# 동적인 계층적 웹 검색 시스템

윤광호*, 이우기**, 김창민*

*성결대학교 컴퓨터공학부 khyoon@gmail.com, kimcm@sungkyul.edu
**인하대학교 산업공학전공 정보공학연구실 trinity@inha.ac.kr

# Implementing Dynamic Web Hierarchical Structuring System

Kwangho Yoon*, Wookey Lee**, Changmin Kim*

*Computer Science, Sungkyul University, Anyang, Korea
**Informatics Engineering Lab., Div. Industrial Engineering, Inha University, Incheon, Korea

## 초록

웹은 유향그래프로 표현할 수 있으며, 이때 웹 페이지를 노드로 페이지 사이의 하이퍼링크를 아크로 나타낸다. 본 연구에서는 이러한 웹 그래프로부터 검색트리를 유도하여 이를 브라우징할 수 있는 시스템으로 구현하였다. 본 시스템은 사용자가 임의의 노드를 동적으로 검색의 시작노드 즉, 루트노드로 결정할 수 있고 이를 변경하여 새로운 검색을 할 수 있으며, 그 루트노드로부터 도달 가능한 노드들의 집합을 도메인이라 하고, 노드 및 링크에 대하여 일정한 가중치를 부여함에 따라 검색트리를 구성할 수 있으며 이를 검색트리 시스템으로 구현하였다.

## 1. 서 론

웹의 폭발적 성장은 양적 질적으로 괄목할 만하다. 양적으로 웹은 현재 조 단위의 페이지를 넘어선지 오래되었다. 질적으로는 단순한 HTML, XML 등의 웹 페이지뿐만 아니라, 전자화된 매뉴얼, 이메일, 그림, 동영상, 심지어 웹 서비스까지 포괄한다. 이렇게 커질수록, 웹 사용자는 점점 더 원하는 정확한 정보를 찾기가 더욱 어려워진다. 따라서 웹 정보를 구조화하여 표현하는 것은 의미 있는 접근법이 될 수 있다. 이때 구조화란 웹을 그래프 관점으로 나타내는 것으로 이를 웹 그래프(Web as a graph)라고 부른다[2,6].

웹을 유향그래프(a directed graph) G = (N, A)로 인식하되, 유향그래프 G의 N과 A는 웹 페이지를 의미하는 노드집합과 노드사이의 아크 집합을 각각 의미하고, 이러한 웹 그래프에서 구조화 검색트리(structured search tree)를 형성하는 것이 본 연구의 목표이다. 또한 우선검색트리의 형성이 웹 사용자의 선택에 따라 동적으로 이루어질 수 있어서, 검색의 시작을 어느 노드에서 하느냐에 따라 그 위치를 루트로 하는 검색트리를 만들 수 있도록 하는 것이다. 또한 하위그래프G* = (N*, A*)는 그래프G로부터 유도되는데, 만일 아크A*⊆A 의 각 시작노드 및 끝 노드가 각각 N*⊆N에 속해야 한다. 하위 그래프G*는 검색 대상 즉, 구조적 검색도메인이라고 부르게 된다. 여기서 얻어질 수 있는 검색트리는 구체적으로는 루트를 가진 유향 스패닝 트리(rooted directed spanning tree) 혹은 루티드 아보레슨스(rooted arborescence)라고 부르는데, 여기서는 검색트리라고 일반화하여 줄여 부르도록 한다. 이러한 검색트리에서 개별노드로 입력되는 아크는 하나로서 모두 n-1개의 아크가 n개의 노드에 연결되며 사이클이 없는 유향그래프가 되어야 하고, 결과적으로 검색트리는 단 하나의 루트로부터 모든 노드에 유일한 경로가 존재하는 경우를 의미한다[5].

검색대상 도메인 즉, 검색트리를 생성하게 되는 범위문제의 경우, 이러한 도메인은 선정된 루트노드로부터 도달 가능한(reachable) 웹 페이지들의 집합으로써 정의된다. 잠정적으로 인덱스 "0"을 가지는 노드를 특정 웹 사이트의 시작페이지(default page)로서 index.html이나

default.php 등을 검색대상의 루트로 결정하는 경우 루트에서부터 도달가능한 웹 사이트 전체가 대상 도메인이 될 수 있다. 다른 극단으로는 단 하나의 단말페이지(dangling page)가 대상도메인이 될 수도 있다. 이러한 양 극단을 포함하는 일반적인 여러 웹 페이지가 하이퍼링크로 연결되어 있는 검색도메인을 잘 표현할 수 있는 검색트리를 찾는 것이 본 연구의 대상이 된다.

본 논문은 다음과 같이 구성되어 있다. 제2장은 가중치평가 방식에 대해 설명하며, 제3장에서는 웹구조화모델에 대해 기술하고 그 중에서 계층적 모델링에 대해 설명한다. 제4장에서 계층적 웹 검색시스템의 구현절차를 단계별로 나타내고 제5장에서 그 구현내용을 예시하며, 제6장에서는 관련연구에 대해 기술하며, 마지막으로 제7장을 통해 논문의 결론을 기술한다.

## 2. 가중치 평가방식

웹은 유방향 그래프 $G(N, A)$로 표현할 수 있다. 이 때, $N$은 웹페이지를 표현하는 웹 노드 집합으로 $A$는 웹 노드를 연결하는 하이퍼링크들을 표현하는 웹 아크 집합을 의미한다. 그래프의 행렬 표현인 인접 행렬(adjacency matrix) $M = [m_{ij}]$는 웹페이지 $i$에서 $j$로의 아크가 존재하면 $m_{ij} = 1$로, 존재하지 않으면 $m_{ij} = 0$으로 하여 표현된다. 인접 행렬은 PageRank와 HITS(Hypertext Induced Topic Search) 및 Social Network 등에서 웹페이지들에 정량적인 순위를 제공하는데 사용되어 왔다.

웹 노드($N$) 및 웹 아크($A$)를 표현하는 일반적인 가중치(weight) 혹은 거리기준(distance measure)은 정보검색 연구그룹에서는 글로벌 가중치와 로컬 가중치로 표현된다[1,3]. 이 두 가지 가중치는 어떤 의미에서 상호 독립적이다. 즉, 글로벌가중치는 개별적 검색 키워드를 고려하지 않고 하이퍼링크 정보만을 이용하여 구해지며, 한편 로컬가중치는 하이퍼링크를 배제하고 키워드 및 질의어를 사용하여 가중치를 구하기 때문이다. 현재 많이 쓰이는 글로벌 가중치 방식에는 페이지랭크 방법[6], J. Kleinberg의 HITS방법[4] 등이 있다. 로컬가중치로서는 벡터공간모형(VSM: Vector Space Model)에 기반한 코사인가중치, 확률적가중치, 퍼지가중치 및 흔히 쓰이는 tf-idf(term frequency and inverse document frequency)방법[1,2,3] 등이 있다. 노드의 내용(키워드)에 기반한 노드 가중치 평가방법과 달리 아크정보를 활용한 아크기반 노드가중치 평가방법이 최근 구글의 성공과 함께 주목받고 있다. 이 방법에서는 앞서 언급한 인접행렬$M$에 대해 출력링크 수를 기준으로 정의하고, 이때 $i, j \in N^*$이 된다. 이러한 인접행렬에 대하여 행렬의 수렴조건을 강제하기위한 전처리를 위해 감쇠요소(demping factor) 등의 조건을 부여하고난 다음, 마코프 체인에 따른 노드별 가중치를 구하게 된다. 이러한 페이지랭크의 성공에 기인하여 여러 가지 성능향상 및 보완을 시도하는 다양한 가중치 계산기법들이 개발되고 있다[2,3]. 한편 HITS는 웹 페이지들을 중요페이지를 권위있는 노드(Authority node)로 분류하는 가정에서 출발하여, 다른 페이지에 연결기능이 강화된 노드를 허브노드(Hub node)로 정의하면, 이 경우 중요한 허브노드는 권위있는 노드에 많이 연결되고, 역으로 권위있는 노드는 중요한 허브노드에 연결되는 경향이 강하다는 가정 하에 개발되었다. 이 개념은 Social Network이나 Citation reference 등의 분야에 본원적인 알고리즘이 되어 주목받고 있으나, 정작 처음 의도했던 정보검색 및 검색시스템 개발로는 이어지지 못하고 있다[11. 12].

이때 페이지랭크 및 HITS와 같은 방법은 개별적인 키워드와는 무관하게 웹페이지 사이의 하이퍼링크 숫자 만을 사용하여 가중치를 구하므로 글로벌 가중치라고 부를 수 있다. 그러나 글로벌 가중치 만 가지고는 개별적 검색결과에 대한 지원되지 않으므로 키워드에 기반한 전통적인 방법들 즉, 로컬가중치(local weight)가 보완적으로 필요하다. 그 대부분이 전통적 정보검색에서 사용하는 벡터공간모델(VSM: vector space model)에 기반하고 있고, 그 중에서도 가장 많이 사용하는 방법의 하나가 *tf-idf*이다. 그러므로 로컬 가중치는 노드 기반의 노드 가중치 평가방식이라 명명될 수 있으며, 글로벌 가중치는 링크기반 노드 가중치라 부를 수 있다. 본 연구에서는 글로벌가중치와 로컬가중치의 곱을 사용하여 가중치로 사

용하였지만, 가중치방식의 선정과 그 결합방식이 본 연구에 제약요인이 되지는 않는다. 또한 본 연구에서 구현한 내용은 [8]의 연구와 출발점은 유사하나, Social Network이나 웹아카이브(Web Archive)를 대상으로 한다는 점에서 차이를 보인다.

### 3. 웹 구조화 모델

여기에서의 유향그래프는 루트노드를 가진다고 가정하는데, 이 노드는 시작노드 집합에서 얻고, 이 시작노드 집합을 얻는 방법은 여러 가지가 가능한데 가장 일반적인 방법은 기존의 검색엔진에 사용자가 질의하여 얻어지는 URL 리스트를 사용하는 것이다. 그러면 해당 유향그래프와 그로부터 얻어지는 검색트리는 루트노드와 하이퍼텍스트 링크 즉, 아크로 연결된 한 웹 사이트내의 다른 노드들의 집합으로 구성된다. 그리고 이 집합이 검색을 위한 구조화의 대상도메인이 된다.
따라서 본 연구에서는 웹사이트를 구조화하기 위해, 아크 중요도의 합을 최대화하는 계층적인 구조를 만들어내는 것이 목표가 된다. 이러한 구조는 최대 중요도 트리생성 문제로 치환할 수 있으며, 이는 기본적으로 최소비용트리 문제는 다음과 같은 트리생성 알고리즘으로 생성할 수 있다. 본 방법은 [5]에 기반하여 구현하였으며, 주어진 그래프에서 모든 아크의 중요도 및 루트 노드 r을 입력으로 하여 다음 알고리즘은 중요도 합을 최대로 하는 검색트리구조를 생성하는 것이다.

### 4. 계층적 웹 검색 시스템

여기서는 검색트리를 브라우져로 표현하는 것으로 그 절차는 다음과 같다. 대상이 되는 웹사이트를 선정하는 방법은 우선 하나의 검색엔진에서 키워드를 입력하여 얻어진 결과를 가지고 검색트리로 구현하는 것이다. 본 연구에서 구현한 시스템의 수행절차는 다음과 같다.

단계1: 시작집합을 얻는 단계이다. 시작집합이란 사용자가 검색을 시작할 수 있는 대상이 되는 URL을 지칭하는데 이를 얻는 방식은 몇 가지가 가능한데, 우선 웹에서 랜덤서치를 하는 방법과 WHOIS 등의 메타정보로부터 얻는 방법, 어떤 검색엔진에 키워드를 입력하고 얻어지는 URL들을 대상으로 하는 방법 등이 있다. 여기서는 마지막 방법을 채택하기로 하였다.

단계2: 이 단계에서는 앞에서 얻어진 시작노드 집합에서 특정 URL을 하나 선정해 루트노드라고 이름하며, 이 루트노드를 시스템에 입력하여 도달하는 웹 페이지 및 그 웹페이지와 연결된 도달 가능한 웹 페이지들을 선정하기 시작한다. 물론 이때 모든 URL들은 각각 구조적 검색에서 루트노드가 될 수 있으며 이중 사용자가 임의로 혹은 Top 1위의 URL을 루트노드로 선정할 수 있다.

단계3: 앞 단계에서 선정된 루트노드로부터 도달가능한 URL집합($N^*$)이 검색대상 도메인($G^*$)이 된다. 루트노드를 결정하면 사용자의 브라우저에는 해당 URL이 지칭하는 웹 페이지의 내용이 제시된다. 또한 웹사이트의 노드와 아크가 분석된다. 이때, 검색가능(indexable) 웹페이지들의 범위를 결정해주는데 검색에 불요한 잡음(Noise)를 제거하는 절차를 동시에 수행하게 된다. 여기에는 Hidden Web 및 redirection 기능 등을 배제하고 검색가능 웹페이지들을 결정해주게 되는데, 이 과정은 확장자에 기반하여 제거하는 절차로 이루어진다. 예컨대, png, img, jpg, c, java, txt, doc, xls, ppt, hwp, 등이 해당된다.

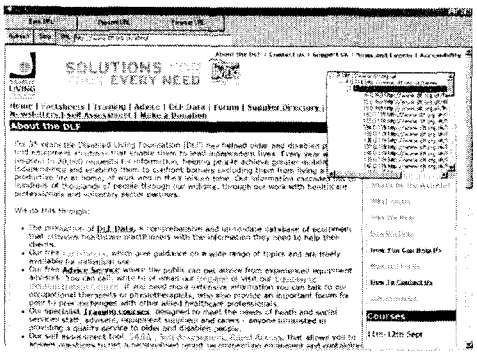
단계4: 노드 및 아크 집합을 확정하고 난 다음 아크의 가중치를 구하게 된다. 여기서는 글로벌 가중치로서 PageRank, 로컬 가중치로서 tf-idf를 구하고 이들 양자의 곱으로 아크 가중치를 정의하며, 이에 따라 가중치를 계산하였다.

단계5: 상기 방식에 따라 얻어지는 자료를 대상으로 본 연구에서는 [5]의 알고리즘에 기반하여 검색트리를 유도하였다. 이때 다양한 검색요구기준을 적용하여 필요한 트리를 구한다. 또한 사용자가 어떤 URL을 선정하는가에 따라 루트를 바꿀 수 있고, 그 루트로부터 상

기 언급된 단계를 적용하여 새로운 검색결과를 얻을 수 있으며 이러한 강점을 따서 동적(dynamic)하다는 이름을 붙였다.

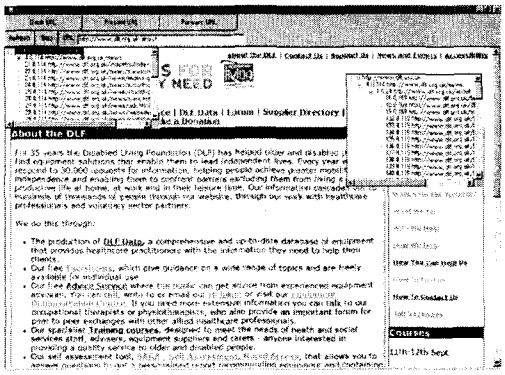
### 5. 검색시스템 구현

앞서 언급한 절차를 따라 특정 도메인에 대한 웹 페이지 정보들을 수집하여 그래프 형태의 구조를 트리 형태로 변환하고 웹 브라우저에 구현한 결과를 다음 그림들에서 확인할 수 있다.



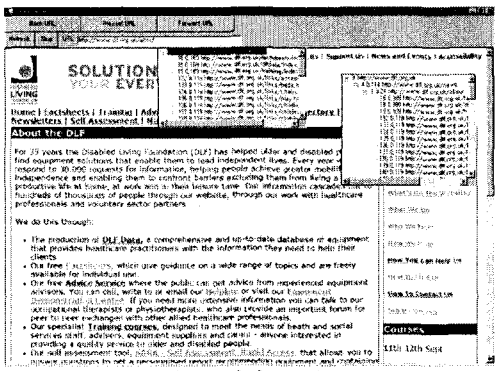
[그림 1] 화면 구성

우선 그림 1은 http://www.dlf.org.uk/about이 루트노드가 된 것을 알 수 있으며, 이들의 구조 정보가 웹 브라우저에 나타난 상태이다. 특정 도메인 내의 웹 페이지의 계층 구조는 오른쪽 상단에 있는Tree View 컨트롤에 나타난다. 이 컨트롤은 위치의 이동과 크기의 축소와 확장이 가능하다. 컨트롤 내에서 웹 페이지는 웹 페이지 번호와 함께 Tree Node로 표현된다. Tree Node는 한 개의 부모 노드와 여러 개의 자식 노드를 갖는다. 현재 페이지의 노드 번호는 1번이며, 웹 페이지 번호가 4번인 노드 http://www.dlf.org.uk/news의 자식 노드이다.



[그림 2] 상위구조 정보

상위구조정보는 Tree View 컨트롤로 화면에 표시된다. 이 Tree View는 현재 페이지와 현재 페이지의 하위 페이지를 제외한 웹 페이지 노드로 구성되어 있다. 그림 2에서 표시되어 있는 1번 노드는 오른쪽의 Tree View에는 표시가 되어있지만 왼쪽의 Tree View에는 1번 노드와 1번 노드의 자식 노드들은 표시되어 있지 않다.



[그림 3] 하위구조정보

하위구조정보의 경우 위 그림에서 보이는 Forward 버튼을 누르면 Back 버튼을 눌렀을 때와 같이 Tree View 컨트롤이 화면에 표시된다. 하지만 Back 버튼을 눌렀을 때와 다르게 Forward를 눌렀을 때 나타나는 Tree View 컨트롤은 현재 페이지 노드와 현재 페이지 노드의 자식 웹 페이지 노드들을 표시한다. 그림 3에 표시된 웹 페이지의 번호는 2번이다. 가운데 Tree View 컨트롤의 최상위 노

드는 1번이며 오른쪽 Tree View 컨트롤에서 1번 노드가 가진 자식 노드를 똑같이 갖고 있는 것을 볼 수 있다.

## 6. 관련 연구

최근 웹의 구조를 분석하는 접근법은 관련된 정보의 검색의 효과성 및 효율성을 제고하는데 중요한 요소로 간주되고 있다. 웹의 구조적 검색 관점에서 하이퍼링크의 구조를 적용하는 연구는 크게 보아 웹 객체 가중치평가, 검색엔진에서의 웹 페이지 순위부여방식, 토폴로지 및 시각화 방식 등이 있다[3,8]. 웹 객체 가중치평가에 있어서는 정보검색 연구자 및 개발자들의 경우 페이지랭크 및 HITS가 가장 많이 활용되고 있다 [4,6]. 웹의 아크 분석에서 HITS의 경우 매우 주목을 받은 알고리즘으로서 연결성 노드(hubs)와 의미를 가진 노드(authorities)를 매우 효율적인 반복연산을 통해 제시했다. 의미를 가진 노드란 특정 주제에 대해 좋은 내용을 담고있는 노드를 의미하며 좋은 연결성 노드에 아크를 많이 가지며, 또한 좋은 연결성 노드는 의미를 가진 노드에 많이 연결된다는 선순환구조에 가정하고 있다. 이 알고리즘은 여러 분야에 활용되었는데 상호강화 관계(mutually reinforcing relationships)라고 부르는 social network을 찾는 문제에 주로 응용되고 있지만, 웹 페이지 검색에는 사용되지 않는다[12]. 로컬 가중치는 기존의 정보검색 연구 진영에서 매우 많은 시도를 해왔으며, 전통적인 벡터공간모형(vector space model)에 입각한 코사인 가중치, 확률 및 퍼지방식, 그리고 본 연구에 도입한 tf-idf등의 방식이 그것들이다. 그런데 본 연구에서 제안하는 글로벌 가중치 및 로컬 가중치의 통합 그리고 이의 아크 가중치로의 전환 등은 최초의 시도이다. 글로벌 및 로컬 가중치 통합의 또다른 시도는 Hou and Zhang[9]이 개발한 LLI가 있으며 이는 링크를 singular value decomposition 방식을 활용한 것으로, 본 연구와는 다르다.

한편으로 웹 검색을 지원하는 시각화 도구를 개발하는데 있어서는 Hyperbolic 브라우저[7]의 경우 2차원 hyperbolic 공간에 트리를 표현하였고, 또한 Cybermap 및 tree graph 방식[10]도 폴더형태로 방문 웹페이지를 표현하려고 하였으며, Toyota & Kitsuregawa [8]의 경우에도 사회망 및 웹 아카이브에 대한 시각화 도구를 제안하였다. 이러한 시각화 도구들은 그래프적 표현에 초점을 맞추었지, 웹 검색의 구조를 고려한 연구는 아니었다.

## 7. 결 론

본 연구에서는 웹 검색에 있어서 새로운 접근법으로서 동적인 계층화 트리를 제안하고 구현한 결과를 제시했다. 구조화 웹 검색은 우선 웹 객체에 있어서 노드와 아크의 유향 그래프 구조에서 아크의 가중치를 구하는 방법과 이러한 트리에서 사용자 검색을 위해 가중치를 평가하는 방식을 포함한다. 그 절차에서 중요한 점은 우선 시작단계에서 검색 대상을 결정한 다음, 루트노드 선정하되, 사용자가 임의로 결정할 수 있으므로 동적인 특성을 가진다. 다음으로 루트노드로부터 도달가능한 URL 집합 즉, 검색대상 도메인에 따라 해당 URL 웹사이트의 노드와 아크가 분석하고, 최종적으로 절차에 따라 검색요구기준을 바꿀 수 있는 검색트리를 메타정보로서 구현한 것이다.

본 연구는 여러 방향으로 확장될 수 있다. 본 논문은 유방향 웹그래프를 일반적인 그래프 구조로 간주하였다. 따라서, 하이퍼링크의 분포와 같은 웹 그래프의 특성을 활용한 특화된 방법론에 대한 개발을 통해 보다 효율적인 검색과 가이드가 가능할 수 있을 것이다. 또한 본 연구를 통해 생성된 웹사이트 구조의 유효성을 입증을 위한 실제 사용자에 대한 비교분석이 필요하다.

## 참고문헌

[1] O. Etzioni, M. Cafarella, D. Downey, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates, "Methods for Domain-Independent Information Extraction from the web: An Experimental Comparison," *AAAI*, pp. 391-398, 2004.

[2] E.J. Glover, D.M. Pennock, S. Lawrence, and R. Krovetz, "Inferring Hierarchical Descriptions," *Proc. CIKM*, pp. 507-514, 2002.

[3] M. Hammami, Y. Chahir, and L. Chen, "WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis," *IEEE TKDE*, vol. 18, no. 2, pp. 272-284, 2006.

[4] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM*, vol. 46, no. 5, pp. 604-632, 1999.

[5] W. Lee, "Hierarchical Web Structuring from the Web as a Graph Approach with Repetitive Cycle Proof," APWeb Workshops, pp. 1004-1011, 2006.

[6] G. Pandurangan, P. Raghavan, and E. Upfal, "Using PageRank to Characterize web Structure," *Proc. COCOON*, pp. 330-339, 2002.

[7] P. Pirolli, S.K. Card, and M.M. Wege, "The Effects of Information Scent on Visual Search in the Hyperbolic Tree Browser," *ACM TOCHI*, vol. 10, no. 1, pp. 20-53, 2003.

[8] M. Toyota, and M. Kitsuregawa, "A System for Visualizing and Analyzing the Evolution of the Web with a Time Series of Graphs," *Proc. Hypertext*, pp. 151-160, 2005.

[9] J. Hou and Y. Zhang, "Effective Finding Relevant web Pages from Linkage Information," *IEEE TKDE*, vol. 15, no. 4, pp. 940-951, 2003.

[10] P. A. Gloor, and S.B. Dynes, "Cybermap - Visually Navigating the Web," *Journal of Visual Languages and Computing*, vol. 9, no. 3, pp. 319-336, 1998.

[11] J.M. Kleinberg, "Patterns of Influence in a Recommendation Network," *Proc. PAKDD* pp. 380-389, 2006.

[12] L. Singh, L. Getoor, and L. Licamele, "Pruning Social Networks Using Structural Properties and Descriptive Attributes." *Proc. ICDM* pp. 773-776, 2005.