링크 빈도와 클릭 빈도를 이용하는 메타 검색엔진의 설계

유태명○ 김준태
동국대학교 컴퓨터공학과
{yoo4001, jkim}@dgu.ac.kr

A Design of Meta Search Engine that Uses Link and Click Frequencies

TaeMyung Yoo○ JunTae Kim
Dept. of Computer Engineering, Dongguk University

요 약

대부분의 검색엔진들이 사용하는 내용 기반 검색 방법은 웹 페이지에 있는 단어의 빈도만을 이용하여 순위를 결정함으로써 비슷한 단어 빈도를 가지고 있는 방대한 양의 검색 결과로부터 참조할만한 가치가 있는 중요한 페이지를 찾아내기가 매우 어렵다. 중요한 페이지를 구분해 내는 한가지 방법은 얼마나 많은 웹 페이지들이 참조하고 있는가 또는 얼마나 많은 사용자들이 그 웹 페이지에 접속하는가를 보는 것이다. 본 논문에서는 링크 빈도와 클릭 빈도를 이용하여 웹 페이지의 중요도를 계산하는 메타 검색엔진의 프로토타입을 구현하였다. 링크 빈도는 검색엔진에 해당 웹 페이지의 URL을 질의로 던져 구하고 클릭 빈도는 servlet을 이용하여 사용자의 클릭 행위를 감시하여 얻어내도록 하였다. 메타 검색엔진은 이 두 값의 가중치 합으로 각 페이지의 중요도를 계산하고 중요도 순으로 검색 결과를 재배열하여 사용자에게 보여 준다.

1. 서론

인터넷의 급속한 팽창에 따라 온라인 문서의 양도 폭발적으로 증가하고 있다. NEC 연구소의 보고서에 따르면 98년 현재 웹에는 약 3억 2천만개의 웹 페이지들이 존재한다고 한다. 정보의 양이 증가할수록 사용자들은 원하는 정보를 얻기까지 더 많은 노력을 필요로 하게 된다. 일반적으로 방대한 웹에서부터 원하는 정보를 얻기 위해서 알타비스타나 야후, 심마니 등의 검색엔진이 사용되고 있지만, 검색 결과가 대부분 상당한 양이어서 사용자는 다시 여기서 중요한 페이지를 찾기 위해 많은 시간과 노력을 들여야만 한다.

이러한 문제점을 해결하기 위해 다양한 연구가 수행되어 왔다. 검색 질의에 대해 다양한 검색 결과를 종합적으로 분석하여 정보를 제공하는 메타 검색엔진[1][4], 사용자를 대신하여 사용자의 취향을 분석하고 웹 페이지를 수집하고 여과하여 알맞은 문서를 추천하는 웹 에이전트에 관한 연구 등이 그 예이다. 대부분 이러한 시스템들은 웹 페이지의 검색 및 순위 결정에 내용 기반 방법을 사용하고 있다. 내용 기반 방법은 문서 내의 단어 빈도와 단어 위치를 이용하는 것이다. 그러나 많은 경우에 사용자는 단순히 주어진 단어가 한 두 개 더 많은 페이지가 아니라, 다수의 사용자에 의해 검증된 대표성이 높거나 인기도가 높은 웹 페이지를 원한다.

본 논문에서는 이러한 접근 방법으로 다른 웹 페이지로부터의 링크 빈도와 다른 사용자들로부터의 클릭 빈도를 이용하여 웹 페이지의 중요도를 순위를 결정하는 기법을 제안하고, 이러한 방법을 적용한 메타 검색엔진의 프로토타입을 구현하였다. 다른 웹 페이지들로부터의 링크가 많거나 사용자들로부터 클릭이 많다는 것은 그 페이지가 다수의 사용자들에게 중요한 페이지로 인식되어진다고 할 수 있기 때문이다.

2. 관련연구

내용 기반 방식의 추천 방식에 한계를 극복하기 위하여 새로운 검색 기법 및 순위 결정 방법을 도입한 이른바 '제 2세대 검색엔진'이라 불리는 새로운 검색 시스템에 대한 연구가 활발하게 진행되고 있다. 이러한 시스템들은 하이퍼링크 구조를 분석하거나 사용자의 행위를 관찰하여 얻어진 정보를 이용함으로써 기존의 검색엔진과는 전혀 다른 순위 결정 개념으로 주목받고 있다.

2.1 클릭 정보의 이용

DirectHit[6] 검색엔진은 사용자의 질의에 대한 결과 페이지 중 클릭이 많이 된 페이지를 기준으로 순위를 매겨 사용자에게 보여준다. 페이지의 순위결정에 사용자들 사이의 인기도를 반영하여 선호도가 높은 페이지를 먼저 보여주게 된다. 검색엔진이 보여준 결과를 사용자가 클릭했을 경우 이 클릭 정보가 다시 검색엔진 내의 인기도 엔진으로 전달되어 그 페이지의 클릭 정보를 누적하게 된다. 이 정보는 다음 검색 시에 반영되어져서 페이지의 인기도를 표시한다. DirectHit 검색엔진은 명시적인 사용자의 클릭뿐 아니라 암시적인 방법도 함께 사용한다. 사용자에게 페이지를 보여준 후 얼마나 오랫동안 그 페이지를 읽고 있는지에 대한 시간을 계산한 후 이를 점수화하여 인기도 엔진에서 선호도 계산에 반영한다.

2.2 하이퍼링크 정보의 이용

하이퍼링크 정보를 이용하는 방법은 문서들 간의 연결 구조에 따른 페이지들 간의 상호 관계를 이용하는 방법이다. A 페이지가 B 페이지를 링크 시킨다면 A 페이지는 B 페이지를 가치 있는 페이지로 판단한다고 볼 수 있다. 또 B 페이지가 A 페이지뿐만 아니라 다른 페이지들로부터 많은 링크를 받고 있다면 B 페이지는 페이지들로부터 중요한 페이지라고 생각되어 지는 것이다. 하이퍼링크를 이용하여 페이지들간의 가중치를 결정하는 방법은 Kleinberg의 HITS 알고리즘에 잘 소개되어져 있다[2].

HITS 알고리즘을 적용하여 연구되고 있는 검색엔진은 Google[7]과 Clever[3][5] 검색엔진을 들 수 있다. 스탠포드 대학의 Google은 web에 있는 문서들을 로봇을 통해 모아 온 후 하이퍼링크 구조를 분석하고 HITS 알고리즘을 이용하여 각 페이지들의 가중치를 계산한 다음 이를 기준으로 검색 결과의 순위를 결정한다. IBM에서 연구되고 있는 Clever 검색엔진에서는 향상된 HITS 알고리즘을 사용하며, 하이퍼링크 정보 이외에 텍스트를 포함한 다양한 정보를 함께 이용하여 순위를 결정한다.

3. 링크 빈도와 클릭 빈도를 이용한 메타 검색엔진

본 논문에서 구현한 링크 빈도와 클릭 빈도를 이용하여 순위를 결정하는 메타 검색엔진의 구조는 그림 1과 같다.

사용자가 메타 검색엔진에 질의어를 주면 메타 검색엔진은 우선 이 질의어를 기존의 검색엔진들에 던져 결과를 얻어 온다. 메타 검색엔진은 이 결과를 분석하여 각 페이지들에 대한 URL, 제목, 요약부분 등의 정보를 추출한다. 그리고 각 페이지에 대한 링크 빈도와 클릭 빈도 정보를 가중치로 변환한 후 이 값들의 가중치 합으로 각 페이지의 중요도를 결정한다. 페이지들을 이렇게 계산된 중요도에 따라 재배열한 후 링크 빈도, 클릭 빈도 정보와 함께 사용자에게 제시한다.

다음 식은 링크 빈도와 클릭 빈도에 의한 페이지의 가중치 결정 공식이다.

$$w_{t,A} = \alpha w_{ih,A} + \beta w_{c,A}$$

$w_{t,A}$: A페이지의 중요도
$w_{ih,A}$: A페이지의 인 링크에 의한 가중치
$w_{c,A}$: A페이지의 클릭에 의한 가중치
α, β : 상수, $\alpha+\beta=1$

3.1 링크 빈도의 수집

다른 페이지들로부터의 링크인 인 링크(in-link)의 결과를 얻기 위해 각 검색 결과 페이지의 URL을 추출한 후 먼저 링크 DB를 검색하여 링크의 개수를 구하고, 자료가 없을 경우에는 URL을 다시 검색엔진에 질의로 던져 그 URL을 포함하고 있는 웹 페이지의 개수를 알아내어 사용하였다. 각 검색 결과에 대한 인 링크를 알아내는 시간을 단축하기 위하여 멀티 스레드를 사용하였으며, 한번 검색엔진으로부터 인 링크의 개수를 구하면 이 정보를 링크 DB에 저장하여 같은 URL에 대해서는 차후에 검색 시간을 단축할 수 있도록 하였다. 링크 빈도에 대한 가중치 공식은 다음과 같다.

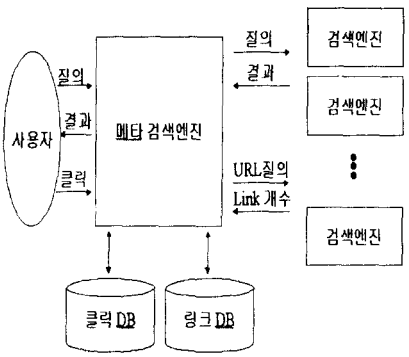

그림 1 메타 검색엔진의 구조

$$w_{ih,A} = \log_2\left[\frac{n_A}{n_{max\_inlink}} + 1\right]$$

$w_{ih,A}$: A 페이지의 인 링크에 의한 가중치
n_A : A 페이지의 인 링크 개수
n_{max_inlink} : 인 링크 개수중 가장 큰 값

3.2 클릭 빈도의 수집

사용자에게 검색 결과를 보여 준 후 사용자가 제시된 결과 중에서 관심 있는 페이지를 클릭 할 경우 이 정보를 시스템에서 받아 들여 그 페이지의 클릭 빈도를 증가시킨다. 이 클릭 빈도는 클릭 DB에 저장되어 다음에 검색 결과로 그 페이지가 추출되었을 때 반영되게 된다.

사용자가 클릭을 할 경우 이를 메타 검색엔진에서 받아들이는 방법은 자바 서블릿(Servlet)을 이용하여 구현하였다.

그림 2는 한글 알타비스타에 '천리안'이라는 질의를 주었을 때 나온 상위 10개의 페이지로써, 일반적으로 이 질의어에 대하여 중요한 페이지인 천리안의 홈페이지가 상위에 나타나지 않는 것을 볼 수 있다. 그림 3은 본 논문에서 구현한 메타 검색엔진에 같은 질의어를 주었을 때 검색결과로써, 천리안 비즈 홈페이지와 천리안 홈페이지가 상위에 보여지고 있음을 알 수 있다.

4. 가상 데이터에 의한 실험

본 논문에서 제시한 검색 방법에 대한 정량적인 성능 측정을 위해서는 대량의 데이터가 필요하지만, 현재 웹에서의 클릭 정보에 대한 실험용 데이터 집합이 없고 각 웹 페이지에 대한 중요도도 객관적인 자료가 없다. 따라서 본 논문에서는 임의로 설정한 질의어에 대하여 가상의 데이터를 만들어 본 논문에서 제안한 검색 방법을 적용해보았다.

4.1 실험 데이터 및 방법

'자바'와 '코스닥'이라는 질의어를 임의로 한글 알타비스타와 한글 야후에서 상위 200개씩의 웹 페이지를 수집하였다.

링크 빈도는 각 페이지의 URL을 다시 한글 알타비스타

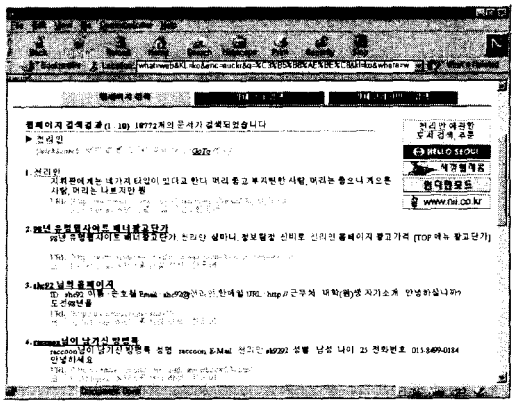

그림 2 '천리안'에 대한 알타비스타 검색 결과

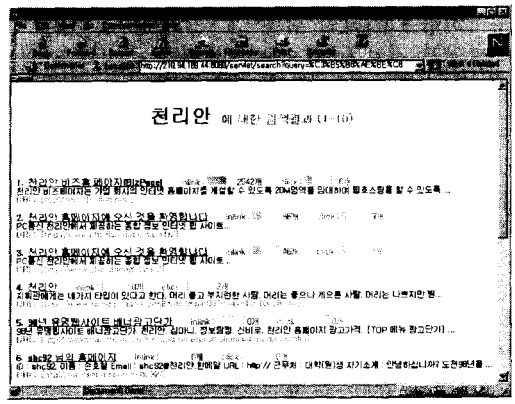

그림 3 '천리안'에 대한 메타 검색엔진 결과

에 질의어로 던져 구하였고, 클릭 빈도는 각 400개의 검색 결과를 섞어서 웹에 제시한 다음 동국대학교 학부 1학년들을 대상으로 관심 있는 페이지를 클릭하도록 하여 구하였다. 실험에 참가한 사용자는 총 130명이고 두 질의어에 대한 클릭 숫자는 총 2284개였다.

각 페이지의 중요도는 각 검색 결과를 섞어서 하나의 페이지로 만든 후 학부, 대학원생 19명에게 "아주 나쁨(1)", "나쁨(2)", "좋음(3)", "아주 좋음(4)"의 네 단계로 제시 후 페이지의 가중치를 구하였다.

4.2 실험 결과

다음 표 1은 질의어들에 대해 각각의 방법으로 구한 상위 10개 페이지들의 평균 중요도를 구한 결과로써 링크 빈도와 클릭 빈도를 함께 사용한 방법이 평균적으로 한글 알타비스타와 한글 야후의 결과보다 더 높은 중요도를 나타내는 것을 볼 수 있다.

검색 방법		질의어		평균
		'자바'	'코스닥'	중요도
링크 빈도와 클릭 빈도사용 (α , β)	(1, 0)	2.43	2.41	2.42
	(0, 1)	2.69	3.21	2.95
	(.5,.5)	2.24	3.17	2.71
한글 알타비스타		2.43	2.55	2.49
한글 야후		1.9	3.01	2.46

표 1 질의어 '자바'와 '코스닥'의 실험 결과

이 실험은 매우 기초적인 것으로써 객관적이고 정량적인 실험은 아니지만 링크 빈도와 클릭 빈도를 이용한 중요도 계산이 어느 정도 의미 있는 방법이 될 수 있음을 보인다고 할 수 있다. 앞으로 이러한 검색엔진을 웹 상에서 다수의 사용자가 사용할 수 있도록 하면 누적된 자료로부터 보다 객관적인 평가를 할 수 있을 것이다.

5. 결론 및 향후 연구과제

본 논문에서는 링크 빈도와 클릭 빈도를 이용하여 웹 페이지의 중요도를 계산하는 메타 검색엔진을 구현하였다. 링크 빈도는 기존의 검색엔진으로부터 인 링크 수를 구해서 사용하며, 클릭 빈도는 서블릿을 이용하여 사용자의 클릭을 받도록 하였다. 이 두 값의 가중치 합으로 중요도를 계산하여 이를 기준으로 검색된 페이지들을 재배열 한 후 검색 결과를 보여준다. 가상적인 실험 데이터로 기초적인 실험을 수행하여 본 논문에서 제안하는 검색 방법을 적용해보았다.

본 논문에서는 질의어에 관계없이 각 페이지에 대한 총 클릭 빈도를 이용하도록 하였는데, 각 질의어에 따라 각각 다른 클릭 빈도를 유지하여 계산하는 방법은 향후 연구 대상이다. 모든 질의어 조합에 대하여 이러한 정보를 유지할 수는 없으나 워드넷(WordNet)[8]을 이용하여 질의어들을 소수의 상위 개념 중 하나로 바꾸어 이들에 대해 각각의 클릭 빈도를 저장하는 것은 가능할 것이다.

6. 참고 문헌

[1] E. Selberg and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the Web", *IEEE Expert*, 1997

[2] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Proceeding of 9th ACMSIAM symposium on discrete Algorithms*. 1998.

[3] S. Chakrabarti, B. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Spectral filtering for resource discovery", *ACM SIGIR workshop on Hypertext Information Retrieval on the Web*, 1998.

[4] S. Lawrence and C. L. Giles, "Inquirus, the NECI meta search engine", *7th International World Wide Web Conference*, 1998

[5] Clever project, http://www.almaden.ibm.com/cs/k53/clever.html

[6] Direct Hit, http://www.directhit.com/

[7] Google, http://www.google.com/

[8] WordNet ,http://www.cogsci.princeton.edu/~wn