클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석

An Analysis of Query Types and Topics Submitted to Naver

박 소 연(Soyeon Park)*
이 준 호(Joon-Ho Lee)**
김 지 승(Ji Seoung Kim)***

목 차

초 록

웹 검색 분야의 대부분의 선행 연구들은 검색 질의를 살펴본 연구자의 판단에 근거하여 질의의 주제를 분석하였다. 그러나 웹 검색 질의의 주제 분야가 방대하고 다양하여서 이용자가 검색 결과에서 실제로 조회한 문서를 모르는 상태에서 연구자의 판단에 근거하여 질의의 주제를 분류하기에는 한계가 있다. 이에 본 연구에서는 1년 동안 네이버 이용자들이 입력한 질의를 기록한 질의로그와 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. 또한 이용자들이 전반적으로 가장 많이 검색한 주제는 컴퓨터/인터넷, 엔터테인먼트, 쇼핑, 게임, 교육 순으로 나타났다. 본 연구의 결과는 인터넷 포탈 업체들의 효과적인 컨텐츠 구축 및 효율적인 검색 시스템 개발에 기여할 것으로 기대된다.

ABSTRACT

This study examines web query types and topics submitted to Naver during one year period by analyzing query logs and click logs. Query logs capture queries users submitted to the system, and click logs consist of documents users clicked and viewed. This study presents a methodology to classify query types and topics. A method for click log analysis is also suggested. When classified by query types, there are more site search queries than content search queries. Queries about computer/internet, entertainment, shopping, game, education rank hightest. The implications for system designers and web content providers are discussed.

키워드: 웹 검색, 질의 형태, 질의 주제, 클릭 로그 분석
Web Search, Query Type, Query Topic, Click Log Analysis

---

* 덕성여자대학교 문헌정보학과 조교수(sypark@duksung.ac.kr), 교신 저자
** 숭실대학교 정보과학대학 컴퓨터학부 부교수(joonho@computing.ssu.ac.kr)
*** 숭실대학교 정보과학대학 컴퓨터학과 박사과정(jskim89@naver.com)
논문접수일자 2005년 2월 15일
게재확정일자 2005년 3월 16일

1. 연구 목적

1990년대 중반 이후 인터넷 사용과 보급의 폭발적 증가는 인터넷을 통한 정보의 접근을 지원하기 위한 웹 검색 서비스들을 활성화시켰다. 이에 따라 국내외 웹 검색에 관한 연구들이 여러 학문 분야에서 다양한 연구 방법을 이용하여 수행되고 있다. 이들 중 장기간에 걸쳐 수집된 웹 검색 질의의 전반적인 형태 및 주제에 대한 분석은 웹 검색 분야에서 매우 중요한 연구 주제라고 할 수 있다. 특히 이러한 연구의 결과는 인터넷 포탈 업체들의 효과적인 컨텐츠 구축 및 효율적인 검색 시스템 개발에 있어서 중요한 자료로서 활용될 수 있다.

한편, 이용자와 검색 시스템 사이의 모든 상호 작용을 기록한 검색 트랜잭션 로그는 이용자의 실제 검색 행위를 사실적으로 반영한다. 따라서 이러한 로그의 분석은 웹 검색 질의의 형태 및 주제 분석을 위한 적절한 연구 방법으로 판단된다. 지금까지 로그 분석을 이용한 대다수 선행 연구들은 하루나 일주일 동안의 단기간에 걸쳐 생성된 로그를 분석하였다. 그러나 이용자들의 보다 현실적이고 지속적인 검색 성향 분석을 위해서는 장기간에 걸쳐 수집된 트랜잭션 로그에 대한 분석이 요구된다.

또한, 질의 로그 분석을 이용한 대다수 선행 연구들은 연구자가 질의를 살펴보고, 그 질의의 주제를 분류하였다(Jansen, Spink, and Saracevic, 2000; Ross and Wolfram, 2000; Spink, et al., 2002; Jansen and Spink, 2005). 그러나 일반적으로 사용자들은 광범위한 정보 요구를 지니고 검색 시스템에 접근하여 그 요구를 질의로서 표현하기 때문에, 연구자가 질의만을 살펴본 후 그 질의의 주제를 분류하기에는 한계가 있다. 따라서 검색 질의의 형태 및 주제를 분석하기 위해서는 그 질의가 내포하는 정보 요구의 파악이 선행되어야 한다.

이에 본 연구에서는 1년 동안 네이버에 입력된 검색 질의들의 표본과 각각의 질의에 대하여 이용자가 조회한 문서들에 근거하여 국내 웹 검색 질의의 형태 및 주제를 분석하고자 한다. 국내외 선행 연구들 중 본 연구에서처럼 장기간에 걸쳐 웹 검색 포탈의 검색 질의를 수집한 경우는 드문 실정이며, 또한 이용자가 실제로 조회한 자료, 즉 클릭 로그에 근거하여 검색 질의가 내포하는 정보 요구를 파악한 사례를 발견하기 어렵다. 본 연구는 클릭 로그 분석이라는 새로운 연구 방법을 도입하였다는 점에서 웹 검색 분야에 학문적으로 기여하는 바가 클 것으로 기대된다.

2. 선행 연구

웹 검색 분야에서 트랜잭션 로그를 활용한 대다수 국내외 선행 연구들은 하루나 일주일 동안의 단기간에 걸쳐 생성된 로그를 분석하였으며, 장기간에 걸쳐 생성된 로그를 분석한 연구는 드문 실정이다. Silverstein et al.(1999)은 1998년 8월 2일부터 9월 13일까지 6주 동안 알타비스타 이용자들이 남긴 2억 8천 5백만 개 이상의 이용자 세션, 9억 9천만 개 이상의 질의를 분석하였다. 이 연구는 지금까지 트랜잭션 로그를 이용한 연구들 중 가장 방대한 자료를 대상으로 하였고, 세션 정의 방법 등과 같은 로그 분석 방법을 제시하였다는데 의의가

있다. 국내에서는 박소연, 이준호(2002)와 이준호, 권혁성, 박소연(2003)이 하루와 일주일 동안 생성된 대규모 트랜잭션 로그에 근거하여, 네이버 이용자의 검색 행태를 분석하였다. 그러나 이러한 연구들은 검색 질의의 주제에 대한 분석을 수행하지 않았다. 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. 그러나 이 연구는 이용자 계층이 제한된 대학교 웹 사이트에 입력된 질의만을 분석하였기 때문에, 이 연구의 결과를 일반 웹 이용자들의 검색 행태로 일반화하기에는 한계가 있다.

한편, 검색 트랜잭션 로그를 분석한 많은 국외 선행 연구들은 검색 질의를 살펴본 연구자의 판단에 근거하여 질의의 주제를 분류하였다. Spink et al.(2001)은 1997년 9월 16일 익사이트 이용자들이 남긴 약 100만개의 질의들로부터 2,414개를 무작위로 추출한 후, 이들을 엔터테인먼트, 성/성인, 상업/여행/고용/경제, 컴퓨터/인터넷, 건강/과학, 사람/장소/사물, 사회/문화/인종/종교, 교육/인문학, 예술, 정부, 불분명과 같은 11개 범주들로 구성된 분류 체계를 도출하였다.

이후 Spink et al.(2001)이 개발한 분류 체계에 근거하여 이용자 질의의 주제를 분류한 연구들이 다음과 같이 수행되었다. Spink et al.(2002)은 1997년부터 2년에 한번씩 하루를 선정하고, 그날 익사이트에 입력된 질의들 중 무작위로 추출된 약 2,500개를 분류하였다. 그 결과 이용자들이 주로 검색하는 주제가 엔터테인먼트와 성 관련으로부터 전자 상거래 관련으로 변화하였음을 보고하였다. Jansen, Spink, Pedersen(in press)은 2002년 9월 8일 알타비스타에서 생성된 약 100만개의 질의들로부터 2,603개를 무작위로 추출한 후 이들의 주제를 분류하였다. 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. 그리고 Jansen과 Spink(in press)는 이러한 연구들을 비교하고 요약하였다.

Ross와 Wolfram(2000)은 익사이트에서 하루 동안 생성된 질의들 중 2개 이상의 검색어로 구성된 질의들을 추출하고, 이들 중 가장 많이 출현하는 1,054개 검색어 쌍들의 유형을 조사하였다. 그 결과 검색어 쌍들을 성, 집단, 장소, 그림, 기관, 교육, 무료, 무역, 컴퓨팅, 인물, 웹/네트워크, 직업/경영, 멀티미디어, 음악, 참고, 커뮤니케이션, 뉴스, 출판물, 정부/법, 게임, 스포츠, 여행, TV/영화, 시각예술, 건강/의학, 역사, 이야기, 동물, 과학, 게임/복권 등 30개의 주제 영역으로 분류하였다. 이들은 검색어 쌍들로부터 주제 범주를 도출하는 귀납적 방법을 사용하였으며, 질의를 보고 주제를 이해할 수 없는 경우에는 익사이트에 질의를 직접 입력하여 출력된 결과물에 근거해 주제를 분류하였다.

이처럼 많은 국외 선행 연구들은 연구자가 질의만을 살펴본 후 주제를 분류하거나, 연구자가 질의로부터 사용자의 정보 요구를 파악하기 어려운 경우, 검색 시스템에 질의를 입력한 후 그 결과에 근거하여 주제를 분류하였다. 즉, 이용자가 검색 결과로 출력된 문서들 중 실제

로 조회한 문서를 모르는 상태에서 연구자들의 판단에 근거하여 질의의 주제를 분류하였다. Spink et al.(2002)은 이용자가 실제로 조회한 문서를 알려주는 클릭 로그를 확보하지 못하였기 때문에 질의만으로 주제를 분류하였음을 언급하였으며, 이는 클릭 로그를 이용한 정보 요구의 파악이 합리적 수단으로 인식되고 있음을 시사한다. 이에 본 연구에서는 1년이라는 장기간에 걸쳐 네이버에 입력된 검색 질의들의 표본과 각 질의에 대한 클릭 로그에 근거하여 국내 웹 검색 질의의 형태 및 주제를 분석하고자 한다.

3. 연구 방법

본 연구에서 네이버를 선택한 이유는 네이버가 대중성이나 인지도면에서 국내 주요 검색 포탈로 인정받고 있기 때문이다. 즉, 네이버는 한국생산성 본부에서 실시한 2003년 1/4분기 국가고객만족도(NCSI)(http://www.ncsi.or.kr) 조사에서 검색 포탈 서비스 부분 1위를 차지하였다. 또한 한국표준협회에서 실시한 한국서비스품질지수(http://www.servqual.or.kr) 인터넷 검색 포탈 부문에서 2003년과 2004년 모두 1위를 차지하였고, 한국경제신문과 브랜드스톡에서 주관한 2004 대한민국 100대 브랜드 중 인터넷 포탈사이트 부문에서 1위를 차지하였다(http://www.brandstock.co.kr/brand/report/brand100). 네이버는 디렉토리 검색, 웹문서 검색, 백과사전 검색, 지식인 검색, 뉴스 검색, 이미지 검색 등을 개별적으로 지원하고 있고, 이들 검색 결과들을 통합하여 보여주는 통합 검색을 제공하고 있으며, 이러한 검색 서비스들 중 통합 검색이 가장 큰 비중을 차지하고 있다.

본 연구에서는 2003년 7월 1일부터 2004년 6월 30일까지 1년 동안 네이버 이용자들이 통합 검색창에 입력한 질의를 기록한 질의 로그와 각각의 질의에 대하여 이용자가 조회한 문서를 기록한 클릭 로그를 분석 대상으로 하였다. 이용자들의 검색 행태가 주중과 주말, 평일과 공휴일 간에 변화할 수 있고, 요일별로 변화할 수 있다는 사실을 염두에 두고, 1년 동안의 주중, 주말, 평일, 공휴일의 분포에 맞추어 격주로 표본 날짜를 선택하였다. 이렇게 선택된 날짜의 질의 로그들로부터 하루에 700개씩의 질의를 무작위로 선정하였다. 2003년 하반기의 경우 하루 동안 네이버에 입력되는 통합 검색 질의는 대략 1,000만개 이상으로 추정된다. 이러한 모집단의 규모를 감안할 때 표본 오차 95% 신뢰 수준 ±4%포인트와 ±5%포인트를 허용할 경우, 필요한 표본의 크기는 각각 600개와 384개로 통계학 관련 문헌에서 제시되고 있다(Arkin and Colton, 1963). 본 연구에서는 이러한 요소를 고려하여 하루에 700개의 질의를 무작위로 선정할 것을 결정하였다. 그리고 총 18,200개 질의의 형태 및 주제를 파악하기 위하여, 연구자들과 문헌정보학 전공 졸업생, 재학생 3명이 공동으로 클릭 로그를 분석하였다.

질의의 형태를 분석한 결과, 사이트 검색, 내용 검색, 사이트와 내용 동시 검색이라는 범주가 도출되었다. 사이트 검색은 이용자가 찾고자 하는 대상이 웹 사이트인 경우로서, "단일 사이트 검색"과 "다수 사이트 검색"으로 세분

화될 수 있다. 단일 사이트 검색은 "네이버", "다음" 등의 질의를 입력한 후 검색 결과로서 노출된 이들 사이트의 URL을 클릭한 경우이며, 다수 사이트 검색은 "병원", "꽃배달" 등의 질의를 입력한 후 검색 결과로부터 다수의 사이트 URL들을 클릭한 경우이다. 그리고 내용 검색은 특정한 주제에 관한 신문 기사, 게시판 글, "지식인"에 올라간 글들을 클릭한 경우이다. 또한 사이트 검색과 내용 검색을 동시에 수행하는 경우를 별도의 범주로 분리하였다. 따라서 질의를 형태에 따라 분류하면 다음과 같다.

- 사이트 검색
  - 단일 사이트 검색
  - 다수 사이트 검색
- 내용 검색
- 사이트 검색 + 내용 검색
  - 단일 사이트 검색 + 내용 검색
  - 다수 사이트 검색 + 내용 검색

한편, 듀이 십진 분류법, 미의회 도서관 분류법, 한국 십진 분류법 등과 같은 전통적인 분류 체계는 학문 분류를 위해 개발되었기 때문에, 이들을 실용적이고 유동적인 성격이 강한 웹 검색 질의들의 분류에 적용하기에는 한계가 있다. 즉, 웹 이용자들은 엔터테인먼트, 컴퓨터, 인터넷, 게임 등과 관련된 질의들을 가장 많이 입력하는 것으로 알려져 있으며(Cacheda & Vinã, 2001; Ross & Wolfram, 2000; Silverstein et al., 1999; Spink et al., 2001; Spink et al., 2002), 이러한 질의들을 전통적인 분류 체계에 의해 분류하기에는 어려움이 있다. 웹 검색 분야의 해외 선행 연구들은 웹 검색 질의의 주제를 분류하기 위해 귀납적인 내용 분석 방법을 사용하였다. 즉, 미리 정해진 전통적 분류 체계를 적용하기보다는 이용자들이 입력한 질의들의 주제를 분석한 후, 그 결과에 근거하여 분류 체계를 도출하였다.

따라서 본 연구에서도 네이버에 입력된 질의들의 주제에 근거하여 분류 체계를 도출하였으며, 이때 해외 선행 연구들이 개발한 분류 체계와 네이버, 야후(한국, 미국), 구글(한국, 미국), 엠파스와 같은 국내외 주요 웹 검색 디렉토리 서비스의 대분류 및 중분류 항목을 참고하였다. 또한 전체 표본에서 3% 이상을 차지하는 주제 범주만을 분류 체계에 포함시키는 것을 원칙으로 하였으며, 3% 미만이지만 주제의 성격상 다른 어떤 주제 범주에도 포함되기 어려운 "성인," "건강," "과학" 등을 별도의 주제로 독립시켰다. 그 결과 다음과 같은 전체 16개의 주제 범주가 도출되었다.

- 건강
- 게임
- 과학
- 교육/학문(교육기관 포함)
- 금융/경제
- 기관(정부기관, 사회단체)
- 기업
- 뉴스/미디어
- 라이프스타일(생활정보, 레져, 스포츠, 취미, 요리, 미용, 애견, 교통정보 등)
- 문화/예술
- 사회(정치, 법, 행정, 종교)
- 성인
- 쇼핑

- 엔터테인먼트
- 지역/여행(지역정보, 숙박시설, 세계정보)
- 컴퓨터/인터넷

이 주제 범주들 중 가장 큰 비중을 차지하는 엔터테인먼트, 컴퓨터/인터넷, 쇼핑 등은 다음과 같은 하위 범주로 세분화되었다.

- 쇼핑
  - 가격비교, 기념일쇼핑, 단일사이트, 상품정보, 특정상품명, 일반상품명
- 엔터테인먼트
  - 드라마, 방송, 사진, 애니메이션, 연예인, 영화, 운세, 유머, 음악, 일반
- 컴퓨터/인터넷
  - 다운로드, 커뮤니티, 포탈사이트, 일반

쇼핑의 하위 범주 중 "가격비교"는 "에누리," "베스트바이어" 등의 질의를 입력한 후 검색 결과로부터 노출된 이러한 가격 비교 사이트들의 URL들을 클릭하는 경우이며, "기념일 쇼핑"은 "발렌타인 데이," "크리스마스"와 같은 질의를 입력한 후 검색 결과에서 출력된 다수의 쇼핑 사이트를 클릭하는 경우이다. "단일사이트"는 "lgeshop," "cjmall," "신세계몰" 등의 질의를 입력한 후 이들 사이트의 URL들을 클릭하는 경우이며, "상품정보"는 상품에 대한 품평이나 상세한 정보가 수록된 신문 기사, 게시판 글들을 클릭한 경우이다. "특정상품명"은 "버버리," "나이키"와 같은 질의를 입력한 후 검색 결과를 클릭하는 경우이며, "일반상품명"은 "옷," "핸드폰," "꽃배달" 등의 질의를 입력한 후 검색 결과를 클릭하는 경우이다.

본 연구에서는 이처럼 도출된 형태 및 주제 범주로의 분류 작업에 대한 상세한 가이드라인을 작성하였으며, 문헌정보학과 졸업생과 재학생으로 구성된 세 명의 평가자들은 이 가이드라인에 따라 질의를 수작업으로 분류하였다. 평가자들은 한 달 이상 연구자들로부터 질의 분류에 관한 교육을 받고 실습을 수행하였으며, 철저히 클릭 로그에 근거하여 분류 작업을 수행하였기 때문에, 분류에 있어 평가자들의 주관이 개입할 여지는 매우 적다고 할 수 있다. 평가자들 사이의 분류 일치성은 평균 약 97%로 매우 높은 것으로 나타났으며, 분류가 불일치하는 경우 클릭 로그의 재검토와 토론을 통하여 합의에 이르는 과정을 거쳤다.

한편 기존의 분석 방법과 클릭 로그 분석 방법을 비교하기 위하여 본 연구에서 분석된 질의 로그들 중 하루의 질의 로그에서 선택된 700개 질의에 대해 문헌정보학과 졸업생과 재학생 세 명으로 하여금 질의만을 살펴본 후 주제를 분류하게 하였다. 분석 결과 평가자들이 평균적으로 전체의 42%의 질의에 대해 주제를 파악하지 못하였고, 18%의 질의에 대해서는 클릭 로그를 참고로 하였을 때와는 다른 주제를 선택하였다. 또한 기존의 방법을 이용한 선행 연구에서 주제가 불분명한 질의의 비중이 클릭 로그 분석을 이용한 본 연구에서보다 높은 것으로 나타났다. 곧 Ross와 Wolfram(2000)의 연구에서 주제가 불분명한 검색어 쌍은 전체 검색어 쌍의 약 8%이었고, Spink et al.(2001)의 연구에서 주제가 불분명한 질의는 전체 질의의 4.1%이었고, Spink et al.(2002)의 연구에서 1999년과 2001년 익사이트 데이터의 경우 주

제가 불분명하거나 비영어로 입력된 질의가 각각 9.3%, 11.3%로 나타났다. 반면 본 연구에서 주제를 파악하기 어려운 질의와 3개 이상의 복수 주제를 포함하는 질의는 모두 합하여 전체 질의의 3.6%인 것으로 나타났다. 이러한 사실은 클릭 로그 분석 방법이 이용자의 정보 요구를 파악하는 데에 있어서 불확실성을 감소시키며 보다 정확한 주제 분석을 위한 도구로 활용될 수 있음을 시사한다.

4. 연구 결과

4. 1 형태 분석

본 연구에 포함된 18,200개 질의의 클릭 로그 상태를 분석한 결과, 클릭 로그가 존재하는 질의가 15,645개(86%)로 나타났다. 즉, 이용자가 질의 입력 후 검색 결과로서 화면에 노출된 문서들 중 1개 이상을 클릭한 질의의 수가 전체의 86%이였다. 그리고 클릭 로그가 존재하지 않는 질의는 2,555개(14%)이였으며, 이들은 다음과 같은 두 가지 범주로 구분될 수 있다.

첫째, 네이버로부터 검색 결과가 출력됨에도 불구하고, 클릭 로그가 생성되지 않는 경우이며, 이는 전체 질의의 12.3%인 2,231개로 나타났다. 이에 대한 이유로는 다음과 같은 사항들을 고려할 수 있다: (i) 클릭 로그 시스템의 오류로 인하여 클릭 로그가 기록되지 않을 수 있다. (ii) 검색 결과로부터 이용자가 원하는 문서를 발견하지 못 할 경우, 이용자는 검색 결과로 출력된 문서를 조회하지 않을 수 있다. (iii) 이용자가 기업명을 질의로 입력할 경우, 주가 등과 같은 그 기업에 관한 기본적인 정보를 검색 결과 화면으로부터 즉시 얻을 수 있으며, 또한 연예인명을 질의로 입력할 경우, 그 연예인에 대한 신상 정보, 사진 등과 같은 간단한 정보를 검색 결과 화면으로부터 얻을 수 있기 때문에, 이용자는 검색 결과로 출력된 문서를 조회하지 않을 수 있다. 본 연구에서는 이들 중 출력된 검색 결과에 근거하여 형태를 분류할 수 없는 20개 질의를 제외한 2,211개 질의의 형태를 분류하였다.

둘째, 검색 결과에 포함된 문서가 0건이기 때문에 클릭 로그가 존재할 수 없는 경우이며, 이는 전체 질의의 1.7%인 324개로 나타났다. 이들 중 0.8%인 152개 질의에 대해서는 "t스캔들," "목동우체군"의 예처럼 질의의 형태 및 주제를 분명히 파악할 수 있었으며, 0.9%인 172개 질의에 대해서는 그 형태와 주제를 파악할 수 없었다.

따라서 본 연구에서는 18,200개 질의 중 192개를 제외한 18,008 질의의 형태를 분석하였으며, 그 결과는 〈표 1〉과 같다. 단일 사이트와 다수 사이트를 포함하여 사이트를 검색하는 경우가 거의 절반에 가까운 48.9%, 내용을 검색하는 경우는 42.6%, 내용과 사이트를 동시에 검색하는 경우는 7.5%로 나타났다.

4. 2 주제 분석

본 연구에 포함된 18,200개 질의 중 형태가 불분명한 192개 질의에 대해서는 그 주제도 파악이 불가능하였다. 또한, 699개(3.6%) 질의는 오타 등으로 인하여 주제를 파악하기 어려운 경우, 질의가 3개 이상의 주제를 포괄하여

〈표 1〉 질의의 형태 분석

형태	빈도	비율
사이트 검색	8,893	48.9%
단일 사이트	6,299	34.6%
다수 사이트	2,594	14.3%
내용 검색	7,758	42.6%
사이트와 내용 동시 검색	1,357	7.5%
단일 사이트와 내용	496	2.7%
다수 사이트와 내용	861	4.7%
형태 불분명	192	1.1%
총계	18,200	100.0%

특정한 주제로 분류하기 어려운 경우 등으로 나타났다. 따라서 본 연구에서는 18,200개 질의 중 주제를 파악하기 어려운 891개(4.7%)를 제외한 17,309개 질의의 주제를 1년 동안 생성된 네이버 클릭 로그에 근거하여 수작업으로 분류하였으며, 그 결과는 〈표 2〉와 같다. 이 결과는 17,309개 질의를 중복 분류한 것이므로, 〈표 2〉의 총계는 주제 분류가 가능한 질의의 총합인 17,309개보다 약간 많다. 네이버 이용자가 1년 동안 가장 많이 검색한 주제는 컴퓨터/인터넷(15.7%), 엔터테인먼트(15.6%), 쇼핑(9.4%), 게임(9.3%), 교육(8.7%), 기업(7.5%), 라이프스타일(7.0%), 금융/경제(6.0%) 순으로 나타났다. 인터넷 초창기에 많이 검색하였던 주제 분야인 "성인" 관련 질의는 약 1%에 불과해 최하위에 머물렀다.

〈표 3〉은 상위 3개 주제인 컴퓨터/인터넷, 엔터테인먼트, 쇼핑의 하위 주제의 순위를 보

〈표 2〉 네이버 이용자가 많이 검색한 주제 순위

주제	빈도	비율
컴퓨터/인터넷	2,797	15.7%
엔터테인먼트	2,776	15.6%
쇼핑	1,682	9.4%
게임	1,666	9.3%
교육/학문	1,550	8.7%
기업	1,340	7.5%
라이프스타일	1,241	7.0%
금융/경제	1,131	6.3%
기관	627	3.5%
사회	568	3.2%
문화/예술	541	3.0%
지역/문화	484	2.7%
뉴스/미디어	479	2.7%
건강	390	2.2%
과학	366	2.1%
성인	209	1.2%
총계	17,847	100.0%

〈표 3〉 컴퓨터/인터넷, 엔터테인먼트, 쇼핑의 하위 주제 순위

상위 주제	하위 주제	빈도	비율
컴퓨터/인터넷	포털사이트	1,074	6.0%
	일반	1,038	5.8%
	커뮤니티	524	2.9%
	다운로드	160	0.89%
엔터테인먼트	연예인	829	4.6%
	음악	568	3.2%
	영화	532	3.0%
	일반	186	1.0%
	애니메이션	182	1.0%
	유머	151	0.8%
	드라마	112	0.6%
	방송	107	0.6%
	사진	69	0.4%
	운세	41	0.2%
쇼핑	일반상품명	658	3.7%
	단일 사이트	470	2.6%
	특정 상품명	274	1.5%
	상품정보	233	1.3%
	가격비교	35	0.2%
	기념일 쇼핑	12	0.06%

여준다. 컴퓨터/인터넷 관련 질의의 하위 주제를 살펴보면, 인터넷 포털 사이트를 검색하는 경우가 가장 많았으며(6%), 컴퓨터/인터넷과 관련된 일반적인 질의(5.8%), 커뮤니티 관련 질의(2.9%)도 많은 편이었다. 엔터테인먼트 관련 질의의 하위 주제의 경우, 연예인 관련 질의가 가장 많았으며(4.6%), 음악, 영화 관련 주제 범주가 뒤를 이었다. 쇼핑 관련 질의의 하위 주제에 있어서는 일반 상품명(3.7%), 쇼핑 관련 단일 사이트(2.6%), 특정 상품(2.5%), 상품 정보(1.3%) 순으로 검색하였다.

한편 이러한 결과는 국외 선행 연구의 결과와 상당한 차이를 보이고 있다. 〈표 4〉는 이용자 질의의 주제 분류를 수행한 Spink et al.의 2002년과 Jansen, Spink, Pederson의 연구 결과(in press)를 보여준다. 이들의 연구에 따르면 1997년에 익사이트 이용자들이 주로 검색하던 주제는 엔터테인먼트, 레크리에이션(19.9%), 성, 포르노그래피(16.8%), 전자상거래, 여행, 고용, 경제(13.3%), 컴퓨터, 인터넷(12.5%) 순이었다. 그리고 1999년 익사이트 이용자들이 주로 검색하던 주제는 전자상거래, 여행, 고용, 경제(24.5%), 인물, 장소, 사물(20.3%), 컴퓨터, 인터넷(10.9%), 건강, 과학(7.8%) 순이었으며, 2001년의 경우는 전자상거래, 여행, 고용, 경제(24.7%), 인물, 장소, 사물(19.7%), 불분명, 비영어권 질의(11.3%), 컴퓨터, 인터넷(9.6%) 순으로 나타났다. 또한 2002년 알

〈표 4〉 Spink et al.의 주제 분류와 순위

순위	1997 익사이트 데이터		1999 익사이트 데이터		2001 익사이트 데이터		2002 알타비스타 데이터	
1	엔터테인먼트, 레크리에이션	(19.9%)	전자상거래, 여행, 고용, 경제	(24.5%)	전자상거래, 여행, 고용, 경제	(24.7%)	인물, 장소, 사물	(49.27%)
2	성, 포르노그래피	(16.8%)	인물, 장소, 사물	(20.3%)	인물, 장소, 사물	(19.7%)	전자상거래, 여행, 고용, 경제	(12.52%)
3	전자상거래, 여행, 고용, 경제	(13.3%)	컴퓨터, 인터넷	(10.9%)	불분명, 비영어	(11.3%)	컴퓨터, 인터넷	(12.40%)
4	컴퓨터, 인터넷	(12.5%)	불분명, 비영어	(9.3%)	컴퓨터, 인터넷	(9.6%)	건강, 과학	(7.49%)
5	건강, 과학	(9.5%)	건강, 과학	(7.8%)	성, 포르노그래피	(8.5%)	교육, 인문학	(5.07%)
6	인물, 장소, 사물	(6.7%)	성, 포르노그래피	(7.5%)	건강, 과학	(7.5%)	엔터테인먼트, 레크리에이션	(4.57%)
7	사회, 문화, 인종, 종교	(5.7%)	엔터테인먼트, 레크리에이션	(7.5%)	엔터테인먼트, 레크리에이션	(6.6%)	성, 포르노그래피	(3.26%)
8	교육, 인문학	(5.6%)	교육, 인문학	(5.3%)	교육, 인문학	(4.5%)	사회, 문화, 인종, 종교	(3.11%)
9	예술	(5.4%)	사회, 문화, 인종, 종교	(4.2%)	사회, 문화, 인종, 종교	(3.9%)	정부	(1.57%)
10	불분명, 비영어	(4.1%)	정부	(1.6%)	정부	(2.0%)	예술	(0.69%)
11	정부	(3.4%)	예술	(1.1%)	예술	(1.1%)		
총계	2,414	(100.0%)	2,539	(100.0%)	2,453	(100.0%)	2,603	(100.0%)

타비스타 이용자들이 주로 검색하던 주제는 인물, 장소, 사물(49.27%), 전자상거래, 여행, 고용, 경제(12.52%), 컴퓨터/인터넷(12.40%), 건강, 과학(7.49%), 교육/인문학(5.07%) 순으로 나타났다. 즉, 1997년과 1999년, 2001년, 2002년의 질의를 비교하면 엔터테인먼트, 레크리에이션 관련 질의는 대폭 감소하고, 컴퓨터, 인터넷 관련 질의는 비슷한 수준을 유지하고 있으며, 전자상거래, 여행, 고용, 경제 관련 질의가 증가하였음을 알 수 있다. 또한 인물, 장소, 사물에 관한 질의가 급증하였음을 알 수 있다.

따라서 국내 웹 이용자들이 국외 웹 이용자들보다 컴퓨터/인터넷과 엔터테인먼트, 특히 인터넷 포탈 사이트, 커뮤니티 사이트, 연예인에 대한 관심도가 매우 높다고 할 수 있으며, 이는 양국간의 문화적 차이를 반영한다고 볼 수 있다. 또한 성인 사이트 관련 주제가 차지하는 비중이 국내에서 훨씬 낮은 점을 발견할 수 있는데, 이러한 현상은 네이버 이용자의 "성인" 관련 사이트에 대한 관심이 인터넷 초기보다 감소하였거나, 성인 사이트를 찾는 이용자들이 검색 엔진을 이용하지 않고, 이러한 사이트의 URL을 직접 입력해서 방문하기 때문인 것으로 추정된다.

〈표 5〉는 네이버 이용자가 주로 검색하는 주제의 순위를 형태별로 보여준다. 이 표로부터 이용자가 주로 검색하는 주제에 있어서 사이트 검색 질의와 내용 검색 질의 사이에 차이가 있음을 알 수 있다. 즉, 사이트 검색 질의의 경우에는 "컴퓨터/인터넷," "쇼핑," "게임," "기업" 등의 주제가 우세하고, 내용 검색 질의의 경우에는 "엔터테인먼트", "교육," "라이프스타일" 등과 같은 주제가 우세하다. 한편, 형태 분류는 가능하나 주제가 불분명한 699개 질

〈표 5〉 형태별 주제 순위

순위	사이트 검색			내용 검색			사이트와 내용 동시 검색		
1	컴퓨터/인터넷	1,953	(21.3%)	엔터테인먼트	1,607	(22.2%)	엔터테인먼트	278	(19.6%)
2	쇼핑	1,162	(12.7%)	교육/학문	820	(11.3%)	쇼핑	151	(10.6%)
3	게임	1,092	(11.9%)	라이프스타일	777	(10.7%)	기업	141	(9.9%)
4	기업	982	(10.7%)	컴퓨터/인터넷	717	(9.9%)	컴퓨터/인터넷	127	(8.9%)
5	엔터테인먼트	891	(9.7%)	금융/경제	464	(6.4%)	게임	124	(8.7%)
6	교육/학문	656	(7.1%)	게임	450	(6.2%)	라이프스타일	117	(8.2%)
7	금융/경제	598	(6.5%)	사회	446	(6.2%)	교육/학문	74	(5.2%)
8	기관	494	(5.4%)	문화/예술	409	(5.6%)	기관	73	(5.1%)
9	뉴스/미디어	445	(4.8%)	쇼핑	369	(5.1%)	금융/경제	69	(4.9%)
10	라이프스타일	347	(3.8%)	과학	320	(4.4%)	지역/여행	66	(4.6%)
11	지역/여행	201	(2.2%)	건강	271	(3.7%)	사회	55	(3.9%)
12	성인	112	(1.2%)	기업	217	(3.0%)	건강	44	(3.1%)
13	문화/예술	91	(1.0%)	지역/여행	217	(3.0%)	문화/예술	41	(2.9%)
14	건강	75	(0.8%)	성인	71	(1.0%)	과학	28	(2.0%)
15	사회	67	(0.7%)	기관	60	(0.8%)	성인	26	(1.8%)
16	과학	18	(0.2%)	뉴스/미디어	27	(0.4%)	뉴스/미디어	7	(0.5%)
총계		9,184	(100.0%)		7,242	(100.0%)		1,421	(100.0%)

의들 중 604개 질의의 형태가 내용 검색으로 나타났다. 즉, 내용 검색 질의에서 주제가 불분명한 질의의 비중이 상대적으로 높으며, 이러한 질의에 대하여 이용자들은 주로 특정한 주제 범주로 분류되기 어려운 네이버 지식인의 답변을 조회하였다.

한편, 내용 검색 질의를 입력한 후, 많은 이용자들이 "지식인" 데이터베이스로부터 검색된 문서들을 조회하는 특징을 보였으며, 내용 검색에서 큰 비중을 차지하는 상위 5개 주제 범주로 분류된 질의들에 대해 이용자들이 조회한 문서들의 특징은 다음과 같다. 첫째, "엔터테인먼트"의 하위 범주들 중 "영화"가 큰 비중을 차지하였다. 영화에 대한 검색은 대부분 관심 있는 영화에 대한 정보를 찾는 것으로, 이용자들은 네이버 디렉토리의 "영화" 카테고리와 지식인의 답변에서 영화의 평이나 주인공, 영화에 삽입된 노래, 영화를 상영하는 장소 등에 관한 자료를 조회하였다. "엔터테인먼트"의 하위 범주인 "음악"으로 분류된 질의의 대부분은 궁금한 노래의 제목 또는 그 노래를 부른 가수 등을 검색하기 위한 것으로, 이용자들은 네이버 지식인에서 정보를 발견하거나 음악 관련 홈페이지, 팬사이트, 블로그 등에서 정보를 발견하기도 하였다. "엔터테인먼트"의 또다른 하위 범주인 "애니메이션"에 관한 질의에 대하여, 많은 이용자들은 네이버 지식인의 답변을 조회하였다. 또한 팬사이트와 홈페이지에 포함된 내용을 조회하는 경우도 비교적 큰 비중을 차지하였다. "유머"로 분류된 질의에 대하여 이용자들은 유머를 전문적으로 다루는 사이트, 개인 카페, 블로그 등에 포함된 글을 조회하였고, "운세"로 분류된 질의에 대하여 운세를 전문적으로 다루는 사이트의 글을 조회하였다.

둘째, 내용 검색에서 "엔터테인먼트"에 이어 네이버 이용자들이 많이 검색하는 주제는 "교

육"이었으며, 교육에서도 네이버 지식인의 답변을 조회하는 경우가 역시 큰 비중을 차지하였다. 즉, 직업명, 학교명 등의 단어형 질의와 '-과 취직?'과 같은 서술형 질의를 입력한 후 지식인의 답변을 조회하는 경우가 많았다. 또한 선생님 사이트, 개인 교육 사이트, 학교 홈페이지, 학원 홈페이지 등에서 초·중·고등학교의 국어, 수학, 영어, 과학과 관련된 간단한 교육적 내용을 검색하는 질의도 발견되었다.

셋째, "라이프스타일"로 분류된 질의는 패션, 요리, 교통 등과 같이 생활에 관련된 잡다한 내용을 검색하기 위한 것으로, 이용자는 주로 지식인의 답변을 조회하였다. 그리고 이용자는 열차 시간과 같은 생활 정보를 얻기 위하여 지식인보다 철도청 사이트에서 제공되는 자료를 선택하는 경향을 보였다. 또한 일부 질의에 대하여 네이버 디렉토리의 하위 카테고리에서 제공되는 자료를 조회하는 경우도 있었다.

넷째, 내용 검색에서 4위를 차지한 주제 범주인 "컴퓨터/인터넷"에 대한 검색은 주로 컴퓨터 일반에 대한 내용을 찾는 것으로, 이용자들은 "지식인"의 답변에서 컴퓨터 부품 사용법, 작동 오류 대처법, 모델 정보 등을 조회하였다.

다섯째, "금융/경제"와 관련된 질의에 대하여 이용자는 주로 네이버 디렉토리의 "경제" 카테고리에서 제공되는 주가나, 환율, 회사 등에 관한 자료를 조회하거나, 네이버 디렉토리의 "뉴스/미디어" 카테고리에 포함된 신문 기사 등을 조회하였다.

4. 3 외국어 검색

네이버 이용자가 외국어 질의를 입력하는 경우는 드물었으며, 전체 18,200개 질의 중 영어 질의는 1,279개(7.0%), 한문 질의는 5개(0.0003%), 일본어 질의는 2개(0.0001%)에 불과하였다. 영어 질의들 중 연예인 이름, 가수 이름, 영화 제목, 컴퓨터 용어 등은 511개(2.8%), cgv, kbs, nhn, nate, msn, mbc, lg, lgeshop, ebs 등과 같은 기관명은 364개(2%)를 차지하였다. 또한 의미를 네이버 사전이나 지식인에서 검색하기 위하여 영어 단어를 입력한 질의가 115개(0.6%), "www.korail.go.kr," "www.daum.net," "http://www.naver.com" 등과 같이 URL을 직접 입력한 질의가 62개(0.3%)로 나타났다. 그 외에 영어 단어들로 구성된 질의 227개(1.2%)는 영화 또는 노래를 검색하기 위한 것으로 추정된다. 따라서 영어 질의의 대부분이 국내 사이트나 국내에서 생성된 문서를 검색하기 위한 것으로 판단된다.

4. 4 오타

전체 질의 중에서 386개(2.1%)의 오타를 발견하였다. 이 중 영문 입력 모드에서 한글을 입력하여 오타가 발생한 경우, 예를 들어 "다음"을 "ekdma"로, "야후"를 "ㅑgn"으로 입력하는 경우가 192개(1.0%), 한글 입력 모드에서 영어를 입력하여 오타가 발생한 경우, 예를 들어 "ktf"를 "ㅏㅅㄹ"로, "nhn"을 "ㅜㅗㅜ"로 입력하는 경우가 19개(0.1%), 질의 입력 시 문자의 삽입, 삭제, 교체, 전치로 인한 오타가 발생한 경우, 예를 들어 "네이버"를 "넹이버"로, "포토샵"을 "포토샤"으로, "스타크래프트"를 "스타크레프트"로 입력하는 경우가 156개(0.86%)로 조사되었다.

5. 결 론

본 연구에서는 각각의 질의에 대하여 이용자가 조회한 문서를 기록한 클릭 로그에 근거하여, 1년 동안 네이버 이용자들이 입력한 질의의 형태 및 주제의 전반적인 특징을 조사하였다. 본 연구의 조사 결과, 첫째, 네이버 이용자가 입력한 질의를 형태별로 분류할 경우 사이트 검색이 거의 절반에 가까운 48.9%로 가장 많았고, 내용 검색이 42.6%, 사이트와 내용 동시 검색이 7.5%로 나타났다. 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. 둘째, 내용 검색 질의를 입력한 후, 많은 이용자들이 "지식인" 데이터베이스로부터 검색된 문서들을 조회하는 특징을 보였다. 그러나 검증되지 않은 답변과 주관적인 답변이 다수에 의해 지식으로 선택되고, 유해 정보가 질문이나 답변으로 게시되는 경우를 발견하였으며, 지식인 서비스의 개선을 위하여 이러한 문제점의 개선이 요구된다.

한편, 네이버 이용자가 전반적으로 가장 많이 검색한 주제는 컴퓨터/인터넷, 엔터테인먼트, 쇼핑, 게임, 교육, 기업, 라이프스타일, 금융/경제 순으로 나타났다. 그리고 네이버 이용자가 주로 검색하는 주제의 순위를 형태별로 살펴 본 결과, 사이트 검색 질의의 경우에는 컴퓨터/인터넷, 쇼핑, 게임, 기업 등의 주제가 우세하고, 내용 검색 질의의 경우에는 엔터테인먼트, 교육, 라이프스타일 등과 같은 주제가 우세하였다. 이러한 결과는 이용자들의 정보 요구를 반영한다고 볼 수 있으므로 웹 검색 포털 업체들이 컨텐츠 구축의 우선 순위를 결정하는데 활용될 수 있을 것이다. 예를 들어 사이트 검색 질의에서의 주제 순위는 디렉토리의 컨텐츠 구축 시 활용될 수 있을 것이다. 본 연구에서 네이버 이용자가 외국어 질의를 입력하는 경우는 드물었으며, 외국어 질의를 입력하더라도 국내 기관이나 국내에서 생성된 자료를 검색하는 경우가 대부분이었다. 이러한 결과는 국내 웹 포탈에서 국외보다는 국내에서 생성된 자료에 집중하는 것이 현재로서는 중요함을 시사한다. 마지막으로, 본 연구에서 분석한 오타의 유형을 심층 분석하여 검색 엔진이 오타의 유형을 자동으로 파악하게 한다면, 이용자가 오타를 입력하더라도 적합한 문서를 제공함으로써 검색 결과의 효율성을 증대시킬 수 있을 것이다.

본 연구의 수행 결과 향후 연구가 요구되는 사항들은 다음과 같다. 첫째, 클릭 로그가 생성되지 않은 경우에 대한 심층적인 분석이 필요하다. 둘째, 본 연구에서 제시한 클릭 로그 분석 방법론과 주제 분류 체계에 대한 검증과 보완 작업이 요구된다. 셋째, 국내 웹 이용자 검색 행태의 추이를 장기간에 걸쳐 추적하는 연구가 요구된다. 즉 날짜별, 요일별, 계절별 검색 행태 비교, 주중과 주말의 검색 행태 비교 등에 대한 분석이 가능할 것이다.

참 고 문 헌

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

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 Proceedings of the 6th IEEE symposium on computers and communications*, 72-79.

Jansen, B. J., and Spink, A. in press. "How are we searching the World Wide Web? A comparison of nine search engine transaction logs." *Information Processing and Management.*

Jansen, B. J., Spink, A., and Pedersen, J. in press. "A temporal comparison of AltaVista web searching." *Journal of the American Society for Information Science and Technology.*

Jansen, B. J., and Spink, A. 2005. "An analysis of Web searching by European AlltheWeb.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.

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.