검색 포털의 클릭 집중 문서 분석 평가*

Analysis and Evaluation of Most Clicked Documents of Korean Search Portal

박 소 연(Soyeon Park)**

〈 목 차 〉

초 록

본 연구에서는 국내 주요 검색 포털인 네이버 통합 검색의 클릭 집중 문서의 특징을 조사, 분석하였다. 즉 이 연구에서는 클릭 집중 문서들을 대상으로 클릭 집중 비율, 컬렉션별 분포, 작성 연도별 분포를 조사하고, 문서의 적합도, 최신성, 신뢰도 등을 평가하였다. 이를 위하여 이용자들이 입력한 통합 검색 질의들로 구성된 질의 로그와 질의에 대한 검색 결과에서 이용자들이 조회한 문서를 기록한 클릭 로그를 분석하였다. 연구 결과, 클릭 집중 문서가 가장 많이 발생한 컬렉션은 블로그였으며, 질의별로 클릭의 절반가량이 한 문서에 집중되고 있는 것으로 나타났다. 또한 클릭 집중 문서의 적합도와 최신성은 상당히 높지만, 신뢰도는 보통 수준인 것으로 나타났다. 본 연구의 결과는 향후 포털의 효과적인 검색 알고리즘 및 인터페이스 개발에 활용될 수 있을 것으로 기대된다.

키워드: 웹 검색, 클릭 집중 문서, 검색 포털, 로그 분석

ABSTRACT

This study aims to investigate characteristics of most clicked documents of Naver's universal search service. In particular, this study analyzed characteristics of most clicked documents such as click ratio, collection distribution, and yearly distribution. Also, clicked documents were evaluated in terms of relevance, credibility, and currency. In conducting this study, query logs and click logs of unified search service were analyzed. The results of this study show that most clicks occurred in blog collection and average click concentration rate reached almost 50%. Also, the relevance and currency of most clicked documents were quite high, but credibility of these documents were on average level. The results of this study can be implemented to the portal's effective development of searching algorithm and interface.

Keywords: Web Searching, Most Clicked Documents, Search Portals, Log Analysis

* 본 연구는 NHN(주)의 지원을 받았음.
** 덕성여자대학교 문헌정보학과 부교수(sypark@duksung.ac.kr)
• 접수일: 2011년 2월 25일 • 최초심사일: 2011년 3월 9일 • 최종심사일: 2011년 3월 28일

Ⅰ. 서 론

웹 검색 포털들은 문서의 순위를 결정하기 위하여 검색 알고리즘을 활용하는데, 검색 알고리즘은 문서 내 검색어의 등장 빈도, 등장 위치, 문서의 조회 수와 같은 기준을 적용한다. 다양한 컬렉션들을 동시에 검색하는 포털들의 통합 검색 서비스의 경우, 1차적으로 문서의 순위를 도출하고 2차적으로 컬렉션의 순위를 도출한 후 이들을 종합하여 전체적인 검색 결과를 제공한다. 컬렉션 순위 결정 시 중요한 역할을 하는 요소들로는 컬렉션에 대한 이용자들의 클릭 빈도, 컬렉션 내에서의 이용자의 체류 시간 등을 들 수 있다.[1)2)] 즉 문서 검색 알고리즘과 컬렉션 검색 알고리즘에서 모두 결정적인 역할을 수행하는 요소가 문서와 컬렉션에 대한 클릭 빈도라고 할 수 있다. 조회 수가 많은 문서는 문서 순위의 상위에 배치되고, 조회 수가 많은 컬렉션은 전체 검색 결과에서 상위에 배치될 확률이 매우 높다. 이는 클릭 빈도가 특정 문서나 컬렉션의 인기도와 중요도를 반영하기 때문이다. 이러한 검색 알고리즘은 클릭 빈도와 문서의 품질, 또는 컬렉션의 품질 간에 밀접한 관계가 있다는 가정을 전제로 한다고 볼 수 있다. 한편 웹 검색 분야에서 다양한 연구가 수행되어 왔지만, 이러한 전제를 검토하기 위하여 클릭 문서의 특징 및 품질을 조사한 연구는 드문 실정이다.

이에 본 연구에서는 국내 주요 검색 포털인 네이버의 클릭 집중 문서의 특징을 조사, 분석하고자 한다. 좀 더 구체적으로 이 연구에서는 통합 검색 질의에 대해 클릭이 집중되는 문서들을 대상으로 클릭 집중률, 문서의 컬렉션별 분포, 문서의 작성 연도별 분포와 같은 특징을 분석하고, 적합도, 최신성, 신뢰도 등과 같은 문서의 품질을 평가하고자 한다. 또한 클릭 집중 문서들 중 오류 문서가 포함되어 있을 경우, 오류 문서의 유형 및 특징도 조사하고자 한다. 이를 위하여 포털의 검색 서비스 중 가장 이용도가 높은 통합 검색 질의들로 구성된 질의 로그와 이용자들이 조회한 문서들로 구성된 클릭 로그를 활용하고자 한다.

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

---

1) 유태명, 김준태, "링크 빈도와 클릭 빈도를 이용하는 메타 검색엔진의 설계," 한국정보과학회 봄 학술발표논문집, 제27권, 제1호(2000), pp.292-294.

2) S. Park, and J. Lee, "Unified search service of NAVER, a major Korean search engine," In: *the 31st SIGIR annual international ACM SIGIR Workshop on Aggregated Search*, edited by M. Lalmas, and V. Murdock, Singapore, 2008, pp.17-19.

Ⅱ. 선행 연구

국내외 웹 검색에 관한 연구는 다양한 분야에서 다양한 연구 방법을 활용하여 수행되어 왔다. 이들 중 클릭 데이터와 관련된 상당수의 국외 연구들은 전산학 분야에서 수행되었으며, 실험 참가자들이 클릭한 클릭 데이터로부터 적합성 피드백 정보를 수집하여 검색 성능 개선에 활용하는 실험 연구에 집중되어 왔다.[3][4][5][6] 또한 일반적인 웹 환경보다는 특정한 기관의 검색 시스템이나 검색 환경을 대상으로 수행되어 왔다. 예를 들어 Jung et al.은 오레곤 주립대학의 포털 사이트인 SERF(System for Electronic Recommendation Filtering)로부터 179개의 검색 세션을 추출하고 이 세션 내의 질의와 클릭 문서들을 대상으로 적합성 피드백의 기능 수행을 위한 클릭 데이터의 유용성을 조사하였다. 이들은 클릭 데이터가 검색의 정확률과 재현률을 향상시키고, 특히 가장 마지막으로 클릭된 문서가 적합 문서를 발견하는데 가장 유용하게 작용함을 발견하였다. 이러한 연구들에서 클릭 로그는 종종 클릭 쓰루(Clickthrough), 클릭 쓰루 데이터, 또는 클릭스트림 데이터로도 불린다.

Jansen과 Spink 등은 1990년대 후반부터 트랜잭션 로그 분석을 통하여 웹 이용자들의 검색 행태를 조사하는 일련의 연구를 수행하여 왔는데, 이들의 연구는 이용자들의 클릭 행태보다는 질의 입력 행태 분석에 치중해 왔다. 즉, 이들은 익사이트, 도그파일, 올더웹 등의 검색 엔진에 입력된 질의들을 대상으로 질의의 주제, 질의의 길이, 질의별 검색어 수 분포, 세션의 길이, 질의별로 조회한 결과 화면 수, 변경된 질의 수, 불리안 연산자 사용 행태 등을 분석하여 왔다.[7][8][9][10]

국내 선행 연구 중 클릭 로그를 활용한 연구로는 박소연, 이준호[11]의 연구를 들 수 있다. 이들은

3) T. Joachims, "In optimizing search engines using clickthrough data," In: *The 8th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, edited by* D. Hand, D. Keim, and R. Ng, Edmonton, Alberta, Canada, 2002, pp.133-142.

4) T. Joachims et al., "Accurately interpreting clickthrough data as implicit feedback," In: *the 28th annual international ACM SIGIR conference on Research and development in information retrieval,* edited by R. A. Baeza-Yates, N. Ziviani, G. Marchionini, Moffat, Al, and J. Tait, Salvador, Brazil, 2005, pp.154-161.

5) T. Joachims et al., "Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search," *ACM Transactions of Information Systems,* Vol.25, No.2(2007), doi: 〈http://doi.acm.org/10.1145/1229719.1229181〉.

6) S. Jung, J. L. Herlocker, J. L, and J. Webster, "Click data as implicit relevance feedback in web search," *Information Processing and Management,* Vol.43, No.3(2007), pp.791-807.

7) B. J. Jansen, and A. Spink, "An analysis of Web searching by European AlltheWeb.com users," *Information Processing and Management,* Vol.41, No.2(2004), pp.361-381.

8) B. J. Jansen, and A. Spink, "How are we searching the World Wide Web?: An analysis of nine search engine transaction logs," *Information Processing and Management,* Vol.42, No.1(2005), pp.248-263.

9) A. Spink, D. Wolfram, and B. J. Jansen, "Searching the Web: The public and their queries," *Journal of the American Society for Information Science and Technology,* Vol.52, No.3(2001), pp.226-234.

10) D. Tjondronegoro, A. Spink, and B. J. Jansen, "A study and comparison of multimedia Web searching: 1997-2006," *Journal of the American Society for Information Science and Technology,* Vol.60, No.9(2009), pp.1756-1768.

2003년 7월 1일부터 2004년 6월 30일까지 1년 동안 네이버에 입력된 통합 검색 질의들의 표본과 각 질의에 대한 검색 결과에서 이용자가 조회한 문서를 기록한 클릭 로그에 근거하여 국내 웹 검색 질의의 주제 및 형태를 분석하였다. 좀 더 구체적으로 이 기간 동안 격주로 수집된 질의들과 클릭 로그의 분석을 통하여 검색 질의의 주제와 형태를 계절별, 주중과 주말, 요일별로 비교하였다. 국내 포털 이용자들의 멀티미디어 검색 행태를 분석한 2010년 연구[12]에서는 멀티미디어 자료를 포함하고 있는 이미지, 음악, 동영상 컬렉션에서 발생한 클릭 행태의 분포와 특징을 분석하였다.

이처럼 국외 선행 연구들의 경우는 소수의 클릭 문서를 이용하여 검색 성능을 향상시키는 실험 연구가 수행되어 왔고, 국내 선행 연구들은 웹 이용자들의 전반적인 클릭 행태를 분석하거나, 질의의 주제나 형태 등을 분석하는데 클릭 로그를 활용하는 연구들이 수행된 바 있다. 즉, 국내외 선행 연구들 중에서 클릭 문서의 특성 및 품질을 상세하게 분석한 연구는 찾아보기 어려운 실정이다. 이에 이 연구에서는 국내 주요 검색 포털인 네이버 통합 검색의 클릭 집중 문서들을 대상으로 이들의 전반적인 특징을 분석하고, 적합도, 최신성, 신뢰도 등과 같은 문서의 품질을 평가하고자 한다.

Ⅲ. 연구 방법

1. 분석 데이터 수집

본 연구를 수행하기 위하여서는 질의의 수집과 이 질의들에 대한 클릭 문서의 수집이 필요하다. 첫째, 질의를 수집하기 위하여 국내 주요 검색 포털인 네이버에 입력된 통합 검색 질의들을 분석 대상으로 하였다. 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 점유율의 평균이 71%인 것으로 나타났다. 또한 2011년 1월 첫째 주 기준으로 검색 시간 점유율에 있어서도 네이버가 76%를 차지하는 것으로 나타났다. 따라서 네이버에 입력된 질의들이 국내 웹 이용자들이 입력한 질의들에 대한 대표성

11) 박소연, 이준호, 김지승, "클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석," 한국문헌정보학회지, 제39권, 제1호(2005. 6), pp.265-278.
12) 박소연, "국내 포털 이용자들의 멀티미디어 검색 행태 분석," 한국문헌정보학회지, 제44권, 제1호(2010. 3), pp.101-115.

을 지니고 있다고 판단되었다.

좀 더 구체적으로 이 연구에서는 2010년 6월부터 8월 첫째 주까지 네이버의 인기 검색어 사이트(http://searchc.naver.com/ntk/)에서 제공된 500개의 질의들을 분석대상으로 하였다. 네이버의 인기 검색어 사이트는 일정한 기간 동안의 통합 검색 질의들 중 이슈 검색어, 인기 검색어, 급상승 검색어 등을 수집하여 순위를 제공하는 서비스이다. 이 사이트로부터 질의를 수집한 이유는 그 속성상 이러한 질의들에 대해서 이용자들의 결과 문서 클릭 행위가 활발하게 수행될 것으로 기대되었기 때문이다. 다음의 검색 트렌드 사이트(http://trend.search.daum.net/SearchTrend/index.html)는 네이버의 인기 검색어 사이트와 유사한 성격의 서비스인데, 같은 기간 동안 두 사이트에서 제공되는 질의들이 대부분 일치하는 것으로 나타났다. 500개의 질의를 선택한 이유는 하루에 네이버에 입력되는 통합 검색 질의의 수를 고려할 때,[13] 신뢰 수준 95% 표본 오차 ±4.4%를 허용할 경우 필요한 표본의 크기가 약 500개인 것으로 통계학 문헌에서 제시되고 있기 때문이다.[14] 한편 500개 질의의 주제를 선행 연구에서[15] 도출된 주제 분류 체계에 따라 분석한 결과, 전체 질의 중 엔터테인먼트와 관련된 주제가 66.7%로 가장 많았고, 이어서 라이프스타일(8.8%), 스포츠(7.8%), 뉴스/미디어(6.8%), 쇼핑(3.2%)순으로 나타났다. 500개 질의의 예로는 "갤럭시U", "배다해", "앙드레김 아들", "박칼린", "로또400", "부부젤라", "아이폰4 출시일", "금값 하락", "가인 닮은 꼴", "인기가요 방송사고" 등을 들 수 있다.

본 연구에서는 이렇게 선택된 질의들에 대하여 질의 수집 직후인 2010년 8월 13일 하루 동안 네이버 이용자가 클릭한 문서들을 수집하였다. 특히 특정 질의에 대해 이용자가 클릭한 문서들 중 클릭이 가장 많이 발생한 클릭 집중 문서를 분석 대상으로 하였다. 네이버의 클릭 로그를 분석 대상으로 한 이유는 첫째, 위에서도 언급되었듯이 네이버의 위상 때문이며, 둘째, 이용자들의 검색 행태가 기록된 로그가 대외비로 간주되어 확보되기 어려운 현실에서, 네이버로부터 클릭 로그가 제공되었기 때문이다. 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.

---

13) 네이버에 따르면 2011년 1월 기준으로 하루에 입력되는 질의 수는 1억 3000만개이며, 이 중 통합 검색 질의가 가장 큰 비중을 차지하고 있다. NHN Home page, 〈http://www.nhncorp.com/nhn/service/naver.nhn〉.

14) H. Arkin, and R. Colton, *Tables for Statisticians*(New York : Barnes & Noble Inc., 1963).

15) 박소연, 이준호, 김지승, 전게논문, pp.269-270.

2. 평가 기준

본 연구에서는 클릭 집중 문서들을 적합도, 신뢰도, 최신성 등의 기준에 따라 평가하였다. 적합도는 정보 검색 분야에서 검색 성능을 평가하는 중요한 척도로 오랜 기간 동안 사용되어 왔으며, 신뢰도와 최신성은 특히 웹 문서 평가에 있어서 중요한 평가 기준이기 때문에 본 연구에서는 이 세 가지 기준을 적용하였다. 적합도 평가 방식은 '적합', '부적합'의 이분법적인 방식이 아닌 '적합', '보통', '부적합'의 3점 척도를 사용하였다. 적합도 평가 시 3점 척도나 5점 척도를 사용할 경우 적합도의 정도를 보다 상세하고 섬세하게 표현할 수 있어서 많은 정보 검색 분야 선행 연구들에서 3점 척도나 5점 척도를 통한 적합도 평가를 수행하여 왔다.[16)17)] 또한 3점 척도 사용 시 아래 〈표 1〉과 〈표 2〉와 같이 점수별로 상세한 기준을 설정하는 것이 가능하고, 평가자들 간의 평가의 일관성을 유지할 수 있다는 점에서 3점 척도 평가 방식을 적용하였다. 동일한 맥락에서, 이 연구에서는 신뢰도와 최신성 평가에도 '높음', '보통', '낮음'의 3점 척도를 적용하였다. 적합도와 신뢰도 평가를 위하여 선행 연구들에서 사용된 평가 기준을[18)19)] 수정, 보완하여 〈표 1〉과 〈표 2〉와 같은 평가 기준을 도출하였다. 최신성의 경우, 연구에 사용된 질의들이 2010년 6월부터 8월 중순까지 인터넷 상에서 이슈가 되었던 질의라는 점을 고려하여, 문서에 수록된 내용의 시점이 2010년인 경우 최신성이 높은 문서로, 2009인 경우 보통으로, 2008년 및 그 이전인 경우 최신성이 낮은 문서로 평가하였다. 문서의 작성 연도 분석과 별도로 최신성 분석을 수행한 이유는 문서의 작성일과 최신성이 일치하지 않는 경우가 있기 때문이다. 예를 들어, 문서의 작성 연도는 2010년이지만, 그 안에 수록된 내용이 2000년대 초반이나 중반 시점인 경우가 존재하기 때문이다.

〈표 1〉 적합도 평가 기준

평가	내 용
적합	- 결과의 전체적인 내용이 질의와 일치하거나 밀접한 관련이 있는 경우 예) 질의 : 서울대학병원 결과 : 서울대학 병원 위치, 진료시간, 의료진 등의 내용 포함 - 결과 문서에 질의가 모두 등장하지는 않지만, 핵심 질의어가 등장하고, 결과 문서가 질의에 부합하는 내용을 충실히 담고 있는 경우

16) 노정순, "Invisible Web 탐색도구의 성능 비교 및 분석," 정보관리학회지, 제21권, 제3호(2004. 9), pp.203-225.
17) 맹성현 등, "정보 검색 시스템 평가를 위한 균형 테스트 컬렉션 구축," 정보관리학회지, 제16권, 제2호(1999. 6), pp.135-148.
18) 박소연, 이준호, "주요 검색 포탈들의 통합 검색 서비스 비교 평가," 한국도서관·정보학회지, 제39권, 제1호(2008. 3), pp.265-278. 이 연구에서는 통합 검색 결과의 적합도 평가 기준을 제시하였음.
19) 박소연, 이준호, 전지운, "지식 검색 서비스 개선을 위한 문서의 적합도 및 신뢰도 분석," 한국문헌정보학회지, 제40권, 제2호(2006. 6), pp.299-314. 이 연구에서는 지식 검색 서비스를 구성하는 지식 문서의 적합도 평가 기준과 답변의 신뢰도 평가 기준을 제시하였음.

평가	내 용
보통	- 결과의 내용이 질의와 일부 관련이 있는 경우 예) 질의 : 제모 비용 변화 결과 : 제모 비용에 관한 결과만 제공될 때(변화에 관한 정보는 제공 안 됨) - 상품명이 질의인 경우 상품명과 일부만 일치하는 결과가 제공되는 경우 예) 질의: 슬래드 1300 결과: "슬래드"에 관한 일반 정보만 제공될 때(슬래드 1300 버전에 관한 구체적인 정보가 제공 안 됨) - 결과 문서에 질의가 모두 등장하지만, 질의와 관련된 내용이 매우 간략하게 처리된 경우(즉 전체 문서 내용 중 질의와 관련된 내용의 비중이 매우 작은 경우) - 결과 문서에 질의의 일부만 등장하고, 문서의 내용이 질의의 의도를 직접적으로 충족시키지는 않으나, 관련된 정보를 제공하는 경우
부적합	- 결과의 내용이 질의와 전혀 일치하지 않거나 관련이 없는 경우 예) 질의 : 주온 결과 : 주전층 - 결과 문서에 질의의 일부가 등장하지만, 문서의 전반적인 내용이 질의와 관련 없고, 부가적인 정보만 수록된 경우

〈표 2〉 신뢰도 평가 기준

평 가	내 용
높음	- 공신력 있는 정확한 출처가 있는 경우 → 신문기사의 경우 신문사명, 날짜, 작성자가 표기되어 있을 때 → 논문, 레포트의 경우 공신력 있는 기관이나 사이트의 자료일 때 예) DBPia → URL이 표기되어있을 때 - 객관적으로 확실한 근거가 있는 경우(이론적, 학문적인 예시 등) - 이미지, 동영상 자료의 경우 출처가 명확하게 제시된 경우(로고가 표시된 경우 포함) → 캡쳐 사진의 경우 방송 로고가 이미지에 표시되었거나 해당 방송 정보가 언급되어 있을 때
보통	- 출처 정보가 불완전한 경우 → 신문기사의 경우 신문사명, 날짜, 작성자 중 누락된 정보가 있을 때 - 논문, 레포트의 경우 권위가 없거나 공신력이 부족한 기관이나 사이트의 자료일 때 예) 해피캠퍼스 - 정확한 출처가 나오지 않고 작성자의 의견에 의존하나 어느 정도 논리적인 경우 - 근거가 부족한 답변을 한 경우 - 학문적 근거는 없지만 속담, 격언, 생활지식 등과 같은 상식적인 수준의 정보를 제공하는 경우 - 팬픽 글 - 이미지, 동영상 자료의 경우 출처가 표시되어 있지 않으나 추측이 가능한 경우 - 출처가 표시된 이미지와 표시되지 않은 이미지가 공존할 경우
낮음	- 근거가 없는 개인 의견 - 추측성 답변 - 비방, 욕설, 음란한 글 - 명예 훼손성 글 - 상황문답놀이 글 - 이미지, 동영상 자료의 출처가 없는 경우

이 연구에서는 이처럼 도출된 평가 기준에 대한 상세한 가이드라인을 작성하였으며, 문헌정보학과 전공자들로 구성된 두 명의 평가자들이 이 가이드라인에 따라 2011년 1월과 2월 동안 클릭 집중 문서에 대한 평가 작업을 수행하였다. 평가자들 간의 평가 일치성은 약 95%로 매우 높은 것으로 나타났으며, 평가가 불일치하는 경우 문서의 재검토와 토론을 통하여 합의에 이르는 과정을 거쳤다.

Ⅳ. 연구 결과

1. 클릭 집중 문서의 분포

본 연구에서 분석된 500개의 문서에 대한 클릭 집중률의 평균은 48%로 나타났다. 이는 질의별로 발생하는 클릭의 절반가량이 한 문서에 집중되고 있음을 의미하며, 이용자들의 조회 행태에 있어서 쏠림 현상이 심하다는 것을 시사한다. 클릭 집중률의 표준 편차는 30%로 질의별로 클릭 집중률의 편차가 매우 크다는 사실을 알 수 있다.

〈표 3〉은 통합 검색 클릭 집중 문서가 소속된 컬렉션들의 분포를 보여 준다. 이 표에서 자체 제작 컨텐츠란 "인물 정보", "공연 정보", "영화 정보", "방송 정보", "기업 정보" 등과 같이 미리 지정해 놓은 키워드에 대해 네이버와 같은 포털들이 직접 제작하는 컨텐츠를 총칭하며, 일반적으로 검색 결과 화면의 상단에 배치된다. 조사 결과, 클릭 집중 문서가 가장 많이 발생한 컬렉션은 블로그이며, 이어서 뉴스, 지식iN, 이미지, 사이트 순으로 나타났다. 특히 블로그는 전체 클릭 집중 문서의 약 43%를 차지하며, 이 중 네이버의 블로그는 전체 문서의 29%를 차지하는 것으로 나타났다. 이 연구에 사용된 질의들이 특정한 시기에 이슈가 되었던 질의라는 점을 고려할 때 뉴스 컬렉션보다 블로그 컬렉션에 클릭이 집중되었다는 사실은 주목할 만하다. 집중적으로 조회된 블로그 문서들의 절반 이상은 독자적인 내용 없이 질의와 관련된 이미지나 동영상 자료만을 모아놓은 경우였으며, 그 다음으로는 특정한 이슈에 대해 자신의 의견만을 올려놓는 경우도 큰 비중을 차지하였다. 블로그에 이어 전체 클릭 집중 문서 중 큰 비중을 차지하는 뉴스의 경우 지명도와 권위가 부족한 인터넷 매체들이 큰 비중을 차지하고 있었다. 즉 조회된 91개의 뉴스 문서들 중에서 지면 신문을 발행하는 신문사의 신문 기사는 11%에 불과한 반면, 온라인으로만 뉴스를 제공하는 뉴스 사이트들의 기사는 89%에 해당하는 81개로 나타났다. 한편, 스포츠 관련 질의들의 경우, 동영상 컬렉션의 자료가 주로 클릭되는 것으로 나타났다.

〈표 3〉 클릭 집중 문서의 컬렉션 별 분포

컬렉션	빈도	%
블로그	216	43.2
뉴스	91	18.2
지식iN	62	12.4
이미지	35	7.0
사이트	32	6.4
카페	32	6.4
동영상	9	1.8
바로가기	4	0.8
음악	3	0.6
자체 제작 컨텐츠	3	0.6
사전	2	0.4
지식백과	1	0.2
지식쇼핑	1	0.2
책	1	0.2
총 계	700	100

클릭 집중 문서의 작성 연도 별 분포는 〈표 4〉와 같다. 문서의 작성 연도가 2010년인 경우가 70%로 대부분이지만, 2009년인 경우가 4%, 2008년 이전인 경우도 5%를 차지하는 것으로 나타났다. 문서의 작성 연도를 파악하기 어려운 경우도 상당 수 존재하였는데, 이는 문서 내부나 결과 화면에 작성 연도가 명시되지 않거나, 클릭된 문서가 오류 문서여서 작성 연도를 파악하기 어렵기 때문이다. 4.3절에서 상세히 논의되겠지만 전체 클릭 집중 문서 중 오류 문서는 14.8%인 74개로 조사되었다. 이처럼 작성 연도가 불분명하여 검색 결과에 노출되지 않을 경우 이용자들이 문서의 최신성을 파악하는데 어려움이 있으므로, 이에 대한 개선이 요청된다.

〈표 4〉 클릭 집중 문서의 작성 연도 별 분포

작성 연도	빈도	%
2003	2	0.4
2004	4	0.8
2005	4	0.8
2006	3	0.6
2007	6	1.2
2008	8	1.6
2009	21	4.2
2010	352	70.4
불분명	100	20
총 계	700	100

2. 적합도, 최신성, 신뢰도 평가 결과

3장의 연구 방법에서 제시된 평가 기준에 따라 평가된 클릭 집중 문서의 적합도, 최신성, 신뢰도 평가 결과는 〈표 5〉와 같다.

〈표 5〉 적합도, 최신성, 신뢰도 평가 결과

	적합도	최신성	신뢰도
평균	2.78	2.82	2.00
표준편차	0.53	0.52	0.87

클릭 집중 문서의 적합도 평균은 3점 척도에 2.78점으로 상당히 높은 것으로 나타났다. 그러나 전체 문서들 중 부적합한 문서들도 4.4.%(n=22)를 차지하고 있었다. 예를 들어 "보들보들 계란찜"이란 질의에 대해 클릭이 집중된 문서는 푸딩을 만들다가 실수로 계란찜처럼 되었는데 표면이 고운 푸딩을 어떻게 만들어야 하는지를 문의하는 지식iN 문서로 질의와는 무관하다고 볼 수 있다. 또한 "듀퐁 김아중"이라는 질의는 배우 김아중이 프랑스 명품 브랜드 듀퐁의 홍보대사가 된 사실과 관련이 있는데, 클릭 집중 문서는 드라마 '자이언트'에서 배우 주상욱이 '듀퐁' 브랜드를 입고 나온 사실을 소개하는 블로그로 질의에 부적합한 문서로 평가되었다. 또한 "신봉선 남자친구"라는 질의의 클릭 집중 문서는 가수 홍진영에 관한 신문 기사로 '신봉선'과 '남자친구'라는 질의어가 문서에 등장하기는 하지만, 질의와 관련이 없는 부적합한 문서였다. 한편 클릭 집중 문서가 10개 이상 소속된 주요 컬렉션들의 적합도 평균을 비교한 결과는 〈표 6〉과 같다. 위의 예를 통해서 나타나듯이 지식iN 소속 문서들의 적합도가 다른 컬렉션보다 낮음을 알 수 있다.

〈표 6〉 주요 컬렉션들의 적합도 비교

컬렉션	뉴스	블로그	사이트	이미지	지식iN	카페
적합도	2.82	2.81	2.78	2.85	2.56	2.68
문서 수	86	184	28	35	59	22

클릭 집중 문서들의 최신성은 2.82로 매우 높은 것으로 나타났다. 그러나 최신성이 낮은 경우도 일부 존재하였는데(5.4%, n=27), 예를 들어 "주말 TV 영화"라는 질의에 대해 집중적으로 클릭된 문서는 2005년 3월 4일의 주말 TV 영화를 소개하는 동아일보 기사였다. 또한 "원빈 삭발"이라는 질의의 정보 요구는 영화 "아저씨"에서 원빈이 삭발을 한 사실과 관련된 것인데, 클릭된 문서는 2004년에 영화와 CF 등에서 원빈의 반삭 스타일을 다룬 2004년 8월 6일의 지식iN 문서였다. 주요

컬렉션들의 최신성 평균을 비교한 결과는 〈표 7〉과 같다. 자료의 특성 상 뉴스의 최신성이 가장 높았고, 이어서 블로그, 사이트, 이미지, 카페 순으로 나타났다. 위의 예와 같이 최신성이 떨어지는 일부 지식iN 문서들로 인하여 지식iN의 최신성이 가장 낮았다.

〈표 7〉 주요 컬렉션들의 최신성 비교

컬렉션	뉴스	블로그	사이트	이미지	지식iN	카페
최신성	2.99	2.94	2.89	2.85	2.12	2.59
문서 수	86	184	28	35	59	22

문서의 신뢰도 평균은 2.0으로 보통 수준인 것으로 나타났다. 이는 블로그나 지식iN, 카페 등에서 출처가 불분명하거나 불완전한 글이 많고, 정확한 근거 없이 본인의 개인적인 의견이나 추측을 올려놓은 경우가 많기 때문인 것으로 보인다. 예를 들어, "유재석 출연료"란 질의의 경우, 2010년 중반을 기준으로 유재석이 KBS, SBS, MBC에서 각각 받는 출연료를 알려달라는 지식iN 질문이었는데, 질문자가 선택한 답변은 정확한 근거를 제시하지 않은 채 유재석의 2010년 출연료가 올랐을 것으로 추정하는 추측성 답변이었다. 또한 "선풍기 에어컨 효과"나 "백두산 폭발 가능성"과 같은 질의에 대해 집중적으로 클릭된 문서들은 이러한 이슈들에 대한 뉴스나 신문 기사를 언급하고 있으나, 이에 대한 어떠한 출처도 명시되지 않아 문서의 신뢰도를 저하시키고 있었다. 또한 이미지나 동영상 자료를 게시한 블로그들의 경우에도 출처가 없거나 불완전한 경우가 많아 문서의 신뢰도가 낮게 평가 되었다. 주요 컬렉션들의 신뢰도 평균을 비교한 결과는 〈표 8〉과 같다. 뉴스의 신뢰도가 가장 높았고, 블로그, 지식iN, 카페의 신뢰도가 낮게 조사되었다.

〈표 8〉 주요 컬렉션들의 신뢰도 비교

컬렉션	뉴스	블로그	사이트	이미지	지식iN	카페
최신성	2.96	1.70	2.25	2.03	1.52	1.45
문서 수	86	181	28	29	59	22

3. 클릭 집중 문서 중 오류 문서 분석

오류 문서란 정상적인 웹 문서가 아니며, 스팸 문서, 접속 시 오류가 발생하는 문서, 문서 수집 시 오류가 발생한 문서, 컨텐츠가 훼손된 문서 등과 같이 문제가 있는 문서를 의미한다. 본 연구에서 분석된 클릭 집중 문서들 중에는 오류 문서도 포함되어 있었는데, 〈표 9〉는 클릭 집중 문서들에 포함된 오류 문서의 유형 및 분포를 보여준다. 웹 오류 문서의 유형은 문서에 대한 상세한 분석과

네이버 실무자와의 논의를 통해 도출되었다.

〈표 9〉 클릭 집중 문서 중 오류 문서의 유형 별 분포

오류 문서 유형	빈도	%
접속 에러 및 본문 삭제	52	10.4
스팸(홍보성, 광고성 사이트) 문서	1	0.2
문서 수집 오류	22	2.4
문서의 내용은 없고, 목록만 존재하는 경우	1	0.2
본문 외의 영역에서 질의가 등장하는 경우	1	0.2
문서가 네이버 검색 결과 화면인 경우	5	1.0
문서가 네이버 메인 페이지인 경우	3	0.6
본문 글이 여러 개 있으며 질의에 적합한 결과가 없는 경우	1	0.2
문서 내 이미지, 동영상 등의 컨텐츠가 훼손된 경우	5	1.0
본문 외의 영역에서 질의가 등장하며, 문서 내 이미지, 동영상 등의 컨텐츠가 훼손된 경우	1	0.2
문서의 내용은 없고 링크만 있는 경우	2	0.4
로그인 에러: 비공개 카페들로 카페 운영자가 초대한 사람만 멤버 가입이 가능하여 문서 열람이 매우 제한적인 경우	2	0.4

전체 클릭 집중 문서들 중 오류 문서는 14.8%인 74개였으며, 조회된 오류 문서 중 가장 큰 비중을 차지하는 유형은 접속 시 에러가 발생하거나 본문이 삭제된 경우였다(10.4%). 이들 중 상당수는 블로그나 카페 문서로, 이러한 오류가 발생한 것은 자료 수집 시점과 평가 시점 간의 시차에 기인한 것으로 보인다. 즉, 블로그나 카페 문서의 큰 특징인 유동성과 휘발성으로 인하여, 자료 수집 시점 이후에 해당 문서나 게시 글이 삭제되거나 이동되었을 가능성이 크다. 두 번째 유형은 문서 수집 오류로 전체 문서의 2.4%인 22개로 나타났다. 이 유형은 웹 문서 수집 제외 대상에 해당함에도 불구하고, 오류로 인하여 수집된 경우에 해당하며, 문서의 내용은 없고 목록만 존재하는 경우, 네이버 메인 페이지가 저장된 경우 등을 예로 들 수 있다. 또한 문서 내 이미지, 동영상 등의 컨텐츠가 훼손된 경우, 문서의 내용은 없고 링크만 있는 경우, 카페의 로그인 에러는 각각 1%, 0.4%, 0.4%로 나타났다. 복수의 오류가 동시에 발생한 경우도 있었는데, 본문 외의 영역에서 질의가 등장하며, 문서 내 이미지, 동영상 등의 컨텐츠가 훼손된 경우도 있었다. 검색 결과에 오류 문서가 포함될 경우 검색의 성능이 저하되고 이용자의 만족도가 감소되므로 오류 문서 발생에 대한 상세한 분석을 통하여 이들을 최소화시킬 수 있는 방안이 모색되어야 할 것이다.

4. 기타

클릭 집중 문서들 중에 개별 문서가 아닌 특정한 컬렉션의 검색 결과 화면이 클릭된 경우는 전체

문서의 2.8%인 14개로 조사되었다. 예를 들어, '재범 프로필 사진'과 '강인 입대'와 같은 연예인 관련 질의의 경우, 이미지 검색 결과 화면에 클릭이 집중되고, '아이폰 수화기'와 같은 상품 관련 질의의 경우 쇼핑 검색 결과 화면에 클릭이 집중되었다. 이처럼 특정한 문서를 선택하기 이전에 검색 결과 화면에서 클릭 행태가 중단되는 현상은 이용자들의 검색 행태가 매우 단순화되고 있음을 시사한다.

한편, 질의의 입력 횟수와 해당 질의의 결과에 대한 조회 수 간에는 매우 강한 상관관계를 보이고 있었다, $r=0.96$, $p<0.001$. 즉 많이 검색되는 질의일수록 결과도 많이 조회되고 있었다. 그런데 질의 입력 횟수에 비해 문서의 조회 수가 적은 경우가 존재하였는데, "로또397", "로또400", "로또395"와 같은 로또 관련 질의의 경우 검색은 많이 되는 반면, 결과 클릭 횟수는 매우 낮았다. "로또397"의 경우 하루 동안 질의가 281회 입력된 반면 클릭은 단 1회 발생하였다. 이용자들이 로또 관련 질의를 입력하는 이유는 대부분 해당 회차의 당첨 번호를 알기 위해서인데, 현재 네이버의 검색 결과 화면에는 '나눔 로또 당첨번호 정보'의 형태로 해당 회차의 당첨 번호를 노출하고 있다. 따라서 최초 검색 결과 화면에서 정보 요구가 충족되기 때문에 더 이상의 클릭이 발생하지 않는 것으로 판단된다.

V. 결 론

본 연구에서는 국내 주요 검색 포털인 네이버에서 조회된 클릭 집중 문서의 특징을 조사, 분석하였다. 이를 위하여 2010년 6월부터 8월 초까지 수집된 이슈성 질의들에 대하여 2010년 8월 13일 하루 동안 클릭된 문서들을 수집하고, 특히 클릭된 문서들 중 클릭이 가장 많이 발생한 클릭 집중 문서들을 분석하였다. 연구 결과, 이 연구에서 분석된 문서들에 대한 클릭 집중률의 평균은 48%로, 질의별로 발생한 클릭의 절반 정도가 한 문서에 집중되는 것으로 나타났다. 또한 클릭 집중 문서가 가장 많이 발생한 컬렉션은 블로그이며, 이어서 뉴스, 지식iN, 이미지, 사이트 순으로 나타났다. 클릭 집중 문서의 적합도와 최신성 평균은 상당히 높은 것으로 나타났다. 그러나 일부 부적합한 문서들과 최신성이 떨어지는 문서들도 존재하였다. 문서의 신뢰도 평균은 보통 수준인 것으로 나타났으며, 이는 블로그나 지식iN, 카페 등에서 출처가 불분명하거나 불완전한 글이 많고, 정확한 근거 없이 본인의 개인적인 의견이나 추측을 올려놓은 경우가 많기 때문인 것으로 보인다. 지식iN 소속 문서들의 경우, 적합도, 최신성, 신뢰도가 타 컬렉션들보다 낮은 것으로 나타났다. 전체 클릭 집중 문서들 중 오류 문서의 경우 접속 시 에러가 발생하거나 본문이 삭제된 경우, 문서 수집 오류가 큰 비중을 차지하였다.

본 연구의 결과는 포털들의 보다 효율적인 검색 알고리즘 개발 및 인터페이스 개발에 기여할 것

으로 기대된다. 질의 입력 후 이용자들의 클릭의 절반 정도가 한 문서에 집중된다는 사실을 고려할 때, 통합 검색 결과 화면을 보다 간결하게 제공할 필요성이 있을 것으로 보인다. 현재 대부분의 포털들의 통합 검색에서는 십여 개의 컬렉션들로부터 검색된 결과가 제공되며, 이용자들이 결과 화면을 여러 번 스크롤해야하는 상황이 발생하기도 한다. 따라서 중복되는 결과를 제외하고, 스크롤 길이를 줄이는 등 결과 제공 방식을 간결하게 만들 필요성이 있을 것으로 보인다. 또한 최초 검색 결과 노출 후 이용자들이 클릭을 중단하거나, 개별 문서를 클릭하기 이전에 특정 컬렉션의 검색 결과 화면만을 클릭한 후 클릭 행태가 중단되는 등 이용자들의 클릭 행태가 매우 단순화되고 있음을 고려하여, 검색 결과의 첫 페이지에 최대한 적합한 결과를 제공하는 것이 중요할 것으로 보인다. 포털들이 제공하는 자체 제작 컨텐츠들이 이 연구에 사용된 이슈성 질의에는 유용하게 작용하는 것으로 보이며, 이러한 자체 제작 컨텐츠의 품질을 강화하는 작업이 중요할 것으로 사료된다.

둘째, 이 연구에서 조사되었듯이 클릭 집중 문서들의 품질이 완벽하지는 않기 때문에 검색 알고리즘에서 클릭 빈도만을 무조건적으로 반영하는 것에 대해서는 검토가 요청된다. 특히 자주 입력되는 질의들에 대한 클릭 문서들의 경우, 신중한 품질 분석 및 평가가 바람직할 것으로 보인다. 또한, 검색 결과의 신뢰도, 적합도와 최신성에 있어서 개선의 여지가 있으며, 특히 지식iN, 블로그 컬렉션 검색 결과의 신뢰도 개선이 시급한 것으로 보인다. 마지막으로, 자주 발생하는 웹 오류 문서 유형들을 검색 결과에서 제외할 수 있는 방안을 모색함으로써, 검색 결과의 성능을 향상시키고 이용자의 만족도를 제고하도록 할 필요성이 있다. 특히 문서 수집 오류의 비중이 큰 것으로 나타났기 때문에 웹 문서 수집 시스템의 개선이 요청된다.

특정한 검색 세션 내에서의 이용자의 검색 행태는 일반적으로 클릭 행태로 완료되며, 클릭 빈도는 검색 포털들이 문서 순위와 컬렉션 순위 결정 시 매우 중요한 역할을 수행한다. 따라서 클릭 문서 분석을 비롯한 클릭 행태에 관한 지속적인 연구가 요청된다. 한편 이 연구에서는 통합 검색 질의들 중 이슈성 질의들에 대한 클릭 집중 문서들을 대상으로 분석과 평가를 수행하였다. 향후 연구에서는 성격이 다른 질의와 클릭 문서들에 대한 평가 및 비교 작업이 필요할 것으로 보인다. 둘째, 클릭이 집중된 문서와 클릭이 집중되지 않은 그 외의 문서들 간의 비교 작업이 요구된다. 셋째, 클릭 빈도의 반영 정도와 같은 검색 포털의 정책 및 전략이 클릭 행태에 미치는 영향에 대한 연구가 필요할 것으로 보인다. 넷째, 보다 장기간에 걸쳐 수집된 질의와 문서들에 대한 분석 작업이 바람직하다. 마지막으로, 후속 연구에서는 본 연구에서 제시한 방법론에 대한 검증 및 보완 작업이 요청된다.

〈참고문헌은 각주로 대신함〉