러나 기존 연구들이 로그분석 데이터에 대한 정확도를 높이기 보다는 분석 데이터를 활용한 웹 사이트 설계 방안을 제시하였다면, 본 논문에서는 분석 데이터의 정확도를 높임으로써 웹 사이트 구조 개선시 보다 정확한 이용자의 의견을 반영했다는데 의미가 있을 것 이다.

3. 로그분석과 웹 정보구조와의 관계

웹 사이트의 규모가 점차 커짐에 따라 사이트 관리자 입장에서는 증가하는 컨텐츠를 관리하는 것이 커다란 부담으로 대두되고 있으며, 이용자 측면에서도 복잡한 사이트를 방문할 때마다 그 사이트만의 메뉴체계를 이해하고, 여러번의 오류와 학습과정을 거치며 찾고자 하는 정보를 위해 많은 시간을 할애해야 하는 불편을 않게 된다[8].

이는 이용자의 정보접근 행태 파악을 위한 로그분석에서도 적용될 수 있다. 명확하지 않은 레이블링, 메뉴구조, 산만한 컨텐츠 배치는 디렉토리와 파일이름을 기반으로 처리하는 분석 툴의 한계로 데이터로서의 효용가치가 떨어질 수 있다[5].

실제로 a기관은 개편 전 동일한 로그를 몇 개의 상용화된 분석 툴을 이용하여 테스트 한 결과 종합분석에서의 데이터와 물리적 개념의 디렉토리별 분석에서의 데이터 값이 동일하다는 데 착안, 다음과 같은 방안을 적용하여 사이트의 정보구조를 개선하게 되었다.

첫째, 물리적 개념의 디렉토리와 논리적 개념의 메뉴이름을 동일하게 처리하였다.

둘째, 모든 페이지 단위 파일들은 어느 한 디렉토리/메뉴에 위치토록하고, DB호출을 통해 생성되는 페이지 파일은 코드값을 통해 메뉴분석이 용이토록 구조화 했다.

셋째, 컨텐츠가 방대한 메뉴의 경우는 웹서버의 가상디렉토리 설정을 통해 별도의 공간에 로그를 저장토록 하였다.

a기관에서는 실제 상기와 같은 방법이 컨텐츠에 대한 관리를 용이하게 할 수 있고, 사이트 이용자의 메뉴별 분석에 있어 보다 효과적일 것이라는 가정을 바탕으로 웹 사이트 구조를 개선하였다.

4. 실험 및 분석

4.1 종합분석

[표1][표2]는 사이트 개편전후의 종합분석현황으로

웹 사이트의 전반적인 상황해석에 필요한 Hits/Page View/ Visitors Session/Visitor를 보여준다.

[표1] 개편전

	전체	평균(일)	평균(시간)
접속횟수	17,275,403	575,847	23993.62
페이지뷰횟수	1,778,376	59,279	2469.97
방문횟수	313,802	10,460	435.84
방문객수	158,118	5,271	219.61
방문소요시간	28382시간 48분 22초	946시간 5분 37초	39시간 25분 14초

[표2] 개편후

	전체	평균(일)	평균(시간)
접속횟수	17,808,287	574,461	23935.87
페이지뷰횟수	1,306,656	42,150	1756.26
방문횟수	256,200	8,265	344.35
방문객수	124,769	4,025	167.7
방문소요시간	30309시간 28분 59초	977시간 43분 31초	40시간 44분 19초

4.2 페이지분석

페이지 분석은 사용자가 가장 많이 찾는 페이지, 가장 오래 읽은 페이지, 가장 최근에 방문한 페이지, 가장 먼저 또는 끝으로 읽은 페이지 등을 디렉토리별 메뉴별로 보여준다.

4.2.1 디렉토리 분석

[표3][표4]은 사이트개편전후의 디렉토리 분석을 보여주고 있다. 웹서버의 모든 페이지 파일은 사이트 내 물리적 개념의 디렉토리에 저장되며 페이지가 저장된 디렉토리로 구분하여 분석하는 방법이 디렉토리 분석이다.

[표3]개편전

디렉토리	접속횟수	방문소요시간	최근접속일자
a	1		2004-08-31 20:01
b	1,709	1시간 1분 30초	2004-08-31 23:21
c	154	1시간 20분 18초	2004-08-31 22:02
d	2	2초	2004-08-18 16:15
e	2	2초	2004-08-18 16:15
f	6,829	9시간 21분 46초	2004-08-31 18:41
g	63	11분 55초	2004-08-31 12:23
h	1,279	17시간 44분 22초	2004-08-31 23:01
i	2	4초	2004-08-21 8:11
합계	17,808,287	30309시간 25분 15초	0

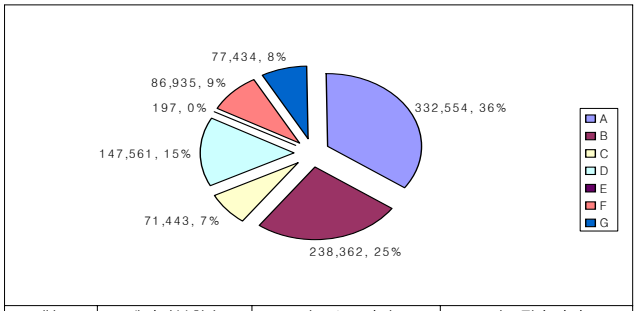
[표4] 개편후

디렉토리	접속횟수	방문소요시간	최근접속일자
a‘	1		2004-06-25 14:58
b‘	4,015	2시간 26분 53초	2004-06-30 21:30
c‘	484,985	3012시간 43분 8초	2004-06-30 23:59
d‘	52	1시간 54분 45초	2004-06-30 15:06
e‘	3	2초	2004-06-03 11:25
f‘	306	17분 9초	2004-06-29 4:46
g‘	185	10분 12초	2004-06-29 4:46
h‘	35,685	93시간 32분 25초	2004-06-30 23:47
i‘	7,900	22시간 42분 49초	2004-06-30 20:57
j‘	22	4초	2004-06-23 19:23
합계	17,275,403	28382시간 47분 28초	0

디렉토리 분석을 통해 웹 관리자는 보다 효율적이고 합리적으로 웹서버의 컨텐츠를 관리할 수있다. 디렉토리 분석을 보다 유용하게 활용하기 위해서는 컨텐츠를 일관성 없이 관리하는 것이 아니라 관심있는 카테로리별로 나누어 관리하는 것이 중요하다.

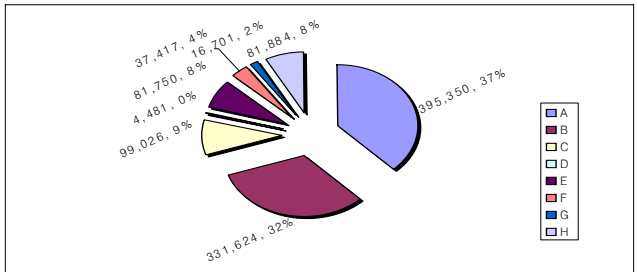
4.2.2 메뉴분석

[그림2][그림3]는 사이트 개편전후의 메뉴분석을 보여준다. 모든 웹 서버는 메뉴를 사용하며, 주 메뉴는 하위 메뉴를 가지고 있기도 하여 방문객이 웹 서버가 만들어 놓은 메뉴를 따라 웹 서버를 검색하게 된다.



매뉴	페이지뷰횟수	방문소요시간	최근접속일자
A	332,554	2404시간 20분 45초	2004-06-30 23:59
B	238,362	2787시간 35분 11초	2004-06-30 23:59
C	71,443	555시간 13분 19초	2004-06-30 23:59
D	147,561	1485시간 24분 48초	2004-06-30 23:59
E	197	6시간 8분 22초	2004-06-30 17:23
F	86,935	1237시간 20분 19초	2004-06-30 23:57
G	77,434	502시간 40분 25초	2004-06-30 23:59
합계	954,486	8978시간 43분 9초	0

[그림 2] 개편전 메뉴분석



매뉴	페이지뷰횟수	방문소요시간	최근접속일자
A‘	395,350	3913시간 41분 24초	2004-08-31 23:59
B‘	331,624	3461시간 26분 57초	2004-08-31 23:53
C‘	99,026	622시간 54분 52초	2004-08-31 23:59
D‘	4,481	47시간 19분 22초	2004-08-31 23:40
E‘	81,750	725시간 57분 27초	2004-08-31 23:59
F‘	37,417	579시간 30분 17초	2004-08-31 23:51
G‘	16,701	80시간 34분 40초	2004-08-31 23:51
H‘	81,884	1422시간 10분 9초	2004-08-31 23:58
합계	1,048,233	10853시간 35분 8초	0

[그림 3] 개편후 메뉴분석

이러한 메뉴는 디렉토리 또는 페이지와는 별도로 관리된다. 디렉토리 또는 페이지가 물리적 분류라면, 메뉴는 그러한 물리적 분류위에 형성된 논리적 분류인 것이다. 그러나 실제 대형포탈사이트를 제외한 대부분의 중소형 사이트의 경우 메뉴별 분석 있어 디렉토리와 혼동하는 경우가 종종 있으며, 이로 인해 메뉴별 분석의 한계에 봉착하는 경우가 있다.

본 실험에서는 웹 사이트내 정보구조의 개편 전과 후의 메뉴별 비교분석을 통해 도출된 노출횟수를 측정하였다.

먼저 개편전의 종합분석[표1]에서 전체 페이지를 노출한 페이지뷰가 1,778,376이고, 메뉴분석[그림2]에

서의 페이지뷰는 954,486회로 종합분석의 전체 페이지뷰의 54%밖에 추출하지 못했으며, 이로인해 실제로 46%정도의 분석 손실이 발생하였다.

개편 후의 종합분석[표2]에서는 전체 페이지뷰가 1,306,656회 이고, 메뉴분석[그림3]에서의 페이지뷰는 1,048,233회로 종합분석의 전체 페이지뷰의 80%로 20%정도의 손실이 발생하였다. 개편전 분석데이터와 비교하여 26%의 손실이 줄어든 것으로 나타났다. 물론 손실률 감소를 위해 모든 페이지를 상위의 개념인 메뉴별로 정리한다면 더욱 손실률을 감소시킬수 있지만 최소 수백, 수만개의 페이지를 매번 업데이트 할 때마다 메뉴별로 정리해야한다는 것은 상당히 비효율적인 일 일 것이다.

로그분석에 있어 페이지 또는 컨텐츠 분석은 사이트를 방문하는 방문자의 페이지 검색 성향을 파악할 수 있는 중요한 단서로 분석의 정확을 위해서는 정보구조 개선이 선행되어야 함을 실험을 통해 보여주고 있다.

5. 결론 및 추후연구

본 논문의 실험을 통해 제안된 웹사이트 정보구조 설계방안이 로그분석의 정확성을 높일 수 있다는 것을 확인하였으며, 이를 토대로 다음과 같은 방안들이 웹 사이트구조 설계시 고려되어야 함을 알 수 있다.

첫째, 모든 페이지파일은 어느 한 메뉴에 소속하게 한다. 이럴 경우 메뉴분석에 있어 메뉴에 소속되지 않는 페이지 파일로 인한 분석 데이터의 손실을 줄일 수 있다.

둘째, 물리적 개념의 디렉토리와 논리적 개념의 메뉴를 일치하는 방법이다. 물론 메뉴가 많은 사이트의 경우에는 비효율적일 수 있으나, 중소형 사이트의 경우 분석손실을 줄 일수 있는 방법이 될 수 있다.

셋째, 대형 포탈사이트의 경우는 메뉴 또는 디렉토리별로 가상 디렉토리를 잡아서 별도로 로그를 기록케 하는 방법으로 메뉴별 로그가 별도로 저장되기 때문에 이 또한 분석손실을 줄일 수 있을 것이다.

넷째, 로그분석 툴들은 대부분 웹 사이트를 구성하는 디렉토리와 파일 이름을 기반으로 분석을 수행하기 때문에 웹 사이트 디렉토리나 파일이름들을 구조화함으로써 분석 툴의 활용성을 높일 수 있다[4].

웹사이트 구조란 웹사이트를 구성하는 각종 컨텐츠가 정리되어있는 체계라 할 수 있다. 이러한 체계는 처음 사용자들이 웹 사이트를 방문했을 때 웹 사이트에 대해 이해하는 전반적인 틀로 웹 사이트의 활용을 위한 출발점이 되며, 이러한 의미에서 잘 정립된 체계를 지닌 웹 사이트는 사용자로 하여금 자신이 원하는 정보에 쉽게 접근할 수 있도록 할 수 있도록 하고, 재 방문을 유도함으로써 사이트 활성화에 도움을 줄 수 있다.

본 논문에서는 효율적인 웹사이트 정보구조 설계를 통해 로그분석의 손실률을 감소시켜 분석의 정확성을 높이는 실험을 하였다. 그러나 웹 사이트의 정보구조 설계에 있어 로그파일 분석이 절대적인 것은 아니며[12], 로그분석의 한계로 인한 오차범위는 항상 존재하는 등의 문제점을 갖고 있다.

향후에는 파악되지 않은 분석상의 손실을 줄이기 위한 추가 연구와 더불어 웹 사이트 성격과 특성, 또는 분석 목적에 맞는 특화된 의미의 로그분석에 대한 연구가 필요할 것이다.

참고문헌

[1] 김진강, 여행사 인터넷 마켓팅을 위한 웹 로그 파일 분석에 관한 연구, 관광레저학회 Vol.13, No.2 2001.

[2] 정선경, 웹 로그 분석을 적용한 웹 사이트내의 웹 컨텐츠 분석 연구 결과, 한국정보과학회 2003년 춘계학술대회 Vol. 30 , No.1. 2003.

[3] Kallepalli, C. and Tian, J, Measuring and Modeling Usage and Reliability for Statistical Web Testing, IEEE Transaction on software engineering, Vol. 27, No.11. 2001.

[4] 아이비즈넷, 인터넷비즈니스, 21세기북스, 2001.

[5] 김형택, 민옥길 공저, 효과적인 인터넷 마켓팅을 위한 웹로그 분석, 비비컴 2001.

[6] 안옥부, 불완전한 웹 로그 자료에 관한 웹 접속 신뢰성 추정, 대구카톨릭대 석사학위논문, 2003.

[7] 문경선, 효과적인 사이트구현을 위한 로그분석에 대한 연구, 세종대학교 석사학위논문, 2002.

[8] 전은용, 웹 사이트를 위한 인포메이션 아키텍쳐, HCI학회 칼럼, 1999.

[9] Borges, J. and Levene(1999), M., "Data mining of user navigation patterns", WebKDD'99, 1999.

[10] 이화영, "표준 로그파일을 이용한 웹마이닝에 관한 연구", 한국과학기술원 석사학위논문, 1999.

[11] 정강용・박나연, "웹서버의 로그파일 분석에 의한 웹 서비스 활용에 관한 연구", 『한국OA학회』, 논문집 13, 2000.

[12] 서진완, 로그화일을 이용한 공공기관의 홈페이지 분석과 정책적 함의, 2001.