# 웹 검색 행태의 추이 및 변화 분석*

## Trends and Changes of Web Searching Behavior

박 소 연(Soyeon Park)**

목 차



초 록

이 연구에서는 국내 주요 검색 포털인 네이버 이용자들의 검색 행태 추이를 조사, 분석하였다. 즉 1년 동안 분기별로 네이버에 입력된 질의들을 대상으로 질의의 입력 행태, 오타 입력 행태, 멀티미디어 검색 행태, 결과 문서 클릭 행태 등의 추이를 조사하였다. 이를 위하여 이용자들이 입력한 통합 검색 질의들로 구성된 질의 로그와 질의에 대한 검색 결과에서 이용자들이 조회한 문서를 기록한 클릭 로그를 분석하였다. 연구 결과, 입력된 질의의 길이 및 주제, 멀티미디어 질의의 특징 및 비율, 오타의 비율 등에 있어서는 1년 동안 큰 변화 없이 일정한 것으로 나타났다. 반면, 질의별로 발생되는 클릭 횟수는 시간이 지남에 따라 점진적으로 증가하는 것으로 나타났다. 본 연구의 결과는 향후 포털의 효과적인 콘텐츠 구축 및 검색 알고리즘 개발에 활용될 수 있을 것으로 기대된다.

ABSTRACT

This study aims to investigate trends of internet searching behavior of users of NAVER, a major Korean search portal. In particular, this study analyzed trends of query submission behaviors, behaviors related to typos, multimedia searching behaviors, and click behaviors. In conducting this study, query logs and click logs of unified search service were analyzed. The results of this study show that there were little changes in the topic and length of queries, the pattern of typos, and multimedia seeking behavior over a year's period. However, click counts of documents have gradually increased over time. The results of this study can be implemented to increase the portal's effective development of internet contents and searching algorithms.

키워드: 웹 검색 행태, 검색 포털, 로그 분석, 종단 연구
Web Searching Behavior, Search Portals, Log Analysis, Longitudinal Study

* 본 연구는 NHN(주)의 지원을 받았음.
** 덕성여자대학교 문헌정보학과 부교수(sypark@duksung.ac.kr)
논문접수일자: 2011년 1월 18일 최초심사일자: 2011년 1월 19일 게재확정일자: 2011년 1월 31일
한국문헌정보학회지, 45(1): 377-393, 2011. 〔DOI:10.4275/KSLIS.2011.45.1.377〕

# 1. 서 론

웹 검색이 활성화되기 시작한 90년대 중반 이후, 다양한 학문 분야에서 웹 검색과 관련된 연구들이 수행되어 왔다. 웹 검색 분야의 중요한 연구 주제 중 하나는 특정한 기간 동안 이용자의 웹 검색 행태의 추이를 분석하는 주제 분야이다. 질의의 주제, 질의의 입력 행태, 결과 문서 조회 행태와 같은 특징이 시간이 지남에 따라 어떻게 변화하는지에 대한 연구는 연구자와 시스템 개발자에게 모두 시사하는 바가 클 것으로 기대된다. 즉, 이용자의 검색 행태가 장기간 동안 안정적인지 또는 가변적인지는 검색 시스템 개발의 중요한 변수로 작용할 것으로 보인다. 또한 검색 행태 추이에 대한 연구 결과는 이용자의 향후 검색 행태와 정보 요구를 예측하는데 활용될 수 있을 것이다. 한편 웹 검색 분야에서 다양한 연구가 수행되어 왔지만, 웹 검색 행태의 추이를 분석한 국외 연구들의 경우 비정기적으로 수집한 질의를 대상으로 하였으며, 질의의 주제 분석에 집중해 온 상황이다. 국내 선행 연구들 중 웹 이용자들의 검색 행태 추이를 조사한 최근 연구는 드문 실정이다.

이에 이 연구에서는 국내 주요 검색 포털인 네이버 이용자의 검색 행태 추이를 조사, 분석하고자 한다. 좀 더 구체적으로 이 연구에서는 1년 동안 분기별로 네이버 에 입력된 통합 검색 질의를 대상으로 질의의 길이, 질의의 주제, 오타 입력 행태 등과 같은 검색 행태의 특징 및 추이를 조사하고자 한다. 또한 멀티미디어 검색이 웹 검색의 중요한 요소임을 고려하여 멀티미디어 검색 행태의 추이도 조사하고자 한다. 마지막으로, 이용자들의 검색 결과 조회 행태의 경향 및 추이도 조사하고자 한다. 이를 위하여 이용자들이 입력한 질의들로 구성된 질의 로그와 이용자들이 조회한 문서들로 구성된 대규모 클릭 로그를 활용하고자 한다. 또한 포털의 검색 서비스 중 가장 이용도가 높은 통합 검색의 로그를 활용하고자 한다.

본 연구의 결과는 웹 이용자들의 검색 행태 및 정보 요구에 대한 이해를 심화시킬 것으로 기대된다. 본 연구는 검색 행태의 특징 및 추이를 분석하기 위한 방법론을 제시함으로써 웹 검색 분야에 학문적으로 기여할 수 있을 것으로 기대된다. 또한 이 연구의 결과는 향후 포털의 검색 서비스의 개선에 활용될 수 있을 것으로 기대된다. 즉 본 연구의 결과는 포털 업체들의 효과적인 콘텐츠 구축 및 검색 알고리즘 개발에 중요한 자료로서 활용될 수 있을 것으로 기대된다.

# 2. 선행 연구

국내외 웹 검색에 관한 연구는 전산학, 문헌정보학, 심리학, 경영학, 신문방송학 등 다양한 분야에서 수행되어 왔다. 이 장에서는 이들 중 웹 검색 행태 추이 분석에 관한 연구에 초점을 두고자 한다. 국외 웹 이용자들의 검색 행태의 추이를 조사한 선행 연구들로는 Jansen과 Spink 등의 연구를 들 수 있다. Jansen과 Spink 등은 1990년대 후반부터 트랜잭션 로그 분석을 통하여 웹 이용자들의 검색 행태를 조사하는 일련의 연구를 수행하여 왔다. 이들은 이러한 연구의 일환으로 웹 검색 엔진 이용자들의 검색 행태의 추이를 분석해 왔다. Spink et al.(2002)은

1997년 9월 16일, 1999년 12월 20일, 2001년 5월 4일 익사이트에 입력된 질의들을 대상으로 세션의 길이, 질의의 길이, 질의별 검색어 수 분포, 질의별로 조회한 페이지 수, 변경된 질의 수, 불리안 연산자 사용 등을 비교, 분석하였다. 또한 날짜별로 전체 질의들로부터 무작위로 추출된 질의를 대상으로 질의의 주제를 분석하였다. 그 결과 이용자들이 주로 검색하는 주제가 엔터테인먼트와 성 관련으로부터 전자 상거래 관련으로 변화하였으나, 세션의 길이, 질의의 길이와 같은 전반적인 검색 행태는 별로 변하지 않았음을 보고하였다.

Jansen과 Spink(2004)는 2001년 2월 6일과 2002년 5월 28일 유럽 검색 엔진인 올더웹 로그에서 질의들을 추출하여 위의 연구와 유사한 분석을 수행하였다. 또한 전체 질의들 중 무작위로 추출된 질의를 대상으로 질의의 주제를 비교, 분석하였다. 이들은 올더웹 이용자들의 검색 행태가 시간이 지남에 따라 점점 단순해지고 있다고 보고하였다. 즉 시간이 경과함에 따라 질의의 길이와 세션의 길이가 더 감소하는 경향이 있음을 발견하였다. 반면 2002년의 이용자들이 검색하는 질의의 주제가 2001년보다 더 다양해졌고, 성과 관련된 질의들이 감소하였다고 기술하였다. Jansen과 Spink의 2005년 연구에서는 자신들의 선행 연구들과 다른 연구자들의 연구를 포함한 9개 연구를 요약, 비교하고, 1997년부터 2002년까지의 연구에서 드러난 웹 검색 행태의 변화를 보고하였다.

한편 이들은 멀티미디어 검색 행태의 추이도 조사하였는데, Ozmutlu, Spink, Ozmutlu(2003)는 1997년 9월 16일, 1999년 12월 20일, 2001년 5월 4일 익사이트 로그를 대상으로 멀티미디어 질의들을 추출한 후, 세션별 질의 수, 질의별 검색어 수, 멀티미디어 질의의 분포, 멀티미디어 질의의 주제 등에 대한 분석을 수행하고, 1997년, 1999년과 2001년도 데이터의 특징 및 추이를 비교·분석하였다. Tjondronegoro, Spink, Jansen(2009)은 2006년 5월 15일 하루 동안 메타검색엔진인 도그파일에서 생성된 트랜잭션 로그 중 일부를 사용하여 도그파일 이용자들의 멀티미디어 검색 행태의 특징을 분석하였다. 또한 이러한 결과를 1997년과 2001년 익사이트 데이터로부터 도출된 결과와 비교하였다.

Spink와 Jansen 등이 수행한 이러한 일련의 연구들은 몇 년에 한번 씩 비정기적으로 수집한 질의를 대상으로 하였으며, 이용자들의 클릭 행태는 분석하지 못하였다는 점에서 한계가 있다고 볼 수 있다. 또한 이들은 상이한 검색엔진 이용자들의 검색 행태를 직접 비교하기도 하였다.

국내 선행 연구 중 로그 분석을 통해 웹 이용자의 검색 행태의 추이를 분석한 연구로는 박소연, 이준호(2005년)의 연구를 들 수 있다. 이들은 2003년 7월 1일부터 2004년 6월 30일까지 1년 동안 네이버에 입력된 통합 검색 질의들의 표본과 각 질의에 대한 클릭 로그에 근거하여 국내 웹 이용자들의 검색 행태 추이를 분석하였다. 좀 더 구체적으로 이 기간 동안 격주로 수집된 질의들과 클릭 로그의 분석을 통하여 검색 질의의 주제와 형태를 계절별, 주중과 주말, 요일별로 비교하였다.

웹 검색 행태의 추이를 분석한 국내외 선행 연구들은 대체로 질의의 주제 분석에 집중해왔다. 또한 국외 선행 연구들의 경우 이용자가 입력한 질의가 주요 분석 대상이었으며, 이용자

가 클릭한 문서들까지 분석한 연구는 드문 실정이다. 즉 선행 연구들 중에서 검색 행태의 다양한 특성을 종합적으로 분석한 연구는 찾아보기 어려운 실정이다. 또한 2006년 상반기 이후의 데이터에 대해서는 분석이 수행되지 않은 실정이다. 이에 본 연구에서는 국내 주요 검색 포털인 네이버 이용자들의 검색 행태의 추이를 입력된 질의의 길이, 주제, 오타 입력 행태, 클릭 행태, 멀티미디어 검색 행태와 같은 다양한 측면에서 분석하고자 한다.

## 3. 연구 방법

### 3.1 자료 수집

본 연구에서는 국내 주요 검색 포털인 네이버에 입력된 검색 질의들의 특징 및 추이를 조사, 분석하였다. 네이버를 선택한 이유는 국내외 검색 포털 분야에서의 네이버의 위상과 인지도 때문이다. 네이버는 2000년대 초반 이후 국내 검색 포털들 중 시장 점유율 조사, 방문자 수 조사, 검색 시간 점유율 조사 등에 있어서 지속적으로 1위를 차지하고 있다. 즉, 네이버는 웹 사이트 평가 및 트래픽 분석업체인 Korean Click(http://www.koreanclick.com)과 인터넷 메트릭스(http://www.metrixcorp.com) 등의 방문자 수 조사에서 1위를 차지하여 왔다. Korean Click의 조사에 따르면, 2010년도 1년 동안 네이버의 국내 검색 시장 점유율이 70% 이상을 유지하고 있는 것으로 나타났다. 또한 2011년 1월 첫째 주 기준으로 검색 사이트별 검색 시간 점유율에 있어서도 네이버가 76.18%를 차지하고 있는 것으로 나타났다. 따라서 네이버에 입력된 질의들이 국내 웹 이용자들이 입력한 질의들에 대한 대표성을 지니고 있다고 판단되었다.

국내 웹 이용자들의 검색 행태의 특징 및 추이를 조사하기 위하여 본 연구에서는 네이버 통합 검색 질의 로그와 클릭 로그를 분석하였다. 웹 검색 서비스의 로그는 일반적으로 이용자들이 입력한 질의를 기록한 질의로그와 질의에 대한 검색 결과에서 이용자가 조회한 문서를 기록한 클릭 로그로 구성된다. 본 연구에서는 2006년 8월 14일 월요일, 2006년 11월 8일 수요일, 2007년 2월 10일 토요일, 2007년 5월 17일 목요일과 2007년 8월 10일 금요일에 이용자들이 네이버 통합검색 창에 입력한 전체 질의 중에서 각각 700개씩 총 3,500개의 질의를 무작위로 선정하여 분석하였다. 질의 선정 이전에 로그 정제 작업을 통하여 로봇이나 프로그램이 자동으로 입력하는 질의, 클릭 횟수를 조작하는 클릭 어뷰즈(click abuse) 등을 제외하였다. 한편 하루에 입력되는 통합 검색 질의들 중 인기 질의들은 여러 번 반복하여 출현한다는 특징이 있다. 즉 “싸이월드”, “다음”, “옥션”과 같은 질의나 인기 연예인과 관련된 질의들의 경우 수십만 회에서 백만 회 이상까지 입력되기도 한다. 이러한 모집단의 특징을 부분적으로 반영하기 위하여 질의 추출 시 질의가 중복되어 포함되는 것을 허용하였다.

위 날짜를 선택한 이유는, 이용자들의 검색 행태가 주중과 주말, 요일별로 변화할 수 있다는 사실을 염두에 두고, 다른 요일의 날짜들을 선택하고자 하였기 때문이다. 또한 이용자들의 구체적인 검색 행태가 기록된 가장 최신 로그

가 포털들의 대외비로 간주되어 확보되기 어려운 현실에서, 연구를 수행하던 시점에서 비교적 최신의 질의들을 구할 수 있는 날이기 때문이다. 실제로 검색 포털의 로그를 분석한 국외 선행 연구들의 경우 현실적인 제약 때문에 7년 이상 경과된 로그를 분석하는 경우도 존재한다(Jansen 2008). 1년 동안 약 3개월의 간격을 두고 질의를 추출한 것은 이 기간 동안 분기별로 이용자들의 멀티미디어 이용 행태의 변화나 추이를 분석하기 위해서이다. 한 날짜에서 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).

본 연구에서는 이렇게 선택된 질의들에 대하여 이용자가 조회한 문서들로 구성된 클릭 로그도 분석하였다. 개별 질의에 대한 클릭 로그는 일정 기간 동안 이용자들이 조회한 문서들로 구성되며, 질의별로 클릭 횟수가 100만 회 이상에 달하는 경우도 있으므로, 대다수 이용자들의 정보 요구가 집대성된 것으로 간주될 수 있다. 본 연구에서 활용한 클릭 로그에는 개별 질의에 대해 이용자가 조회한 문서의 URL, 조회한 문서의 클릭 빈도, 문서가 소속된 컬렉션에 대한 세부 정보가 날짜별로 축적, 저장되어 있다. 날짜별로 수집된 700개의 질의에 대해 2006년 8월 14일에는 25,468개, 2006년 11월 8일에는 21,558개, 2007년 5월 17일에는 23,297개, 2007년 8월 10일에는 3,1487개의 문서가 조회되었으며, 본 연구에서는 이렇게 조회된 문서들의 분포 및 특징을 분석하였다. 한편 2007년 2월 10일의 클릭 로그는 기술적인 문제로 저장되지 않아서 클릭 행태 분석에서 제외되었다.

### 3.2 검색어 분석 방법

본 연구에서 제공받은 질의 로그는 띄어쓰기가 되어 있지 않은 채 검색어들이 연속적으로 결합되어 있는 형태였기 때문에 질의별로 별도의 검색어 수 분석이 필요하였다. 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(박소연, 이준호 2002). 한글은 복합 명사를 구성하는 단일 명사들 사이의 띄어쓰기를 비교적 자유롭게 규정하고 있는 상황이다. 따라서 한글 질의 로그 분석에서 검색어 분석 시 영어의 경우와 유사하게 띄어쓰기 단위로 검색어를 인식, 분석하는 어절 단위 분석과 어절을 의미의 최소 단위인 형태소 단위로 분리한 후, 각각의 형태소를 검색어로 인식하여 분석하는 형태소 분석이 모두 가능하다. 한편 어절 단위로 검색어 분석 시 동일한 질의가 이용자의 띄어쓰기 방법에 따라 다르게 분석될 수 있다는 문제점이 존재한다. 즉, "정보검색"과 "정보 검색"은 동일한 질의임에도 불구하고, 이용자의 띄어쓰기 방법에 따라 각각 1개의 검색어와 2개의 검색어로 간주되어 분석된다.

따라서 본 연구에서는 보다 일관성 있고 정확한 분석을 위하여 질의별로 의미의 최소 단위인 형태소 단위로 분리하여 분석하는 방법을 선택하였다. 질의를 형태소 단위의 검색어로 분석하기 위하여 국립국어원의 띄어쓰기 표준

및 국어대사전을 참고하였다. 즉, 복합 명사 중 완전히 하나의 단어로 굳어진 합성어의 경우 하나의 검색어로 분석하였고, 합성어로 정착되지 않은 상태에서 띄어쓰기가 적절하지만 붙여쓰기가 허용되는 경우, 형태소 단위로 분리하여 분석하였다. 한편 웹 검색 질의들의 경우 신조어, 이모티콘, 줄임말, 숫자, 시리즈물, 오타, 한글과 외국어의 조합, 외국어를 한글 발음대로 입력한 경우 등과 같이 특이한 질의들이 다양하게 존재한다. 본 연구에서는 검색어 분석의 전반적인 과정 및 이런 특이한 사례들에 대해서 국어국문학과 교수의 자문을 받았다.

### 3.3 주제 및 오타 분석 방법

웹 검색 질의들의 주제를 분석하기 위하여, 본 연구에서는 웹 검색 관련 국내외 선행 연구들에서 개발된 분류 체계를(박소연, 이준호 2005; Ross and Wolfram 2000; Jansen and Spink 2005; Spink et al. 2002) 참고하였다. 특히 박소연, 이준호의 2005년 연구에서 제시된 주제 분류 체계를 참고하였는데, 이들은 귀납적 내용 분석 방법을 사용하여 네이버에 입력된 질의들의 주제에 근거한 주제 분류 체계를 도출하였다. 이들의 연구에서는 “건강”, “게임”, “과학”, “교육/학문”, “금융/경제”, “기관”, “기업”, “뉴스/미디어”, “라이프스타일”, “문화/예술”, “사회”, “성인”, “쇼핑”, “엔터테인먼트”, “지역/여행”, “컴퓨터/인터넷”의 16개의 주제 범주로 구성된 주제 분류 체계를 제안하였다. 또한 이들의 연구에서는 웹 검색 질의의 주제 분석 시 이용자가 입력한 질의뿐만 아니라 질의에 대한 검색 결과에서 이용자가 조회한 문서를 함께 분석하는 방안을 제시한 바 있다. 웹 검색 질의의 주제 분야가 방대하고 다양하여서 이용자가 실제로 조회한 문서를 모르는 상태에서 연구자의 판단에 근거하여 질의의 주제를 분류하기에는 한계가 있기 때문이다. 이에 본 연구에서는 이들의 방법론을 참고하여 질의 로그와 클릭 로그에 근거하여 주제 분석을 수행하였다.

또한 이들의 연구와 전산학 분야의 오타 관련 선행 연구들(Damerau 1963; Zobel and Dart 1996)에서 제시된 오타 분류 체계를 참고하여, 전체 오타를 “교체”, “삭제”, “삽입”, “전치”, “영문 모드 한글 입력”, “한글 모드 영문 입력” 등의 하위 범주로 세분화하였다. 또한 이러한 개별 오타 유형들이 동시에 발생하는 경우도 가능하기 때문에 “삭제와 교체”처럼 오타 유형이 결합된 형태도 오타 분류 체계에 포함시켰다. 이러한 오타 분석을 통하여 자주 발생하는 오타유형을 파악하고, 검색 엔진이 이들을 자동으로 교정하여 적합한 검색 결과를 제공한다면 검색 결과의 성능을 높이고, 이용자의 만족도를 제고할 수 있을 것이다.

### 3.4 멀티미디어 질의 분석 방법

멀티미디어 검색은 전통적인 검색과 웹 검색을 차별화시키는 중요한 요소 중 하나이며, 최근에 멀티미디어 검색에 대한 관심이 증대되고 있기 때문에 멀티미디어 검색 행태에 대한 연구가 필요하다고 할 수 있다. 국내 멀티미디어 검색에 관한 선행 연구는 대학생들의 멀티미디어 검색 행태를 분석한 연구(정은경 2010)와 특정한 검색 시스템을 대상으로 멀티미디어 검

색 기법을 구현 개발하고, 실험을 통해 검색 기법의 성능 평가를 수행한 연구들을 들 수 있다(김용, 소민호 2009; 노승민, 황인준 2003; 박창섭 2007; 백우진 외 2008). 웹 이용자들의 멀티미디어 검색 행태를 조사하기 위하여, 이용자들이 입력한 질의들 중 멀티미디어성 질의를 파악하는 작업이 필요하다. 본 연구에서는 멀티미디어 질의를 파악하기 위하여 국내외 관련 선행 연구에서 사용된 기준을(박소연 2010; Jansen, Goodrum and Spink 2000; Ozmutlu, Spink and Ozmutlu 2003) 참고하고, 본 연구에서 사용된 네이버 질의로그와 클릭로그를 상세히 분석하였다. Ozmutlu, Spink, Ozmutlu (2003)는 특정 멀티미디어 검색어 포함 여부를 멀티미디어 질의 기준으로 활용하였다. 즉 오디오 질의는 "audio", "cd", "concerts"와 같은 27개의 오디오 관련 검색어, 이미지 질의는 "image", "picture" 등의 30개의 이미지 관련 검색어, 비디오 질의는 "video", "animated"와 같은 13개의 비디오 관련 검색어를 포함하는 질의로 정의하였다.

한편 이들의 멀티미디어 검색어 기준은 90년대 후반과 2000년대 초반 익사이트 로그를 대상으로 하였으므로, 이들의 기준을 업데이트하고, 한국적 상황을 반영하는 멀티미디어 검색어 기준이 요청된다. 이에 본 연구에서는 이미지 질의를 "그림", "이미지", "사진"과 같은 31개의 이미지 관련 검색어와 21개의 이미지 파일 형식을 포함하는 질의로, 음악 질의를 "가요", "노래", "뮤직"과 같은 36개의 음악 관련 검색어와 10개의 음악 파일 형식을 포함하는 질의로, 동영상 질의를 "동영상", "비디오", "뮤직비디오"와 같은 21개의 동영상 관련 검색어 및 18개의 동영상 파일 형식을 포함하는 질의로 정의하였다. 멀티미디어 검색어 선정 시 모호한 경우에는 클릭 로그를 참고하여, 멀티미디어 자료에 대한 정보 요구가 존재하는지를 확인하였다.

멀티미디어 성 질의에는 질의에 멀티미디어 검색어를 직접적으로 포함하는 경우도 있지만, 멀티미디어 컬렉션에서 이용자들의 클릭이 많이 발생한 질의들도 고려할 수 있다. 따라서 질의별로, 멀티미디어 컬렉션인, 이미지, 음악, 동영상 컬렉션에서 조회된 문서들을 수합하여 그 분포 및 특징을 분석하였다.

## 4. 연구 결과

### 4.1 검색어 수 분석

〈표 1〉은 1년 동안 네이버에 입력된 질의별 검색어 수를 분석한 결과를 보여 준다.

질의별 평균 검색어 수는 분기별로 2개를 약

〈표 1〉 질의별 검색어 수의 기술통계

| | 2006년 8월 14일 | 2006년 11월 8일 | 2007년 2월 10일 | 2007년 5월 17일 | 2007년 8월 10일 |
|---|---|---|---|---|---|
| 평균 | 2.11 | 2.15 | 2.06 | 2.19 | 2.11 |
| 표준편차 | 1.09 | 1.11 | 1.14 | 1.18 | 1.09 |
| 최소값 | 1 | 1 | 1 | 1 | 1 |
| 최대값 | 9 | 8 | 13 | 11 | 8 |

간 넘는 수치를 유지하고 있어서, 1년 동안 그 수치가 대체로 일정한 것을 알 수 있다. 이를 통하여 웹 이용자들은 매우 적은 수의 검색어로 구성된 단순한 질의를 수행하고 있음을 알 수 있다. 한편 본 연구의 이러한 결과는 이준호, 박소연, 권혁성(2003)의 연구 결과와 유사한데, 이들은 2003년 1월 5일부터 1월 11일까지 네이버에 입력된 400만개 이상의 질의를 분석하였으며, 형태소 단위로 질의별 검색어 수 분석 시 평균 검색어 수가 2.03개임을 발견하였다. 반면, 2002년 6월 24일 입력된 네이버의 웹 검색 질의 54만 여개를 분석한 연구에서는(박소연, 이준호 2002) 질의별 평균 검색어 수가 2.54개임을 보고하였다. 분석된 전체 질의 수는 상이하지만, 네이버의 질의별 검색어 수를 동일한 방법으로 분석 시 2002년 이후 2003년 초까지는 질의별 평균 검색어 수가 감소하였고, 그 이후 2007년 8월까지는 질의별 평균 검색어 수가 큰 변화가 없이 대체로 일정한 수준을 유지하고 있음을 알 수 있다.

〈표 2〉는 질의별 검색어 수의 분포를 분기별로 보여준다. 오타나 오류 등으로 인하여 질의별 검색어 수 분석이 불가능한 질의가 일부 존재하였는데, 이 표에는 이처럼 결측값으로 처리된 질의들은 포함되어 있지 않다.

〈표 2〉를 통하여 전체 질의들 중 2개의 검색어로 구성된 질의가 가장 많으며, 대부분의 질의가 2개 이하의 검색어로 구성되어 있음을 알 수 있다. 6개 이상의 검색어로 구성된 긴 질의들을 조사한 결과, 이들 대부분이 클릭을 거의 발생시키지 않는 특수하고 비대중적인 질의인 것으로 나타났다. 즉 6개 이상의 검색어들로 구성된 45개의 질의들 중에서 31%에 해당하는 14개 질의에서 클릭이 전혀 발생하지 않았고, 87%에 해당하는 39개 질의에서 클릭이 100회 미만으로 발생하였다. 날짜별 평균 클릭 횟수를 초과하는 질의는 2개에 불과하였으며, 이들은 "어느 날 갑자기 두 번째 이야기 네 번째 층"과 "만원의 행복 강인 vs 강은비 2부"로 모두 엔터테인먼트 관련 질의였다.

### 4.2 주제 분석

날짜별로 네이버 이용자가 가장 많이 검색한 상위 5개 주제의 순위는 〈표 3〉과 같다.

〈표 3〉에 따르면, 분기별로 네이버 이용자가 가장 많이 검색한 주제는 엔터테인먼트이고, 1위와 2위 간의 격차가 상당히 큰 것으로 나타났다. 또한 쇼핑, 게임, 컴퓨터 관련 주제가 지속적으로 상위 5위에 포함되고 있었으며, 대학 입시가 수행되는 11월에는 교육, 학문 관련 주제의 비중이, 휴가철인 8월에는 라이프스타일 관

〈표 2〉 질의별 검색어 수 분포

| | 1개 | 2개 | 3개 | 4개 | 5개 | 6개 이상 |
|---|---|---|---|---|---|---|
| 2006년 8월 14일 | 217 (31.5%) | 276 (39.4%) | 126 (18.0%) | 50 (7.1%) | 11 (1.6%) | 8 (1.1%) |
| 2006년 11월 8일 | 207 (29.6%) | 283 (40.4%) | 138 (19.7%) | 41 (5.8%) | 15 (2.1%) | 10 (1.4%) |
| 2007년 2월 10일 | 243 (34.7%) | 279 (39.8%) | 109 (15.6%) | 45 (6.4%) | 18 (2.6%) | 6 (0.8%) |
| 2007년 5월 17일 | 204 (29.1%) | 283 (40.4%) | 135 (19.3%) | 45 (6.4%) | 19 (2.7%) | 10 (1.4%) |
| 2007년 8월 10일 | 235 (33.6%) | 262 (37.4%) | 126 (18.0%) | 40 (5.7%) | 18 (2.6%) | 11 (1.6%) |

〈표 3〉 날짜별 질의 주제 순위 비교

| | 2006년 8월 14일 (월) | | 2006년 11월 8일 (수) | | 2007년 2월 10일 (토) | | 2007년 5월 17일 (목) | | 2007년 8월 10일 (금) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 주제 | % | 주제 | % | 주제 | % | 주제 | % | 주제 | % |
| 1 | 엔터테인먼트 | 23.0 | 엔터테인먼트 | 20.9 | 엔터테인먼트 | 33.1 | 엔터테인먼트 | 23.1 | 엔터테인먼트 | 27 |
| 2 | 라이프스타일 | 10.6 | 교육, 학문 | 13.9 | 게임 | 12.7 | 교육, 학문 | 13.1 | 게임 | 10.9 |
| 3 | 쇼핑 | 8.9 | 쇼핑 | 11.3 | 쇼핑 | 10.7 | 컴퓨터 | 10.7 | 라이프스타일 | 9.9 |
| 4 | 게임 | 8.7 | 게임 | 8.4 | 컴퓨터 | 8.4 | 쇼핑 | 9.7 | 쇼핑 | 9.6 |
| 5 | 컴퓨터 | 8.1 | 컴퓨터 | 8.3 | 라이프스타일 | 7.7 | 게임 | 8.1 | 컴퓨터 | 9.3 |

련 주제의 비중이 높은 것을 알 수 있다. 토요일이며 활동성이 떨어지는 겨울인 2007년 2월 10일에 엔터테인먼트와 게임의 비중이 가장 높다는 사실을 주목할 만하다. 검색 포털 업체들은 이러한 결과를 콘텐츠 구축에 반영할 수 있을 것이다. 즉 이러한 결과는 엔터테인먼트 관련 콘텐츠는 상시적으로 요구되며, 계절별로 변화하는 정보 요구를 반영하여, 입시철에는 교육, 학문 관련, 휴가철에는 여행을 포함하는 라이프스타일 관련 콘텐츠를 강화하는 것이 필요함을 시사한다.

### 4.3 오타 분석

날짜별 오타의 빈도는 〈표 4〉와 같고, 오타 세부 유형의 분포는 〈표 5〉와 같다.

전체 질의 중 오타의 비중은 2% 내외로 전반적으로 매우 낮은 수준임을 알 수 있다. 또한 오타 중 가장 빈번하게 발생하는 유형은 교체, 삽입인 것으로 나타났다. 질의의 의미 파악이 안 되고, 검색 결과도 노출되지 않아서 오타 유형 분석이 불가능한 경우는 전체 질의의 0.5% 정도의 비중을 차지하는 것으로 나타났다.

〈표 4〉 오타의 빈도

| 2006년 8월 14일 | 2006년 11월 8일 | 2007년 2월 10일 | 2007년 5월 17일 | 2007년 8월 10일 |
|---|---|---|---|---|
| 11 (1.6%) | 18 (2.6%) | 20 (2.9%) | 15 (2.1%) | 19 (2.7%) |

〈표 5〉 오타 세부 유형 분포

| | 2006년 8월 14일 | 2006년 11월 8일 | 2007년 2월 10일 | 2007년 5월 17일 | 2007년 8월 10일 |
|---|---|---|---|---|---|
| 교체 | 3 (0.42%) | 6 (0.85%) | 11 (1.57%) | 5 (0.71%) | 2 (0.28%) |
| 삭제 | 1 (0.14%) | 1 (0.14%) | 3 (0.42%) | | |
| 삭제, 교체 | | | | 1 (0.14%) | 1 (0.14%) |
| 삽입 | 4 (0.57%) | 4 (0.57%) | 4 (0.57%) | 4 (0.57%) | 5 (0.71%) |
| 전치 | | | | 1 (0.14%) | |
| 영문모드한글입력 | 2 (0.28%) | 2 (0.28%) | 2 (0.28%) | 1 (0.14%) | 3 (0.42%) |
| 오류분석불가능 | 1 (0.14%) | 5 (0.71%) | | 3 (0.42%) | 8 (1.14%) |

### 4.4 클릭 행태 분석

〈표 6〉은 중복 질의 제거 시, 질의별로 발생한 클릭 횟수에 대한 기술통계를 보여 준다. 본 연구에 사용된 클릭 로그의 생성 단위는 중복 질의 제거 후의 고유한 개별 질의이다. 즉, 질의 로그에는 질의의 인기도 및 분포를 분석하기 위하여 하루 동안 자주 검색되는 질의들이 반복되어 저장되지만, 클릭 로그에는 특정 질의에 대해 이용자들이 하루 동안 조회한 문서들을 모두 수합하여 한 번만 저장하게 된다. 따라서 클릭 행태에 대한 분석은 이러한 클릭 로그의 형태를 반영하여 중복되는 질의 제거 후 개별 질의를 단위로 수행하는 것이 적절하다.

〈표 3〉을 통하여 시간이 지남에 따라 질의별 평균 클릭 횟수가 점진적으로 증가하는 추세임을 알 수 있다. 또한 클릭이 발생하지 않는 질의와 클릭 횟수가 수십만 회가 넘는 질의가 공존함으로 인하여 질의별 클릭 횟수의 편차가 매우 큰 것을 알 수 있다. 한편 질의별 클릭 횟수가 2006년 11월 8일에 다소 감소하였다가 이후 다시 증가하는 양상을 보이고 있다. 조회된 총 문서 수도 2006년 11월 8일이 가장 낮은 양상을 보이고 있다. 11월 8일이 주중 한 가운데인 수요일인데 비해, 2006년 8월 14일은 공휴일 전날이고, 2007년 8월 10일은 금요일이라는 점에서, 요일이 클릭 행태에 영향을 미친 것으로 추정된다.

전체 질의들 중 클릭이 전혀 발생하지 않은 질의는 2006년 8월 14일에 65개, 11월 8일에 79개, 2007년 5월 17일에 78개, 8월 10일에 67개 존재하였다. 이용자가 질의 입력 후 결과 문서를 조회하지 않는 이유로는 다음과 같은 사항을 고려할 수 있다. 첫째, 이용자가 오타나 특이한 질의를 입력하여 검색 결과가 출력되지 않는 경우이다. 이 연구에서 클릭이 0회인 질의들 중 30% 이상은 이 경우에 해당되는 것으로 보인다. 둘째, 이용자가 연예인명, 기업명, 영화명, 드라마명, 게임명 등을 입력할 경우, 특정 연예인이나 기업에 관한 기본 정보가 네이버의 자체 제작 콘텐츠 형태로 검색 결과에 제공된다. 이용자는 이처럼 간단한 정보를 검색 결과 화면으로부터 얻을 수 있기 때문에, 검색 결과로 출력된 문서를 조회하지 않을 수 있다. 셋째, 검색 결과에서 원하는 문서를 발견하지 못하여 검색 결과로 출력된 문서를 조회하지 않을 수 있다. 이 경우에는 검색 결과의 개선이 필요하기 때문에, 검색 결과가 출력되는데 조회하지

〈표 6〉 클릭 횟수의 기술통계

| | 2006년 8월 14일 (월) | 2006년 11월 8일 (수) | 2007년 5월 17일 (목) | 2007년 8월 10일 (금) |
|---|---|---|---|---|
| 평균 | 9,549 | 9,225 | 10,903 | 12,577 |
| 표준편차 | 38,290 | 59,451 | 73,659 | 55,662 |
| 최소값 | 0 | 0 | 0 | 0 |
| 최대값 | 538,035 | 893,001 | 1,460,820 | 762,180 |
| 조회된 문서 수 | 25,468 | 21,558 | 23,297 | 31,487 |
| 총 질의 수 | 635 | 645 | 642 | 662 |

않는 이유를 분석하기 위한 후속 연구가 필요할 것으로 보인다.

### 4.5 멀티미디어 검색 행태 분석

멀티미디어 검색어 기준에 따라 네이버의 멀티미디어 질의를 분석한 결과는 〈표 7〉과 같다. 〈표 7〉에 따르면, 2006년 8월 14일, 11월 8일, 2007년 2월 10일, 2007년 5월 17일과 8월 10일 전체 질의 중 멀티미디어 질의가 차지하는 비중은 각각 6.1%, 7.0%, 5.3%, 6.6%와 6%로 나타났다. 또한 모든 날짜에서 멀티미디어 질의 중 음악 질의가 차지하는 비중이 가장 높았다. 1년 동안 전반적인 멀티미디어 질의의 분포나 세부 유형의 분포는 큰 변화 없이 유사한 것으로 나타났다. 한편 전체 700개 질의들 중 중복되는 질의를 제거한 후, 멀티미디어 질의의 기술통계를 분석한 결과는 〈표 8〉과 같다. 멀티미디어 질의의 비중이 0.3%에서 0.7%까지 증가한 것을 제외하면, 〈표 8〉의 결과는 〈표 7〉의 결과와 유사하다고 할 수 있다. 〈표 7〉과 〈표 8〉에서 2007년 5월 17일과 2007년 8월 10일의 멀티미디어 질의에 대한 기술 통계는 박소연의 연구(2010)에서 보고된 바 있다. 한편 멀티미디어 질의의 길이를 일반 질의와 비교한 결과는 〈표 9〉와 같다.

전반적으로 동영상 질의의 길이가 가장 길고, 이어서 음악 질의, 이미지 질의, 일반 질의의 순으로 나타났다. 이처럼 멀티미디어 질의의 길이가 긴 것은 질의에 노래 제목, 드라마 제목, 영화 제목, 프로그램 제목 등이 포함되는 경우가 있기 때문인 것으로 판단된다. 3장의 연구 방법에서 논의된 주제 분류 체계에 따라 멀티미디어 질의의 주제를 살펴 본 결과, 전체 멀티미디어 질의 중 약 77%가 엔터테인먼트에 관한 질의로, 엔터테인먼트의 비중이 일반 질의보다 훨씬 더 큰 것을 알 수 있다.

한편, 2006년 5월 15일의 도그파일의 로그를

〈표 7〉 멀티미디어 질의의 기술통계(중복 질의 포함 시)

| | 이미지 질의 | 음악 질의 | 동영상 질의 | 총계 | 전체 질의수 |
|---|---|---|---|---|---|
| 2006년 8월 14일 | 11 (1.6%) | 25 (3.6%) | 7 (1.0%) | 43 (6.1%) | 700 |
| 2006년 11월 8일 | 8 (1.1%) | 27 (3.9%) | 14 (2.0%) | 49 (7.0%) | 700 |
| 2007년 2월 10일 | 8 (1.1%) | 19 (2.7%) | 11 (1.6%) | 38 (5.3%) | 700 |
| 2007년 5월 17일 | 12 (1.7%) | 26 (3.7%) | 8 (1.1%) | 46 (6.6%) | 700 |
| 2007년 8월 10일 | 6 (0.9%) | 25 (3.6%) | 11 (1.6%) | 42 (6.0%) | 700 |

〈표 8〉 멀티미디어 질의의 기술통계(중복 질의 제거 시)

| | 이미지 질의 | 음악 질의 | 동영상 질의 | 총계 | 전체 질의수 |
|---|---|---|---|---|---|
| 2006년 8월 14일 | 11 (1.7%) | 25 (3.9%) | 7 (1.1%) | 44 (6.8%) | 635 |
| 2006년 11월 8일 | 8 (1.2%) | 27 (4.2%) | 12 (1.9%) | 47 (7.3%) | 645 |
| 2007년 2월 10일 | 8 (1.3%) | 19 (3.0%) | 11 (1.8%) | 38 (6.0%) | 626 |
| 2007년 5월 17일 | 12 (1.9%) | 26 (4.0%) | 8 (1.2%) | 46 (7.2%) | 642 |
| 2007년 8월 10일 | 6 (0.9%) | 25 (3.8%) | 11 (1.7%) | 42 (6.3%) | 662 |

〈표 9〉 멀티미디어 질의의 길이 분석(중복 질의 포함 시)

| | 이미지 질의 | | 음악 질의 | | 동영상 질의 | | 일반 질의 | |
|---|---|---|---|---|---|---|---|---|
| | 검색어 수 평균 | 질의 수 | 검색어 수 평균 | 질의 수 | 검색어 수 평균 | 질의 수 | 검색어 수 평균 | 질의 수 |
| 2006년 8월 14일 | 2.18 | 11 | 2.52 | 25 | 3.00 | 7 | 2.09 | 645 |
| 2006년 11월 8일 | 2.75 | 8 | 2.63 | 27 | 2.79 | 14 | 2.11 | 646 |
| 2007년 2월 10일 | 2.87 | 8 | 2.68 | 19 | 3.36 | 11 | 2.01 | 662 |
| 2007년 5월 17일 | 2.83 | 12 | 3.00 | 26 | 3.00 | 8 | 2.13 | 650 |
| 2007년 8월 10일 | 1.50 | 6 | 3.12 | 25 | 3.18 | 11 | 2.06 | 650 |

분석한 Tjondronegoro, Spink, Jansen(2009)의 연구 결과에 따르면, 네이버와는 달리 전체 멀티미디어 질의들 중 이미지 질의의 비중이 가장 높았고, 이어서 음악, 동영상 질의 순으로 나타났다. 또한 멀티미디어 질의가 일반 질의보다 길이가 긴 것으로 조사되었는데, 질의의 길이에 있어서는 오디오 질의의 평균 검색어 수가 3.1로 가장 길고, 이미지 질의와 비디오 질의의 검색어 수 평균은 모두 2.3으로 나타났다. 상이한 주제 분류 체계를 사용하였기 때문에 직접적인 비교는 어렵지만, 이들은 멀티미디어 질의의 주제에 있어서 엔터테인먼트의 비중이 높았지만, 의학이나 스포츠, 기술 등의 주제도 이전보다 증가하는 추세라고 보고되었다. 따라서 멀티미디어 질의의 분포, 멀티미디어 질의의 길이, 멀티미디어 질의의 주제에 있어서 국내와 국외 연구 간에 차이가 있는 것으로 보인다.

#### 4.5.1 멀티미디어 클릭 로그 분석

2006년 8월 14일의 경우 이미지 컬렉션에 소속된 이미지 문서들에 대한 총 조회 수는 270,977회, 동영상 컬렉션에 소속된 동영상에 대한 총 조회 수는 207,174회, 음악 컬렉션에 대한 소속된 음악에 대한 총 조회 수는 184,259회로 나타났다. 2006년 11월 8일의 경우 동영상 컬렉션에서 발생된 총 조회 수는 1,137,078회, 이미지 컬렉션에서 발생된 총 조회 수는 20,250회, 음악 컬렉션에서 발생된 총 조회 수는 3,669회로 나타났다. 2007년 5월 17일의 경우 이미지 컬렉션에서 총 112,315회, 동영상 컬렉션에서 총 102,166회, 음악 컬렉션에서 총 15,330회의 조회가 발생하였다. 또한 2007년 8월 10일의 경우 동영상 컬렉션에서 총 122,829회, 이미지 컬렉션에서 총 103,986회, 음악 컬렉션에서 총 9,710회의 조회가 발생하였다. 모든 날짜의 클릭 로그를 종합한 결과, 멀티미디어 관련 컬렉션들 중 가장 많은 클릭이 발생한 컬렉션은 동영상 컬렉션이었고, 근소한 차이로 이미지 컬렉션이 뒤를 이었고, 음악 컬렉션에서 가장 적은 수의 클릭이 발생하였다.

멀티미디어 질의 분석 시 음악 관련 질의가 가장 큰 비중을 차지하였으나, 클릭 로그 분석 시에는 음악 컬렉션의 클릭 수가 가장 적었다. 조사 결과, 이는 이용자들이 음악 관련 질의를 입력 후 노출된 결과에서 음악 컬렉션보다는 블로그 컬렉션이나 카페 컬렉션 등을 선택하여 블로그나 카페에서 제공되는 음악을 감상하였기 때문인 것으로 보인다.

〈표 10〉 컬렉션별 클릭 수 1,000회 이상 질의 (중복 질의 포함 시)[1)]

| | 이미지 컬렉션 | 음악 컬렉션 | 동영상 컬렉션 | 총계 | 전체질의 수 |
|---|---|---|---|---|---|
| 2006년 8월 14일 | 30 (4.2%) | 16 (2.2%) | 36 (5%) | 82 (11.5%) | 715 |
| 2006년 11월 8일 | 6 (0.9%) | 1 (0.1%) | 18 (2.6%) | 25 (3.6%) | 703 |
| 2007년 5월 17일 | 49 (6.6%) | 3 (0.4%) | 42 (5.7%) | 94 (12.7%) | 738 |
| 2007년 8월 10일 | 31 (4.3%) | 3 (0.4%) | 24 (3.3%) | 58 (8.0%) | 720 |

〈표 10〉은 이미지, 음악, 동영상 컬렉션에서 각각 클릭 횟수가 1,000회 이상인 질의에 관한 기술통계를 보여 준다. 모든 날짜의 질의별 총 클릭 횟수의 중간값이 1,000회 이하이고, 이미지, 음악, 동영상 컬렉션에서의 평균 클릭 횟수가 모두 500회 이하임을 고려할 때, 클릭 횟수가 1,000회 이상인 질의는 멀티미디어에 관한 정보요구가 강한 인기 질의라고 볼 수 있다. 클릭 횟수가 1,000회 이상인 질의들은 음악 컬렉션보다 이미지 컬렉션과 동영상 컬렉션에서 더 많이 나타났다. 이 표에서 2007년 5월 17일과 2007년 8월 10일의 질의에 대한 기술 통계는 박소연의 연구(2010)에서 보고된 바 있으며, 1년 동안의 추이를 파악하기 위한 목적으로 이 표에 인용되었다. 이러한 질의들의 빈도는 중복 질의 포함 시에는 2007년 5월 17일이 가장 높고, 이어서 2006년 8월 14일, 2007년 8월 10일, 2006년 11월 8일 순으로 나타났다. 2006년 11월 8일의 경우 전체 클릭 횟수뿐만 아니라, 멀티미디어 컬렉션의 클릭 횟수도 가장 낮게 나타났는데, 이 날이 주중 수요일이라는 점에서 요일이 영향을 미친 것으로 추정된다.

또한 멀티미디어 컬렉션에서 이용자들이 많이 조회한 질의들의 주제도 대부분 엔터테인먼트와 관련되어 있는 것으로 나타났다. 즉 중복 질의 포함 시에는 멀티미디어 컬렉션에서 클릭이 발생한 질의들 중 74%가, 중복 질의 제거 시에는 56%가 엔터테인먼트에 대한 질의들이었다.

## 5. 결 론

이 연구에서는 국내 주요 검색 포털인 네이버 이용자들의 1년 동안의 검색 행태의 추이를 조사, 분석하였다. 이를 위하여 2006년 8월 14일부터 2007년 8월 10일까지 분기별로 수집한 네이버 통합 검색 질의 로그와 클릭 로그에 근거하여, 이용자의 질의 입력 행태, 오타 입력 행태, 클릭 행태, 멀티미디어 검색 행태 등과 같은 검색 행태의 추이를 분석하였다. 연구 결과, 질의의 길이 및 주제, 오타의 비율, 멀티미디어 질의의 비율 및 특징 등은 1년 동안 큰 변화 없이 일정한 것으로 나타났다. 반면, 질의별로 발생되는 클릭 횟수는 시간이 지남에 따라 점진적으로 증가하는 추세로 나타났으며, 주중보다 주말 또는 주말이나 공휴일 직전의 클릭 횟수가 높은 것으로 나타났다.

1) 한 질의가 동시에 여러 컬렉션에서 조회되는 경우, 표에 중복되어 포함되기 때문에 〈표 10〉의 질의 총계는 분석된 질의 수보다 약간 많다.

좀 더 구체적으로, 질의의 주제에 있어서는 분기별로 네이버 이용자가 가장 많이 검색한 주제는 엔터테인먼트였으며, 쇼핑, 게임, 컴퓨터 관련 주제가 지속적으로 상위 5위에 포함되고 있었다. 계절별로 주제의 순위에 약간의 변동은 있었지만 전반적인 주제 순위의 변화는 미약하였다. 질의별 평균 검색어 수에 있어서는 분기별로 약 2개의 수치를 유지하고 있어서, 네이버 이용자들이 매우 단순한 질의를 입력하고 있음을 보여준다. 전체 질의들 중 오타의 비율은 꾸준히 낮은 수준을 유지하고 있었다. 멀티미디어 질의에 있어서는 음악 질의가 차지하는 비중이 가장 높았으며, 이미지 질의와 동영상 질의의 비중은 비슷한 것으로 나타났다. 멀티미디어 성 질의는 일반 질의보다 길이가 길었으며, 이용자가 입력하는 멀티미디어 질의의 주제에 있어서는 엔터테인먼트 관련된 주제의 비중이 압도적으로 높았다. 한편 멀티미디어 관련 컬렉션들 중 이용자가 가장 많이 조회하는 컬렉션은 동영상 컬렉션이었고, 이어서 이미지 컬렉션, 음악 컬렉션 순으로 나타났다.

본 연구의 결과는 첫째, 웹 이용자들의 검색 행태 및 정보 요구에 대한 이해를 심화시킬 것으로 기대된다. 또한 본 연구는 검색 행태의 특징 및 추이를 분석하기 위한 방법론을 제시함으로써 웹 검색 분야에 학문적으로 기여할 수 있을 것으로 기대된다. 웹 검색 행태의 추이를 분석한 국내 선행 연구의 경우(박소연, 이준호 2005), 주제의 추이 분석에 집중하고, 질의의 주제 및 형태 분석 방법론을 제시한 반면, 이 연구에서는 검색어 수 분석 방법론 및 클릭 로그 분석 방법론 등을 제시하였다. 또한 주제뿐만 아니라, 질의의 길이, 오타 입력 행태, 클릭 행태, 멀티미디어 검색 행태와 같은 다양한 측면에서 검색 행태의 추이를 분석하였다. 둘째, 이 연구의 결과는 이용자의 향후 검색 행태와 정보 요구를 예측하는데 활용될 수 있으며, 포털들의 보다 효과적인 콘텐츠 구축 및 효율적인 검색 알고리즘 개발에 기여할 것으로 기대된다. 검색 시스템 개발 시 클릭 행태를 제외한 대부분의 검색 행태가 1년 동안 큰 변화 없이 안정적이라는 사실을 고려할 수 있을 것이다. 좀 더 구체적으로 검색 포털들의 콘텐츠 구축 시, 이용자들이 검색하는 질의에서 엔터테인먼트의 비중이 높고, 특히 대부분의 멀티미디어 질의가 엔터테인먼트와 관련되어 있다는 사실을 반영할 수 있을 것이다. 또한 계절별로 변화하는 정보 요구도 반영할 수 있을 것이다. 셋째, 검색 알고리즘이나 테스트컬렉션 개발 시, 웹 검색 질의의 대부분이 2개 이하의 검색어로 구성되어 있으며, 이용자의 클릭 행태도 짧은 질의에 집중된다는 사실을 반영할 수 있을 것이다. 일반적으로, 검색 포털이나 시스템들은 검색 시스템과 검색 기술의 평가를 위하여 테스트 컬렉션을 구축할 수 있는데, 검색 포털들이 테스트 컬렉션을 구축할 경우에도 이러한 특징을 반영할 수 있을 것이다. 넷째, 이용자가 자주 입력하는 오타 유형에 대한 분석을 통하여 이러한 유형에 대해서 자동 오타 교정 기능을 제공하고, 적합한 검색 결과를 제공함으로써, 검색 결과의 성능을 향상시킬 수 있을 것이다. 한편, 스마트폰 등의 등장으로 이용자들의 웹 자료에 대한 요구와 이용도가 더욱 증가할 것으로 예상되므로, 향후 이용자들의 검색 행태 추이에 관한 지속적인 연구가 요청된다.

본 연구의 수행 결과 향후 연구가 요구되는

사항들은 다음과 같다. 첫째, 후속 연구에서는 본 연구에서 제시한 방법론에 대한 검증 및 보완 작업이 요청된다. 둘째, 웹과 관련된 정보 환경이 급변한다는 점을 고려하여 보다 최신 로그 자료에 근거한 분석 작업이 바람직하다. 셋째, 본 연구에서는 1년에 걸친 검색 행태 추이를 분석하였으며, 향후 연구에서는 보다 장기간에 걸친 웹 검색 행태의 추이 분석을 통하여 본 연구의 연구 결과를 비교, 확인하는 작업이 요구된다. 마지막으로 국외 웹 이용자와 국내 웹 이용자의 검색 행태에 대한 심층적인 비교 분석이 요청된다.

## 참 고 문 헌

[1] 김용, 소민호. 2009. XML 기반의 동영상콘텐츠 검색 시스템 설계 및 구현.『정보관리학회지』, 26(4): 113-128.

[2] 노승민, 황인준. 2003. 멀티미디어 검색 시스템의 설계 및 구현.『한국정보과학회논문지: 데이타베이스』, 30(5): 494-506.

[3] 박소연. 2010. 국내 포털 이용자들의 멀티미디어 검색 행태 분석.『한국문헌정보학회지』, 44(1): 101-115.

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

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

[6] 박창섭. 2007. 의미적 연관성을 이용한 멀티미디어 정보 검색.『인터넷정보학회논문지』, 8(5): 67-79.

[7] 백우진, 정선은, 김기영, 안의근, 신문선. 2008. Content-based image retrieval using data fusion strategy.『정보관리학회지』, 25(2): 49-68.

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

[9] 정은경. 2010. A preliminary examination on the multimedia information needs and web searches of college students in Korea.『한국문헌정보학회지』, 44(4): 95-114.

[10] Arkin, H. & Colton, R. 1963. *Tables for statisticians*. New York: Barnes & Noble Inc.

[11] Damerau, F. 1963. "A technique for computer detection and correction of spelling errors." *Communications of the ACM*, 7(3): 171-176.

[12] Jansen, B. J. 2008. "Searching for digital images on the Web." *Journal of Documentation*, 64(1): 81-101.

[13] Jansen, B. J., Goodrum, A. & Spink, A. 2000. "Searching for multimedia: Analysis of audio, video, and image Web queries." *World Wide Web*, 3(4): 249-254.

[14] Jansen, B. J. & Spink, A. 2004. "An analysis of web searching by European AlltheWeb.com users." *Information Processing and Management*, 41(2): 361-381.

[15] Jansen, B. J. & Spink, A. 2005. "How are we searching the world wide web?: An analysis of nine search engine transaction logs." *Information Processing and Management*, 42(1): 248-263.

[16] Ozmutlu, S., Spink, A., & Ozmutlu, H.C. 2003. "Multimedia web searching trends: 1997-2001." *Information Processing and Management*, 39(4): 611-621.

[17] 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.

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

[19] Tjondronegoro, D., Spink, A., & Jansen, B. J. 2009. "A study and comparison of multimedia web searching: 1997-2006." *Journal of the American Society for Information Science and Technology*, 60(9): 1756-1768.

[20] Zobel, J. & Dart, P. 1996. "Phonetic string matching: Lessons from information retrieval." *Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 166-172. New York: Association for Computing Machinery.

**• 국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

[1] Kim, Yong & So, Min-Ho. 2009. "A study on implementation of XML-based information retrieval system for video contents." *Journal of the Korean Society for Information Management*, 26(4): 113-128.

[2] Rho, Seung Min & Hwang, Een Jun. 2003. "Design and implementation of multimedia retrieval a system." *Journal of KIISE: Database*, 30(5): 494-506.

[3] Park, Soyeon. 2010. "The multimedia searching behavior of Korean portal users." *Journal of the Korean Society for Library and Information Science*, 44(1): 101-115.

[4] Park, Soyeon & Lee, Joon Ho. 2002. "Investigating Web search behavior via query log

analysis." *Journal of the Korean Society for Information Management*, 19(3): 111-122.

[5] Park, Soyeon & Lee, Joon Ho. 2005. "Trends of search behavior of Korean web users." *Journal of the Korean Society for Library and Information Science*, 39(2): 147-160.

[6] Park, Chang-Sup. 2007. "Multimedia information retrieval using semantic relevancy." *Journal of the Korean Society for Information Managemen*t, 8(5): 67-79.

[7] Paik, WooJin, Jung Sun-Eun, Kim, Gi-Young, Ahn, Euigun, & Shin, Moon-Sun. 2008. "Content-based image retrieval using data fusion strategy." *Journal of the Korean Society for Information Managemen*t, 25(2): 49-68.

[8] Lee, Joon Ho, Park, Soyeon, & Kwon, Hyuk-Sung. 2003. "Information seeking behavior of the NAVER users via query log analysis." *Journal of the Korean Society for Information Managemen*t, 20(2): 27-40.

[9] Chung, Eunkyung. 2010. "A preliminary examination on the multimedia information needs and Web searches of college students in Korea." *Journal of the Korean Society for Library and Information Science*, 44(4): 95-114.