질의 로그 분석을 통한 네이버 이용자의 검색 행태 연구

Information Seeking Behavior of the NAVER Users via Query Log Analysis

이준호(Joon-Ho Lee)*, 박소연(Soyeon Park)**, 권혁성(Hyuk-Sung Kwon)***

초 록

이용자와 검색 서비스 시스템의 모든 검색 과정을 기록한 질의 로그는 이용자의 실제 검색 행위를 사실적으로 반영한다. 따라서, 웹 검색 이용자들의 검색 행태를 이해하기 위하여 웹 검색 서비스 시스템이 생성한 질의 로그를 분석하는 방법이 널리 사용되고 있다. 본 연구는 네이버 이용자의 웹 검색 행태를 파악하기 위하여 기존의 질의 로그 분석 방법론을 보완하여 제시한다. 또한, 본 연구는 통합 검색, 디렉토리 검색, 웹 문서 검색과 같은 다양한 검색 유형에 대하여 일주일 동안 생성된 질의 로그를 분석함으로써, 네이버 웹 검색 이용자들의 전반적인 검색 행태를 파악하였다. 본 연구의 결과는 보다 효과적인 웹 검색 시스템 개발과 서비스 구축에 기여할 것으로 기대된다.

ABSTRACT

Query logs are online records that capture user interactions with information retrieval systems and all the search processes. Query log analysis offers an advantage of providing reasonable and unobtrusive means of collecting search information from a large number of users. In this paper, query logs of NAVER, a major Korean Internet search service, were analyzed to investigate the information seeking behavior of NAVER users. The query logs were collected over one week from various collections such as comprehensive search, directory search and web document search. 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, Query logs analysis, information seeking behavior, web search systems

* 숭실대학교 정보과학대학 컴퓨터학과 부교수(joonho@computing.ssu.ac.kr)
** 계명대학교 문헌정보학과 조교수(sypark@kmu.ac.kr)
*** 숭실대학교 정보과학대학 컴퓨터학과 석사과정(khscello@naver.com)

■ 논문 접수일 : 2003. 4. 14
■ 게재 확정일 : 2003. 6. 4

1 서 론

웹 검색 이용자들의 검색 형태 연구를 위해 기존에 많이 사용되었던 설문 조사 또는 인터뷰 자료를 분석하는 방법은 이용자의 실제 검색 행위와 다소 차이가 있는 결과를 생성할 수 있다. 이러한 문제를 해결하기 위하여, 국외에서는 웹 검색 서비스 시스템이 생성한 질의 로그를 분석하는 방법을 사용하고 있다. 이용자와 검색 서비스 시스템의 모든 검색 과정을 기록, 저장한 질의 로그는 이용자의 실제 검색 행위를 사실적으로 반영한다. 따라서 이러한 질의 로그의 분석은 웹 검색 이용자들의 검색 행태 연구를 위한 합리적이고 객관적인 방법으로 인정 받고 있다(Jansen and Pooch 2001). 질의 로그 분석을 통하여 웹 검색 이용자들의 검색 행태를 조사한 국외 연구들로는 익사이트 연구(Jansen, Spink, and Saracevic 2000; Spink et al. 2001; Spink et al. 2002), 파이어볼 연구(Hoelscher 1998), 알타비스타 연구(Silverstein et al. 1999) 등을 들 수 있다.

Jansen, Spink & Saracevic(2000)은 1997년 3월 9일 익사이트에서 생성된 51,473개의 질의를 분석하였다. 그리고 Spink et al.(2001)은 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. Hoelscher(1998)는 1998년 7월 한 달 동안 독일의 웹 검색 엔진인 파이어볼에서 생성된 약 1,600만개의 질의 로그를 분석하였다. Silverstein et al.(1999)은 1998년 8월 2일부터 9월 13일까지 6주간 알타비스타 이용자들이 남긴 2억 8천 5백 만개 이상의 이용자 세션, 9억 9천 만개 이상의 질의를 분석하였다. 이 연구는 지금까지 질의 로그를 분석한 연구 중 가장 장기간에 걸쳐 가장 방대한 자료를 연구 대상으로 했다는 점에서 의미가 있다.

질의 로그 분석을 이용한 이들 국외 연구들이 공통적으로 발견한 것은 웹 검색에 있어서 검색 방식의 단순성이다. 즉 웹 검색 이용자들은 복잡한 검색식이나 연산자를 사용하지 않고, 적은 수의 검색어로 구성된 단순한 질의를 통해 정보 검색을 수행하는 경향이 있었다(Spink et al. 2002; 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, 국내 웹 검색에 관한 연구는 전산학, 경영학, 문헌정보학, 심리학, 신문방송학 등 다양한 분야에서 수행되고 있다. 그리고, 이용자 검색 행태를 분석한 연구들은 주로 문헌정보학 분야에서 수행되어

왔으며, 이들 중 대부분의 연구들은 전공, 연령, 검색 경험에 있어서 비교적 동질적인 소수의 이용자들을 대상으로 이루어져 왔다. 즉, 오경묵, 황상규, 이용현(1999)는 전자과 대학원생 19명을 대상으로 수행한 실험을 통해 이용자들의 행동 양식을 조사하고, 4개의 검색 시스템을 비교하였으며, 이명희(1998)는 교육학 분야의 주제 전문가 10명을 주제 전문가와 검색 전문가로 나누어 이들의 검색 행태를 비교하였다. 또한, 오삼균, 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, 설문 조사 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.

국내에서도 질의 로그를 상용 로그 분석 도구를 이용하여 분석함으로써, 검색 서비스 시스템에 입력된 질의 수를 파악하는 작업은 많이 수행되고 있다. 그러나, 질의에 포함된 단어 수, 이용자가 검토하는 검색 결과 페이지 수 등과 같은 웹 검색 이용자들의 검색 행태를 분석하기 위하여 질의 로그 형식을 설계하고, 생성된 질의 로그를 분석하는 연구는 드문 실정이다. 한편 박소연, 이준호(2002)는 웹 검색 서비스 네이버 이용자들의 검색 행태를 조사하기 위하여 로그 형식을 직접 설계하고, 세션 정의 방법, 로그 정제 및 질의 유형 분류 방법, 검색어 정의 방법 등 한글 웹 검색 서비스 시스템의 질의 로그 분석에 필요한 방법론을 제시하였다. 또한 이들은 웹 검색 서비스 네이버에서 하루 동안 생성된 질의 로그 중 일부를 분석하여, 이용자들의 웹 문서 검색 행태를 조사하였다.

본 연구는 네이버 이용자의 웹 문서 검색 행태에 대한 박소연, 이준호의 2002년 연구(이하 2002년 연구)의 후속 연구로서 2002년 연구에서 이용한 질의 로그 분석 방법론을 보완하여 제시하고자 한다. 또한 본 연구는 질의 로그 수집의 기간을 하루에서 일주일로 연장하고, 수집 범위를 웹 문서 검색과 더불어 통합 검색, 디렉토리 검색과 같은 다양한 검색 유형들로 확대하여 네이버 웹 검색 이용자들의 전반적인 검색 행태를 파악하고자 하였으며, 특히 재검색 질의의 세부적인 분석을 통하여 이용자들의 질의 변경 행태를 조사한다.

2 연구 방법

본 연구에서는 웹 검색 이용자들의 검색 행태를 파악하기 위하여 질의 로그를 분석하였다. 분석 대상이 된 질의 로그는 2003년 1월 5일부터 11일까지 웹 검색 서

비스 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 매트릭스(http://www.internetmatrix.co.kr)에 의하면 네이버는 2003년 2월 1개월 동안 국내 네티즌들이 가장 많이 방문하는 사이트 중 3위를 차지하였으며, 네이버는 2002년 12월 한국인터넷기업협회와 한국기자협회가 선정한 올해의 인터넷기업 대상을 수상하였다. 따라서, 네이버 질의 로그를 분석함으로써 국내 네티즌들의 전반적인 웹 검색 행태를 추측할 수 있다.

네이버는 디렉토리 검색, 웹 문서 검색, 지식iN 검색, 일본 웹 검색, 백과사전 검색, 뉴스 검색, 이미지 검색 등을 개별적으로 지원하고 있으며, 또한 이들 검색 결과들을 통합하여 보여주는 통합 검색을 제공하고 있다. 본 연구에서는 이들 중에서 가장 큰 비중을 차지하는 통합 검색, 디렉토리 검색, 웹 문서 검색에 대한 질의 로그를 분석 대상으로 하였다. 네이버의 초기 페이지에서 기본 검색은 통합 검색으로 설정되어 있으며, 이러한 설정은 이용자가 풀다운 메뉴에서 웹 문서 검색이나 디렉토리 검색을 선택함으로써 웹 문서 검색과 디렉토리 검색으로 변경될 수 있다.

질의 로그는 온라인 정보 검색 시스템과 이를 사용하는 이용자의 상호 작용에 대한 전자적인 기록으로서, 질의 로그 분석은 정보 검색 분야에서 이용자의 검색 행태 연구를 위한 합리적이고 객관적인 방법으로 인정 받고 있다(Jansen and Pooch 2000). 질의 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 질의 로그 자료를 이용하고 있다(Peter 1993, 37).

질의 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 질의 로그 분석에 적용할 경우 문제점이 발생한다. 따라서 다음에서는 일반적인 질의 로그 분석에 필수적인 세션 정의 방법, 로그 정제, 질의 유형 분류, 그리고 한글 검색 질의 로그 분석에 필수적인 검색어 정의에 대하여 기술한다.

2.1 세션 정의 방법

일반적으로 세션은 한 명의 이용자가 단일한 정보 요구를 지니고 처음 검색을 시작하여 검색을 종료하기까지의 일련의 과정으로 정의된다. 선행 연구들은 세션에 관한 일반적인 정의에는 동의하나, 검색 질의들을 세션으로 구분함에 있어서 상이한 방법을 적용하고 있으며, 익사이트의 질의 로그를 분석한 Spink et al.(2001)의 연구, 알타비스타의 로그를 분석한 Silverstein et al.(1999)의 연구와 로이터, 알타비스타, 익사이트 로그를 분석한 He & Goker(2000)의 연구가 그 대표적인 예이다.

Spink et al.(2001)은 세션의 정의를 위하여 익사이트 서버가 할당한 이용자 식별자를 이용하였다. 즉, 임의의 컴퓨터가 브라우저를 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 브라우저들이 모두 종료될 때 소멸된다. 따라서 이 연구에서 세션은 일 초가 될 수도 있고, 몇 시간이 될 수도 있다. Spink et al.이 사용한 세션 설정 방법은 공공 장소에 위치한 하나의 컴퓨터를 다수의 사용자가 이용할 경우 또는 다수의 컴퓨터들이 하나의 프록시로 설정되어 있는 경우, 하나의 세션에 포함되는 질의 수가 과다해지는 문제점을 지니고 있다 .

He & Goker(2000)는 로이터와 익사이트 로그를 대상으로 세션의 시간 간격을 1분부터 50분까지 달리한 분석을 수행한 후, 10분에서 15분이 웹 검색 로그의 최적의 세션 간격이라고 제시하였다. 그러나 He & Goker의 세션 정의 방법은 비교적 소규모의 로그에만 사용되었고 대규모의 웹 트랜잭션 로그를 이용하여 검증되지 않았으며, 이들이 제시한 결론은 엄격한 통계 분석에 기초한 것이 아니므로 본 연구에 적용시키지 않았다.

한편 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.

2.2 로그 정제

질의 로그들로부터 이용자가 정상적으로 입력한 질의라고 판단하기 어려운 다수의 로그들을 발견하였다. 대부분의 국외 선행 연구들은 로그 정제 과정을 수행하지 않고, 이러한 비정상적인 질의들을 분석 대상에 포함시킴으로써 로그 분석 결과에 오류를 포함하고 있다. 예를 들어 Silverstein et al.(1999)은 질의와 검색어 분석 시 이러한 극단치(outliers)들을 포함시킨 결과 편차와 평균이 지나치게 과대 평가되는 문제점이 발생되었다.

비정상적인 질의 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 로그가 생성되며, 동일한 질의에 대한 로그가 연속해서 파일에 기록된다. 본 연구에서는 이용자가 검색 결과를 받지 않았다는 사실을 기록하도록 질의 로그를 설계함으로써, 이러한 질의에 대한 로그들을 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, 질의 로그들을 살펴보면 프로그램이 자동으로 입력한 질의라고 판단되는 다수의 로그들이 존재한다. 예를 들면, 하나의 세션에 수백 수천 개의 입력 질의, 문서 질의, 반복 질의가 포함된 경우를 발견할 수 있다. 본 연구에서는 질의 로그에 대한 수작업 검토와 더불어 세션 내에서 연속된 질의 쌍에 대하여 질의가 입력된 시간의 간격을 계산하고, 이러한 시간 간격들에 대한 통계 분석을 수행하였다. 그리고, 이러한 분석을 기반으로 프로그램에 의해 자동으로 생성되었을 가능성이 높다고 판단되는 세션들을 로그 파일로부터 제거하였다.

2.3 검색어 정의

검색어는 질의를 구성하는 기본 단위로서, 영어의 경우 빈칸, 마침표, 쉼표, 개행문자 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 문

제점이 발생한다. 즉, 정보검색시스템과 정보 검색 시스템은 동일한 질의임에도 불구하고, 서로 다른 색인어들을 포함하고 있는 것으로 인식된다. 따라서 한글 질의 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.

2.4 질의 유형 분류

질의는 세션의 구성 요소로서, 하나 이상의 검색어들로 구성된다. 본 연구에서는 전체 질의들을 최초 질의와 재검색 질의로 구분하고, 또한 재검색 질의를 다음과 같이 4가지 유형으로 분류하였다. 따라서 전체 질의 수는 최초 질의 수, 재검색 질의 수의 합과 동일하며, 재검색 질의 수는 입력 질의 수, 전환 질의 수, 문서 질의 수, 반복 질의 수의 합과 동일하다.

입력 질의: 이용자가 검색창에 직접 입력한 스트링으로서 검색어들로 구성된다. 하나의 세션에 포함된 재검색 입력 질의들은 검색어 추가 질의, 검색어 삭제 질의, 검색어 추가 및 삭제 질의, 이전 질의와 중복된 검색어를 포함하지 않는 변경 질의, 그리고 동일 질의로 구분될 수 있다. 또한, 동일 질의는 이전 질의에 포함된 검색어들만으로 구성된 질의를 의미하며, 띄어 쓰기가 변경된 질의, 검색어 순서가 변경된 질의, 그리고 불용어 제거 후 이전 질의와 동일한 검색어들을 포함하는 질의로 세분화될 수 있다. 여기에서 검색어는 형태소 단위로 인식되었다.

전환 질의: 전체 질의는 검색 유형별 분류에 의하여 통합 검색, 디렉토리 검색, 웹 문서 검색으로 구분될 수 있다. 네이버 서비스가 제공하는 검색 결과 화면에는 다른 검색 유형으로 전환할 수 있는 탭 버튼이 존재하며, 전환 질의는 사용자가 이러한 탭 버튼을 클릭한 경우에 생성된다. 예를 들면, 사용자가 초기 검색으로 디렉토리 검색을 수행하고, 그 결과 화면에서 웹 문서 탭 버튼을 클릭한 경우, 디렉토리 검색에서 웹 문서 검색으로의 전환 질의가 생성된다.

문서 질의: 네이버 서비스에서 웹 문서 검색을 수행한 후, 그 결과 화면에는 관련문서검색으로 표기된 링크가 존재하며, 문서 질의는 사용자가 이 링크를 클릭한 경우에 생성된다. 문서 질의는 검색어를 포함하지 않는 질의로서, 이용자가 지정한 웹 문서 전체가 질의로 사용된다.

반복 질의: 일반적으로 웹 검색 시스템들은 디렉토리와 웹 문서에 대한 검색을 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 로그 분석 결과

3.1 질의 수 분석

본 연구는 2003년 1월 5일부터 11일까지 웹 검색 서비스 네이버에서 생성된 질의 로그들 중 일부를 분석하였다. 분석 대상이 된 로그 파일에 포함된 전체 세션 수는 22,562,531개이고, 전체 질의 수는 41,221,606개이었다. 전체 질의들을 최초 질의와 재검색 질의로, 재검색 질의는 입력 질의, 전환 질의, 반복 질의, 문서 질의로 구분한 후, 각 유형별로 질의 수를 살펴보면 〈표 1〉과 같다. 즉, 전체 질의 중 최초 질의 수는 22,562,531개(55.38%)이었고, 재검색 질의 수는 18,183,642개(44.62%)이었다. 재검색 질의 중 입력 질의 수는 15,037,296개(82.7%), 반복 질의 수는 2,471,414개(13.59%), 전환 질의 수는 433,026개(2.38%), 문서 질의 수는 241,906개(1.33%)이었다. 전체 질의 중 최초 질의 수가 절반 이상이며, 재검색 질의 중 입력 질의가 가장 큰 비중을 차지함을 알 수 있다.

〈표 1〉 질의 유형별 질의 수 분석

구 분		질의 수	비율(%)
최초질의		22,562,531	55.37
재검색 질의	입력 질의	15,037,296	36.91
	전환 질의	433,026	1.06
	반복 질의	2,471,414	6.07
	문서 질의	241,906	0.59
총 계		40,746,173	100.00

〈표 2〉 최초 질의의 검색 유형별 질의 수 분석

구 분	질의 수	비율(%)
통합검색 (디렉토리 검색) (웹 문서 검색)	22,549,768 (9,770,489) (11,304,742)	99.94
디렉토리 검색	3,536	0.02
웹 문서 검색	9,227	0.04
총 계	22,562,531	100.00

〈표 3〉 전체 질의의 검색 유형별 질의 수 분석

구 분	질의 수	비율(%)
통합검색 (디렉토리 검색) (웹 문서 검색)	36,841,650 (15,408,275) (19,311,804)	90.42
디렉토리 검색	239,067	0.58
웹 문서 검색	3,665,456	9.00
총 계	40,746,173	100.00

최초 질의들을 검색 유형별로 구분해 보면 〈표 2〉와 같다. 전체 질의 중 통합검색은 22,549,768개(99.94%), 디렉토리 검색은 3,536개(0.02%), 웹 검색은 9,227개(0.04%) 이었다. 네이버 서비스의 초기 페이지에서 기본 검색은 통합 검색으로 설정되어 있으며, 최초 질의 중 통합 검색이 가장 많은 비중을 차지한다는 사실은 이용자들이 수동적이어서 시스템에 의해 설정된 환경을 변경하지 않거나, 또는 통합 검색의 결과에 가장 만족하기 때문일 것으로 추정된다. 〈표 3〉은 전체 질의들을 검색 유형별로 구분한 결과이며, 역시 통합 검색이 가장 큰 비중을 차지하고 있다. 그리고, 〈표 3〉은 통합 검색이나 웹 문서 검색에 비해 비교적 양질의 정보들

〈표 4〉 재검색 질의의 검색 유형별 질의 수 분석

구 분	질의 수			
	입력 질의	전환질의	반복 질의	문서 질의
통합 → 통합	13,981,286	0	0	0
통합 → 디렉토리	7,089	136,168	0	0
통합 → 웹 문서	17,112	164,486	803,211	100,243
디렉토리 → 통합	21,590	20,584	0	0
디렉토리 → 디렉토리	55,484	0	11,152	0
디렉토리 → 웹 문서	214	32,719	10	0
웹 문서 → 통합	211,794	56,628	0	0
웹 문서 → 디렉토리	3,197	22,441	0	0
웹 문서 → 웹 문서	739,530	0	1,657,041	141,663

을 선택하여 분류해 놓은 디렉토리 검색의 이용률이 상대적으로 낮음을 보여준다.

한편, 이용자가 통합 검색을 수행할 경우, 디렉토리 검색 결과에 10건 이상의 문서가 포함되면 웹 문서 검색을 수행하지 않으며, 또한 디렉토리 검색 및 웹 문서 검색에 의해 0건의 문서가 검색될 수 있다. 따라서 통합 검색 결과의 일부로서 디렉토리 검색 또는 웹 문서 검색 결과가 포함되지 않는 경우가 존재한다. 〈표 2〉, 〈표 3〉에서 괄호 안의 숫자는 통합 검색들 중에서 디렉토리 검색과 웹 문서 검색의 결과가 출력된 질의 수이다. 즉, 통합 검색들 중에서 디렉토리 검색과 웹 문서 검색의 결과가 출력된 비율이 최초 질의의 경우 각각 43.33%, 50.13%이며, 전체 질의의 경우 각각 41.82%, 52.42%임을 알 수 있다.

〈표 4〉는 재검색 질의를 검색 유형별로 분석한 결과를 보여 준다. 이용자가 질의를 다시 입력하는 입력 질의의 경우, 통합 검색이 절대적인 비중을 차지하고 있다. 질의를 변경하지 않고 탭 버튼을 클릭하여 검색하는 전환 질의들 중에서 동일한 검색 유형으로의 전환 질의는 이용자에게 이전 검색 결과와 동일한 검색 결과를 제공한다. 그러나, 〈표 4〉는 통합 검색에서 통합 검색으로의 전환이 다수를 차지하고 있으며, 디렉토리 검색에서 디렉토리 검색, 웹 검색에서 웹 검색으로의 전환도 많이 수행되었음을 보여준다. 이처럼 유용하지 않은 동일 검색 유형으로의 전환이 빈번하게 수행되는 이유는 이용자가 전환 질의의 성격을 정확히 이해하지 못하고 있기 때문인 것으로 추정된다.

〈표 5〉는 전제 질의의 세션당 질의 수

〈표 5〉 전체 질의의 세션당 질의 수에 대한 기술 통계

구 분	평 균	중간값	표준편차	최소값	최대값	질의 수
전체 질의	1.8	1	2.03	1	147	40,746,173

〈표 6〉 질의 유형별 세션당 질의 수에 대한 기술 통계

구 분		평 균	중간값	표준편차	최소값	최대값	질 의 수
최초 질의		1.00	1	0.00	1	1	22,562,531
재검색 질의	입력 질의	0.67	0	1.56	0	140	15,037,296
	전환 질의	0.02	0	0.18	0	28	433,026
	반복 질의	0.11	0	0.95	0	133	2,471,414
	문서 질의	0.01	0	0.25	0	103	241,906

〈표 7〉 재검색 질의 중 입력 질의의 유형별 세션당 질의 수에 대한 기술 통계

구 분	평 균	중간값	표준편차	최소값	최대값	질의 수
검색어 추가	0.066	0	0.30	0	28	1,500,401
검색어 삭제	0.039	0	0.22	0	23	888,761
검색어 추가 및 삭제	0.067	0	0.37	0	80	1,517,102
변경	0.478	0	1.19	0	134	10,791,403
동일(띄어 쓰기 변경)	0.006	0	0.09	0	20	131,411
동일(검색어 순서 변경)	0.001	0	0.03	0	4	14,529
동일(불용어 제거 후)	0.009	0	0.12	0	32	193,689

에 대한 기술 통계를 보여 준다. 이용자들은 세션당 평균 1.8개의 질의를 수행하였다. 〈표 6〉은 전체 질의를 질의 유형별로 구분한 후, 각 질의 유형별 세션당 질의 수에 대한 기술 통계를 보여 준다. 최초 질의는 세션 시작 질의이므로 세션당 평균이 1회이며, 최초 질의의 총계는 전체 세션 수와 동일하다. 재검색 질의들 중에서 관련문서검색으로 지칭되는 문서 질의와 전환 질의는 각각 세션당 평균 0.02회, 0.04회 수행되었으며, 이로부터 이용자들이 문서 질의와 전환 질의를 드물게 사용하고 있음을 알 수 있다. 반복 질의는 세션당 평균 0.11회 수행되었으며, 이는 이용자들이 평균적으로 출력하는 결과 화면이 약 1.11개 페이지임을 의미한다. 입력 질의와 반복 질의의 표준 편차가 비교적 큰 이유는 일부 이용자가 많은 수의 입력 질의 및 반복 질의를 수행하기 때문이다. 〈표 7〉에는 재검색 질의 중 입력 질의의 유형별 세션당 질의 수에 대한 기술 통계가 요약되어 있다. 이 표로부터 이용자는 검색어를 추가 또는 삭제하기보다는 이전 질의를 전적으로 변경하는 경우가 많음을 알 수 있다. 한편, 〈표 8〉은 검색 유형별 세션당 질의 수에 대한 기술 통계를 보여준다.

〈표 9〉는 전체 질의의 세션당 질의 수의 분포를 보여 주며, 〈표 10〉은 전체 질의를 질의 유형별로 구분한 후, 각 질의 유형별 세션당 질의 수의 분포를 보여 준다. 전체 세션의 83.52%가 2개 이하의 질

〈표 8〉 검색 유형별 세션당 질의 수에 대한 기술 통계

구 분	평 균	중간값	표준편차	최소값	최대값	질의 수
통 합	1.65	1	1.51	0	143	36,841,650
디렉토리	0.01	0	0.16	0	89	239,067
웹 문서	0.16	0	1.23	0	143	3,665,456

〈표 9〉 전체 질의의 세션당 질의 수 분포

	1개	2개	3개	4개	5개	6개 이상
전체 질의	15,163,729	3,681,717	1,581,043	796,304	447,179	892,559

〈표 10〉 질의 유형별 세션당 질의 수 분포

		1개	2개	3개	4개	5개	6개 이상
최초 질의		22,562,531	0	0	0	0	0
재검색 질의	입력 질의	3,726,846	1,550,794	734,567	383,547	219,799	394,001
	전환 질의	262,662	53,099	12,285	3,502	1,261	952
	반복 질의	360,157	158,579	86,298	54,285	35,205	109,903
	문서 질의	77,182	19,371	7,900	4,356	2,744	7,075

〈표 11〉 검색 유형별 세션당 질의 수 분포

	0개	1개	2개	3개	4개	5개 이상
통 합	11,099	15,707,812	3,715,739	1,500,044	707,055	920,782
디렉토리	22,398,753	105,611	38,429	10,291	4,740	4,707
웹 문서	21,240,146	398,551	493,164	121,450	79,950	234,270

의를 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. 한편, 〈표 11〉은 세션당 질의 수의 분포를 검색 유형별로 보여 준다.

〈표 12〉 검색 유형별 질의당 어절 단위 검색어 수에 대한 기술 통계

구 분		평 균	중간값	표준편차	최소값	최대값
최초 질의	통 합	1.09	1	0.38	1	71
	디렉토리	1.12	1	0.47	1	10
	웹 문서	1.17	1	0.62	1	16
재검색 입력 질의	통 합	1.18	1	0.55	1	49
	디렉토리	1.19	1	0.52	1	10
	웹 문서	1.50	1	0.93	1	28
최초 질의 + 재검색 입력 질의		1.13	1	0.48	1	71

〈표 13〉 검색 유형별 질의당 형태소 단위 검색어 수에 대한 기술 통계

구 분		평균	중간값	표준편차	최소값	최대값
최초 질의	통 합	1.83	1	2.13	1	930
	디렉토리	1.92	1	2.59	1	100
	웹 문서	2.21	2	3.93	1	210
재검색 입력 질의	통 합	2.26	2	2.92	1	665
	디렉토리	2.29	2	2.45	1	100
	웹 문서	3.38	2	5.10	1	280
최초 질의 + 재검색 입력 질의		2.03	2	2.56	1	930

3.2 검색어 수 분석

본 절에서는 최초 질의와 재검색 질의 중 입력 질의에 포함된 검색어 수에 대한 분석 결과를 기술한다. 〈표 12〉, 〈표 13〉은 검색 유형별 질의당 검색어 수에 대한 기술 통계를 각각 어절 단위와 형태소 단위로 보여 준다. 검색어가 어절 단위로 정의되는 경우 질의당 평균 검색어 수는 1.13개이었고, 검색어가 형태소 단위로 정의되는 경우 질의당 평균 검색어 수는 2.03개이었다. 이러한 결과를 통하여 웹 검색 이용자들은 매우 적은 수의 검색어로 구성된 단순한 질의를 수행하는 경향이 있음을 알 수 있다. 또한, 〈표 12〉, 〈표 13〉은 이용자가 다른 검색 유형에 비하여 웹 문서 검색을 위해 보다 많은 수의 검색어를 입력하는 경향이 있음을 보여준다.

한편 〈표 14〉, 〈표 15〉는 검색 유형별 질의당 검색어 수의 분포를 어절 단위와 형태소 단위로 보여준다. 검색어가 어절 단위로 정의되는 경우 하나의 검색어만을 포함하는 질의가 90.42%이었으며, 검색어가 형태소 단위로 정의되는 경우 하나의 검색어만을 포함하는 질의가 47.67%이었다. 따라서 질의에 대한 형태소 분석의

〈표 14〉 검색 유형별 질의당 어절 단위 검색어 수 분포

구 분		1개	2개	3개	4개	5개	6개 이상
최초 질의	통합	21,021,711	1,206,047	239,493	53,041	16,584	12,892
	디렉토리	3,238	225	47	18	5	3
	웹 문서	8,210	676	232	67	21	21
재검색 입력 질의	통합	12,392,989	1,340,589	345,206	87,038	27,900	20,948
	디렉토리	55,987	7,712	1,606	325	91	49
	웹 문서	517,033	153,293	57,669	17,750	6,363	4,748

〈표 15〉 검색 유형별 질의당 형태소 단위 검색어 수 분포

구 분		1개	2개	3개	4개	5개	6개 이상
최초 질의	통합	11,332,779	8,625,673	985,866	928,405	25,764	651,281
	디렉토리	1,871	1,196	159	169	5	136
	웹 문서	4,420	3,320	417	503	34	533
재검색 입력 질의	통합	6,319,872	5,074,650	880,147	978,034	38,556	923,411
	디렉토리	29,095	21,340	4,790	5,594	312	4,639
	웹 문서	236,224	201,809	64,133	104,096	5,314	145,280

수행 여부와 관계없이 하나의 검색어로 구성된 질의를 수행하는 검색 서버의 별도 구축이 바람직함을 알 수 있다.

4 결론 및 제언

본 연구에서는 웹 검색 이용자들의 전반적인 검색 행태를 이해하기 위하여 국내에서 널리 사용되고 있는 웹 검색 서비스 네이버에서 생성된 일주일간의 질의 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 질의 로그 분석에 필요한 세션 정의 방법을 설명하고, 로그 정제 및 질의 유형 분류 방법을 제시하였으며, 한글 질의 로그 분석에 필수적인 검색어 정의 방법을 기술하였다.

본 연구의 수행 결과, 향후 연구가 요구되는 다음과 같은 사항들을 발견하였다. 첫째, 향후 연구에서는 네이버가 제공하는 다른 기능의 이용에 대한 추가적인 분석이 가능할 것이다. 곧 불리안 연산, 자연어 검색, 인접 연산, 부정어 제외 기능 등과 같은 고급 검색 기능의 이용에

대한 분석, 이용자가 실제로 클릭하여 선택한 문서들에 대한 분석, 본 연구에 포함되지 않은 지식iN 검색, 일본 웹 검색, 뉴스 검색, 이미지 검색 등에 대한 분석을 통하여 이용자들의 보다 다양한 검색 행태를 파악할 수 있을 것이다. 둘째, 곧 본 연구에서는 일주일 동안 생성된 질의 로그를 분석하였다. 그러나 이용자들의 보다 현실적인 검색 성향 분석을 위해서는 장기간에 수집된 질의 로그의 분석이 요구된다. 보다 장기간의 질의 로그를 연구 대상으로 할 경우, 시간대별 이용자들의 검색 행태 비교나, 요일별 검색 행태 비교, 주중과 주말의 검색 행태 비교, 그리고 이용자들이 검색하는 주제의 변화 등 다양한 연구의 수행이 가능할 것이다. 셋째, 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 등을 통한 별도의 보완 작업이 요구된다. 넷째, 본 연구에서 사용된 세션 정의 방법, 로그 정제 방법에도 약간의 한계가 있으므로, 웹 로그 분석 방법론에 대한 계속적인 수정, 보완 작업이 필요할 것이다. 마지막으로, 질의 로그 분석을 이용한 국외 연구와의 비교를 통하여 국내 웹 검색 행태의 특수성을 밝혀내는 것도 향후 과제라고 할 수 있다.

참 고 문 헌

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

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

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

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

Abdulla, G., B. Liu, and E. Fox. 1998. "Searching the World-Wide Web: Implications from studying different user behavior." *The World Conference of the World Wide Web, Internet, and Intranet*, Orlando, FL.

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

He, D., and A. Goker. 2000. "Detecting session boundaries from Web

user logs." *22nd Annual Colloquium of IR research*, Cambridge, UK.

Jansen, B.J., and U. Pooch. 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.

Hoelscher, C. 1998. "How Internet experts serach for information on the web." *The World Conference of the World Wide Web, Internet, and Intranet*, Orlando, FL.

Jansen, B. J., A. Spink, and T. Saracevic. 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. 1993. "The history and development of transaction log analysis." *Library Hi Tech*, 11(2): 41-66.

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

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