검색 포탈에서 사용자 질의분석을 통한 검색형태 연구

다음커뮤니케이션 이소영 · 조영환

1. 서 론

정보화 사회가 성숙되면서 인간의 정보 요구는 계속 다양화, 전문화, 개인화 되어왔으며, 한편 정보의 양과 다양성은 기하급수적으로 증가하여 왔다. 인터넷은 이러한 변화의 중심에서, 여러 가지 새로운 사회 현상을 적나라하게 표현하는 한편, 새로운 현상이 생기는 바탕이 되어왔다. 클리어링 하우스의 개념으로 시작된 인터넷 검색 포탈은, 다각적인 방법으로 실험 되어졌고, 검색 효율적 또 상업적인 목적에 의하여 괄목한 성장을 보여 왔다.

기존의 검색 엔진의 성능 측정이나, 이용자의 이용행태 연구에서는, 믿을 수 있는 결과를 얻기 위하여 변수가 통제된 실험 환경에서 수행되는 것이 일반적이다. 예를 들어, 검색 엔진의 성능을 측정하기 위해서는 정제된 질의어나 문헌세트를 이용하여 검색 결과의 정확한 정확율과 재현율을 측정해낼 수 있다. 또한, 이용자의 이용행태를 파악하기 위하여, 피실험자의 정보 요구를 다양한 방법으로 제한하거나, 피실험자의 검색 행위를 실험자가 밀접하게 관찰하면서 이루어지는 경우가 많다. 소수의 피실험자 집단을 선정하여 그들의 이용행태를 밀착 관찰하는 실험을 통하여, 다양한 평가 기준을 개발하여 웹 검색 엔진을 평가하는 기준을 만드는 작업은 많이 수행되고 있다[1]. 이를 통하여 여러 평가 기준이 마련되었으며, 이용자의 행태도 간접적으로 파악되고 있다.

한국에서의 인터넷 검색 포탈은 2000년부터 웹 문서 검색이라는 단일범주의 컨텐츠를 대상으로 하던 관념에서 벗어나서 홈페이지명, 이미지, 사전, 뉴스 등 20여 가지의 이질적인 컨텐츠를 하나의 결과 화면에 제공하고 있다. 또한 서비스적인 측면이 강하여 평가의 결과치는 궁극적으로 사용자의 만족도가 되어야 한다. 여기에는 그래픽 처리와 검색 속도, 컨텐츠의 다양성, 과거의 경험 등 다양한 심리적인 요소들이 존재하므로, 현재의 인터넷 검색 포탈의 경우에 검색성능 평가를 위한 단일의 통제된 실험 환경을 구축하는 것이 불가능하다. 따라서 현재 인터넷 검색 포탈에서 가장 많이 사용되는 평가의 방법은 다양한 이용자 설문 방법을 이용하는 것이다. 이용자의 설문을 이용하는 경우에는 이용자의 주관이 개입되는 변수를 제어할 수 없고, 이는 실제의 검색행태와 어느 정도의 차이를 의미한다.

검색 포탈에서의 이용자 행태를 정확하게 파악하기 위하여 입력된 질의를 분석하는 방법이 있다. 질의는 이용자가 검색 포탈에서 정보를 검색하기 위하여 입력한 모든 질의 및 행태를 사실 그대로 반영한다. 따라서 실제의 이용자를 관찰하거나 설문하지 못하여 생기는 문제에도 불구하고, 질의 로그를 분석하는 방법은 합리적이고 객관적인 방법이다[2]. 국내의 네이버의 웹 문서 검색 질의를 분석한 연구[3]에서 로그 분석에 필요한 세션의 정의 방법을 실험하는 등의 보다 합리적인 접근을 시도하였다. 이 연구들은 세션에 중심을 두고, 한 세션 내에서의 이용자의 검색행태를 분석하였다. 해외에서는 익사이트의 대용량 질의 분석 연구[4,5,6], 알타비스타 [7], 파이어볼[8]을 이용한 연구 등이 있다. Spink 등의 일련의 연구는 특정일의 대용량 로그를 분석하여, 이용자들이 검색하는 주제는 시대에 따라 바뀌나, 검색행태 자체는 크게 변하지 않은 것을 발견하였다. Sil-verstein은 6주간의 10억개에 가까운 대용량 질의를 분석하였다.

본 연구에서는 인터넷 검색 포탈 다음검색의 검색 서비스 이용자들이 직접 입력한 로그를 분석하여 이용자의 이용행태를 파악하였다. 사전 조사로서 검색 서비스의 사용자를 대상으로 한 설문 조사의 결과를 소개한다. ***특히, 주로 검색하는 질의어를 분석하여, 이용자들이 어떤 정보 요구를 가지고 검색 포탈을 이용하는가를 검색 로그와 선행된 설문 조사를 비교하여 분석하였다.***

2. 사용자 설문 조사

2.1 설문 조사의 목적과 방법

조사의 목적은 인터넷 이용의 주 행동 중 하나인 검색 행동에 대해서 네티즌의 생각, 태도 등을 파악하는 데 있다. 조사의 방법은 1) 설문지를 제작하여 2) 네티즌

중에 분포에 적합한 대상을 선정하여 3) 이메일을 통한 설문지 작성 요청을 한 후에 5) 발송된 설문지 중에서 응답된 자료와 무응답 자료에 대한 정규화 작업을 거쳐서 최종 분석을 하였다.

표 1 검색 사용자 설문조사 방식

모집단	만 13~45세의 일반 네티즌 남녀
조사지역	전국
표본크기	1003 (유효표본)
자료수집방법	구조화된 질문지(Structured Questionnaire)를 이용한 인터넷조사
표본오차	± 3.09% (95% 신뢰수준)
조사일시	2003년 8월

2.2 인터넷 설문 조사의 결과

가) 인터넷 검색 사이트를 주로 이용하게 되는 상황, 목적에 대해 알아본 결과, "원하는 사이트를 찾기 위해"가 39.9%로 가장 많았으며, 그 다음으로 "모르는 단어나 내용이 있을 때"(19.6%), "숙제/리포트/논문을 준비할 때"(13.0%) 등의 순으로 나타났다.

나) 검색 사이트의 이용 목적은 응답자 특성별로 살펴보면, 30세 전후의 남성 층은 희망 site를 탐색하는 것에, 30세 전후 연령층의 여성은 모르는 단어 혹은 내용을 찾기 위하여, 중고생의 경우 숙제와 연예/문화 정보 탐색을 위해 검색 서비스를 상대적으로 많이 이용하는 것으로 나타났다.

다) 인터넷 검색 사이트를 이용하는 방식을 알아본 결과, 응답자 10명 중 7명 정도가 현재 이용하는 사이트에서 1차로 검색하고, 검색 내용에 불만이 있을 때 전문 검색 사이트를 이용하는 것이 일반적인 것으로 나타났다.

라) 인터넷 검색 만족도에 있어서는 응답자의 73%가 원하는 정보를 찾았다고 응답하였으며, 반면, 희망 정보를 못 찾은 경험이 있는 사람은 27%로 나타났다. 이를 응답자 특성별로 살펴보면, 여성보다는 남성 층에서, 30세 전후의 연령층에서 성공률이 높다고 응답하였다.

마) 검색 결과가 만족할 만한 수준이 안 될 경우 취하는 방법은 포기한다(25.6%), 책을 찾아본다(22.6%), 다른 사람에게 문의(14.0%) 등의 순으로 나타났다.

바) 최근에 이용한 검색 사이트에서의 키워드에 대해 알아본 결과, 다양한 응답 가운데 연예(인), 스포츠 스타 관련 내용이 가장 많았으며, 교육, 최근 뉴스와 관련된 내용이 그 뒤를 이었다.

사) 최근 인터넷 검색의 종류를 구분하여 보면, 통합검색이 69.8%로 가장 많았으며, 그 다음으로 뉴스(12.2%), 지식 D/B(4.3%), 지도/위치 등의 순으로 나타났다.

표 2 이용자별 검색 목적 설문 결과

		사례수(명)	원하는 사이트를 찾기 위해	모르는 단어나 내용 탐색	숙제/리포트/논문 준비	최근 관심사 내용검색	뉴스 탐색	연예/문화정보 탐색	커뮤니티 탐색
전체		1,003	39.9	19.6	13.0	10.8	5.5	4.8	3.3
성별	남성	517	**44.0**	17.7	10.7	10.8	6.7	4.0	3.0
	여성	486	35.4	**21.6**	15.5	10.8	4.2	5.7	3.5
연령별	13~18세	202	30.4	9.5	**29.9**	8.0	3.0	**11.2**	5.8
	19~24세	209	38.9	21.1	17.2	10.1	2.6	4.5	3.9
	25~34세	303	**47.7**	**25.2**	4.8	9.7	5.4	3.7	0.8
	35~45세	289	39.0	19.5	6.9	**14.4**	9.3	1.8	3.6
직업별	자영업	38	58.2	7.4	3.8	26.8	3.8	-	-
	블루칼라	36	52.0	11.8	1.8	10.5	-	1.6	16.1
	화이트칼라	381	46.1	23.8	3.5	12.5	7.0	2.6	1.4
	학생	388	32.5	15.3	25.8	9.0	2.9	7.7	4.6
	가정주부	88	41.6	20.6	12.9	7.4	10.1	3.1	4.3
	무직/기타	71	28.3	29.2	5.5	7.3	9.1	7.2	-

3. 인터넷 검색 사용자 로그분석

3.1 설문 조사의 목적과 방법

조사의 목적은 검색 서비스의 실제 검색 행동에 대한 동향을 파악하는데 있다. 조사의 방법은 검색 서비스에 투입중인 모든 장비의 검색 로그를 수집하여 중앙에 데이터베이스화 하고, 이를 특정 항목에 따라 재분류 혹은 일정 기간별 비교를 수행하는 자동화된 검색 로그분석

시스템을 사용하였다.

3.2 인터넷 질의 분석의 결과

아) 인터넷 검색의 사용 시간대를 구분하여 보면, 오전 2시부터 7시까지가 5.67% (시간당 0.95%), 8시부터 12시까지가 25.13% (시간당 5.03%), 13시부터 18시까지가 39.43% (시간당 6.57%), 19시부터 1시까지가 29.78% (시간당 4.25%)의 사용량을 보였다. 하루 중에는 보통 2번의 최대 사용량을 갖는 시간대가 발견되었는데, 15시(시간당 6.5%)와 21시(시간당 6.4%)였다.

자) 인터넷 검색 컨텐츠의 검색량은 통합 검색 37.80%, 이미지 검색 26.92%, 커뮤니티 검색 12.35%, 사전 검색 5.43%, 디렉터리 검색 3.49%, 기타 컨텐츠 검색 14.01% 등으로 나타났다.

차) 컨텐츠별 사용자의 분포는 통합 검색 사용자를 100%로 하였을 경우에, 커뮤니티 검색이 41.54%, 디렉터리 검색이 12.19%, 이미지 검색 11.12%, 뉴스 검색 4.58%, 사전 검색 4.57%, 웹문서 검색 1.96% 등으로 나타났다.

4. 인터넷 검색 사용자 질의분석

질의분석에 이용할 데이터는 2004년 2월의 1달간 검색 포탈 사이트인 다음 검색(http://search.daum.net)의 검색창을 통하여 입력된 로그를 이용하였다. 따라서 다음 검색을 통하여 수행되는 검색 질의를 분석하여 한국 이용자들이 검색 포탈을 이용하여 수행하는 검색의 행태를 미루어 짐작할 수 있다. 객관적인 분석을 위하여, 전체 로그를 대상으로 수치 분석을 수행하였다. 또 검색 질의의 특성을 파악하기 위하여, 질의를 입력 빈도로 소팅하여, 상위 1천위에 속하는 키워드를 추출하여 분석 대상으로 삼았다

본 논문에서는 이용자의 행태를 패턴화하기 위하여, 질의의 특성을 검색을 수행하는 목적에 따라 다음과 같이 4개로 분류하였다.

4.1 사이트 접속 질의

고유한 특정 사이트를 검색하는 경우

이용자가 특정 사이트를 염두에 두고, 그 사이트를 찾아내려는 경우에 해당

주로 고유명사를 포함한 질의를 의미한다. (바로가기 정보가 대개 적절한 응답이 됨)

4.2 사이트 검색 질의

특정 주제나 컨텐츠에 대해 다루는 사이트를 검색하는 경우

특정 사이트를 염두에 두지 않고, 일반적이고 광범위한 정보를 검색하고자 하는 경우

4.3 컨텐츠 검색 질의

사전, 웹 문서, 지식인 등 특정 사실의 구체적인 답을 필요로 하는 질의

사이트의 구조나 다양성과 관계없이, 질의에 대한 구체적인 정보를 필요로 하는 질의

4.4 미디어성 질의

연예인 등과 같이 뉴스, 프로필 정보 등 유흥적이고 시기적 성격이 강한 정보를 요구하는 질의

이미지 정보나 뉴스, 혹은 문화 정보 등 개인적으로 생산된 정보나 시기성이 강한 정보를 필요로 하는 질의

이러한 분류는 현재 검색 포탈 서비스에서 일반적으로 받아들여지는 분류이며, 이미 서비스에 반영하고 있음이 확연하게 드러나고 있다. 현재 대부분의 검색 포탈은 정보의 본질에 따라 여러 개의 탭으로 나누어 검색 결과를 분류하고 있다. 한국의 검색 포탈에서만 특징적으로 나타나는 "통합 검색"의 기능은 이 다양한 결과군 중 대표적인 검색 결과만 뽑아서 보여주는 기능을 한다. 이는 이용자에게 다양한 정보를 한 페이지 안에 모아서 보여주고, 추가적으로 원하는 정보를 한번의 클릭으로 확장할 수 있게 한다. 또한, 이 "통합 검색"의 내용은 질의어의 성격 및 이용자의 추가 행태 분석에 의하여 달라짐을 볼 수 있다. 즉, 사이트 접속 질의의 경우, 바로가기나 사이트 검색 결과가 상위에 노출되고, 미디어성 질의의 경우에는 엔터테인먼트 정보, 뉴스 정보가 상위에 노출되는 경향이 있다.

위의 기준에 따른 분류는 질의의 성격을 실험자가 질의어의 성격과 그 결과를 자의적으로 판단하여 수행하였다. 이러한 분류의 목적은 질의의 성격에 따라 이용행태가 달라질 것이라는 가정에 기반을 둔다. 즉, 질의의 성격에 따라 검색의 수행에 의하여 얻어지는 검색 대상 및 결과가 다를 것이라는 것이다. 나아가, 검색 결과에 따른 만족도를 평가하는 기준도 달라질 것으로 예상한다. 예를 들어, 지식 정보의 경우에는 검색 사이트의 자체 컨텐츠 (사전 등)와 커뮤니티에서 생산되는 정제되지 않은 비공식적 정보 (카페글 검색, 지식인 등)가 주요 검색 대상이 될 것이다. 나아가 이용자들은 컨텐츠의 질과, 커뮤니티의 크기 및 자체적으로 생산되는 정보의 량에 따라 서비스 만족도가 달라질 것이다.

본 실험에서는 상위 10개, 100개, 1,000개 질의어 등 3개의 세트로 나누어 수행하여 비교하였다.

표 3 상위 질의어의 분류 분포표

	상위 10	상위 100	상위 1000
미디어성	100%	65%	42.70%
사이트 접속		13%	24.50%
사이트 검색		16%	23.30%
컨텐츠 검색		6%	9.20%
기타			0.30%

위에 따르면 미디어성 정보검색이 대부분을 차지하는 것을 볼 수 있다. 특히, 극 상위의 분류에 해당하는 상위 10개 및 100개 검색어의 경우 미디어성 검색이 대부분을 차지한다. 이것은 시대의 조류에 매우 민감하게 반응하는 포탈 및 네티즌의 기본적인 속성인 것으로 보인다. 반면, 표 2의 설문 조사에서 실제 이용자의 정보검색 요구를 질문하였을 때 결과와는 약간의 차이를 보였다. 이를 도식으로 그려보면 다음과 같다.

표 4 검색 사용자 설문과 실제 질의의 차이

	질의	설문	차이
미디어	64.11%	33.50%	30.61%
사이트 접속	15.13%	7.90%	7.23%
사이트 검색	15.25%	25.30%	-10.05%
컨텐츠 검색	5.46%	28.60%	-23.14%

즉, 이용자는 정보검색을 정보나 지식을 찾는 것으로 주로 이용한다고 생각하고 있으나, 실제 검색은 미디어 및 사이트 검색에 더 많이 나타나고 있다. 즉, 로그분석의 결과 79.24%가 미디어와 사이트 검색에 치중하고 있으나, 이용자는 53.9%가 정보나 지식을 찾기 위하여 검색 포탈을 이용한다고 인식하고 있다.

5. 결 론

본 연구에서는 로그분석을 통하여, 이용자가 검색 포탈을 이용하는 실제 행태를 짚어보고, 이를 질의의 속성으로 분류하여 보았다. 이용자들이 가장 많이 검색하는 질의어를 중심으로 분석하여 보았을 때, 이용자는 정보나 지식 자체를 검색하려는 요구보다는, 연예인이나, 게임 등의 엔터테인먼트성 정보를 보기 위하여 검색하는 경우가 많았다. 이는 검색 포탈이 전통적인 정보검색 서비스의 요구보다는 전반적인 네티즌의 이용 패턴을 반영해야 한다는 결론에 도달한다.

인터넷의 정보는 첫째 대량의 정보이다. 둘째로, 조직되어 있지 않고, 표준화 되어 있지 않다. 또한, 매우 다양 속성을 띤다. 검색 포탈은 이 모든 속성의 정보를 검색하기 원하며, 검색 포탈은 이를 반영하여, 서비스하여야 한다. 산재되어 있는 다양한 정보를 어떠한 형태로 보여줄 것인가가 가장 연구가 필요한 분야이다. 이를 연구하는데 질의를 이용하는 방법이 가장 현실적으로 객관적이라 할 수 있다.

또 본 연구를 통하여 주목할 만한 점은, 이용자의 인식과 실제 이용행태에 관한 괴리이다. 이용자는 정보와 지식을 검색하기 위하여, 검색 포탈을 이용한다고 대답하였고, 실제로는 미디어성 질의나, 알려진 사이트를 검색하기 위한 질의를 입력한다. 이에는 물론, 모범적인 답안, 즉, 당위성에 의한 대답이라고도 할 수 있다. 그러나 이용자가 실제 "지식" 혹은 "정보"라고 정의하는 인식의 차이일 수 있다. 즉, 미디어성 정보가 서비스적 측면이나, 정보 자체의 속성이 엔터테인먼트성이라 하더라도, 이를 협의의 "지식"과 구분지어 생각하지 않는다는 것이다.

정보의 속성이 다양함에 따라, 질의의 속성별로, 가장 적합한 검색 대상을 효율적으로 보여주는 것에 집중할 필요가 있다. 또한, 정보 및 검색 대상을 효율적으로 클러스터링하여, 조직하고, 검색된 결과를 효율적으로 디스플레이하는 연구도 필요하다.

또한, 질의 자체를 분석할 필요가 있을 뿐만 아니라, 이용자가 검색 포탈 내에서 이동하는 행로를 추적하고, 또 다른 사이트로 빠져나가기까지의 연구도 필요하다. 이를 총체적으로 파악하여 실제 이용자의 행태를 포괄적으로 짚어낼 필요가 있다. 즉, 질의를 입력하는 것 뿐만 아니라, 결과를 살펴보고, 이동하는 것까지의 로그를 체계적으로 분석하여야 한다. 즉, 이용자의 검색 요구의 발생에서부터, 검색행태 자체, 질의의 반복적 입력, 입력 질의의 수정, 결과의 브라우징, 검색 대상의 변경, 그리고, 검색 포탈을 빠져나가기까지의 과정을 체계적으로 연구할 필요가 있다. 이는 로그의 체계적인 분석 뿐만 아니라, 이용자의 클릭을 분석함으로서 가능하다.

질의 로그는 실제 이용자의 검색 요구를 반영할 뿐만 아니라, 이용자의 인식과 행태까지도 반영하고 있으므로, 이용자가 검색 결과 만족하지 못할 때의 연구와 이에 대한 대처 방안의 고려는 검색 서비스를 한 단계 발전시키고, 또한 이용자의 만족도도 크게 제고할 수 있을 것으로 생각된다.

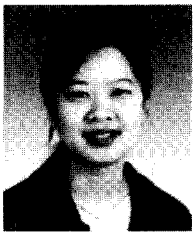
설문 결과에 의하면 이용자는 검색 서비스 결과가 만족스럽지 않거나, 검색에 실패한 이유로 결과가 다양하지 못하다는 점을 들고 있다. 대개의 검색 포탈에서 인터넷 상의 거의 대부분의 정보를 여러 방법으로 나누어 보여주고 있음에도, 결과가 다양하지 못하다고 인식하는 이유에 대한 분석 및 그 대책에 대한 연구가 요구된다.

참고문헌

[1] 오삼균, 박희진, "국내 인터넷 탐색 엔진에 대한 이용자 중심의 평가에 관한 연구 한국 알타비스타

와 네이버를 중심으로", 한국문헌정보학회지, 제34권, 제2호, pp. 117-135, 2000.
[2] Jansen and Pooch, "A review of web searching studies and a framework for future research," Journal of American Society for Information Science and Technology, Vol.52, No.3, pp. 235-246, 2001.
[3] 이준호, 박소연, 권혁성, "질의 로그 분석을 통한 네이버 이용자의 검색 행태 연구", 정보관리학회지, 제20권, 제2호, pp. 27-41, 2003.
[4] Spink et al., "From e-sex to e-commerce : Web search changes," IEEE Computer, Vol.35, No.3, pp. 133-135, 2002.
[5] Spink et al., "Searching the web : The public and their queries," Journal of the American Society for Information Science and Technology, Vol.52, No.3, pp. 226-234, 2001.
[6] Spink, and Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," Information Processing & Management, Vol.36, No.2, pp. 207-227, 2000.
[7] Silverstein et al., "Analysis of a very large web search engine query log," SIGIR Forum, Vol.33, No.1, pp. 6-12, 1999.
[8] Bernard J. Jansen, Amanda Spink and Tefko Saracevic Hoelscher, "How Internet experts search for information on the web," The World Conference of the World Wide Web, Internet and Intranet, Orlando, FL., 1998.

이 소 영



1993 연세대학교 문헌정보학과 학사
1995 Master of Information & Library Studies, Univ. of Michigan
1996~2001 (주)데이콤 천리안사업본부
2001~2002 잉크토미코리아 포탈사업부
2002~현재 연세대학교 문헌정보학과 박사과정 재학중
2003~현재 (주)다음커뮤니케이션 검색본부
관심분야 : 정보검색, 이용자 분석, 검색 품질 평가, 검색 인터페이스
E-mail : qtink@daumcorp.com

조 영 환

1989 연세대학교 전산과학과 학사
1991 한국과학기술원 전산학과 석사
1997 한국과학기술원 전산학과 박사
1997~1998 연구개발정보센터 선임연구원
1999~2000 삼성종합기술원 HCI 연구소 전문연구원
2001 (주)서치솔루션 CEO, CTO
2002 (주)맥시메타 CEO, CTO
2003~현재 (주)다음커뮤니케이션 검색본부 본부장
관심분야 : 정보검색, 대화 시스템, 기계번역, 음성 인터페이스
E-mial : choyh@daumcorp.com