4

웹 검색 분야에서의 로그 분석 방법론의 활용도

박 소 연
덕성여자대학교 문헌정보학과 조교수
sypark@duksung.ac.kr
이 준 호
숭실대학교 정보과학대학 컴퓨터학부 부교수
joonho@naver.com

목차

1. 연구 목적
2. 선행 연구
3. 로그 분석 방법론의 활용도
   3.1 이용자들의 전반적인 검색 행태 분석
   3.2 검색 행태 추이 분석
   3.3 키워드 마케팅 전략 구축
   3.4 서비스 활용도 평가 지표
   3.5 개별 서비스 평가
   3.6 이용자의 항해 경로 조사
   3.7 클릭 어뷰즈(click abuse) 파악
   3.8 클릭 로그 품질 평가
   3.9 지역 관심사 파악, 개인 이용자의 관심사 파악
   3.10 오타 분석
4. 결 론

1. 연구 목적

웹 검색에 관한 연구는 다양한 분야에서 다양한 연구 방법을 적용하여 수행되고 있다. 이 중 이용자와 검색 시스템 사이의 모든 상호 작용을 기록한 검색 트랜잭션 로그는 이용자의 실제 검색 행위를 사실적으로 반영한다. 로그 분석 방법은 트랜잭션 로그에 기록된 시스템의 성능 분석과 이용자의 이용 행태 분석을 통하여 검색 시스템을 개선하는데 그 목적이 있다고 할 수 있다. 로그 분석 방법이 최초로 등장한 시점은 60년대 후반으로 추정되며(Peters, 1993), 웹이 등장하기 이전부터 OPAC 시스템, 온라인 데이터베이스 시스템, 서지 통정 시스템, 실험적 정보 검색 시스템 등 다양한 환경에서 활용되어 왔다.

로그 분석 방법과 비교하여 이용자 연구 분야에서 기존에 많이 사용되었던 설문 조사 또는 인터뷰 자료의 경우 실제 검색 행위와 설문 조사 또는 인터뷰 자료 간에 차이점이 발생할 수 있다. 따라서 로그 분석은 웹 검색 이용자들의 검색 행태 연구를 위한 합리적이고 객관적인 방법으로 인정받고 있다(Jansen and Pooch, 2000). 또한 로그 분석을 통하여 대다수 이용자의 전반적인 이용 행태를 분석할 수 있으므로, 로그 분석 방법은 이용자 계층이 다양하고 이질적인 웹 검색 분야의 연구에 적합한 방법이라고 할 수 있다. 일반적으로 웹 검색 서비스의 로그는 이용자들이 입력한 질의를 기록한 질의 로그와 질의에 대한 검색 결과에서 이용자가 조회한 문서를 기록한 클릭 로그로 구성된다.

한편 본 연구자들은 2002년부터 약 5년간 네이버에서 생성된 대용량 트랜잭션 로그에 근거하여 국내 웹 이용자의 검색 행태를 분석하는 일련의 연구를 수행하여 왔다. 본 연구에서는 연구자들이 수행한 연구들을 중심으로 로그 분석 방법론이 웹 검색 분야에 어떻게 활용되고 기여할 수 있는지를 파악하고 향후 활용 분야를 제시하고자 한다. 국내에서 로그 분석 방법을 활용한 최근 연구들로는 특허 상호 검색 시스템 연구(Lee & Paik, 2006), 청소년 과학 분야 사이트 연구(곽승진, 2003), 국가과학기술전자도서관 시스템 연구(유사라, 2002) 등을 들 수 있다. 본 연구에서는 로그 분석 방법론의 활용도를 이용자 계층이 가장 광범위한 웹 검색 분야에 한정하여 논하고자 한다. 본 연구의 결과는 웹 검색 분야의 학문적 발전과 보다 효율적인 웹 검색 시스템 개발과 서비스 구축에 기여할 수 있을 것으로 기대된다.

2. 선행 연구

국외의 웹 검색에 관한 연구의 경우, 90년대 후반부터 본격화되기 시작하였으며, 국내 연구보다 자료의 규모가 방대하고, 연구 주제나 연구 방법도 더 다양한 실정이다. 국외 연구 중 트랜잭

션 로그 분석을 통하여 웹 검색 이용자들의 검색 행태를 조사한 연구들로는 Jansen, Spink, Saracevic 등의 익사이트 연구(2000, 2001, 2002), 알타비스타 연구(2005), 유러피안 올더웹 연구(2005), Hoelscher의 파이어볼 연구(1998), Silverstein 등의 알타비스타 연구(1999), Ross와 Wolfram의 익사이트 연구(2001) 등을 들 수 있다.

Silverstein 등(1999)은 1998년 8월 2일부터 9월 13일까지 6주간 알타비스타 이용자들이 남긴 2억 8천 5백만개 이상의 이용자 세션, 9억 9천만개 이상의 질의를 분석하였다. 이 연구는 지금까지 트랜잭션 로그 관련 연구 중 가장 장기간에 걸쳐 가장 방대한 자료를 연구 대상으로 하였고, 세션 정의 방법 등과 같은 로그 분석 방법을 제시하였다는데 의미가 있다.

Jansen, Spink, Saracevic 등은 익사이트 엔진을 대상으로 일련의 연구를 수행하였는데, 2000년 연구에서는 1997년 3월 9일 익사이트에서 생성된 검색 트랜잭션 로그 중 일부에 해당하는 51,473개의 질의를 분석하였다. 또한 2001년 연구에서는 1997년 9원 16일 익사이트 엔진의 이용자들이 남긴 100만개 이상의 질의를 분석하였다. 또한, Spink et al.(2002)은 1997년, 1999년, 2001년에 수집된 자료들을 비교, 분석하였는데, 그 결과 이용자들이 주로 검색하는 주제는 연예와 성 관련으로부터 전자 상거래에 관련된 주제로 바뀌었으나, 이용자들의 전반적인 검색 행태는 변하지 않았음을 발견하였다.

Jansen, Spink, Pedersen(2005)은 2002년 9월 8일 알타비스타에서 생성된 약 100만개의 질의들로부터 2,603개를 무작위로 추출한 후 이들의 주제를 분류하고 이를 Silverstein 등(1999)의 연구 결과와 비교하였다. 이들은 2002년에 알타비스타 이용자들이 검색하는 질의의 주제가 1998년보다 더 다양해지고 광범위해졌으며, 성과 관련된 질의들이 감소하고 일반적인 엔터테인먼트성 질의가 증가하였다고 기술하였다.

Wang, Berry, Yang(2003)은 1997년 5월부터 2001년 5월까지 4년 동안 비교적 장기간에 걸쳐 University of Tennesse at Knoxville의 웹 사이트에 입력된 약 54만개의 질의를 분석하였다.

미국이 아닌 국외 선행 연구들은 대부분 유럽 시스템에 집중되어 있다. Hoelscher(1998)는 1998년 7월 한 달 동안 독일의 웹 검색엔진인 파이어볼에서 생성된 트랜잭션 로그에 기록된 약 1,600만개의 질의를 분석하였다. Jansen과 Spink(2005)는 올더웹의 이용자들이 남긴 2001년 2월 6일의 약 45만개, 2002년 5월 28일의 약 96만개의 질의들로부터 무작위로 추출된 약 2500개를 분류한 후 그 결과를 비교하였다. 이들은 올더웹 이용자들의 검색 행태가 시간이 지남에 따라 점점 단순해지고 있다고 보고하였다. 올더웹은 유럽의 검색 엔진이며 대부분의 이용자들은 독일인과 노르웨이인들이다. Cacheda와 Viña는(2001) 스페인의 디렉토리 시스템인 BIWE로부터 16일동안 생성된 로그를 분석하였다. 국내에서는 박소연, 이준호(2002)와 이준호, 권혁성, 박소연(2003)이 하루와 일주일 동안 생성된 대규모 트랜잭션 로그에 근거하여, 네이버 이용자의 검색 행태를 분석하였다.

트랜잭션 로그 분석을 이용한 이들 국내외연구들이 공통적으로 발견한 것은 웹 검색에 있어서 검색방식의 단순성이다. 즉 웹 검색 이용자들은 복잡한 검색식이나 연산자를 사용하지 않고, 적은 수의 검색어로 구성된 단순한 질의를 통해 정보검색을 수행하는 경향이 있었다(Jansen and Pooch 2001). 이러한 검색 행태는 전통적인 정보 검색 시스템 이용자들의 검색 행태와는 매우 상이하다고 할 수 있다.

3. 로그 분석 방법론의 활용도

본 장에서는 연구자들이 네이버에서 생성된 트랜잭션 로그에 근거하여 수행한 연구들을 중심으로 로그 분석 방법론이 웹 검색 분야에 어떻게 활용될 수 있는지를 제시하고자 한다. 연구자들이 네이버를 선택한 이유는 네이버가 대중성이나 인지도 면에서 국내 주요 검색 포탈로 인정받고 있기 때문이다. 즉 네이버가 국내 검색 포탈들 중 고객 만족도나 방문자 수 조사, 시장 점유율 등에서 지속적으로 1위를 차지하고 있기 때문이다. 또한 연구자들이 질의 로그나 클릭 로그의 설계, 로그 정제 등에 관여할 수 있기 때문이다.

위에서 언급되었듯이 웹 검색 서비스의 로그는 질의 로그와 클릭 로그로 구성되며, 대부분의 국내외 선행 연구들은 질의 로그의 분석에 집중되어 왔다. 본 연구자들은 질의 로그뿐만 아니라 클릭 로그에 근거하여 국내 웹 검색 질의의 형태 및 주제를 분석하였다. 국내외 선행 연구들 중 이용자가 실제로 조회한 자료, 즉 클릭 로그에 근거하여 이용자의 정보 요구를 파악한 사례는 찾아보기 드문 실정이다.

3.1 이용자들의 전반적인 검색 행태 분석

로그 분석을 통해 조사된 이용자들의 전반적인 이용 행태의 특징을 웹 검색 시스템 개선에 반영할 수 있다. 본 연구자들이 수행한 2003년 연구(이준호, 박소연, 권혁성)에서는 2003년 1월의 1주일 간 네이버의 통합 검색창에 입력된 질의들을 분석하였다. 연구 결과, 검색어가 어절 단위로 정의되는 경우와 형태소 단위로 정의되는 경우 하나의 검색어만을 포함하는 질의의 비중이 각각 90.42%, 47.67%로 나타났다. 이는 질의에 대한 형태소 분석 수행 여부와 관계없이 하나의 검색어로 구성된 질의를 수행하는 검색 서버의 별도 구축이 바람직함을 시사한다. 또한 대부분의 이용자들이 길이가 짧은 하나의 검색어로만 구성된 질의를 입력한다는 사실은 인기 있는 질의에 대해 테스트 컬렉션을 구축할 필요가 있음을 시사한다.

2005년 연구에서는(박소연, 이준호, 김지승) 1년 동안 네이버 이용자들이 입력한 질의를 기록

한 질의 로그와 질의에 대한 검색 결과에서 이용자가 조회한 문서를 기록한 클릭 로그에 근거하여 국내 웹 검색 질의의 형태 및 주제를 분석하였다. 질의를 형태별로 분류한 결과 사이트 검색 질의가 내용 검색 질의보다 많은 것으로 나타났다[1]. 따라서 검색 시스템 설계 시 사이트 검색 질의와 내용 검색 질의에 서로 다른 검색 알고리즘과 인터페이스의 적용을 고려할 수 있다. 곧 이용자가 입력한 질의의 형태를 파악할 수 있다면, 그 질의의 형태에 적합한 컬렉션들로부터 검색된 문서들을 결과 화면에서 우선 배열할 수 있다. 예를 들어 이용자가 사이트 검색으로 분류된 질의를 입력할 경우 "바로가기," "사이트" 등의 컬렉션에서 검색된 결과를 화면 상단에 배치할 수 있을 것이다. 또한 이용자들이 전반적으로 가장 많이 검색한 주제는 컴퓨터/인터넷, 엔터테인먼트, 쇼핑, 게임, 교육 순으로 나타났다. 이러한 결과는 이용자들의 정보 요구를 반영한다고 볼 수 있으므로 웹 검색 포털 업체들이 컨텐츠 구축의 우선 순위를 결정하는데 활용될 수 있을 것이다.

3.2 검색 행태 추이 분석

장기간에 걸쳐 웹 이용자 검색 행태의 추이를 분석하는 연구는 웹 검색 분야에서 매우 중요한 연구 주제로 인식되고 있다. 즉, 이용자가 검색하는 주제, 이용자가 검색하는 방법, 그리고 이용자가 입력하는 질의의 특성이 어떻게 변화하는지에 대한 분석은 이용자의 향후 검색 행태와 정보 요구를 예측하는데 활용될 수 있다. 이용자들의 검색 행태 추이에 대한 연구 결과는 인터넷 검색 포탈 업체들의 컨텐츠 구축 및 검색 시스템 개발에 있어서 중요한 자료로서 활용될 수 있다.

본 연구자들은 2005년 연구에서(박소연, 이준호) 1년이라는 장기간에 걸쳐 네이버에 입력된 검색 질의들의 표본과 각 질의에 대한 클릭 로그에 근거하여 국내 웹 이용자의 검색 행태 추이를 분석하였다. 조사 결과 웹 이용자들이 입력한 질의의 주제가 계절별, 주중과 주말, 요일별로 변화하는 것으로 나타났다. 질의 주제의 추이를 계절별로 분석한 결과, 활동성이 떨어지는 겨울에는 엔터테인먼트, 게임의 비중이 높았으며, 교육/학문에 관한 질의의 비율은 학기가 시작되는 봄에 가장 높고, 여름에 가장 낮은 것으로 나타났다. 지역/여행에 관한 질의는 많은 이들이 여행을 떠나는 여름에 가장 높고, 가을에 가장 낮은 것으로 나타났다. 컴퓨터에 관한 질의의 비율은 사계절 모두 높게 나타났다.

또한, 질의의 주제를 주중과 주말로 구분하여 비교한 결과, 주말에 엔터테인먼트, 게임에 관한 질의의 비율은 주중보다 현저하게 높았으며, 컴퓨터에 관한 질의의 비율도 주중보다 주말이 높았다. 반면 기업, 경제, 교육, 기관 등과 같은 나머지 주제에 있어서는 주중에 입력된 질의 수가 주말에 입력된 질의 수보다 많은 것으로 나타났으며, 주말에 쇼핑에 관한 질의의 비율은 주중과 비

1) 사이트 검색은 이용자가 찾고자 하는 대상이 웹 사이트인 경우이며, 내용 검색은 특정한 주제에 관한 신문 기사, 게시판 글, "지식인"에 올라간 글들을 클릭한 경우이다.

슷한 것으로 나타났다.

이러한 결과는 웹 이용자들의 정보 요구가 계절별, 주중과 주말, 요일별로 변화함을 보여주며, 따라서 인터넷 검색 포탈 업체들은 이러한 결과를 컨텐츠 구축에 다음과 같이 반영할 수 있다. 즉 겨울에는 엔터테인먼트, 게임, 쇼핑, 봄에는 교육, 기관, 문화, 라이프스타일, 사회, 여름에는 지역/여행, 그리고 가을에는 쇼핑, 경제와 관련된 컨텐츠를 강화하는 것이 바람직하다. 또한 인터넷 검색 포탈 업체들은 주말에 엔터테인먼트, 게임, 컴퓨터에 관한 컨텐츠를 집중적으로 강화하는 것을 고려하고, 이러한 주제와 관련된 광고의 노출 비율을 주말에 높이는 것을 고려할 수 있다.

한편 1년의 조사 기간 동안 계절별 또는 주중과 주말에 따라 주제의 비율이 변화하였으나, 전반적으로는 주제 비율의 변화가 미약함을 알 수 있다. 즉 1년의 기간 동안 특정한 주제와 관련된 질의의 비율이 증가하거나 감소하지는 않았으며, 전체적으로 주제의 비율이 어느 정도 고착화되어 사용자들의 검색 행태에 변화가 적었음을 알 수 있다. 따라서 이러한 결과는 인터넷 검색 포탈 업체들이 지난 3개월이나 6개월의 이용자 검색 행태의 분석을 통해 향후 3개월 또는 6개월의 검색 행태를 예측하는 것이 가능함을 시사한다.

3.3 키워드 마케팅 전략 구축

기업이나 광고주의 입장에서는 검색 행태 추이 분석의 일환으로 장기간의 키워드 추이 분석을 수행할 수 있다. 즉 특정한 키워드의 검색 빈도의 증감 추이에 근거하여 키워드의 검색 빈도가 증가할 것으로 예측되는 시점에 키워드에 대한 마케팅을 강화하는 것을 고려할 수 있다. 예를 들어 "꽃," "쵸콜렛," "크리스마스" 등의 키워드의 추이 분석을 통하여 키워드 입력이 증가할 것으로 예측되는 시점에 이러한 키워드와 관련된 광고의 노출 비율을 높일 수 있다. 또한 TV, 라디오, 신문 등과 같은 타 매체에 광고를 하는 경우 특정한 제품에 대한 광고 횟수와 검색 엔진에 입력된 질의의 빈도 간의 상관관계가 존재하는지에 대한 조사도 수행할 수 있다.

3.4 서비스 활용도 평가 지표

검색 포탈들의 평가 지표 중 가장 일반적이며 단순한 지표는 페이지 뷰(page view)와 방문자 수라고 할 수 있다. 예를 들어 웹 사이트 평가 및 트래픽 분석업체인 인터넷 매트릭스(http://www.metrixcorp.com)의 경우 일평균 방문자수, 방문자수, 도달률, 방문일수, 중복방문 사이트, 방문회수, 페이지뷰, 체류시간 등을 웹 사이트들의 주요 측정 지표로 사용하고 있다. 또한 검색 포탈들도 자체적으로 페이지 뷰와 방문자 수를 파악하고 있다. 페이지 뷰와 이용자 뷰는 질의 로그와 클릭 로그에서 직접적으로 도출될 수 있는 자료이며, 이러한 자료들은 포탈들의 성능 비교, 서비

스의 개선 등에 활용될 수 있다.

3.5 개별 서비스 평가

자료의 검색에 치중하는 국외 검색 포탈들과는 달리 국내 검색 포탈들은 지식인, 블로그, 카페, 뉴스, 디렉토리 등 다양한 서비스들을 제공하고 있다. 검색 포탈들의 경쟁력을 강화하고 서비스의 개선을 위해서는 이러한 서비스들에 대해 주기적이고 객관적인 평가 작업을 수행하는 것이 필수적이다. 또한 연구의 타당도를 높이기 위해 서비스들의 평가 시 실제 이용자들의 실제 질의를 활용하는 것이 바람직하다.

이에 본 연구자들은 국내 주요 검색 포탈들의 백과사전 서비스, 통합 검색 서비스, 지식인 서비스 등의 비교, 평가를 수행하여 왔다(박소연, 이준호, 2006a; 박소연, 이준호, 2006b; 박소연, 이준호, 전지운 2006). 첫째, 백과사전 평가 연구에서는 국내 주요 검색 포탈들인 네이버, 다음, 야후, 엠파스의 백과사전 서비스를 결과의 적합성, 최신성, 멀티미디어 제공 측면에서 분석, 평가하였으며, 평가 기준별로 세부적인 평가 항목과 평가방법론을 제시하였다. 백과사전 서비스 평가 시 본 연구에서는 실제 이용자들이 입력한 실제 질의들을 활용하였다. 즉 2005년 5월 31일 이용자들이 네이버 통합검색 창에 입력한 질의 중에서 무작위로 700개의 질의를 선정한 후, 중복되는 질의와 클린 검색에 의해 접속이 차단된 성인성 질의 등을 제외한 621개의 질의를 분석 대상으로 하였다. 2005년 5월 31일의 질의들을 선택한 이유는 이 날짜가 주중의 평일을 대표할 수 있는 날짜이며, 연구를 수행하던 시점에서 최신의 질의들을 구할 수 있는 날이기 때문이다. 700개의 질의를 선택한 이유는 하루에 네이버에 입력되는 통합 검색 질의의 수를 고려할 때, 표본 오차 95% 신뢰수준 ±4%와 ±5%를 허용할 경우 필요한 표본의 크기가 각각 600개와 384개로 통계학 문헌에서 제시되고 있기 때문이다(Arkin and Colton, 1963). 결과의 적합도와 최신성을 평가 시 질의 로그와 클릭 로그에 나타난 이용자의 정보 요구를 참고하였다.

조사 결과 모든 평가 항목에서 네이버의 백과사전 서비스가 가장 우수하고, 그 다음은 야후, 엠파스, 다음 순으로 나타났다. 첫째, 결과의 적합도에 있어서 4개의 검색 포탈들 중 네이버의 백과사전 커버리지가 가장 광범위하며, 결과의 적합도도 가장 높았다. 둘째, "넷마블", "싸이월드"와 같은 최신 개념의 포함, "한류"와 같이 최근에 의미가 변화된 개념의 포함, 정보의 갱신성이라는 측면에서 네이버 백과사전의 최신성이 가장 높은 것으로 나타났다. 본 연구에 포함된 평가 항목 중 결과의 최신성에 있어서 포탈별 편차가 가장 큰 것으로 나타났다. 셋째, 멀티미디어가 포함된 질의의 비율도 네이버, 야후, 엠파스, 다음 순으로 높게 나타났다. 멀티미디어의 경우 거의 모든 포탈들이 이미지만을 제공하고 있었다. 이러한 연구의 결과는 향후 웹 기반 백과사전 서비스의 개선에 활용되고, 이용자가 우수한 웹 기반 백과사전을 선택하는데 참고자료로 활용될 수

있을 것으로 기대된다.

백과사전 서비스 평가에 이어 통합 검색 서비스의 평가(박소연, 이준호, 2006b)도 수행되었다. 통합 검색 서비스는 포탈들이 제공하는 다양한 서비스들 중 이용도가 가장 높은 서비스이므로 통합 검색 서비스의 개선을 위하여서 통합 검색 서비스에 대한 평가 기준을 개발하고 이에 근거한 평가 작업을 수행하는 것이 필수적이다. 이에 연구자들은 2004년 6월 29일 이용자들이 네이버 통합 검색 창에 입력한 질의 중에서 무작위로 700개의 질의를 선정한 후, 중복되는 질의와 클린 검색에 의해 접속이 차단된 성인성 질의 등을 제외한 591개의 질의를 분석 대상으로 하였다. 통합 검색 서비스 평가 기준을 개발하고 이러한 평가 기준에 근거하여 네이버, 다음, 야후, 엠파스의 통합 검색 서비스를 평가하였다. 조사 결과 네이버의 통합 검색 서비스 커버리지가 가장 광범위하며, 결과의 만족도도 가장 높은 것으로 나타났다. 즉 네이버의 경우 결과가 제공되는 비율이 100%로 가장 높았고, 그 다음은 야후, 엠파스, 다음 순으로 나타났다. 결과의 만족도 비율에 있어서도 네이버가 전체 질의의 94.9%로 가장 높았으며, 그 다음은 엠파스, 야후, 다음 순으로 나타났다. 이러한 결과는 백과사전 서비스 평가 결과와 매우 유사하다고 할 수 있다.

백과사전, 통합 검색 서비스 외에 지식 검색 서비스에 대한 평가도 수행되었다(박소연, 이준호, 전지운, 2006). 곧 이 연구에서는 지식 검색 서비스 개선을 위하여 지식 검색 서비스를 구성하는 문서 평가 기준을 개발하였으며, 이러한 평가 기준에 근거하여 네이버 지식인을 대상으로 지식 문서 질문 제목의 적합도, 전체 질문의 적합도, 지식 문서 답변의 적합도 및 신뢰도를 분석, 평가하였다. 이 연구의 결과는 지식 검색 테스트 컬렉션 구축과 신뢰도 컬렉션 구축에 활용될 수 있을 것으로 보인다.

3.6 이용자의 항해 경로 조사

특정한 세션에서 이용자가 이동한 경로에 대한 연구를 통하여 인터페이스 개선을 도모할 수 있다. 예를 들어 이용자들이 "뒤로" 버튼을 과다하게 사용한다면 이는 인터페이스에 문제가 있음을 시사한다. 또한 이용자들이 동시에 자주 방문하는 서비스들이 있다면 이 서비스들을 나란히 배치함으로써 이용자의 동선을 줄일 수 있다.

3.7 클릭 어뷰즈(click abuse) 파악

현재 국내 검색 포탈들이 제공하는 통합 검색 서비스의 알고리즘에서 고려하는 중요한 요소는 문서의 클릭 횟수이다. 이에 일부 웹마스터들이 자신들이 구축한 페이지가 통합 검색 결과에서 상위에 오르도록 로봇이나 프로그램을 이용하여 자동으로 지속적으로 클릭하는 경우들이 존재하

며 이는 흔히 클릭 어뷰즈(click abuse) 또는 클릭 미스유즈(click misuse)로 불리운다. 이처럼 편법을 이용하여 클릭 횟수를 조작한 문서들이 통합 검색 결과에서 상위에 오르거나 인기 검색어에 오를 경우, 검색 결과의 질이 저하되며, 이용자들은 다른 적합한 문서를 찾는데 시간과 노력을 소모해야 한다. 이러한 클릭 어뷰즈는 클릭 로그에 나타난 정보를 통해 추적이 가능한데, 예를 들어 IP 주소 분석, 문서가 클릭된 시간대, 또는 문서가 클릭된 시간 간격 분석 등을 통해 프로그램에 의해 자동으로 클릭되었을 가능성이 높은 문서 파일들을 추적하고 결과 순위에서 배제할 수 있다.

3.8 클릭 로그 품질 평가

국내 검색 포탈들이 제공하는 다양한 서비스 중 가장 대중적이고 인기가 있는 서비스는 통합 검색 서비스이다. 해외 검색 포탈들의 경우 개별 컬렉션별로 분리하여 검색을 제공할 뿐 통합 검색 서비스를 제공하고 있지 않다. 따라서 급변하는 정보 환경에서 국내 검색 포탈들의 경쟁력을 강화하기 위해서는 우수한 통합 검색 서비스를 이용자에게 제공하는 것이 필수적이다. 한편 현재 국내 검색 포탈들의 통합 검색 서비스 알고리즘은 약간의 차이는 있으나 대부분 서비스별 클릭 횟수에 의존하고 있다. 즉 클릭이 많이 되는 서비스가 통합 검색 결과 노출 시 상위에 노출되고, 클릭이 별로 안 되는 서비스는 하위에 노출되거나 아예 통합 검색 결과에 노출되지 않는 것이다. 이러한 알고리즘은 문서의 질이나 적합도가 클릭 횟수에 비례한다는 것을 전제로 하고 있다. 그러나 이용자들이 클릭한 문서에는 다양한 유형이 존재할 수 있다. 질의에 대해 적합한 문서, 적합해 보이지만 실제 내용은 질의와 무관한 문서, 로봇이나 프로그램이 클릭한 문서 등 다양한 유형이 존재할 수 있다. 따라서 향후 통합 검색 알고리즘 개발 시 서비스별 클릭 횟수뿐만 아니라 클릭되는 문서들의 질도 함께 고려하는 작업이 바람직할 것으로 보인다. 클릭 로그에 포함된 문서의 질을 포함하기 위해서는 특정한 날짜에 이용자들이 입력한 질의를 무작위로 선정한 후 이 질의들에 대한 클릭 로그로터 이용자들이 조회한 문서들을 무작위로 선정하여 이 문서들의 질을 평가할 수 있다.

3.9 지역 관심사 파악, 개인 이용자의 관심사 파악

트랜잭션 로그에 나타난 IP 주소를 통해 이용자들이 어떤 지역에서 질의를 입력하는지 파악이 가능하며, 특정한 질의에 대해 이용자들이 위치한 지역에 대한 정보를 우선적으로 제공함으로써 이용자들의 편의를 도모할 수 있다. 예를 들어 이용자들이 극장, 식당, 날씨 등을 검색할 때 "지역 정보" 서비스에서 지역의 극장이나 식당 등에 대한 결과를 제공하는 것이다. 이러한 지역 정

보 자동 제공은 이미 네이버, 구글 등에서 수행하고 있으며, 이용자들은 지역 정보를 입력하지 않더라도 지역 정보를 제공받을 수 있다는 장점이 있다. 또한 검색 포탈 측에서는 과거 특정한 기간 동안의 개인 이용자의 검색 행태 분석에 근거하여 시기별로 이용자의 관심사에 부합하는 결과를 제공할 수도 있다.

3.10 오타 분석

검색 포탈들이 이용자의 만족도를 증대하기 위해서는 이용자의 편의를 높이는 기능들을 구현하는 것이 필요하다. 예를 들어 이용자가 자주 입력하는 오타를 검색 엔진이 자동으로 교정하여 적합한 검색 결과를 제공한다면 검색 결과의 효율성을 증대하고, 이용자의 만족도를 극대화할 수 있을 것이다. 본 연구자들이 수행한 2005년 연구에 의하면, 1년 동안 수집된 18,200개의 질의들 중 오타는 전체 질의의 2.1%에 해당하는 386개로 조사되었다. 이용자가 흔히 입력하는 오타의 유형은 영문 입력 모드에서의 한글 입력, 한글 입력 모드에서의 영어 입력, 문자의 삽입, 삭제, 교체, 전치로 인한 오타 순으로 나타났다. 김지승은 웹 검색 질의의 오류 교정을 위한 오류 교정 랭귀지 모델을 개발하였으며(2005), 실제로 일부 국내 검색 포탈들은 오타 교정 기능을 제공하고 있다.

4. 결 론

본 연구에서는 로그 분석 방법론이 웹 검색 분야에 어떻게 활용되고 기여할 수 있는지를 제시하였다. 로그 분석 방법론은 이용자들의 전반적인 검색 행태 분석, 검색 행태 추이 분석, 키워드 마케팅 전략 구축, 서비스 활용도 평가, 개별 서비스의 비교, 평가, 이용자의 항해 경로 파악, 클릭 어뷰즈 파악, 클릭 로그 품질 평가, 지역 관심사 파악, 이용자의 관심사 파악, 오타 분석 등에 활용될 수 있다. 이를 통하여 로그 분석 방법론은 궁극적으로 보다 효율적인 검색 시스템 개발과 컨텐츠 구축에 기여할 수 있다.

한편 본 연구의 수행 결과 향후 연구가 요구되는 사항들은 다음과 같다. 첫째, 로그 분석 방법은 이용자들의 이용 행태를 계량적인 방법으로 분석하므로, 이용자들이 특정한 방식으로 행동하는 이유, 이용자들의 웹 검색 서비스에 대한 만족도, 이용자들이 느끼는 웹 검색 서비스의 문제점 등의 분석을 위해서는 심층적인 인터뷰, 포커스 그룹 인터뷰, 관찰 등과 같은 질적 연구 방법을 병행하는 것이 필요하다. 또한 인터페이스 평가, 서비스의 질이나 신뢰도의 평가에도 전문가나 이용자를 대상으로 하는 조사를 수행하는 것이 요구된다. 둘째, 로그 분석을 이용한 국외 연

구와의 비교를 통하여 국내 이용자들의 웹 검색 행태의 특수성을 밝혀내는 것도 향후 과제라고 할 수 있다. 이는 구글과 같은 해외 업체의 국내 진출이 논의되고 있는 시점에서 필수적인 과제이다. 현재까지 발견된 특징은 국내 이용자들의 검색 행태가 국외 이용자들보다 훨씬 단순하다는 점이다. 즉 국내 이용자들의 세션 당 평균 질의 수, 질의 당 평균 검색어 수, 질의 당 평균 조회 페이지 수 등이 국외 이용자들보다 훨씬 낮은 것으로 나타났다(Park, Lee, and Bae, 2005). 셋째, 보다 장기간에 걸쳐 수집, 축적된 질의 로그 및 클릭 로그에 대한 분석 작업이 요구된다. 넷째, 포탈들을 둘러싼 정보 환경이 급변하므로, 포탈들의 서비스에 대한 주기적인 평가가 필요하다. 다섯째, 지금까지의 선행 연구들은 검색 행태의 추이를 질의의 형태와 주제에 국한시켜 조사하여 왔다. 향후 검색 방법의 변화, 즉 세션 길이, 질의 길이, 검색어 길이, 연산자 사용, 오타 비율, 이용자가 조회한 페이지 수 등의 변화에 대한 연구가 요구된다.

참 고 문 헌

곽승진. 2003. 청소년 대상 과학 분야 디지털도서관 구축을 위한 관련 사이트 분석 및 평가에 관한 연구. 『한국문헌정보학회지』, 37(3): 197-215.

김지승. 2005. 『확률 모델에 근거한 검색 질의의 문자열 유사도 계산』. 박사학위 논문, 숭실대학교 대학원, 컴퓨터학부.

박소연, 이준호. 2006a. 국내 주요 검색 포탈들의 백과사전 서비스 비교 평가. 『한국도서관정보학회지』, 37(2): 217-230.

박소연, 이준호. 2006b. 국내 주요 검색 포탈들의 통합 검색 서비스 만족도 비교 평가. NHN 기술보고서.

박소연, 이준호. 2005. 국내 웹 이용자의 검색 행태 추이 분석. 『한국문헌정보학회지』, 39(2): 147-160.

박소연, 이준호. 2002. 로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구. 『정보관리학회지』, 19(3): 111-122.

박소연, 이준호, 김지승. 2005. 클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석. 『한국문헌정보학회지』, 39(1): 265-278.

박소연, 이준호, 전지운. 2006. 지식 검색 서비스 개선을 위한 문서의 적합도 및 신뢰도 분석. 『한국문헌정보학회지』, 40(2): 299-314.

유사라. 2002. 국가과학기술전자도서관 이용자 정보요구와 이용 행태 분석. 『한국문헌정보학회지』, 36(1): 25-40.

이준호, 박소연, 권혁성. 2003. 질의 로그 분석을 통한 네이버 이용자의 검색 행태 연구. 『정보관리학회지』, 20(2): 27-40.

Arkin, H., and Colton, R. 1963. *Tables for Statisticians.* New York: Barnes & Noble Inc.

Cacheda, F., & Vinã, Á.(2001). Experiences retrieving information in the World Wide Web. In K. Jeffay, & R. Steinmetz(Eds.), Proceedings of the 6th IEEE Symposium on Computers and Communications(pp. 72-79). Piscataway, NJ: IEEE.

Jansen, B. J., & Pooch, U. 2001. "A review of web searching studies and a framework for future research." *Journal of the American Society for Information Science and Technology,* 52(3): 235-246.

Jansen, B. J., Spink, A., and Pedersen, J. 2005. "A temporal comparison of AltaVista web searching." *Journal of the American Society for Information Science and Technology,* 56(6): 559-570.

Jansen, B. J., and Spink, A. 2005. "An analysis of Web searching by European Allthe Web.com users." *Information Processing and Management,* 41(2), 361-381.

Jansen, B. J., Spink, A., and Saracevic, T. 2000. "Real life, real users, and real needs: a study and analysis of user queries on the web." *Information Processing and Management,* 36(2): 207-227.

Lee, J. Y., & Paik, W. 2006. Analysis of Korean Patent & Trademark Retrieval Query Log to Improve Retrieval and Query Reformulation Efficiency. 『정보관리학회지』, 23(2): 61-80.

Park, S., Lee, J., & Bae, H. 2005. "End user searching: A web log analysis of NAVER, a Korean web search engine." *Library and Information Science Research,* 27(2), 203-221.

Peters, T. A. 1993. "The history and development of transaction log analysis." *Library Hi Tech,* 11(2), 41-66.

Ross, N. C. M., and Wolfram, D. 2000. "End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine." *Journal of the American Society for Information Science and Technology,* 51(10): 949-958.

Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. 1999. "Analysis of a very large web search engine query log." *SIGIR Forum,* 33(1): 6-12.

Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. 2001. "Searching the web: The public and their queries." *Journal of the American Society for Information Science and Technology*, 52(3): 226-234.

Spink, A., Jansen, B. J., Wolfram, D., and Saracevic, T. 2002. "From e-sex to e-commerce: Web search changes." *IEEE Computer*, 35(3): 133-135.

Wang, P., Berry, M. W., and Yang, Y. 2003. "Mining Longitudinal Web Queries: Trends and Patterns." *Journal of the American Society for Information Science and Technology*, 54(8): 743-758.