국내 웹 이용자의 검색 행태 추이 분석*

Trends of Search Behavior of Korean Web Users

박 소 연(Soyeon Park)**
이 준 호(Joon Ho Lee)***

목 차

초 록

본 연구에서는 1년이라는 장기간에 걸쳐 네이버에 입력된 검색 질의들의 표본과 각 질의에 대한 클릭 로그에 근거하여 국내 웹 이용자의 검색 행태 추이를 분석하였다. 질의의 형태에 대한 조사 결과, 계절별, 주중과 주말, 요일별 질의 형태의 분포에 있어서 유의한 차이가 있는 것으로 나타났다. 또한 웹 이용자들이 입력한 질의의 주제 역시 계절별, 주중과 주말, 요일별로 변화하는 것으로 나타났다. 반면 1년 동안을 전체적으로 살펴볼 때 사이트 검색과 내용 검색의 비율 그리고 주제의 비율이 큰 변화 없이 일정한 상태를 유지하였다. 본 연구의 결과는 인터넷 검색 포탈 업체들의 효과적인 컨텐츠 구축 및 효율적인 검색 시스템 개발에 기여할 것으로 기대된다.

ABSTRACT

This study examines trends of web query types and topics submitted to NAVER during one year period by analyzing query logs and click logs. There was a seasonal difference in the distribution of query types. Query type distribution was also different between weekdays and weekends, and between different days of the week. The log data show seasonal changes in terms of the topics of queries. Search topics seem to change between weekdays and weekends, and between different days of the week. However, there was little change in overall patterns of search behavior across one year. The implications for system designers and web content providers are discussed.

키워드: 웹 검색, 질의 형태, 질의 주제, 종단 연구
Web Search, Query Type, Query Topic, Longitudinal Study

---

* 본 연구는 2004학년도 덕성여자대학교의 연구비 지원으로 이루어졌음.
** 덕성여자대학교 문헌정보학과 조교수(sypark@duksung.ac.kr)
*** 숭실대학교 정보과학대학 컴퓨터학부 부교수(joonho@computing.ssu.ac.kr)
논문접수일자 2005년 5월 15일
게재확정일자 2005년 6월 10일

1. 연구 목적

인터넷 사용과 보급의 폭발적 증가는 인터넷을 통한 정보의 접근을 지원하기 위한 웹 검색 서비스들을 활성화시켰다. 이에 따라 국내외 여러 학문 분야에서 웹 검색에 관한 연구들이 다양한 연구 방법을 이용하여 수행되고 있다. 이들 중 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 연구자와 검색 시스템 설계자 모두에게 시사하는 바가 클 것으로 기대된다. 또한 검색 행태 추이에 대한 연구의 결과는 이용자의 향후 검색 행태와 정보 요구를 예측하는데 활용될 수 있을 것이다.

한편, 이용자와 검색 시스템 사이의 모든 상호 작용을 기록한 검색 트랜잭션 로그는 이용자의 실제 검색 행위를 사실적으로 반영한다. 따라서 이러한 로그의 분석은 웹 검색 행태의 추이 및 변화 분석을 위한 적절한 연구 방법으로 판단된다. 지금까지 로그 분석을 이용한 대다수 선행 연구들은 하루나 일주일 동안의 단기간에 걸쳐 생성된 로그를 분석하거나, 1년 또는 2년에 한번씩 수집한 질의들을 분석 대상으로 하였다. 그러나 이용자들의 보다 현실적인 검색 성향 분석을 위해서는 장기간에 걸쳐 지속적으로 수집된 트랜잭션 로그에 대한 분석이 요구된다.

이에 본 연구에서는 1년 동안 네이버에 입력된 검색 질의들과 각각의 질의에 대하여 이용자가 조회한 문서들을 기반으로 국내 웹 이용자의 검색 행태 추이를 분석하고자 한다. 즉, 본 연구는 박소연, 이준호, 김지승(2005)의 후속 연구로서 선행 연구에서 수집한 자료에 근거하여 검색 질의의 형태와 주제를 계절별, 주중과 주말, 요일별, 날짜별로 비교하고자 한다. 국내외 선행 연구들 중 장기간에 걸쳐 지속적으로 웹 이용자들의 검색 성향을 분석한 경우는 드문 실정이며, 따라서 본 연구의 결과는 웹 검색 분야에 학문적으로 기여하는 바가 클 것으로 기대된다. 또한 본 연구의 결과는 인터넷 검색 포탈 업체들의 효과적인 컨텐츠 구축 및 효율적인 검색 시스템 개발에 있어서 중요한 자료로서 활용될 수 있다.

2. 선행 연구

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. 이 연구들은 웹 검색에 있어서의 검색 방식의 단순성을 공통적으로 발견하였다. 그러나 이러한 국내외 선행 연구들은 단기간에 걸쳐 생성된 로그들을 분석하였다.

Spink et al.(2002)은 1997년부터 2001년까지 2년에 한번씩 하루를 선정하고, 그날 익사이

트에 입력된 질의들 중 무작위로 추출된 약 2,500개의 주제를 분류하였다. 그 결과 이용자들이 주로 검색하는 주제가 엔터테인먼트와 성 관련으로부터 전자 상거래 관련으로 변화하였으나, 이용자들의 전반적인 검색 행태는 변하지 않았음을 보고하였다. Jansen, Spink, Pedersen (2005)은 2002년 9월 8일 알타비스타에서 생성된 약 100만개의 질의들로부터 2,603개를 무작위로 추출한 후 이들의 주제를 분류하고 이를 Silverstein et al.(1999) 의 연구 결과와 비교하였다. 이들은 2002년에 알타비스타 이용자들이 검색하는 질의의 주제가 1998년보다 더 다양해지고 광범위해졌으며, 성과 관련된 질의들이 감소하고 일반적인 엔터테인먼트 성 질의가 증가하였다고 기술하였다.그러나 Silverstein et al. 의 연구에서는 질의들의 주제를 분류하지 않고, 이용자들이 가장 많이 검색하는 검색어들의 순위만 보고하였으므로, 상이한 두 연구의 자료를 비교하는 것은 타당성이 부족한 것으로 보인다.

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. 즉 이용자들이 입력하는 질의의 길이가 더 짧아지고, 이용자들이 조회하는 문서의 수가 더 감소하는 경향이 있음을 발견하였다. 또한 이들은 2002 년의 이용자들이 검색하는 질의의 주제가 2001년보다 더 다양해졌고, 성과 관련된 질의들이 감소하였다고 기술하였다. 그리고 Jansen과 Spink(in press) 는 이러한 연구들을 비교하고 요약하였다. 그러나 Spink와 Jansen이 수행한 이러한 일련의 연구들은 1년 또는 2년에 한번씩 수집한 약 2500개의 질의를 대상으로 하였으므로, 이용자들의 지속적인 검색 행태에 대한 연구로 보기에는 어려움이 있다.

Wang, Berry, Yang(2003) 은 1997년 5월부터 2001년 5월까지 4 년 동안 비교적 장기간에 걸쳐 University of Tennesse at Knoxville의 웹 사이트에 입력된 약 54만개의 질의를 분석하였다. 이들은 1997년과 1998년 학년도에 입력된 총 질의 수와 1998년과 1999년도에 입력된 "취업(career)" 질의와 "풋볼(football)" 질의 수의 분포를 월별로 분석하여 이용자들의 검색 행태가 계절 주기, 학기의 주기에 따라 변화하였다고 기술하였다. 그러나 4년을 전체적으로 살펴볼 때 이용자의 검색 행태나 입력된 질의의 주제에 있어서 거의 변화가 없었다고 보고하였다. 이 연구는 이용자 계층이 제한된 대학교 웹 사이트에 입력된 질의만을 분석하였기 때문에, 이 연구의 결과를 일반 웹 이용자들의 검색 행태로 일반화하기에는 한계가 있다. 또한 특정 질의의 월별 분포에 대한 분석을 이용자 검색 행태의 추이에 대한 연구로 보기에는 어려움이 있다.

이처럼 많은 국내외 선행 연구들로부터 장기간에 걸쳐 지속적으로 수집된 웹 검색 로그를 분석한 연구를 찾아보기가 어려운 실정이다. 이에 본 연구에서는 1년이라는 장기간에 걸쳐 네이버에 입력된 검색 질의들의 표본과 각 질의에 대한 클릭 로그에 근거하여 국내 웹 이용자들의 검색 행태 추이를 분석하고자 하며, 이를 위하여 박소연, 이준호, 김지승(2005) 의 연구에서 수집한 로그 자료를 활용하였다. 이들의 연구 이전에 수행된 대부분의 선행 연구들은 웹 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 박소연, 이준호, 김지승은 1년 동안 네이버 이용자들이 입력한 질의를 기록한 질의 로그와 질의에 대한 검색 결과에서 이용자가 조회한 문서를 기록한 클릭 로그에 근거하여 국내 웹 검색 질의의 형태 및 주제의 전반적인 특징을 분석하였다. 이들의 연구에서 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, 기업, 라이프스타일, 금융/경제 순으로 나타났다.

3. 연구 방법

본 연구는 검색 질의의 형태와 주제를 계절별, 주중과 주말, 요일별, 날짜별로 비교하고자 하며, 이를 위하여 박소연, 이준호, 김지승(2005)의 연구에서 다음과 같이 수집한 로그 자료를 활용하였다. 이들의 연구는 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. 또한 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 총 26일의 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). 이들의 연구에서는 이러한 요소를 고려하여 하루에 700개의 질의를 무작위로 선정하여 총 18,200개의 질의를 분석하였다.

본 연구를 위해서는 검색 질의의 형태 및 주제의 분류가 선행되어야 한다. 박소연, 이준호, 김지승(2005)의 연구에서 질의의 형태를 분석한 결과, 사이트 검색, 내용 검색 사이트와 내용 동시 검색이라는 범주가 도출되었다. 질의의 형태 범주 도출 시 (i) TREC(Text REtrieval Conference) 내 Web Track의 검색 과제와 (ii) 네이버가 제공하는 데이터베이스의 구성을 참고하였다. 첫째, TREC은 1992년부터 시작된 미국 내 정보검색 분야의 주요 학회 중의 하나인데, 새로운 정보검색 시스템 및 기술개발과 평가를 주요 목표로 하고 있다(Voorhess, 2004). TREC안에는 여러 특수 분야가 존재하는데, 가장 최근에 등장한 분야가 Web Track이며, Web track에서는 참가자들에게 검색과제로서 "topic distillation task," "homepage finding task"와 "named page finding task"을 요구하고 있다(Crasswell and Hawking, 2005; Crasswell and Hawking, 2004). "Topic distillation

task"는 광범위한 주제와 관련된 다양한 핵심 자료를 웹에서 찾는 과제이고, "homepge finding task"는 특정한 사이트에 대한 홈페이지를 찾는 과제이며, "named page finding"은 이용자가 입력한 "이름"과 관련된 내용을 포함하며 홈페이지가 아닌 자료를 찾는 과제이다. 둘째, 이용자가 네이버에서 통합 검색 수행 시 제공되는 데이터베이스들은 크게 다음과 같이 두 범주로 나누어질 수 있다.

- 사이트 정보 제공 데이터베이스: 바로가기, 카테고리 사이트 등
- 내용 제공 데이터베이스: 뉴스, 지식인, 전문지식, 책 본문, 이미지, 사전, 카페/블로그 등

Web Track의 "homepage finding"이 사이트를 찾는 과제이며, "named page finding"이 특정한 이름이나 주제와 관련된 내용을 찾는 과제이므로, 박소연, 이준호, 김지승(2005)은이를 참고하여 질의의 형태를 사이트 검색, 내용 검색, 사이트와 내용 동시 검색이라는 범주로 분류하였다. 사이트 검색은 이용자가 찾고자 하는 대상이 웹 사이트인 경우로서, "단일 사이트 검색"과 "다수 사이트 검색"으로 세분화될 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, 지식인에 올라간 글들을 클릭한 경우이다.

또한, 이들의 연구에서는 귀납적 내용 분석 방법을 사용하여 네이버에 입력된 질의들의 주제에 근거한 분류 체계를 도출하였다. 이때 해외 선행 연구들이 개발한 분류 체계와(Ross and Wolfram, 2000; Spink et al., 2001; Spink et al., 2002) 네이버, 야후(한국, 미국), 구글(한국, 미국), 엠파스와 같은 국내외 주요 웹 검색 디렉토리 서비스의 대분류 및 중분류 항목을 참고하였다. 그리고 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 범주를 도출하였다.

- 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

박소연, 이준호, 김지승(2005)의 연구에서는 이처럼 도출된 형태 및 주제 범주로의 분류 작업에 대한 상세한 가이드라인을 작성하였으며, 문헌정보학과 졸업생과 재학생으로 구성된 세 명의 평가자들은 이 가이드라인에 따라 질의의 형태 및 주제를 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.

다음 장에서는 이러한 분류 결과를 계절별, 주중과 주말, 요일별, 날짜별로 비교함으로써, 국내 웹 검색 행태의 추이와 변화를 분석한다. 본 연구에서는 질의가 입력된 날짜를 기준으로 3, 4, 5월을 봄으로, 6, 7, 8월을 여름으로 9, 10, 11월을 가을로, 12, 1, 2월을 겨울로 구분하여, 계절별로 질의들의 주제와 형태를 비교하였다. 본 연구에서 선택된 날짜 중에는 주중이면서 휴일인 날들이 이틀 포함되었는데 이들을 주말에 포함시켜 주중과 주말 간의 검색 행태를 비교하였다. 또한 질의 주제의 분석 시 중복 분류를 허용하였기 때문에, 질의의 수가 아닌 질의의 비율을 비교 대상으로 하였으며, 계절별, 주중과 주말, 요일별 질의 주제의 분포 분석에 있어서 추론 통계를 적용하지 않았다.

4. 연구 결과

4. 1 계절별 검색 행태

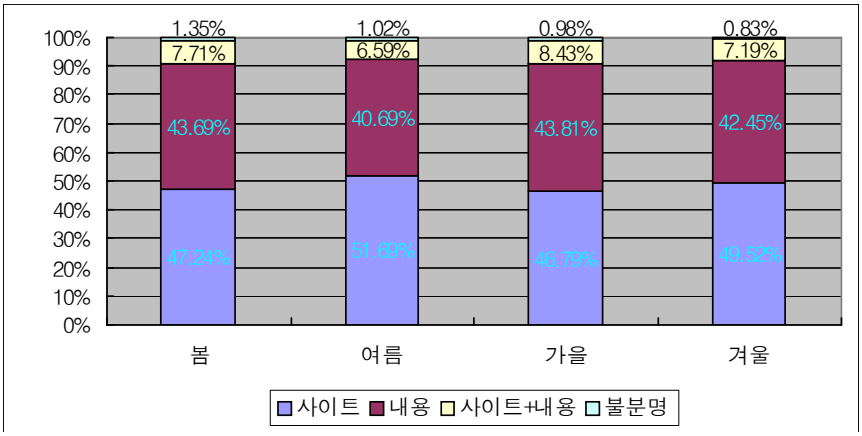
4. 1. 1 질의 형태 비교

계절별로 질의들의 형태를 비교한 결과는 〈그림 1〉과 같다. 사계절 중 사이트 검색 질의의 비중이 가장 큰 계절은 여름으로 여름의 총 질의 중 약 51.7%가 사이트 검색 질의로 나타났다. 반면 내용 검색 질의의 비중이 가장 큰 계절은 가을로서 전체 질의의 43.81%가 내용 검색 질의로 나타났다. 사이트와 내용 동시 검색 질의의 비중이 가장 큰 계절은 역시 가을로 나타났다. 형태가 불분명한 질의를 제외한 후, 계절별 질의 형태 분포의 차이를 검증하기 위하여 카이자승법을 적용하였으며, 그 결과 계절별 질의 형태 분포에 있어서 통계적으로 유의한 차이가 있는 것으로 나타났다. X^2 (6, N=18008) = 32.760, $p \leq 0.001$.

4. 1. 2 질의 주제 비교

네이버 이용자가 입력한 질의의 주제를 계절별로 비교한 결과는 〈표 1〉과 같다. 이 표로부터 활동성이 떨어지는 겨울에는 엔터테인먼트, 게임의 비중이 높다는 사실을 알 수 있다. 즉, 겨울에 입력된 전체 질의의 17.6%가 엔터테인먼트, 11.3%가 게임에 관한 질의였다. 반면 봄에 입력된 전체 질의 중 엔터테인먼트에 관한 질의는 14.3%였으며, 여름에 입력된 질의 중 엔터테인먼트에 관한 질의는 15.2%, 게임에 관한 질의는 8.4%, 쇼핑에 관한 질의는 8.4%로 겨울보다 낮은 것으로 나타났다.

컴퓨터에 관한 질의의 비율은 사계절 모두



〈그림 1〉 계절별 질의 형태 비교

〈표 1〉 계절별 질의 주제 비교

	봄			여름			가을			겨울		
	전체	사이트	내용	전체	사이트	내용	전체	사이트	내용	전체	사이트	내용
컴퓨터/인터넷	15.4%	21.3%	9.9%	15.9%	20.5%	10.0%	16.0%	23.5%	8.3%	15.4%	20.1%	11.3%
엔터테인먼트	14.3%	8.7%	19.4%	15.2%	10.4%	22.2%	15.4%	10.0%	21.3%	17.6%	9.7%	26.4%
쇼핑	9.1%	13.2%	4.3%	8.4%	11.2%	4.1%	10.3%	13.3%	6.3%	10.1%	13.1%	6.0%
게임	8.7%	11.8%	5.5%	8.4%	10.5%	5.5%	9.2%	11.9%	6.1%	11.3%	13.7%	7.9%
교육/학문	9.7%	7.1%	13.6%	7.7%	6.3%	10.6%	8.6%	6.0%	12.1%	8.8%	9.4%	8.6%
기업	7.1%	11.2%	1.9%	8.1%	11.0%	2.9%	7.3%	10.1%	3.4%	7.6%	10.3%	4.0%
라이프스타일	7.9%	4.0%	11.8%	7.3%	4.2%	11.5%	6.1%	3.6%	8.9%	6.4%	3.1%	10.4%
금융/경제	6.2%	6.3%	6.5%	6.5%	6.9%	6.1%	6.9%	6.9%	7.2%	5.8%	6.0%	5.8%
기관	4.0%	6.1%	1.2%	3.9%	5.9%	0.7%	2.9%	4.5%	0.7%	3.1%	4.7%	0.6%
사회	3.8%	0.9%	6.8%	3.0%	1.0%	6.0%	3.6%	0.6%	7.2%	2.3%	0.4%	4.6%
문화/예술	3.4%	1.1%	6.3%	3.2%	1.0%	6.5%	3.4%	1.0%	6.1%	2.1%	0.8%	3.6%
지역/여행	2.6%	1.9%	3.2%	3.4%	2.9%	3.1%	2.2%	1.9%	2.5%	2.5%	1.9%	3.1%
뉴스/미디어	2.3%	4.4%	0.2%	3.4%	5.7%	0.5%	2.7%	5.0%	0.5%	2.3%	4.2%	0.3%
건강	2.2%	0.9%	3.6%	2.2%	0.7%	4.2%	1.9%	0.8%	2.8%	2.5%	0.8%	4.4%
과학	2.5%	0.1%	5.3%	2.1%	0.3%	4.7%	2.6%	0.3%	5.6%	0.9%	0.0%	2.0%
성인	0.8%	1.0%	0.5%	1.5%	1.5%	1.4%	1.0%	0.6%	1.2%	1.4%	1.7%	0.9%

높게 나타났다. 교육/ 학문에 관한 질의의 비율은 학기가 시작되는 봄에 9.7%로서 가장 높고, 여름에 7.7%로서 가장 낮은 것으로 나타났다. 금융/경제에 관한 질의의 비율은 가을에 6.9%로서 가장 높고, 겨울에 5.8% 로서 가장 낮은 것으로 나타났다. 기관에 관한 질의는 봄, 여름에 높고, 가을, 겨울에 낮은 것으로 나타났다. 사회에 관한 질의는 봄에 3.8%로서 가장 높고, 겨울에 2.3%로서 가장 낮은 것으로 나타났다. 지역/여행에 관한 질의는 많은 이들이 여행을 떠나는 여름에 3.4%로서 가장 높고, 가을에 2.2%로서 가장 낮은 것으로 나타났다.

이러한 결과는 웹 이용자들의 정보 요구가 계절별로 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.

4. 2 주중과 주말의 검색 행태

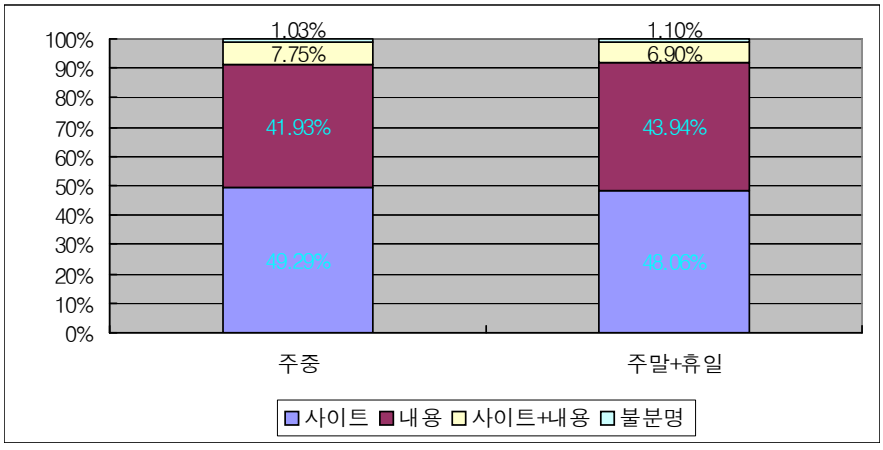
4. 2. 1 질의 형태 비교

〈그림 2〉는 주중과 주말의 질의 형태를 비교한 결과를 보여준다. 휴일인 날들은 주말에 포함되어 분석되었다. 주중에 사이트 검색 질의의 비율은 49.29%로서 주말보다 높았으며, 또한 사이트와 내용 동시 검색 질의의 비율도 7.75%로서 주말보다 높았다. 반면 주말에 내용 검색 질의의 비율은 43.94%로서 주중보다 높았다. 형태가 불분명한 질의를 제외한 후, 주중과 주말 사이의 질의 형태 분포 차이를 검증하기 위하여 카이자승법을 적용하였다. 그 결과 주중과 주말 사이의 질의 형태 분포에 있어서 통계적으로 유의한 차이가 있는 것으로 나타났다. $\underline{X}^2$ (2, N=18008) =9.067, p ≤ 0.05.

이처럼 주중과 주말에 질의 형태의 분포에 차이가 있는 것은 질의 형태별 주제의 순위와 어느 정도 관련이 있는 것으로 보인다. 즉 박소연, 이준호, 김지승(2005)의 연구에 의하면, 내용 검색 질의의 경우에는 엔터테인먼트의 비율이 가장 높았다. 따라서 주말에 엔터테인먼트와 관련된 질의가 증가함에 따라 내용 검색 질의도 증가하는 것으로 추정된다.

4. 2. 2 질의 주제 비교

네이버 이용자가 입력한 질의의 주제를 주중과 주말로 구분하여 분석한 결과는 〈표 2〉와 같다. 주말에 엔터테인먼트, 게임에 관한 질의의 비율은 주중보다 현저하게 높았으며, 컴퓨터에 관한 질의의 비율도 주중보다 주말이 높았다.



〈그림 2〉 주중과 주말의 질의 형태 비교

〈표 2〉 주중과 주말의 질의 주제 비교

	주중			주말		
	전체	사이트	내용	전체	사이트	내용
컴퓨터/인터넷	15.4%	20.8%	9.9%	16.2%	22.2%	9.8%
엔터테인먼트	14.1%	9.1%	19.7%	18.2%	10.8%	26.5%
쇼핑	9.4%	12.6%	5.0%	9.4%	12.7%	5.2%
게임	7.9%	9.9%	5.2%	12.1%	15.7%	8.0%
교육/학문	9.0%	7.4%	12.1%	8.0%	6.6%	10.0%
기업	8.3%	11.4%	3.5%	6.1%	9.3%	2.1%
라이프스타일	7.0%	3.9%	10.9%	6.8%	3.6%	10.5%
금융/경제	6.9%	7.4%	6.6%	5.2%	4.8%	6.1%
기관	4.0%	6.0%	1.0%	2.7%	4.3%	0.5%
사회	3.5%	0.8%	6.9%	2.7%	0.6%	4.9%
문화/예술	2.9%	1.0%	5.4%	3.3%	1.1%	6.1%
지역/여행	2.8%	2.5%	2.9%	2.5%	1.7%	3.2%
뉴스/미디어	2.9%	5.1%	0.5%	2.4%	4.4%	0.2%
건강	2.3%	1.0%	4.0%	1.9%	0.5%	3.4%
과학	2.4%	0.2%	5.3%	1.4%	0.2%	2.9%
성인	1.2%	1.1%	1.2%	1.1%	1.5%	0.6%

주말에 쇼핑에 관한 질의의 비율은 주중과 비슷한 것으로 나타났다. 반면 기업, 경제, 교육, 기관 등과 같은 나머지 주제에 있어서는 주중에 입력된 질의 수가 주말에 입력된 질의 수보다 많은 것으로 나타났다. 따라서 인터넷 검색 포탈 업체는 주말에 엔터테인먼트, 게임, 컴퓨터에 관한 컨텐츠를 집중적으로 강화하는 것을 고려할 수 있다. 또한 이러한 주제와 관련된 광고의 노출 비율을 주말에 높이는 것을 고려할 수 있다.

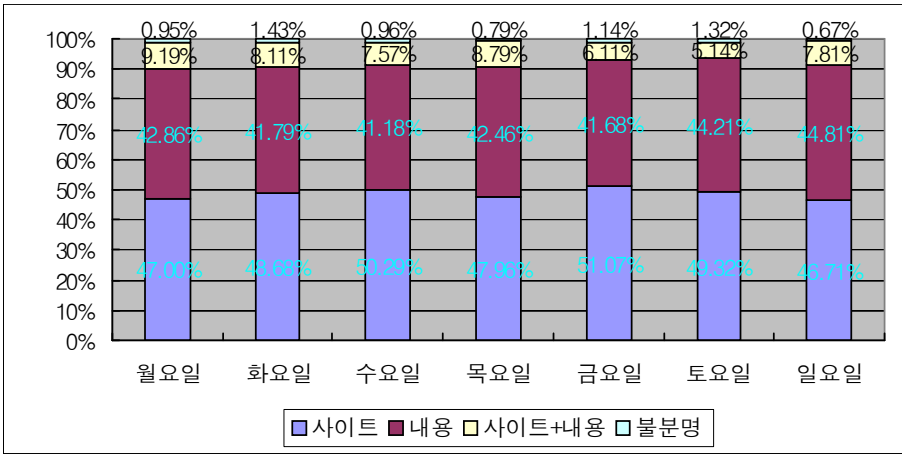
4. 3 요일별 검색 행태

4. 3. 1 질의 형태 비교

전체 질의의 형태를 보다 상세히 요일별로 분석한 결과는 〈그림 3〉과 같다. 일주일 중 사이트 검색 질의의 비율은 금요일에 51.07%로서 가장 높았으며, 내용 검색 질의의 비율은 일요일에 44.81%로서 가장 높은 것으로 나타났다. 그리고 사이트 검색 질의의 비율은 수요일에도 높은 것으로 나타났다. 형태가 불분명한 질의를 제외한 후 요일별 질의 형태 분포의 차이를 검증하기 위하여 카이자승법을 적용하였으며, 그 결과 요일별 질의 형태의 분포에 있어서 통계적으로 유의한 차이가 있는 것으로 나타났다. $\underline{X}^2$ (12, N=18008)=58.453, p≤0.001.

4. 3. 2 질의 주제 비교

네이버 이용자가 입력한 질의의 주제를 요일별로 분석한 결과는 〈표 3〉과 같다. 이 표는 엔터테인먼트(18.6%), 컴퓨터(16.9%), 쇼핑(10.1%) 및 게임(12.8%)에 관한 질의의 비율이 가장 높은 요일은 일요일임을 보여준다. 또한 이 표로부터 라이프스타일, 건강, 자연에 관한 질의



〈그림 3〉 요일별 질의 형태 비교

의 비율은 월요일에 가장 높고, 기업, 미디어, 기관, 사회에 관한 질의의 비율은 화요일에 가장 높고, 교육, 금융/경제, 성인에 관한 질의의 비율은 수요일에 가장 높음을 알 수 있다.

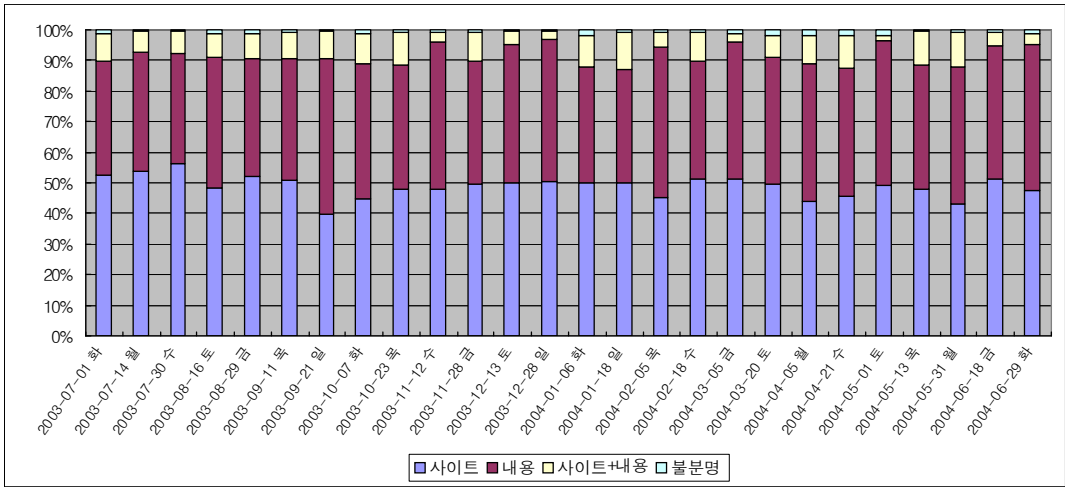
4. 4 날짜별 검색 행태

국내 검색 포탈 중 1999년 말에 서비스를 시작한 엠파스는 인터넷 검색에 있어서 후발 주자임에도 불구하고 "자연어 검색"을 앞세워 기존 업체들에 필적할 만큼의 사용자들을 확보하였다. 또한 네이버는 2002년 말에 "지식 검색" 서비스를 시작함으로써 인터넷 검색 포탈의 선두적 지위를 확보할 수 있는 기반을 마련하였다. 즉, 인터넷 검색 포탈들은 사이트 검색 결과의 차별화에 대한 어려움을 인식하고, 내용 검색 결과의 개선에 많은 노력을 기울여왔다. 따라서 본 연구자들은 1년의 조사 기간 동안 내용 검색의 비율이 점진적으로 증가할 것으로 기대하였다. 그러나 전체 질의의 형태를 날짜별로 분석한 〈그림 4〉로부터 1 년 동안 사이트 검색과 내용 검색의 비율에 변화가 매우 적었으며, 또한 여전히 사이트 검색의 비율이 내용 검색의 비율보다 높음을 알 수 있었다.

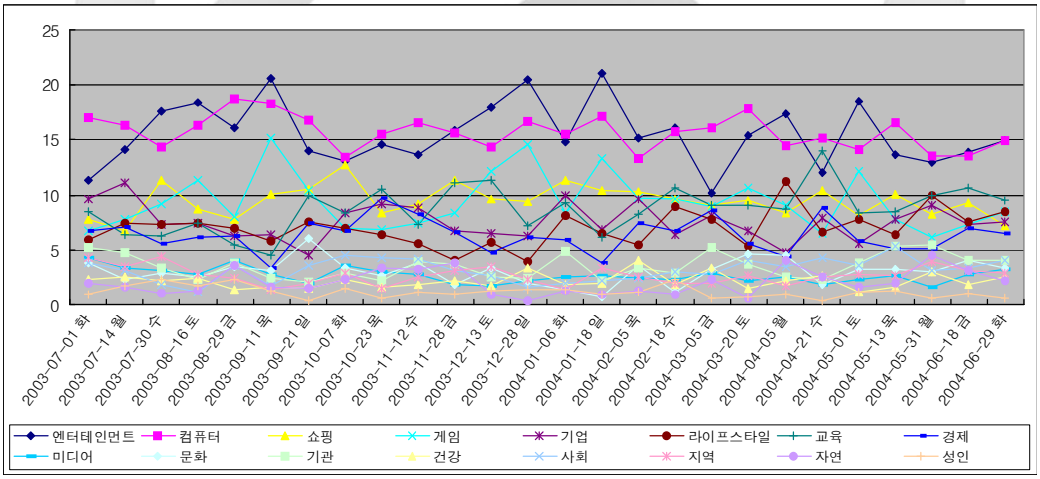
한편, 〈그림 5〉는 전체 질의의 주제를 날짜별로 분석한 결과를 보여준다. 이 그림으로부터 1년의 조사 기간 동안 계절별 또는 주중과 주말에 따라 주제들의 비율이 변화하였으나, 전반적으로는 주제 비율의 변화가 미약함을 알 수 있다. 즉, 1 년의 기간 동안 특정한 주제와 관련된 질의의 비율이 증가하거나 감소하지는 않았으며, 전체적으로 주제의 비율이 어느 정도 고착화되었음을 알 수 있다. 따라서 〈그림 4〉와 〈그림 5〉는 1년의 조사 기간 동안 사용자들의 검색 행태에 변화가 적었음을 보여준다.

5. 결 론

본 연구에서는 1년이라는 장기간에 걸쳐 네이버에 입력된 검색 질의들의 표본과 각 질의에 대한 클릭 로그에 근거하여 국내 웹 이용자들의



〈그림 4〉 날짜별 질의 형태 비교



〈그림 5〉 날짜별 질의 주제 비교

검색 행태 추이를 분석하였다. 질의 형태에 대한 조사 결과, 계절별 질의 형태의 분포에 있어서 통계적으로 유의한 차이가 있는 것으로 나타났다. 또한 주중과 주말 요일별 질의 형태의 분포에 있어서도 통계적으로 유의한 차이가 있는 것으로 나타났다. 반면 1 년 동안을 전체적으로 살펴볼 때, 사이트 검색과 내용 검색의 비율이 큰 변화 없이 일정한 상태를 유지하였고, 여전히 사이트 검색의 비율이 내용 검색의 비율보다 높음을 알 수 있었다.

한편 질의 주제의 추이를 계절별로 분석한 결과, 활동성이 떨어지는 겨울에는 엔터테인먼트, 게임의 비중이 높았으며, 교육/학문에 관한 질의의 비율은 학기가 시작되는 봄에 가장 높

고, 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 비슷한 것으로 나타났다. 이러한 결과를 좀 더 상세히 요일별로 살펴본 결과, 엔터테인먼트, 컴퓨터, 쇼핑 및 게임에 관한 질의의 비율이 가장 높은 요일은 일요일인 것으로 나타났다. 또한 라이프스타일, 건강, 자연에 관한 질의의 비율은 월요일에 가장 높고, 기업, 미디어, 기관, 사회에 관한 질의의 비율은 화요일에 가장 높고, 교육, 금융/경제, 성인 관련 질의의 비율은 수요일에 가장 높음을 알 수 있었다.

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.

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.

한편 본 연구의 이러한 결과는 Wang, Berry, & Yang(2003)의 연구 결과와 유사한데, 이들은 4년 동안 University of Tennesse at Knoxville의 웹 사이트에 입력된 질의를 대상으로 이용자들의 검색 행태가 계절 주기, 학기의 주기에 따라 변화하였으나, 4년을 전체적으로 살펴볼 때 이용자의 검색 행태나 입력된 질의의 주제에 있어서 거의 변화가 없었다고 보고하였다.

본 연구의 수행 결과 향후 연구가 요구되는 사항들은 다음과 같다. 첫째, 본 연구에서는 국내 웹 이용자들의 1년에 걸친 검색 형태 추이를 분석하였으며, 이러한 추이를 5년, 10년과 같이 보다 더 장기간에 걸쳐 추적하는 연구가 필요하다. 둘째, 본 연구에서는 검색 행태의 추이를 질의의 형태와 주제에 국한시켜 조사하였다. 향후 검색 방법의 변화, 즉 세션 길이, 질의 길이, 검색어 길이, 연산자 사용, 오타 비율, 이용자가 조회한 페이지 수 등의 변화에 대한 연구가 요구된다. 이러한 연구는 인터넷 검색 포탈 업체

들의 검색 시스템 개발과 인터페이스 개발에 유용하게 활용될 수 있을 것이다. 마지막으로 본 연구의 결과를 해외의 연구와 비교, 분석하는 연구가 요청된다.

참 고 문 헌

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

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.

Crasswell, N., and Hawking, D. 2005. "Overview of the TREC 2004 Web Track." [Cited 2005.6.17.]. 〈http://trec.nist.gov/pubs/trec13/papers/WEB.OVERVIEW.pdf〉.

Crasswell, N., and Hawking, D. 2004. "Overview of the TREC 2003 Web Track." In E. M. Voorhess & L. P. Buckland (Eds.), The Twelfth Text REtrieval Conference(TREC 2003) (pp. 78-92). Washington D. C. : Government Printing Office.

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. 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 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.

Voorhess, E. M. 2004. "Overview of TREC 2003." In E. M. Voorhess & L. P. Buckland (Eds.), The Twelfth Text REtrieval Conference(TREC 2003) (pp. 1-13). Washington D. C. : Government Printing Office.

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.