2002 년 한국정보처리학회 춘계학술발표논문집 제 9 권 제 1 호

Hyperlink 구조와 Hypertext 분류방법을 이용한 Web Crawler

이동원, 현순주
한국정보통신대학교 공학부
e-mail : {zodiac,shyun}@icu.ac.kr

A Web Crawler using Hyperlink Structure and Hypertext Categorization Method

Dongwon Lee, Soon J. Hyun
School of Engineering Information and Communications University

요 약

웹 정보검색에서 웹 문서를 수집하고, 색인을 구축하는 작업에서 Web Crawler 의 역할은 매우 중요하다. 그러나, 웹 문서의 급속한 증가로 인하여 Web Crawler 가 모든 웹 문서를 수집하는 것은 불가능하며, 웹 정보검색의 정확성을 증가 시키기 위한 방법으로 특정한 영역의 문서를 수집하는 focused web crawler 에 대한 연구가 활발히 진행 되어 왔다. 이와 함께, 웹 문서의 link 구조를 이용하여 문서의 집합에서 중요한 문서를 찾는 연구들이 많이 진행되었다. 그러나, 기존의 연구에서는 문서의 link 구조에만 초점이 맞추어져 있으며, hypertext 전체의 연결 구조를 알아야 한다는 문제점이 있다. 본 연구에서는 hyperlink 의 구조와 hypertext 분류방법을 이용하여 문서에 연결된 다른 문서 중 중요한 문서를 결정하는 방법을 제시하고 이를 이용한 web crawler 를 통하여 특정영역에서 정확한 문서를 수집함을 보였다.

1. 서론

현재 인터넷 사용자들은 인터넷을 통하여 많은 정보를 얻고있다. 그러나, 인터넷상의 정보가 급속도로 증가하여 사용자가 원하는 정보를 찾아주기 위한 웹 정보검색 시스템의 역할은 이미 매우 중요한 위치를 차지 하게 되었다[1]. 현재 서비스 되고 있는 웹 정보검색 시스템은 웹 문서의 색인을 만들기 위하여 Web Crawler 를 사용하고 있다. Web Crawler 는 웹 문서 내부에 포함 된 Hyperlink 를 따라서 이동하면서 웹상의 정보, 즉, HTML 문서, 이미지, 또는 기타 자원을 수집하는 일종의 소프트웨어이다[2,3]. 그러나, 웹 문서는 매우 자주 변경되며, 매우 자주 없어지는 특징을 가지기 때문에 Web Crawler 가 웹상에 존재하는 모든 문서를 방문하여 색인을 만들기 위해서는 많은 어려움이 따른다[4]. 그리고, 현재 웹을 통하여 서비스되는 검색 시스템들은 많은 정답 문서를 사용자에게 제시하는 반면 낮은 정확도의 문제점을 가지고 있다. 이러한, 웹 환경에서 사용자에게 정확한 검색 결과를 제시하기 위하여 웹 정보의 분류와 다양한 검색 기술들이 연구되고 있다. 최근, 웹 문서가 가지는 특징 중 하나인 Hyperlink 의 구조를 이용하여 문서를 분류하고 문서 집합에서 중요한 문서를 추출하는 기술들을 웹 정보 검색 기술에 이용하는 연구가 활발히 진행되고 있다[5]. 특히, 웹 검색 시스템에서 사용되는 Web Crawler 가 기존에 사용되는 DFS, BFS 방식으로 웹의 모든 문서를 탐색하는 방법대신, 특정 영역의 문서들만 수집하는 기술과 웹 문서에 연결된 Link 중 Web Crawler 의 목적에 적합한 문서에 대한 Link 를 결정하는 방법을 Web Crawler 에 적용하여 웹 검색 시스템을 통하여 사용자들이 원하는 양질의 결과를 제공하는 것이 필수적이다[4,6,7,8,9,10]. 본 논문에서는 기존에 연구되어진 특정영역에 관한 Web Crawler 에서 나타난 문제점에 대한 해결 방법을 제시하고 이를 실험을 통하여 알아 보겠다. 2 장에서는 기존에 웹 정보검색 시스템에 적용되고 있는 Hyperlink 구조와 Hypertext 문서 분류방법에 대하여 설명하고, 3 장에서는 이러한 방법들의 문제점과 그 해결책을 제시하여 어떻게 Web Crawler 가 효과적으로 Link 를 결정하는지 설명하겠다. 4 장에서는 제시된 방법을 실험하여 기존의 방법보다 효율적임을 설명하고 마지막으로 5 장에서 결론과 향후과제에 대해서 설명하겠다

2. Hyperlink 구조와 Hypertext 문서분류

현재 웹 정보검색은 사용자에게 너무 많은 결과를 제시하며, 중복된 결과, 그리고 낮은 정확도의 문제점을 가지고 있다. 이러한 문제점을 해결하기 위하여 문서 분류를 통하여 검색 영역을 줄이는 것과 많은 결과 문서 중에서 사용자가 원하는 문서를 높은 순위를 부여하여 사용자에게 쉽게 찾을 수 있도록 하는 것이 필수적이다. 이러한, 문서 분류와 순위 알고리즘을 위하여 다양한 방법들이 제시되고 있으며, 최근에

Hypertext 에 포함된 link 구조를 이용하여 문서를 분류하는 방법과 결과 집합에서 link 구조를 이용하여 사용자가 원하는 문서에 높은 중요도를 부여하는 방법들이 제시되고 있다[11,12]. Hyperlink 를 이용하여 hypertext 문서들 중 중요문서를 결정하는 방법 가운데 대표적으로 RankPage 알고리즘은 Google 에 실제로 적용되어 상당히 높은 검색결과를 보여주고 있다. RankPage 알고리즘은 웹 문서에 포함된 link 들이 상호 참조되는 정도를 이용하여 문서들 중 중요한 역할을 하는 문서를 결정하는 방법으로 다음 식에 의해서 문서의 중요도가 결정된다[8].

$$PR(A) = (1-d) + d(\sum_{i=1}^{n} PR(Ti)/C(Ti))$$

문서의 중요도 PR(A)는 A 문서를 참조하는 다른 모든 Ti 문서의 PR(Ti)에 각각 문서들에 포함된 외부로 나가는 link 의 수로 나눈 값들의 합이 된다. 위와 같은 방법을 통하여 검색 결과 집합에서 다른 문서들과 비교하여 중요한 문서들을 사용자에게 우선적으로 보여줌으로써 사용자들로 하여금 자신이 원하는 문서를 빠르게 찾을 수 있도록 도와주는 역할을 한다. 그러나, 이러한 방법은 웹 페이지간의 모든 link 구조를 알아야 정확한 중요도를 계산할 수 있다는 단점이 있다.

위에서 설명한 link 구조를 이용한 중요 문서 결정 방법과 함께 기존의 문서 분류 방법에서 사용되는 문서의 내용과 함께 Hypertext 에 포함된 부가적인 정보인 Hyperlink, Html 태그, link 로 연결된 이웃 문서들의 문서 분류 결과, 그리고 Hypertext 에 포함된 메타 정보를 이용한 문서 분류 방법이 많이 사용되고 있다 [5,12].

3. Link 결정 모델

웹 정보 검색에서 정보검색에 필요한 색인을 만들기 위하여 웹 상에 존재하는 문서 및 기타 자원을 문서에 포함된 Hyperlink 를 통하여 탐색하여 수집하는 Web Crawler 가 사용되고 있다. 그림 1 은 Web Crawler 의 기본 구조를 설명하고 있다.

현재 대부분의 웹 정보검색 시스템은 Web Crawler 를 이용하여 많은 문서를 가져오는 것에 초점을 맞추고 있다. 그러나, Web Crawler 가 웹 문서전체를 가져오기에는 그 양이 많으며, 주기적으로 다시 방문하여 색인을 변경해야 하는 문제점이 있다. 기존의 연구에서는 Web Crawler 가 link 를 따라 탐색할 때 문서에 포함된 link 들의 순서를 부여하여 중요한 문서가 연결된 link 들부터 우선적으로 방문하는 방법이 있다. 이러한 방법들로는 link 로 연결된 문서들간의 코사인

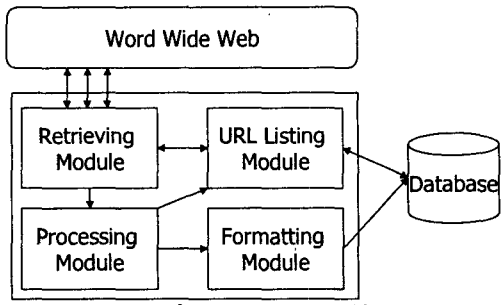

그림 1. Web Crawler 구조

유사도를 이용하여 순서를 결정하는 방법, PageRank 를 이용하여 순서를 결정하는 방법들이 있다[6]. 그리고, 특정한 영역의 문서에 대하여 학습된 Web Crawler 를 이용하여 웹 정보를 수집하게 하여 자신이 원하지 않는 내용을 가진 문서에 대해서는 색인을 하지 않는 방법이 많이 이용되고 있다[4]. 그러나, PageRank 를 이용한 방법은 웹 문서들 사이의 모든 link 구조를 알아야 하는 단점이 있고, 내용만을 이용하는 경우 Hypertext 의 특징인 link 구조를 반영할 수 없는 단점이 있다. 따라서, 본 논문에서는 Web Crawler 가 탐색할 link 를 결정할 때 PageRank 의 단점을 보완하여 웹 문서들 사이의 link 구조를 적용하고, 문서분류방법을 이용하여 문서의 내용을 함께 적용할 수 있는 방법을 제시하였다. 웹 문서는 그림 2 와 같이 link 들로 연결된 tree 형태로 표현될 수 있다.

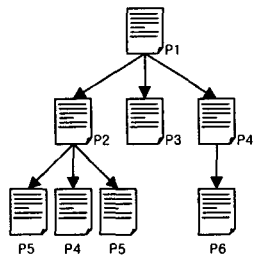

그림 2. Hypertext 구조

위와 같은 Hypertext 구조에서 Web Crawler 가 링크를 결정하는 모델은 다음과 같다.

$P = \{ p_i \mid p_i = p_k$가 지나온 문서 $\}$

$L = \{ p_{k_i} \mid p_{k_i} = p_k$에서 연결된문서 $\}$

$C(p_k) = p_k$문서가 특정영역으로 분류될 확률

$d(p_i)_k = p_k$와 p_i사이의 link 수

문서 p_k 에 연결된 p_{k_i} 들의 중요도는 아래 식에 의해서 계산된다.

$$IPL(p_{k_i}) = C(p_{k_i}) \times \left(\frac{\#\ of\ outgoing\ links}{Max\ \#\ of\ outgoing\ links}\right) + \sum_{p_i \in P_k} \frac{(\gamma^{d(pi)k}\ (IPL(p_i) - C(p_{k_i}))}{\#\ of\ outgoing\ links\ of\ p_i}$$

문서 p_k 에 포함된 link 로 연결된 문서 p_{k_i} 의 중요도 IPL(p_{k_i})은 p_{k_i} 문서의 분류확률에 외부로 연결된 link 의 수를 반영하여 만약 문서가 외부로 연결된 link 가 많다면, 다른 문서들로 연결되는 허브 역할을 하는 문서이므로 중요도를 높여주고, p_{k_i} 가 지나온 다른 문서들의 중요도를 반영하여 만약 다른 문서들이 p_{k_i} 의 중요도 보다 높은 중요도를 가진 문서였다면, 현재 문서의 분류 확률이 낮더라도 적당한 정도의 값을 보상하여 준다. 그리고,0 과 1 사의 값 γ 계수를 곱하여 가까운 문서의 값을 많이 적용하여 link 의 중요도를 결정한다. 위의 식을 적용하여 Web Crawler 가 탐색할 문서를 결정할 때 기존의 link 만 이용하는 방식이나 내용만 이용하는 방식보다 더욱 정확히 Web Crawler 가 원하는 문서를 가져올 수 있다.

4. 구현 및 실험

3 장에서 설명한 Link 결정모델을 적용한 Web Crawler 의 구조는 그림 3 과 같다.

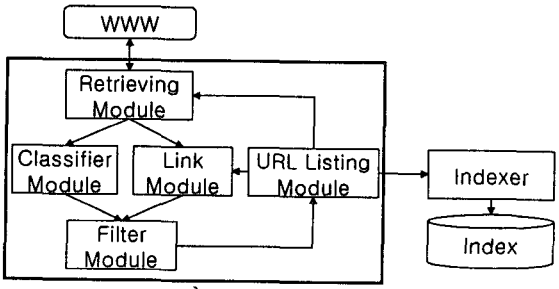

그림 3. Web Crawler 구조

Index DB 에 들어있는 기존의 URL 정보를 통하여 URL 정보의 수집을 시작하게 되며, Retrieving Module 을 통하여 수집된 link 정보는 Link Module 로 입력되고, 문서의 내용정보는 Classifier Module 로 입력된다. Link Module 은 link 정보를 입력 받아서 out link 의 개수와 crawler 가 지나온 link 의 정보를 이용하여 계산하며, Classifier Module 은 각 문서의 내용정보를 이용하여 문서를 분류하게 된다. 이렇게 계산된 값들은 Filter Module 로 입력되어 문서에 포함된 link 들 중 중요도가 높은 순서로 URL Listing Module 로 입력되고 저장되며, 특정 값을 넘지 못한 link 는 버려지게 된다.

실험은 인공지능에 관련된 200 개의 문서를 수집하였다. 이때, 문서에 포함된 상위 20%~100%의 중요도를 가지는 link 를 수집하였으며, 수집된 200 개의 문서에 대하여 인공지능에 관련된 페이지와 유사도를 계산하여 0.3 이 넘은 경우 관련 있는 문서로 인정하였다. 위 실험에서 사용된 문서 분류 알고리즘은 SVM 이며 JAVA1.3 을 기반으로 구현되었다.

위 실험의 결과 그림 4, 5 와 같다.

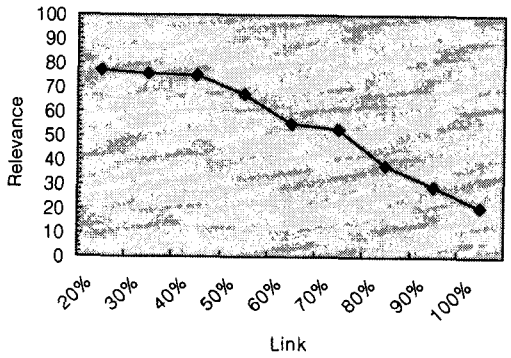

그림 4. 실험결과 1

그림 4 에서 보는 바와 같이 상위 20%의 중요도를 가지는 link 에 대해서만 탐색한 경우는 200 개의 문서 가운데 156 개의 문서가 인공지능과 관련된 문서였으며 BFS 로 탐색한 경우는 약 40 개의 문서만 인공지능과 관련된 문서를 수집한 것을 보이고 있다.

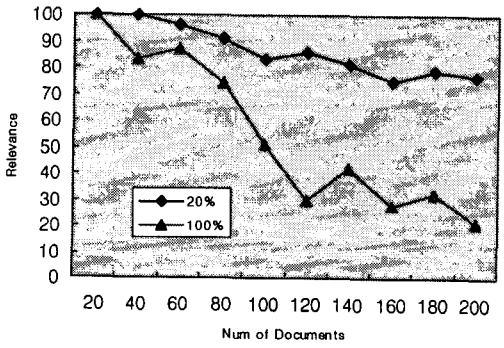

그림 5. 실험결과 2

그림 5 는 문서를 수집하는 동안 문서의 유사도 변화를 보여주는 것으로, 상위 20% 중요도를 가지는 link 를 통하여 수집한 문서는 200 의 문서를 수집하는 동안 70% 이상의 문서가 인공지능에 관련된 문서를 수집하였으며 BFS 를 이용한 문서 수집의 경우는 계속적으로 유사도가 감소하여 200 개의 문서가 수집되었을 경우는 20% 의 문서만 인공지능에 관련된 것을 보이고 있다.

위 결과에서 상위 20%와 BFS 의 경우 모두 문서를 수집하는 동안 외부로 나가는 link 가 많은 페이지들이 연결된 부분에서는 유사도가 감소하지만 이후 다

시 증가 하는 것을 확인 할 수 있었다.

5. 결론 및 향후과제

지금까지 살펴본 바와 같이 기존의 웹 정보검색 시스템에서 사용하고 있는 문서 수집 방법인 BFS 방법을 통하여 문서를 수집하고 색인 한 정보 검색 시스템의 사용자에게 너무 많은 결과집합과 낮은 정확도의 결과를 제시하여 사용자가 원하는 문서를 쉽게 찾는데 많은 어려움이 있었다. 그러나, 웹 문서를 색인하기 위하여 사용되는 Web Crawler 가 제한된 영역의 웹 문서를 수집할 수 있게 하여 웹 정보 검색 시스템의 검색 영역을 줄일 수 있다. 본 연구에서는 Web Crawler 가 link 의 구조와 문서의 내용을 통하여 중요한 link 를 우선적으로 탐색하는 방법을 사용하여 검색 영역의 감소와 검색 결과의 정확성 향상의 측면에서 효과적임을 보였다. 본 논문에서 제시한 방법은 특정한 영역에 대해서만 검색을 지원하는 검색 시스템, 예를 들어, 의료 정보 검색 시스템, 여행 정보 검색 시스템, 생물학 정보 검색 시스템 등에 적용할 경우 효과적 일수 있다. 그러나, 본 연구에서 제시한 방법은 BFS 방법을 이용한 정보검색 시스템에 비하여 문서의 수집속도가 느린 단점을 가지고 있으며 이에 대한 해결책으로는 멀티 에이전트 시스템을 이용한 문서 수집과 단순한 web crawler 가 아닌 web robot 을 이용하여 문서를 검색 시스템으로 옮겨오지 않고 분석하는 방법 등이 있다. 그리고, 웹 문서를 가져오는 Web Crawler 를 강화 학습 방법을 이용하여 학습시켜서, 중요도가 떨어지는 link 의 경우라도 계속 탐색할 경우 중요도가 높은 문서가 수집 될 것을 예측하는 방법등에 대한 연구가 필요할 것이다.

참고문헌

[1] S.Lawrence and C. L. Giles. Accessibility of information on the web. Nature, 400:107-109,July,1995

[2]Brain Pinkerton, Finding What People Want: Experiences with the Web Crawler,SDG,IT94

[3]George Chang, Marcus J. Healey, James A. M. McHugh and Jason T.L.Wang, Mining the World Wide Web: An Information Search Approach, Kluwer Academic Publishers,2000

[4]Soumen Chakrabarti, Martin van den Berg, and Byron Dom, Focused Crawling: a new approach to topic-specific Web resource discovery, 8th WWW Conference, 1999

[5]Hyo-jung Oh, Sung-Hyon Myaeng and Mann-Ho Lee, A Practical Hypertext Categorization Method using Links and Incrementally Available Class Information, 23th SIGIR, 2000

[6]Junghoo cho, Hector Garcia-Molina, Lawrence Page, Efficient Crawling Through URL Ordering,1998

[7]Lawrence Page, Sergey Brin. PageRank, an Eigenvector based Ranking Approach for Hypertext. Submitted to the 21th Annual ACM/SIGIR International Conference on Research and Development in Information Retrieval. 1998.

[8]Chris Ridings, PageRank Explained, http://www.google.com/technology/,2001

[9]Byoung-Tak Zhang and Young-Woo Seo, Personalized Web-Document Filtering Using Reinforcement Learning

[10]Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment,1998

[11]Ben Kao, Joseph Lee, Chi-Yuen Ng, and David Cheung, Anchor Point Indexing in Web Document Retrieval, IEEE Transaction on Systems, Man, And Cybernetics-Part C: Applications and Reviews, Vol.30, No 3, 364-373, 2000

[12]Yiming Yang, Sean Slattery, And Rayid Ghani, A Study of Approaches to Hypertext Categorization,Journal of Intelligent Information Systems,2000