로그 분석을 통한 이용자의 웹 문서 검색 행태에 관한 연구*

Investigating Web Search Behavior via Query Log Analysis

박소연(Soyeon Park)**, 이준호(Joon-Ho Lee)***

초 록

본 연구에서는 웹 검색 이용자들의 전반적인 검색 행태를 이해하기 위하여 국내에서 널리 사용되고 있는 웹 검색 서비스 네이버에서 생성된 검색 트랜잭션 로그를 분석하였다. 본 연구에서는 웹 검색 트랜잭션 로그 분석에 필요한 세션 정의 방법을 설명하고 로그 정제 및 질의 유형 분류 방법을 제시하였으며, 한글 검색 트랜잭션 로그 분석에 필수적인 검색어 정의 방법을 제안하였다. 본 연구의 결과는 보다 효과적인 국내 웹 검색 시스템 개발과 서비스 구축에 기여할 것으로 기대된다.

ABSTRACT

In order to investigate information seeking behavior of web search users, this study analyzes transaction logs posed by users of NAVER, a major Korean Internet search service. We present a session definition method for Web transaction log analysis, a way of cleaning original logs and a query classification method. We also propose a query term definition method that is necessary for Korean Web transaction log analysis. It is expected that this study could contribute to the development and implementation of more effective Web search systems and services.

키워드: 웹 검색, 질의 로그 분석, 검색 행태 분석, information seeking behavior, log analysis, query classification method, query term definition

---

* 이 논문은 2001년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2001-003-C00426)

** 계명대학교 문헌정보학과 전임강사(sypark@kmu.ac.kr)

*** 숭실대학교 컴퓨터학부 조교수(joonho@computing.soongsil.ac.kr)

■ 논문 접수일 : 2002. 8. 19

■ 게재 확정일 : 2002. 9. 9

1. 연구의 목적 및 필요성

인터넷의 사용과 보급이 폭발적으로 증가함에 따라 인터넷을 통한 정보의 접근을 지원하기 위해 웹 검색 서비스들이 활성화되었다. 1990년대 중반의 초기 웹 검색 시스템들은 질의와 문서 사이의 유사도를 계산하는 전통적인 정보 검색 이론들을 기반으로 구축되었다. 그러나, 웹 검색 환경과 전통적인 정보 검색 환경 사이의 차이점이 인식되었으며, 그 결과 1990년대 후반에는 유사도 계산 이외에 페이지의 중요도를 문서의 순위 결정에 반영하는 시스템이 출현하게 되었다. 따라서, 기존의 정보 검색 환경과 매우 상이한 웹 검색 환경에서 보다 효율적인 검색을 지원하기 위하여 새로운 정보 검색 환경에 대한 이해 및 분석이 절실히 요구된다.

즉, 웹 검색 환경과 전통적인 정보 검색 환경 사이에는 다음과 같은 차이점을 발견할 수 있다. 첫째, 웹 정보 환경은 문서의 생성과 소멸이 빈번한 매우 유동적인 환경이기 때문에, 검색 효과의 척도로 사용되어 왔던 재현율과 정확률의 적용에 어려움이 있다. 둘째, 전통적인 정보 검색 시스템의 이용자들은 직업, 전공, 검색 경험 등에 있어서 비교적 동질적인 집단에 속한다. 그러나 웹 검색 환경에서는 이용자들이 다양한 계층 및 연령층으로 확대되었다. 셋째, 전통적인 정보 검색 환경에서 검색 대상이 되는 문서들은 형식, 주제, 생산자 등에 있어서 비교적 동질적이다. 그러나 웹 검색 환경에서는 검색 대상 문서들이 매우 이질적인 특성을 지니고 있다.

한편, 국내 웹 검색에 관한 연구는 전산학, 경영학, 문헌정보학, 심리학, 신문방송학 등 다양한 분야에서 수행되고 있으며, 특히 문헌정보학 분야의 선행 연구들은 인터넷 검색 엔진들의 특성 및 기능 비교(이란주 2000), 그리고 실험을 통한 성능 평가(정영미, 김성은 1997; 황상규, 오경묵, 변영태 1999), 웹 사이트의 설계 및 모형 개발(곽철완 2001), 이용자 만족도의 조사, 이용자 검색 행태의 분석을 수행하는 연구들로(오경묵, 황상규, 이용현 1999; 오삼균, 박희진 2000; 이명희 1998; 이해영, 이수영 2001) 구분될 수 있다. 이들 중 이용자 검색 행태를 분석한 대부분의 연구들은 전공, 연령, 검색 경험에 있어서 비교적 동질적인 소수의 이용자들을 대상으로 이루어져 왔다. 그러나 웹 검색 이용자들은 매우 다양한 계층 및 연령으로 구성되어 있기 때문에, 기존의 연구를 통한 전체 웹 검색 이용자들의 검색 행태 파악은 어려운 실정이다.

국외의 경우, 전체 웹 검색 이용자들의 검색 형태 연구를 위해 검색 엔진이 생성한 검색 트랜잭션 로그를 분석하는 방법이 사용되고 있다. 이용자와 검색 시스템의 모든 검색 과정을 기록, 저장하는 검색 트랜잭션 로그는 이용자의 실제 검색 행위를 사실적으로 반영한다. 따라서 이러한 로그의 분석은 웹 검색 이용자들의

검색 행태 연구를 위한 합리적이고 객관적인 방법으로 인정받고 있다(Jansen and Pooch 2000). 그러나, 기존에 많이 사용되었던 설문 조사 또는 인터뷰 자료를 분석하는 방법은 검색 트랜잭션 로그 분석 방법과 비교하여 이용자의 실제 검색 행위와 설문 조사 또는 인터뷰 자료에 차이점이 발생할 수 있다.

국내의 경우, 검색 트랜잭션 로그를 상용 로그 분석 도구를 이용하여 분석함으로써, 검색 시스템에 입력되는 질의 수를 파악하는 작업은 많이 수행되고 있다. 그러나, 질의에 포함되는 단어 수, 이용자가 검토하는 검색 결과 페이지 수 등과 같은 웹 검색 이용자들의 검색 행태를 분석하기 위하여 로그 형식을 설계하고, 생성된 로그를 분석하는 연구는 발견하기 어려운 실정이다. 이에 본 연구에서는 국내에서 널리 사용되고 있는 웹 검색 서비스 네이버에서 생성된 검색 트랜잭션 로그 중 일부를 분석하여, 웹 검색 이용자들의 검색 행태를 파악하고자 한다.

2. 선행 연구

국외의 웹 검색에 관한 연구의 경우, 다른 정보학 분야의 연구에 비해 비교적 초기 단계에 있으나, 국내 연구보다 자료의 규모가 방대하고, 다양한 연구 주제가 연구되며 다양한 연구방법이 적용되고 있다고 할 수 있다. 국외 연구 중 트랜잭션 로그 분석을 통하여 웹 검색 이용자들의 검색 행태를 조사한 연구들로는 Jansen, Spink, Saracevic의 익사이트 연구(2000, 2001), Hoelscher의 파이어볼 연구(1998), Silverstein 등의 알타비스타 연구(1999), Ross와 Wolfram의 익사이트 연구(2001) 등을 들 수 있다.

Jansen, Spink, Saracevic은 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. Hoelscher (1998)는 1998년 7월 한 달 동안 독일의 웹 검색엔진인 파이어볼에서 생성된 트랜잭션 로그에 기록된 약 1,600만개의 질의를 분석하였다. 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 했다는 점에서 의미가 있다.

트랜잭션 로그 분석을 이용한 이들 국외연구들이 공통적으로 발견한 것은 웹 검색에 있어서 검색방식의 단순성이다. 즉 웹 검색 이용자들은 복잡한 검색식이나 연산자를 사용하지 않고, 적은 수의 검색어로 구성된 단순한 질의를 통해 정보

검색을 수행하는 경향이 있었다(Jansen and Pooch 2001; Abdulla, Liu, and Fox 1998). 이러한 검색 행태는 전통적인 정보 검색 시스템 이용자들의 검색 행태와는 매우 상이하다고 할 수 있다.

한편, 국내에서 이용자 검색 행태를 분석한 대부분의 연구들은 전공, 연령, 학력, 검색 경험에 있어서 비교적 동질적인 소수의 이용자들을 대상으로 이루어져 왔다. 오경묵, 황상규, 이용현의 연구는(1999) 전자과 대학원생 19명을 대상으로 수행한 실험을 통해 이용자들의 행동양식을 조사하고, 4개의 검색 시스템을 비교하였다. 이명희(1998)는 교육학 분야 주제전문가 10명을 주제전문가와 탐색전문가로 나누어 이들의 검색 행태를 비교하였다. 오삼균과 박희진(2000)은 대학생 30명을 대상으로 실험을 실시하여 한글 알타비스타와 네이버를 검색 효율성, 검색 결과의 정확률, 검색 결과의 갱신성, 이용자의 만족도로 비교, 평가하였다.

이들 연구들은 실험, 설문조사 등을 통하여 웹 검색 이용자들의 검색 행태를 분석하였다는 데에 의의가 있다. 그러나 웹 검색 이용자들은 매우 다양한 계층 및 연령으로 구성되어 있기 때문에, 기존의 국내 연구를 통한 전체 웹 검색 이용자들의 검색 행태 파악은 어려운 실정이다. 따라서 본 연구에서는 웹 검색 서비스 네이버의 트랜잭션 로그 분석을 통하여 보다 다양한 웹 검색 이용자들의 전반적인 검색 행태를 조사하고자 한다.

3. 연구 방법

본 연구에서는 웹 검색 이용자들의 검색 행태를 파악하기 위하여 검색 트랜잭션 로그를 분석하였다. 분석 대상이 된 검색 트랜잭션 로그는 2002년 6월 24일 웹 검색 서비스 네이버에서 생성되었으며, 본 연구에서는 이들 중에서 일부에 대한 분석을 수행하였다. 네이버는 대중성이나 인지도에 있어서 국내 주요 웹 검색 서비스로 인정받고 있다. 즉, 웹사이트 평가 및 트래픽 분석업체인 인터넷 매트릭스에(http://www.internetmatrix.co.kr) 의하면 네이버는 2002년 7월 1개월 동안 국내 네티즌들이 가장 많이 방문하는 사이트 중 3위를 차지하였으며, 네이버는 2001년 12월 조선일보, 한국일보가 주관한 인터넷 포탈 부문에서 대상을 수상하였다. 따라서, 네이버 검색 트랜잭션 로그를 분석함으로써 국내 네티즌들의 전반적인 웹 검색 행태를 추측할 수 있다.

네이버는 웹 검색, 디렉토리 검색, 백과사전 검색, 뉴스 검색, Q/A 검색, 이미지 검색 등을 개별적으로 지원하고 있으며, 또한 이들 검색 결과들을 통합하여 보여주는 통합 검색을 제공하고 있다. 본 연구에서는 이들 중에서 웹 검색 트랜잭션 로그만을 분석 대상으로 하였다. 또한, 본 연구에서 사용된 검색 트랜잭션 로그가 생성될 당시 네이버는 불리안 연산, 근접 연산, 절단 연산을 지원하지 않았기 때문에, 이러한 연산과 관련된 이용자들의 행태

분석은 본 연구에 포함되지 않았다.

검색 트랜잭션 로그는 온라인 정보 검색 시스템과 이러한 시스템을 사용하는 이용자의 상호 작용에 대한 전자적인 기록으로서, 검색 트랜잭션 로그 분석은 정보 검색 분야에서 이용자의 검색 행태 연구를 위한 합리적이고 객관적인 방법으로 인정받고 있다 (Jansen and Pooch 2000). 트랜잭션 로그 분석은 웹이 등장하기 이전부터 도서관의 OPAC(Online Public Access Catalog) 시스템(Ballard 1994), 실험적 정보 검색 시스템 등 다양한 환경에서 활용되어 왔다. 즉, 연구자들은 정보 검색 시스템, 이러한 시스템에 대한 인간의 이용, 그리고 시스템이 어떻게 이용되는가에 대한 인간의 이해를 개선할 목적으로 트랜잭션 로그 자료를 이용하고 있다 (Peter 1993, p.37).

검색 트랜잭션 로그를 분석한 국외 선행 연구들을 검토한 결과, 아직까지 일반화된 용어와 방법론이 정착되어 있지 않다는 문제점을 발견하였다. 예를 들어 동일한 용어가 연구마다 약간씩 다른 의미로 사용되며, 로그를 처리하는 방식에 있어서도 연구마다 차이가 있음을 확인하였다. 또한 이들 연구들은 영어 문서를 검색하는 시스템을 대상으로 하였기 때문에, 이들 연구의 방법론을 한글 검색 트랜잭션 로그 분석에 적용할 경우 문제점이 발생한다. 따라서 다음에서는 일반적인 검색 트랜잭션 로그 분석에 필수적인 세션 정의 방법, 로그 정제, 질의 유형 분류, 그리고 한글 검색 트랜잭션 로그 분석에 필수적인 검색어 정의에 대하여 기술한다.

3.1 세션 정의 방법

세션은 검색 트랜잭션 로그의 기본 단위로서, 일반적으로 한 명의 이용자가 단일한 검색 목적을 지니고 처음 검색을 시작하여 검색을 종료하기까지의 일련의 과정으로 정의된다. 선행 연구들은 세션에 관한 일반적인 정의에는 동의하나, 검색 질의들을 세션으로 구분함에 있어서 상이한 방법을 적용하였는데, 익사이트의 로그를 분석한 Spink et al.(2001)의 연구와 알타비스타의 로그를 분석한 Silverstein et al.(1999)의 연구가 그 대표적인 예이다.

Spink et al.(2001)는 세션의 정의를 위하여 익사이트 서버가 할당한 이용자 식별자(user identifier)를 이용하였다. 즉, 임의의 컴퓨터가 브라우져를 통하여 익사이트에 검색을 요청할 때, 익사이트는 쿠기를 생성하여 검색을 요청한 컴퓨터에게 쿠기의 소멸 시간을 지정하지 않은 상태로 전달한다. 이후부터 이 쿠기는 이용자 식별자로 사용되며, 검색을 요청한 컴퓨터의 브라우져들이 모두 종료될 때 소멸된다. 따라서 이 연구에서 세션은 일 초가 될 수도 있고, 몇 시간이 될 수도 있다. Spink et al.가 사용한 세션 설정 방법은 공공 장소에 위치한 하나의 컴퓨터를 다수의 사용자가 이용할 경우 하나의 세

션에 포함되는 질의 수가 과다해지는 문제점을 지니고 있다 (Jansen and Pooch 2001).

위에서 언급된 문제점을 해결하기 위해 Silverstein et al.(1999)은 세션이 일정 기간 동안의 일련의 검색 과정이기 때문에, 이용자가 특정한 검색 목적을 다른 검색 목적으로 전환하게 되는 경우 시간적 공백이 발생하는 점에 착안하였다. 즉 Silverstein et al.은 검색을 요청한 컴퓨터에게 쿠기를 전달할 때, 이 쿠기가 5분 후에 소멸되도록 지정하였다. 또한, 5분 이내에 다시 검색이 요청될 때, 동일한 쿠기를 5분 후에 소멸되도록 지정하여 전달하였다. 따라서 이용자가 5분 동안 질의를 입력하지 않으면 새로운 세션이 정의되며, 5분 이내에 질의를 입력하면 기존 세션이 연장된다. 다시 말해 이들은 5분 내에 새로운 질의가 입력될 경우 이전 세션을 연장하는 세션 정의 방법을 제시하였다. 본 연구는 Spink et al.이 사용한 세션 정의 방법에 분명한 오류가 있다고 판단하고, Silverstein et al.이 사용한 세션 정의 방법을 사용하였다.

3.2 로그 정제

검색 트랜잭션 로그로부터 이용자가 정상적으로 입력한 질의라고 판단하기 어려운 다수의 로그들을 발견하였다. 이러한 로그들이 생성되는 이유는 다음과 같이 크게 3가지로 구분될 수 있다. 첫째, 이용자가 질의를 입력한 후 네트워크의 속도 저하, 검색 시스템의 과부하 등의 이유로 검색 결과 화면의 생성이 다소 지연되면, 이용자는 새로고침 버튼을 클릭하거나 재검색을 수행하는 경향이 있다. 이때 이용자가 검색 결과를 받지 않은 질의에 대해서도 검색 트랜잭션 로그가 생성되며, 동일한 검색 트랜잭션 로그가 연속해서 파일에 기록된다. 본 연구에서는 동일한 로그가 연속해서 출현할 경우, 1개의 로그만을 남기고 나머지를 제거하였다.

둘째, 3.1절에서 설명된 것처럼 본 연구에서 사용한 세션 정의 방법이 현재까지 알려진 최선의 방법일 지라도, 이 방법 역시 문제점이 있음을 발견하였다. 즉, 일부 이용자들은 5분의 휴식이나 공백 이후, 이전 질의의 검색 결과를 기반으로 문서 질의 또는 반복 질의를 수행하였다. 이러한 경우 새로운 세션이 생성되며, 이 세션은 입력 질의를 포함하지 않고 문서 질의와 반복 질의만으로 구성된다. 본 연구에서는 이러한 세션을 비정상적으로 간주하고, 로그 파일로부터 제거하였다.

셋째, 검색 트랜잭션 로그를 살펴보면 프로그램이 자동으로 입력한 질의라고 판단되는 다수의 로그들이 존재한다. 예를 들면, 하나의 세션에 수백 수천개의 입력 질의, 문서 질의, 반복 질의가 포함된 경우를 발견할 수 있다. 본 연구에서는 질의 파일을 수작업으로 검토하였다. 그 결과 100개 이상의 입력 질의, 21개 이상의 문서 질의, 21개 이상의 반복 질의를 포함하는 세션은 프로그램에 의해 자동으로 생성되었을 가능성이 높음을 확인하였으

며, 따라서 이러한 세션들을 로그 파일로부터 제거하였다.

3.3 질의 유형 분류

질의는 세션의 구성 요소로서, 하나 이상의 검색어들로 구성된다. 본 연구에서는 질의의 개념이 비교적 명확하게 정의된 Spink et al.(2001)의 연구를 참고로 하여, 전체 질의들을 다음과 같이 3가지 유형들로 분류하였다. 전체 질의 수는 입력 질의 수, 문서 질의 수, 반복 질의 수의 합과 동일하다.

입력 질의: 이용자가 검색창에 직접 입력한 스트링으로서 검색어들로 구성된다. 하나의 세션에 포함된 입력 질의들은 최초 질의, 검색어 추가 질의, 검색어 삭제 질의, 검색어 추가 및 삭제 질의, 변경 질의, 동일 질의로 세분화될 수 있다. 여기에서 검색어는 어절 단위로 인식되었고, 변경 질의는 이전 질의와 중복된 어절이 없는 경우이며, 동일 질의는 어절의 순서만이 변화된 질의를 의미한다. 전체 입력 질의 수는 세분화된 6개 유형의 질의 수의 합과 동일하다.

문서 질의: 검색어를 포함하지 않는 질의로서 이용자가 지정한 문서 전체가 질의로 사용되며, 일반적으로 "유사 문서 검색"이라는 기능으로 지칭된다.

반복 질의: 웹 검색 시스템들은 입력 질의와 문서 질의를 수행한 후, 질의에 적합할 가능성이 높은 상위 10개 정도의 문서들을 검색 결과로서 이용자에게 제공한다. 이때 이용자가 상위로부터 11번째 이후의 문서들을 보고자 할 경우, 즉 다음 결과 화면을 보고자 요청할 때 반복 질의가 발생한다.

3.4 검색어 정의

검색어는 질의를 구성하는 기본 단위로서, 영어의 경우 빈칸, 마침표, 쉼표, 개행 문자 등과 같은 공백 문자(blank character)들에 의해 구분되는 일련의 문자 또는 숫자로서 정의된다. 그러나, 한글은 복합 명사를 구성하는 단일 명사들 사이의 띄어쓰기를 자유롭게 규정하고 있기 때문에, 검색어를 영어의 경우에서처럼 정의할 경우 문제점이 발생한다. 즉, "정보검색시스템"과 "정보 검색 시스템"은 동일한 질의임에도 불구하고, 서로 다른 색인어들을 포함하고 있는 것으로 인식된다. 따라서 한글 검색 트랜잭션 로그 분석에서 검색어에 대한 통계를 추출할 경우, 검색어에 대한 정확한 정의가 요구된다. 본 연구에서는 첫째, 영어의 경우와 유사하게 어절 단위로 검색어를 인식한 경우, 둘째, 어절을 형태소 단위로 분리한 후, 각각의 형태소를 검색어로 인식한 경우 모두에 대하여 분석을 수행하였다.

4. 로그 분석 결과

4.1 질의 수 분석

본 연구는 2002년 6월 24일 웹 검색

서비스 네이버에서 생성된 트랜잭션 로그 중 일부를 분석하였다. 분석 대상이 된 로그 파일에 포함된 전체 세션 수는 162,312개이고, 전체 질의 수는 547,532개 이었다. 전체 질의들을 입력 질의, 문서 질의, 반복 질의로 구분한 후, 각 유형별로 질의 수를 살펴보면 〈표1〉과 같다. 즉, 전체 질의 중 입력 질의 수는 344,236개(62.87%), 문서 질의 수는 3,915개(0.7%), 반복 질의 수는 199,381개(36.41%) 이었다.

〈표 1〉 전체 질의의 유형별 질의 수 분석

	질의 수
입력 질의	344,236
문서 질의	3,915
반복 질의	199,381
총 계	547,532

〈표 2〉는 세션별 전체 질의 수, 입력 질의 수, 문서 질의 수, 반복 질의 수에 대한 기술 통계를 보여준다. 이용자들은 세션별 평균 3.3733개의 질의를 수행하였고, 평균 2.1208개의 질의를 직접 입력하였으며, "유사 문서 검색"으로 지칭되는 문서 질의는 세션별 평균 0.02회 수행되었다. 최초 질의는 세션 시작 질의이므로 세션별 평균이 1회이며, 최초 질의의 총계는 전체 세션 수와 동일하다. 한편 세션별 평균 반복 질의 수는 1.2284로서, 이는 이용자들이 평균적으로 출력하는 결과 화면이 약 2개 페이지임을 의미한다. 결과 화면 출력 페이지 수의 편차는 비교적 큰 편으로 이는 일부 이용자들이 많은 수의 결과 화면을 출력하기 때문이다. 또한 이용자들이 질의를 변경할 경우, 검색어를 추가 또는 삭제하기보다는 이전 질의를 전적으로 변경하는 경우가 많음을 알 수 있다.

〈표 2〉 세션별 질의 수에 대한 기술 통계

	평 균	표준편차	최소값	최대값	총 계
세션별 전체 질의 수	3.3733	4.2602	1	99	547,532
세션별 입력 질의 수	2.1208	2.3807	1	82	344,236
최초 질의	1	0	1	1	162,312
검색어 추가 질의	0.0533	0.2851	0	10	8,644
검색어 삭제 질의	0.0449	0.2461	0	7	7,284
검색어 추가 및 삭제 질의	0.1336	0.6669	0	39	21,689
동일 질의	0.0690	0.3249	0	23	11,204
변경 질의	0.8200	1.8332	0	60	133,103
세션별 문서 질의 수	0.0241	0.3992	0	20	3,915
세션별 반복 질의 수	1.2284	2.9082	0	98	199,381

〈표 3〉 세션별 질의 수 분포 (단위: 세션)

	0개	1개	2개	3개	4개	5개 이상
세션별 전체 질의 수		71,913 (44.31%)	28,770 (17.73%)	16,952 (10.44%)	10,553 (6.50%)	34,124 (21.02%)
세션별 입력 질의 수		94,553 (58.25%)	30,860 (19.01%)	14,644 (9.02%)	7,795 (4.8%)	14,460 (8.91%)
최초 질의		162,312 (100%)				
검색어 추가 질의	155,304 (95.68%)	5,859 (3.61%)	860 (0.53%)	186 (0.11%)	56 (0.03%)	47 (0.03%)
검색어 삭제 질의	156,099 (96.17%)	5,393 (3.32%)	648 (0.4%)	119 (0.07%)	35 (0.02%)	18 (0.01%)
검색어 추가 및 삭제 질의	150,753 (92.88%)	6,925 (4.27%)	2,439 (1.5%)	1,014 (0.62%)	480 (0.30%)	701 (0.43%)
동일 질의	153,273 (94.43%)	7,465 (4.6%)	1,202 (0.74%)	257 (0.16%)	69 (0.04%)	46 (0.03%)
변경 질의	105,883 (66.23%)	28,523 (17.57%)	12,583 (7.75%)	6,286 (3.87%)	3433 (2.12%)	6104 (3.76%)
세션별 문서 질의 수	160,890 (99.12%)	749 (0.46%)	248 (0.15%)	132 (0.08%)	66 (0.04%)	227 (0.14%)
세션별 반복 질의 수	108,699 (66.97%)	18,897 (11.64%)	10,467 (6.45%)	6,213 (3.83%)	4,258 (2.62%)	13,778 (8.49%)

〈표 3〉은 질의 유형에 따른 세션별 질의 수의 분포를 보여준다. 62%의 세션이 2개 이하의 질의를 포함하고 있으며, 또한 입력 질의만을 고려할 경우 77%의 세션이 2개 이하의 입력 질의를 포함하고 있다. 이는 이용자들이 정보의 발견을 위하여 웹 검색에 소비하는 시간과 노력이 많지 않음을 시사하며, 이에 대한 이유로서 다음과 같은 사항들을 고려할 수 있다. 첫째, 웹 검색을 통하여 이용자가 접근하고자 하는 정보들이 매우 일반적이기 때문에, 웹 상에 많은 수의 적합 문서들이 존재한다. 따라서 이러한 적합 문서들의 일부가 웹 검색 시스템에 의해 쉽게 검색될 수 있다. 둘째, 웹 상에 적은 수의 적합 문서들이 존재할 지라도, 웹 검색 시스템의 성능이 우수하기 때문에 이용자가 만족하는 검색 결과를 제공한다. 셋째, 웹 검색 이용자는 정보 발견에 있어서 절실한 필요성을 지니고 있지 않기 때문에, 정보의 발견을 도중에 쉽게 포기하는 경향이 있다.

4.2 검색어 수 분석

본 절에서는 입력 질의에 포함된 검색어 수에 대한 분석 결과를 기술한다. 〈표 4〉는 입력 질의별 검색어 수에 대한 기술

〈표 4〉입력 질의별 검색어 수에 대한 기술 통계

	평 균	표준편차	최 소 값	최 대 값
어절 단위 검색어	1.5841	1.4168	0	152
형태소 단위 검색어	2.5450	2.2087	0	163

〈표 5〉 입력 질의별 검색어 수 분포 (단위: 질의)

	0개	1개	2개	3개	4개	5개 이상
어절 단위 검색어	236 (0.07%)	231,700 (67.31%)	68,110 (19.79%)	25,708 (7.47%)	9,604 (2.79%)	8,878 (2.58%)
형태소 단위 검색어	236 (0.07%)	103,358 (30.03%)	107,678 (31.28%)	67,317 (18.10%)	32,951 (9.57%)	32,696 (9.5%)

통계를 보여준다. 검색어가 어절 단위로 정의되는 경우 입력 질의별 평균 검색어 수는 1.5841이었고, 검색어가 형태소 단위로 정의되는 경우 입력 질의별 평균 검색어 수는 2.5450이었다. 이러한 결과를 통하여 웹 검색 이용자들은 매우 적은 수의 검색어로 구성된 단순한 질의를 수행하는 경향이 있음을 알 수 있다.

〈표 5〉는 입력 질의별 검색어 수의 분포를 보여준다. 현재 서비스를 제공하고 있는 웹 검색 시스템들을 살펴보면, 질의에 대하여 형태소 분석을 수행하는 시스템과 형태소 분석을 수행하지 않는 시스템으로 구분될 수 있다. 검색어가 어절 단위로 정의되는 경우 하나의 검색어만을 포함하는 입력 질의가 67.31%이었으며, 검색어가 형태소 단위로 정의되는 경우 하나의 검색어만을 포함하는 입력 질의가 30.03%이었다. 따라서 질의에 대한 형태소 분석의 수행 여부와 관계없이 하나의 검색어로 구성된 질의를 수행하는 검색 서버의 별도 구축이 바람직함을 알 수 있다. 한편, 일부의 입력 질의들은 검색어를 포함하고 있지 않으며, 이러한 질의들은 검색어로 부적합한 의미 없는 문자들로 구성되어 있다.

5. 결론 및 제언

본 연구에서는 웹 검색 이용자들의 전반적인 검색 행태를 이해하기 위하여 국내에서 널리 사용되고 있는 웹 검색 서비스 네이버에서 생성된 검색 트랜잭션 로그들 중 일부를 분석하였다. 즉, 실험 환경이 아닌 현실 환경에서 이루어진 실제 이용자들의 정보 요구와 검색 행위를 조사하였다. 4장에서 기술된 분석 결과는 보다 효과적인 국내 웹 검색 시스템 개발과 서비스 구축에 기여할 것으로 기대된다. 또한, 본 연구에서는 웹 검색 트랜잭션 로그 분석에 필요한 세션 정의 방법을 설명하고 로그 정제 및 질의 유형 분류

방법을 제시하였으며, 한글 검색 트랜잭션 로그 분석에 필수적인 검색어 정의 방법을 제안하였다.

한편, 본 연구의 수행 결과 향후 연구가 요구되는 다음과 같은 사항들을 발견하였다. 첫째, 본 연구에서는 웹 검색 이용자들의 검색 행태를 계량적인 방법으로 분석하였다. 따라서 이용자들이 특정한 방식으로 행동하는 이유, 이용자들의 검색 과정에 대한 만족도 등의 분석을 위해서는 실험이나, 인터뷰, 관찰 등의 별도의 연구가 요구된다. 둘째, 본 연구에서는 하루 동안 생성된 검색 트랜잭션 로그를 분석하였다. 그러나 이용자들의 보다 현실적인 검색 성향 분석을 위해서는 장기간에 수집된 검색 트랜잭션 로그의 분석이 필요하다. 셋째, 3장에서 설명된 바와 같이 본 연구에서 사용된 세션 정의 방법에도 문제점이 있음을 발견하였으며, 따라서 새로운 세션 정의 방법의 개발에 대한 노력이 절실히 요구된다. 마지막으로, 검색 트랜잭션 로그 분석을 이용한 국외 연구와의 비교를 통하여 국내 웹 검색 행태의 특수성을 밝혀내는 것도 향후 과제라고 할 수 있다.

참 고 문 헌

곽철완. 2001. 인터넷 쇼핑몰의 상품 분류 체계에 대한 연구. 『정보관리학회지』, 18(4): 201-216

오경묵, 황상규, 이용현. 1999. 인터넷 이용자의 검색행동 성향에 관한 연구. 『한국문헌정보학회지』, 33(3): 87-108.

오삼균, 박희진. 2000. 국내 인터넷 탐색엔진에 대한 이용자 중심의 평가에 관한 연구 - 한글 알타비스타와 네이버를 중심으로. 『한국문헌정보학회지』, 34(2): 117-135.

이란주. 2000. 메타검색엔진의 특징에 관한 연구. 『정보관리학회지』, 17(2): 85-100.

이명희. 1998. 교육학 분야 주제전문가와 탐색전문가의 인터넷 검색엔진을 사용한 정보탐색 행태 비교연구. 『한국문헌정보학회지』, 32(3): 5-22.

이해영, 이수영. 2001. 인터넷 정보의 탐색, 평가 및 활용: 대학 이공계 연구자의 사례를 중심으로. 『정보관리학회지』, 18(4): 163-182.

정영미, 김성은. 1997. WWW 탐색도구의 색인 및 탐색 기능 평가에 관한 연구. 『한국문헌정보학회지』, 31(1): 153-184.

황상규, 오경묵, 변영태. 1999. 어휘의미 중의성이 인터넷 정보검색효율에 미치는 영향에 관한 연구. 『정보관리학회지』, 16(3): 65-82.

Abdulla, G., Liu, B., & Fox, E. 1998. "Searching the World-Wide Web: Implications from studying different user behavior." Paper presented at the World Conference of the World Wide Web, Internet, and

Intranet, Orlando, FL. <http://www.microsoft.com/usability/webconf.htm>

Ballard, T. 1997. "Comparative searching styles of patrons and staff." *Library Resources and Technical Services*, 38(3): 47-72.

Hoelscher, C. (1998). "How Internet experts search for information on the web." Paper presented at the World Conference of the World Wide Web, Internet, and Intranet. Orlando, FL.

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): 235-24.

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. 1998. "Real life information retrieval: A study of user queries on the web." *SIGIR Forum*, 33(1): 5-17.

Jansen, B. J., Spink, A., & 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.

Peters, T. A. et al. 1993. "The history and development of transaction log analysis." *Library Hi Tech*, 11: 41-66.

Peters, T.A. et al. 1996. "Web server logs as data sources for library and information science research." Paper presented at the Library Researdch seminar I, Talahassee, FL.

Ross, N.C.M. & 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., & Moricz, M. 1999. "Analysis of a very large web search engine query log." *SIGIR Forum*, 33(1): 6-12. http://www.acm.org/sigir/forum/F99/Silverstein.pdf

Spink, A., Jansen, B. J., & Ozmultu. 2000. "Query reformulation and relevance feedback by Excite users." *Internet Research: Electronic Networking Applications and Policies*, 10(4): 317-328.

Spink, A., Wolfram, D., Jansen, M.B.J., & 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.