사용자 뷰 기반 페이지 랭킹 알고리즘의 설계 및 구현

김성후, 이종민, 박규석
경남대학교 컴퓨터공학과

Design and Implementation of a Page Ranking Algorithm Based on User's View

Seong-hoo Kim, Jong-Min Lee, Kyoo-Seok Park
Dept. of Computer Engineering, Kyungnam University
E-mail : arrayiv@csc.ac.kr, w6w@korea.com, kspark@kyungnam.ac.kr

요 약

인터넷의 빠른 성장과 함께 그 속에 존재하는 많은 정보들을 이용하고자 하는 인터넷 이용자들의 욕구를 충족시키기 위해 정보검색 기술이 발달하였고, 그 결과 현재 수많은 검색엔진들이 개발되어 사용되고 있다. 본 논문에서는 검색엔진을 이용한 검색의 경우 페이지 순위 결정시 사용자들에 의해 많이 읽혀진 페이지가 높은 순위에 랭크될 수 있도록 하여 좀 더 빠르고 정확한 결과를 찾을 수 있도록 하는 사용자 뷰 기반 페이지 랭킹 알고리즘을 제안한다.

1. 서론

문자의 발명 이후 축적되어 온 정보의 양은 거의 천문학적인 분량에 달하고 있다. 굳이 인류 역사를 더듬지 않더라도 매일매일 쏟아져 나오는 정보들은 우리가 감당하기에는 너무도 엄청난 분량이어서 과연 정보의 홍수 속에서 살고 있다는 사실을 더욱 실감나게 한다. 이제 어떤 막연한 정보에 대한 요구보다는 '어떤 목적에 필요한 어떤 정보'라는 더욱 구체화된 목적을 달성하기 위한 '정보 찾기'의 노력이 요구되고 있다[1].

이러한 요구들을 해결해 주는 것이 바로 '검색 엔진'이며 링크 정보를 이용한 웹 검색엔진은 최근 들어 인기를 얻고 있다. 링크 정보를 이용한 검색엔진은 텍스트 기반 검색엔진의 문제점을 해결해 주고 있다. 텍스트 기반 검색엔진은 웹페이지내의 텍스트 정보만을 이용하기 때문에 검색결과를 향상시키는데 한계를 가지고 있으며, 또한 키워드의 반복적인 나열 등을 이용한 의도적인 순위 높이기 등의 단점이 있다.

본 논문에서는 링크 정보를 이용한 정보추출 방법에 있어서 사용자들이 많이 보는 페이지에 가산점을 주어 높은 순위에 보여질 수 있도록 하는 사용자 뷰 기반 페이지 랭킹 알고리즘을 설계 및 구현한다. 본 논문의 구성은 다음과 같다. 2장에서는 제안 알고리즘과 관련된 연구분야에 대해 기술하고, 3장에서는 제안 알고리즘을 설계한다. 그리고 4장에서는 구현 및 평가, 5장에서는 본 논문의 결론을 내린다.

2. 관련 연구

2.1 랭킹 알고리즘 종류

1) 위치/빈도 방법(Location/Frequency Method)

웹페이지의 타이틀에 키워드가 나타난다면 검색엔진은 해당 페이지가 다른 것들보다 사용자의 요구에 대한 연관성이 깊다고 판단한다. 검색엔진이 웹페이지의 우선순위를 부여하는데는 키워드와의 연관성이 중요하다.

특히 키워드의 위치 및 나타나는 빈도수가 중요하다. 그래서 이것을 위치/빈도 방법이라고 한다.

- 위치(Location)

위에서 언급한 것처럼 웹페이지의 타이틀에 키워드가 나타난다면 검색엔진은 해당 페이지가 검색자의 요구에 깊은 관련이 있다고 판단한다. 또한 해당 페이지의 첫머리 가까이에 키워드가 나타난다면 이 역시 높은 우선 순위를 받을 수 있는 조건이 된다. 즉 페이

지의 첫머리에 키워드가 나타난다거나 페이지의 시작 몇 줄 이내에서 키워드가 발견된다면 검색엔진은 해당 페이지에 좀 더 높은 우선 순위를 부여하게 된다[2].

- 빈도(Frequency)

키워드가 페이지 내에서 얼마나 자주 나타나는가 하는 것은 검색엔진이 페이지의 우선 순위를 결정하는 또 하나의 중요한 요소이다. 검색엔진은 페이지 내에 키워드가 기타 다른 단어와 비교하여 얼마나 자주 나타나는가에 대한 분석을 한다. 문서 내에 키워드의 나타나는 빈도가 높을수록 해당 페이지는 높은 순위를 부여받을 수 있다.

2) 링크 인기도(Link Popularity)

모든 검색엔진들은 페이지의 우선 순위 결정을 위해 어느 정도까지 위치/빈도 방법을 따르지만 이것에 더하여 각 검색엔진은 나름의 순위 부여 방법을 가지고 있다.

웹크롤러(WebCrawler)와 같은 몇몇 검색엔진들은 우선 순위의 결정을 위해 링크 인기도를 이용한다. 링크 인기도란 해당 웹페이지가 얼마나 많은 다른 외부의 웹페이지에 의해 링크되어 있는가 하는 것이다. 자신을 링크한 외부 호스트의 웹페이지가 많을수록 자신의 링크 인기도는 높아지는 것이다. 웹크롤러는 링크인기도가 높을수록 해당 페이지가 양질의 정보를 가지고 있다고 판단하여 우선 순위에 약간의 가산점을 부여한다[2].

3) 하이브리드 검색엔진에서의 우선 순위

하이브리드 검색엔진들은 나름대로 정리된 디렉토리들을 가지고 있다. 디렉토리에 포함된 웹페이지들은 검색엔진에 의해 검증되고 양질의 정보를 포함하고 있다고 판단된 것들이다. 이러한 페이지들이 검색 결과로 출력될 때 검색엔진들은 해당 페이지의 우선 순위에 있어 약간의 가산점을 부여하게 된다[2].

2.2 페이지랭크(PageRank)

페이지랭크 계산 방법은 다른 웹 문서로부터 많은 링크를 받게 되면 좋은 점수를 얻게 되는 방법으로 페이지랭크는 식 1에 의하여 구할 수 있다[3,4,5].

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn)) \quad (1)$$

PR(A) : 문서 A의 페이지랭크
d : 완충 팩터(dampening factor)
일반적으로 0.85로 둠.
PR(T1) : 문서 A를 링크한 한 문서의 페이지랭크
C(T1) : T1이 링크한 문서의 수

위 식 1과 같은 방법으로 계산되어진 페이지랭크는 사용자들이 검색엔진에서 정보 추출 시 페이지의 순위를 결정하는 다른 값들과 함께 계산되어져서 그림 1과 같은 과정을 통해서 페이지 순위가 정해진다[6].

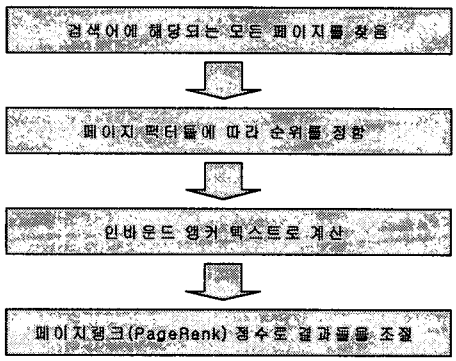

그림 1. 페이지랭크를 적용한 페이지 순위 결정 과정

3. 시스템 설계

사용자 뷰 기반 페이지 랭킹 시스템은 검색되어진 페이지들의 순위 결정 시 이용자들에게 많이 읽혀진 페이지에 높은 점수를 부여하는 시스템으로서 이용자의 인터넷 탐색 히스토리 목록을 수집하는 히스토리 수집 모듈과 수집된 목록을 분석·처리하는 히스토리 분석 모듈로 구성되며 이용자의 컴퓨터에 히스토리 수집 컨트롤이 설치되면 사용자 뷰 기반 페이지 랭킹 모듈은 다음과 같이 수행된다.

3.1 사용자 뷰 기반 페이지 랭킹 시스템의 동작 과정

제안한 사용자 뷰 기반 페이지 랭킹 시스템은 사용자의 컴퓨터에 설치되는 히스토리 수집 모듈과 검색엔진에 설치되는 히스토리 분석 모듈로 구성되며 동작 과정은 다음과 같다.

① 인터넷 이용자가 특정 사이트에 접속하게 되면 이용자의 컴퓨터에 설치된 히스토리 수집 모듈이 실행된다.

② 실행된 히스토리 수집 모듈은 현재 컴퓨터의 히스토리를 수집할 시기가 되었는지를 검사.

③ 히스토리 수집 모듈은 컴퓨터에 기록된 인터넷 탐색 히스토리를 수집.

④ 히스토리 수집 모듈은 수집된 히스토리 정보를 압축하여 히스토리 분석 서버로 전송.

⑤ 히스토리 분석 모듈은 수신된 히스토리 파일의 압축을 해제한 뒤 히스토리 정보를 분석·처리.

⑥ 처리된 히스토리 정보를 이용해서 인덱스 DB에 저장된 사용자 뷰 정보를 변경.

3.2 인터넷 탐색 히스토리 수집 모듈 설계

인터넷 탐색 히스토리 수집 모듈은 인터넷을 이용하는 사용자들의 컴퓨터에 설치되어 사용자가 특정 사이트에 접속할 때 실행되어 사용자의 컴퓨터에 저장된 히스토리 정보를 수집하고, 수집된 정보를 압축하여 지정된 히스토리 분석 서버로 전송하는 역할을 하며, 동작 과정은 그림 2와 같다.

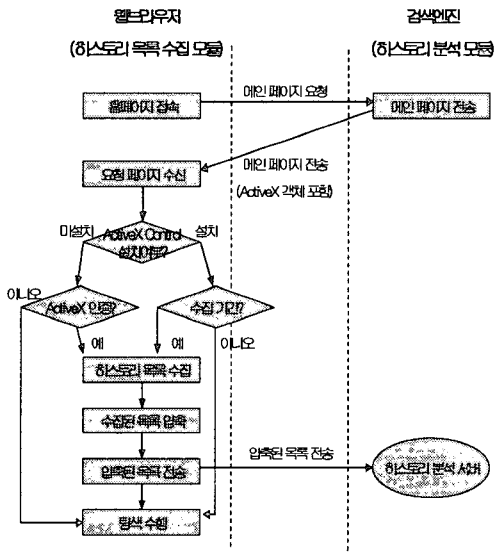

그림 2. 히스토리 수집 모듈의 동작 과정

3.3 인터넷 탐색 히스토리 분석 모듈 설계

인터넷 탐색 히스토리 분석 모듈은 사용자들의 컴퓨터에 설치된 히스토리 수집 모듈로부터 전송되어진 압축파일을 압축해제 하고 해제된 파일의 정보를 읽어서 페이지당 체류 시간 계산을 위해 임시 데이터베이스에 저장한 다음 페이지당 읽혀진 시간을 계산하고 계산된 결과 중 유효한 히스토리 정보만 데이터베이스에 저장한다. 그림 3은 인터넷 탐색 히스토리 분석 모듈의 동작 과정이다.

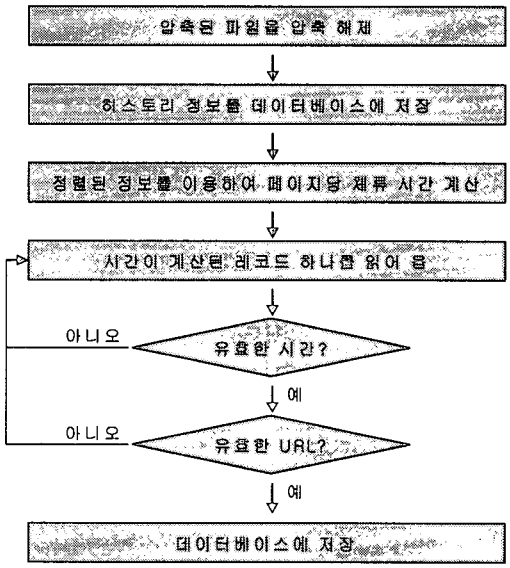

그림 3. 히스토리 분석 모듈의 동작 과정

3.4 사용자 뷰 값 계산

위의 과정을 거쳐 계산된 인터넷 탐색 히스토리 분석 모듈의 결과값은 문서에 대한 사용자 뷰를 계산한 것이기 때문에 특정 키워드와 관련된 문서에 대한 평가값으로는 그대로 적용할 수 없다. 즉, 문서에 대한 사용자 뷰 값에서 특정 키워드가 가지는 비중을 포함한 계산 결과를 사용하여야 하며, 키워드의 비중을 고려한 문서에 대한 사용자 뷰 값 PV는 식 2에 의하여 구할 수 있다.

$$PV(kwd, url) = pv(url) * \frac{pf(kwd, url)}{\sum_{i=1}^{n} pf(url)_i} \quad (2)$$

pv(url) : 문서의 사용자 뷰 횟수.
pf(kwd,url) : 문서내에서 특정 키워드의 분석 값.
pf(kwd) : 문서내의 모든 키워드의 분석 값.

3.5 사용자 뷰 값 적용

위의 과정을 거쳐 계산된 키워드의 비중을 고려한 문서에 대한 사용자 뷰 값은 사용자가 정보검색시 추출된 문서를 보여줄 때의 순위 결정을 위해 적용되며, 사용자 뷰 값이 적용되는 과정은 그림 4와 같다.

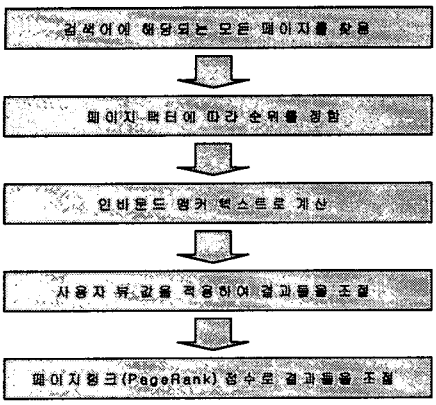

그림 4. 정보검색시 사용자 뷰 값의 적용 과정

4. 구현 및 평가

표 1은 사용자 뷰 기반 페이지 랭킹 기법을 적용한 검색엔진으로 한 웹사이트 내에서 검색어 '로그파일'에 대한 검색을 실시한 결과이고, 표 2는 로그파일을 알고 있는 임의의 사용자 3명이 '로그파일'과 관련성이 높은 순으로 판단한 결과이다.

표 1. '로그파일'에 대한 검색결과

1. http://clinicagent.co.kr/LogAnalysis/Log1.asp
2. http://clinicagent.co.kr/LogAnalysis/Log2.asp
3. http://clinicagent.co.kr/LogAnalysis/Log3.asp
4. http://clinicagent.co.kr/LogAnalysis/Log5.asp
5. http://clinicagent.co.kr/LogAnalysis/Log4.asp
6. http://clinicagent.co.kr/
7. http://clinicagent.co.kr/solution/AnaVi.asp
8. http://clinicagent.co.kr/solution/DrLog.asp

표 2. '로그파일'에 대한 사용자들의 순위 결정 결과

사용자 1	사용자 2	사용자 3
1	1	1
2	2	2
3	3	5
4	5	4
5	4	3
-	-	-

사용자 뷰 기반 페이지 랭킹 기법을 적용한 검색엔진이 검색한 결과와 사용자들이 결정한 결과를 비교하면 1위와 2위에 랭크된 페이지가 모두 동일하고, 6, 7, 8위에 랭크된 페이지에 대해서 사용자들은 순위에서 제외시켰다.

5. 결론

본 논문에서는 웹페이지들이 사용자들에게 읽혀진 정보를 수집하고, 분석하여 검색시 웹페이지 순위 결정에 적용한 결과, 검색어와 관련성이 높으면서 사용자 뷰 회수가 많은 페이지에 대한 순위가 높음을 알 수 있었다.

향후 연구 과제로는 사용자가 웹페이지를 읽은 정확한 시간 측정에 관한 계속적인 연구가 필요하다.

[참고문헌]

[1] http://www.yahoone.com/jungbo/000-4-001.htm

[2] http://www.gngood.net/web/webpromote/wpsearch3.html

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", The Seventh International WWW Conference, 1998

[4] Chris Ridings and Mike Shishigin, "PageRank Uncovered", 2002

[5] Phil Craven, "Google's PageRank and how to make the most of it", http://webworkshop.net/pagerank.html, 2002

[6] http://pr.efactory.de/e-pagerank-implementation.shtml