2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

# 하이퍼텍스트 기반의 정보 지도에 관한 연구*

류 철*, 이강찬**
* 충남대학교, **한국전자통신연구원
e-mail : ryuch@ce.cnu.ac.kr, kangchan@etri.re.kr

# A Study on Information Map based on Hypertext

Cheol Ryu*, Kangchan Lee **
* Chungnam National University
** Electronics and Telecommunication Research Institute

요 약

웹 문서는 하이퍼텍스트의 특성을 가지는 문서 형태를 가지며, 일반적인 문서의 특성 보다는 사용자에 의하여 쉽게 생성, 변경, 삭제되는 특성을 가지고 있다. 본 논문은 WWW 검색 엔진이 WWW의 확장성과 역동성을 반영하지 못하는 단점을 보완하는데 그 의의가 있다. 본 논문에서 제시하는 시스템은 기존의 WWW 검색 엔진을 통하여 얻은 검색 결과를 출발점으로 한 실시간 검색을 통하여 WWW 문서의 현재 상태를 정확하게 파악할 수 있는 장점이 있다. 또한 탐색 결과의 가시화를 통하여 웹 문서에 대한 정보 지도(information map)를 추출할 수 있으며, 이러한 기능을 통하여 기존의 정보 검색 엔진에서 제공하지 못하던 자신의 정보 요구에 맞는 정보 지도를 제공함으로써 새로운 지식의 전달을 꾀할 수 있다.

## 1. 서론

1990 년 초반, 인터넷 환경을 변화를 가져온 WWW 이 개발됨에 따라 지난 10 여년간 인터넷 상의 정보의 양은 기하급수적으로 증가하고 있으며, 매일 새로운 정보가 생산되고 있다. 이러한 웹을 검색하는 방식으로는 크게 에이전트를 이용하여 각 웹사이트의 정보를 수집한 후, 검색할 수 있도록 하는 *키워드 기반 검색 방식*과 각 인터넷 사이트를 지정해 놓은 분류 방식에 따라 분류해 놓은 *디렉토리 기반 검색 방식*으로 나누어 볼 수 있다. 그리고 추가적으로 기존의 검색 엔진을 통합하여 검색할 수 있는 메타 검색도 현재 유용히 사용하고 있는 검색 방식중의 하나 이다. 현재 대부분의 검색 엔진에서는 키워드 기반 검색과 디렉토리 기반의 검색을 기본적으로 제공하고 있으며, 다수의 검색 엔진에서 메타 검색도 지원하고 있다. 즉, 대부분의 검색 엔진에서는 사용자 편의를 위하여 다양한 검색 방식을 지원하고 있다[1].

일반적인 웹 검색 엔진은 WWW 의 문서를 수집하여 전통적인 정보 검색 방법론을 적용하여 수집된 문서를 색인하고 그 결과를 데이터베이스에 저장한 후, 인덱싱 과정 등의 후처리를 거친 후, 사용자가 질의할 때 마다 질의에 맞는 검색 결과를 생성하여 전달하는 방식을 채택한다.

그러나 일반적인 문서와는 다르게 웹 문서는 하이퍼텍스트의 특성을 가지고 있다. 대부분의 정보 검색 엔진들은 전통적인 정보검색 방법론을 채택하여 검색 엔진이 개발되기 때문에 하이퍼 텍스트의 특성을 살린 기능을 제공하고 있지 않다.

본 논문에서는 검색 엔진의 부가적인 기능으로 실시간으로 검색 결과에 대한 검증과 함께 하이퍼텍스트를 기반으로 하는 정보 지도를 제공하는 기능에 대하여 언급한다.

## 2. 실시간 탐색 기법

WWW 의 현재 상황을 반영하기 위해서는 WWW 탐색 엔진이 색인하기 위하여 WWW 문서를 수집할 때의 시점이 아닌 현재의 WWW 문서를 수집하는 것이 필요하다. WWW 에서 실시간 탐색이란 사용자가 질의를 입력하였을 때 비로소 문서를 수집하여 그 수

* 본 연구는 충남대학교 자체연구비(과제명 : 하이퍼미디어를 이용한 연구정보 베이스의 구축에 관한 연구)의 지원으로 수행되었습니다.

집된 문서를 색인하고 관련도가 높은 문서의 URL 을 사용자에게 돌려주는 정보검색 방법이다.

실시간 WWW 문서 수집에 영향을 미치는 변수는 네트워크의 속도와 수집해야 하는 문서의 개수와 문서의 크기이다. 전체 문서를 수집하는데 필요한 시간을 수식으로 나타낼 수 있다.

$$T = \sum_{s \in S} \sum_{i=1}^{n_s} l(d_{s,i}) t_s$$

**수식 1. WWW 문서 수집 시간**

수식 1은 WWW 전체를 탐색할 때 소요되는 시간 T 를 나타낸 것으로, 전체 WWW 에 존재하는 WWW 사이트마다 접근하는 시간이 다를 때, 그 WWW 사이트에 위치하는 WWW 문서들을 수집할 때 걸리는 시간이다. 문제는 $n_s$ 와 S 의 개수가 점점 증가한다는 점에 있다.

현실적으로 사용자의 질의에 대하여 전체 WWW 을 검색하는 것은 불가능하다. 그러나, WWW 탐색 결과에 초점을 맞추어 실시간으로 문서를 수집하는 것은 가능하다. 전체가 아니라, 특정한 정보와 관련되었을 것으로 추측되는 WWW 문서들을 수집하는 것은 탐색 공간을 크게 줄여준다. 일반적인 WWW 문서를 출발점으로 실시간으로 연결된 WWW 문서를 수집하는 것은 다음과 같은 시간이 소요된다.

$$N = \sum_{i=0}^{d} b^i$$

**수식 2. 일반적인 웹 문서에서 출발한 문서 수집 개수**

그러나, WWW 탐색 결과가 잘 연결된 WWW 문서들의 집합이라면 수집해야 할 새로운 문서의 개수를 의미하는 분기 개수 β 는 수식 2에서의 평균적인 분기 개수 $b$ 와 성격이 다르다. 즉, 적절한 깊이 $d$ 의 WWW 문서의 집합에서는 서로 가리키는 문서들이 많기 때문에 분기 개수가 실제적으로는 줄어든다.

$$N = \sum_{i=0}^{d} \beta^i$$

**수식 3. WWW 검색 엔진의 결과에서 시작한 문서의 수집 개수**

β 가 적절히 결정되면 수집해야 하는 WWW 문서의 개수가 적어지므로 WWW 문서 수집시간은 짧아진다. 탐색 깊이가 깊어질수록 수집해야 할 문서의 개수는 기하급수적으로 증가하게 된다. 그림 1의 N 은 일반적인 WWW 문서에서부터 탐색을 시작했을 때, 수집해야 할 문서의 개수를 나타낸 것이다. N'은 WWW 탐색 결과에서부터 탐색을 시작했을 때, 수집해야 할 문서의 개수를 나타낸 것이다.

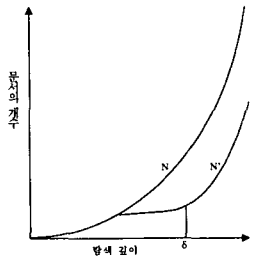


**그림 1. 탐색 공간의 증가 그래프**

그림 1은 WWW 탐색 결과에 포함되어 있는 URL 에서부터 적절한 깊이 만큼의 실시간 탐색을 한다면 전체의 WWW 을 검색하지 않고 관련도가 높은 WWW 문서를 비교적 짧은 시간 안에 검색할 수 있다는 것을 의미한다.

## 3. 하이퍼텍스트를 고려한 관련도

현재 WWW 검색 엔진들은 단어 빈도수(Term Frequency)를 이용하여 관련도를 계산하고 있다. 이것은 WWW 문서가 하이퍼텍스트임을 전혀 고려하지 않고, 문서의 내용을 중심으로 관련도를 계산한 것이다. 그러나, Croft 에 의하면 하이퍼텍스트를 위한 확률 모델에 정보검색 방법론은 적용하면 정확율이 20%가 증가한다고 한다[2]. 또한 그래프를 사용하여 하이퍼텍스트 구조를 비교하고 표현하는 방법이 [3] 고안되어 하이퍼텍스트의 구조를 이용하는 방법이 제시되었으며, 하이퍼링크는 아니지만 유사한 참고문헌 정보를 정보검색의 효율을 높이는 방법[4]으로 사용할 수 있음이 알려져 있다. 본 논문은 간단한 방법으로 정보검색의 효율을 높이는 방법을 소개한다.

질의와 문서의 공통된 단어의 개수를 유사도라고 한다면, 관련도는 유사도가 이웃에 영향을 준 결과로 얻어진다. 한 노드의 관련도는 그 이웃 노드의 관련도에 영향을 미친다. 관련도를 이용하면 하이퍼텍스트의 상호참조의 의존성을 고려한 랭킹을 만들어 낼 수 있다. 여기서 주목할 점은 하이퍼텍스트의 상호참조 자체로는 관련도를 만들어 낼 수 없고, 유사도가 제공되어야 한다는 점이다. 즉, WWW 문서는 많은 하이퍼 링크들이 복잡하게 연결되어 있지만, 두 문서가 링크 되어있을 때에는 서로 내용적 유사성이 존재할 때와 존재하지 않을 때의 두 가지 경우가 있다. 이 두 가지 중에서 후자는 제외하고 전자만 관련도에 영향을 미치고 있다. 아무리 많은 하이퍼 링크가 존재한다고 해도 하이퍼 링크가 가르키는 문서의 유사도가 작다면 그 이웃들은 관련도에 영향을 미치지 못한다.

그림 2는 연결되어 있는 노드를 일직선상에 그린 것으로, 서로에게 영향력을 미치는 것을 나타낸 것이다. 자신의 주위에 연결된 노드의 유사도가 높으면 자신의 유사도도 증가하는 것을 볼 수 있다. 즉, 하이퍼텍스트에서 자신의 관련도는 주위의 유사도에 영향을 받는 것이 반영된 것이다. 실시간으로 검색하기 때문에, 단어의 변별도(discrimination value)를 구하는 것의 비용을 감당해 낼 수 없고, 또 실시간으로 한 문서가

수집되었을 때 바로 관련도를 계산할 필요성이 있기 때문에 주위의 유사도를 이용한 하이퍼텍스트 색인 방법이 유용하다.

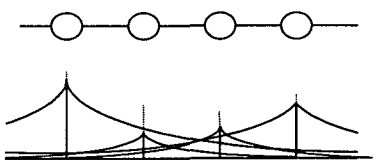

그림 2. 유사도의 이웃 지배도

$$R_d = S_d \cdot (1 + H_d)$$

$$H_d = \sum_{i=1}^{I} H_d^i$$

$$H_d^i = \sum_{n \in N_d^i} wf(i) \cdot S_n$$

수식 4. 이웃 지배도

## 4. 탐색 결과의 가시화

WWW 문서들이 하이퍼 링크로 연결된 네트워크는 문서들의 구조적 관련도를 나타낸다. 이것을 그래프로 나타내면 사용자는 정보의 연결 구조를 파악할 수 있다. 또한 한가지 주제에 대하여 연구하는 연구 그룹이 분산되어 존재할 때, 그들이 구축해 놓은 WWW 서비스로부터 그림 3과 같은 정보지도가 그려질 수 있다.

밀접하게 연결된 부분들은 정보가 동질적인 것을 의미한다. A 와 F 보다는 A 와 D 가 더 밀접하게 관련을 맺다. 즉, 최단 참조 사슬(Shortest Referential Chain)의 길이가 A-D 보다 A-F 가 더 길다. 특정 노드에서 최단 참조 사슬의 깊이를 제한 함으로 정보영역을 정의할 수 있다. 정보영역 I 는 A 를 중심으로 최단 참조 사슬의 깊이를 1 로 제한한 영역이다. 정보영역 II 는 C 를 중심으로 최단 참조 사슬의 깊이를 3 으로 제한한 영역이다. I, II 는 구조적으로 밀접하게 연결된 두 정보 영역을 각각 나타내고 있다. 여기에서 주의해야 할 점은 A 를 중심으로 한 정보영역 I 이 정보의 동질성보다 정보영역 II 의 정보의 동질성이 높다는 점이다.

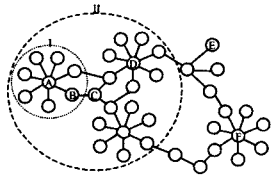


그림 3. 정보 영역

관련도가 E > D > A > C < B < F 의 순서라고 하더라도, 구조적으로 C 는 모든 전체의 노드에서 가장 중심적인 위치를 차지하고 있다. 정보들의 관계를 살펴본다면 E 를 방문하는 것보다 C 를 방문하는 것이 사용자가 원하는 정보를 찾는데 짧은 경로를 제공할 수 있다.

또한 I 와 같은 수준의 정보영역이 노드 D 를 중심으로 하는 영역과 노드 F 를 중심으로 하는 영역이 존재하는 것 처럼 여러 개 존재하고 있음을 볼 수 있다. 이러한 정보영역을 통하여 특정한 주제에 관하여 연구하고 있는 그룹을 가시화하여 볼 수 있다.

이러한 잇점이 있기 때문에 그래프를 통하여 WWW 문서 네트워크로 정보 구조를 파악하면 한 눈에 정보의 특성을 파악할 수 있다.

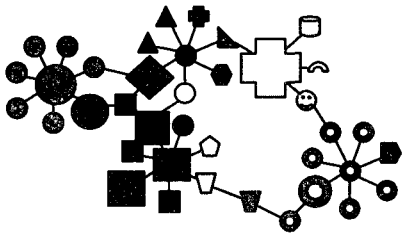

그림 4. 정보 지도

WWW 검색결과를 볼 때, 몇 개의 문서가 같은 서버에 있는 같은 종류의 문서일 경우가 있다. 매뉴얼 같은 문서가 올려져 있을 때 같은 내용이므로 관련도가 높게 정해져서 연속되어 보이게 된다. 같은 매뉴얼에 속해있는 문서들이 높은 순서를 모두 차지하고 있다면 전체적인 WWW 탐색 결과를 파악하는데 오히려 방해가 된다. 매뉴얼 같은 경우 그 매뉴얼에서 대표적인 것만 열거하는 것으로 충분하다. 그런데 이러한 문제는 그래프를 이용하면 쉽게 해결된다. 즉, WWW 문서가 위치하고 있는 WWW 서버 군별로 노드의 모양을 달리해서 그림 4와 같이 보여주면 사용자는 이것을 고려하여 WWW 을 검색할 수 있다.

## 5. 설계 및 구현

WWW 문서 수집공간을 최소하기 위하여 탐색 깊이를 최소화 해야 한다. 그림 1에서 보는 바와 같이 탐색 공간이 기하급수적으로 증가하기 전의 지점에 $d$ 을 결정해야 한다. $d$ 는 WWW 탐색 결과의 응집도에 영향을 받는다. 이 응집도는 WWW 탐색 결과가 가르키는 문서들의 유사도에 영향을 받을 것으로 예상되는데, 본 구현에서는 동적으로 결정하지 않고 직관적으로 예상 가능한 값을 이용하였다.

WWW 탐색 결과 검증기는 wget[5]을 토대로 만들어졌는데, wget 은 순차적으로 열거된 WWW 문서를 깊이 우선(depth first)의 방법으로 수집한다.

문제는 수집시간을 최대한 줄이는 것인데 이것은 다중 연결(multiple connection)과 HTTP1.1 에서 지원하는 keep alive 방법을 이용하면 수집시간을 줄일 수

있을 것으로 보인다.

WWW 문서의 하이퍼 링크 연결 그래프는 양방향 그래프이다. 본 구현에서는 이 양방향 그래프를 이용하지 않고 단방향 그래프로 하이퍼텍스트 구조 그래프를 구현하였다.

$$wf(i) = C \cdot \frac{1}{i}$$

수식 5. 이웃 지배도 함수

사용자가 WWW 검색 엔진에 질의를 하면 WWW 검색 엔진은 WWW 검색 결과를 사용자에게 돌려준다. WWW 검색결과는 최종적으로 사용자의 정보요구를 표현한 결과이다. 이것은 질의 확장의 일종으로 생각할 수 있다. 유사도는 WWW 탐색 결과와 각 WWW 문서의 내적이다. 즉, WWW 탐색 결과의 후처리 결과와 각 WWW 문서의 유사도는 두 문서에 공통적으로 들어있는 단어의 빈도를 곱한 값한다.

유사도가 계산된 후 제 3 장에서 제시한 하이퍼텍스트임을 고려한 방법으로 관련도를 계산한다. 수식 5 를 적용하여 관련도를 구현하였는데 수식 5 에서 C 를 조절함으로 지배도를 바꿀 수 있다. C 는 이웃의 지배도를 나타낸다.

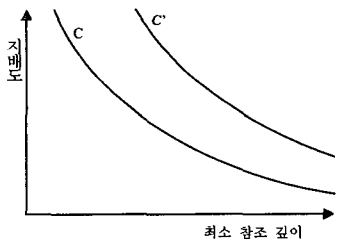


그림 5. 이웃 지배도 변화에 따른 그래프

이웃 문서들의 유사도가 WWW 문서의 관련도에 영향을 미치는데 특정 노드에서 다른 노드로의 경로는 여러 개 있을 수 있기 때문에 최단 경로를 찾아내어야 하며, 이 알고리즘은 최단 경로만이 유사도에 영향을 끼치도록 고안 되었다.

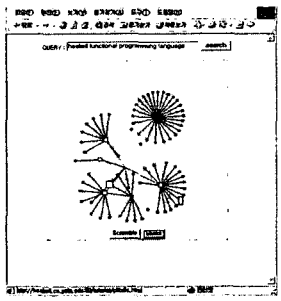

그림 6. 사용자 인터페이스

## 6. 결론

본 연구는 하이퍼텍스트의 상호 참조 정보를 이용하여 WWW 탐색 결과의 정확도를 향상시키고, 일괄 색인 방식이 아닌 실시간 검색을 통하여 WWW 문서의 현재의 상태를 정확하게 파악할 수 있다. 일괄 색인할 때에 존재하지 않았던 문서도 실시간 검색 시에 관련 정도를 판단하여 WWW 탐색 결과에 포함시키며, 일괄 색인할 때에는 존재하였으나 실시간 검색 시에는 없어져 버린 문서는 WWW 탐색 결과에서 제외한다.

사용자 입장에서 WWW 탐색 결과 검증기는 기존의 탐색 엔진과 동일한 기능을 하는 것으로 보이게 되는데, WWW 탐색 결과가 단순 열거형이 아니라 그래프로 문서들의 관련도에 따른 그래프로 보여진다. 사용자는 WWW 탐색 결과를 그래프를 통하여 구조적으로 파악할 수 있다. 탐색 엔진의 단순 나열형의 사용자 인터페이스보다 그래픽 사용자 인터페이스로 나타내어지는 것이 사용자 들이 정보를 보다 직관적으로 파악할 수 있도록 해주는 장점이 있다.

추후의 연구 과제로서는 보다 적중률을 높이는 방법으로 메타 데이터를 이용하는 방법을 모색중이다[6]. 데이터의 데이터라고 정의하는 메타 데이터는 현재 RDF(Resource Description Framework)와 같은 방법의 연구가 진행중이며, 이는 정보 자원에 대한 정보를 XML 형식으로 기술하는 것이다. 이를 이용하면 웹의 페이지들에 대한 의미를 추출하는 시맨틱 웹(Semantic Web)에 대한 연구가 가능하다. 메타 데이터에 룰(rule)을 적용하여 새로운 정보 또는 지식을 도출해 낼 수 있을 것이며, 그러한 정보는 지금의 검색 엔진에서 찾아내는 무의미하고, 부정확한 검색 결과가 아니라 사실(fact)에 기반한 지식이 될 것이다.

## 참고문헌

[1] 웹코리아, "*가자, Web 의 세계로!*", 정보시대, 1995. 10.

[2] W. Bruce Croft, "Retrieval Strategies for HyperText", *Information Processing & Management*, Vol. 29, No. 3, pp. 313-324, 1993.

[3] Jonathan Furner, David Ellis, Peter Willett, "The Representation and Comparison of HyperText Structures using Graphs", Information Retrieval and Hypertext, Kluwer Academic Publishers

[4] Jacques Savoy, "Citation Schemas in Hypertext Information Retrieval", Information Retrieval and Hypertext, Kluwer Academic Publishers

[5] Wget, Hrvoje Niksic, ftp://gnjilux.cc.fer.hr/pub/unix/util/wget

[6] 이강찬, 손홍, 박기식, "XML 표준화 동향", *정보과학회지*, 제 19 권 제 1 호, pp.6-14, 2001.