韓國經營科學會誌
第30권 第2號
2005年 6月

웹의 연결구조로부터 Hub와 Authority를 효과적으로 도출하기 위한 상호강화모델의 확장

황 인 수*

An Extended Mutual Reinforcement Model for Finding Hubs and Authorities from Link Structures on the WWW

Insoo Hwang*

Abstract

The network structures of a hyperlinked environment can be a rich source of information about the contents of the environment and it provides effective means for understanding it. Recently, there have been a number of algorithms proposed analyzing hypertext link structure so as to determine the best authorities for a given topic or query. In this paper, we review the algorithm of mutual reinforcement relationship for finding hubs and authorities from World Wide Web, and suggest SHA, a new approach for link-structure analysis, which uses the relationships among a set of relative authoritative pages, a set of hub pages, and a set of super hub pages.

Keyword : Link-structure Analysis, Hubs and Authorities, Super Hubs, Content Structure

1. 서 론

인터넷을 위한 인프라의 구축이 확대되고 인터넷의 활용이 생활화됨에 따라 월드 와이드 웹(WWW)은 기업의 의사결정과 경쟁우위를 확보하기 위한 가장 중요한 정보원이 되고 있다. 그러나 웹은 비구조화된 정보들이 하이퍼링크를 통해 자발적으로 연결된 구조를 갖고 있기 때문에, 웹이 제공하는 콘텐

논문접수일 : 2004년 4월 4일 논문게재확정일 : 2004년 11월 26일
* 전주대학교 정보기술공학부

츠(contents)뿐만 아니라 연결구조에 관한 정보를 효과적으로 조직하여 활용하는 것이 바람직하다[16].

WWW는 정보를 담고 있는 웹페이지를 하이퍼텍스트로 연결한 것으로서, 콘텐츠가 가장 중요한 정보이지만, 웹페이지간의 연결구조는 Authoritative 웹페이지를 검색하는데 있어서 또 하나의 중요한 정보를 제공한다[16]. 예를 들어, 웹페이지 p에서 q로 연결하는 p→q 링크는 웹페이지 p를 방문한 후에 링크를 따라 q로 방문하도록 유도한다. 그러므로 이들 두 페이지는 동일한 분야의 정보를 담고 있거나, 혹은 웹페이지 p가 정보를 담고 있는 웹페이지들을 연결하는 허브(Hub)의 역할을 수행하고 있음을 알 수 있다. Kleinberg[11]는 이와 같이 정보가 있는 웹페이지로 연결하는 링크를 정보링크(informative link)라고 불렀으며, 최근에는 웹의 정보를 보다 효과적으로 활용하기 위해 웹페이지간의 연결구조를 파악하기 위한 연구들이 활발히 이루어지고 있다[3, 13, 15, 18, 22].

링크구조에 대한 연구는 웹을 구성하는 하이퍼텍스트의 개념이 도입되기 이전에 논문의 인용구조를 분석하는 Bibliometrics 분야에서 연구되어 왔다[10, 19, 20]. 이 분야에서의 주된 연구는 과학저널에 발표된 논문중에서 중요도가 높은 논문을 검색하거나 관련 문서들을 클러스터링하는 것이었다. 이들 연구는 웹페이지가 담고 있는 텍스트 자료뿐만 아니라 웹의 링크구조를 통해 중요도가 높은 웹페이지를 검색하거나, 링크구조로부터 공통된 주제를 다루고 있는 웹페이지들을 클러스터링하는 연구에 활용되어 왔다.

웹의 연결구조를 이용하는 대표적인 접근법으로는 Brin and Page[3]가 제안하여 Google 검색엔진의 검색방법에 적용된 PageRank 알고리즘과 웹페이지간의 상호강화관계(mutually reinforcing relationship)를 통해 웹페이지를 Hub와 Authority로 구분한 HITS(Hyperlink Induced Topic Search) 알고리즘[6, 11]을 들 수 있다. PageRank는 웹의 연결구조에서 Random Walk를 통해 웹페이지 p를 방문할 확률에 따라 웹문서의 중요도를 계산한 것으로서, 각 문서를 참조하는 다른 문서들의 PageRank에 따라 계산된다. 그러므로 다른 많은 문서들로부터 참조될 뿐만 아니라, 중요도가 높은 문서들로부터 참조될수록 중요한 문서로 인정되어 높은 PageRank를 갖게 된다[1]. 결과적으로 PageRank는 전체 웹에서 각 웹문서가 갖는 중요도를 계산하기 때문에 검색엔진에서의 검색질의에 무관한 글로벌한 중요도를 계산하며, Hub와 Authority를 구분하지 않는 특징이 있다.

Kleinberg[11]는 웹페이지의 기능을 Hub와 Authority의 두 가지로 구분하였다. 여기서 Authority는 주제에 대한 직접적인 콘텐츠를 담고 있으며, Hub는 콘텐츠의 구조와 관련하여 좋은 정보가 있는 웹페이지로 안내하는 역할을 담당한다. 좋은 Hub는 좋은 Authority로의 연결구조를 가지며, 좋은 Authority는 좋은 Hub로부터의 연결구조를 갖는다는 가정하에 상호강화관계에 따라 각 웹페이지의 Hub와 Authority의 값을 계산한다. Kleinberg의 알고리즘을 수정/확장한 알고리즘으로는 Lempel and Moran[16]이 웹의 연결구조에 random walk의 확률적 개념을 도입하여 제안한 SALSA(The Stochastic Approach for Link Structure Analysis), Hub로부터 연결되는 Authority들의 평균값을 Hub의 값으로 취하는 Hub-Averaging Kleinberg 알고리즘 등이 있다.

위의 내용을 정리하면, Brin and Page[3]는 모든 웹페이지가 동일한 기능을 수행하는 것으로 보고 PageRank라는 하나의 척도로 평가하였으나, Kleinberg[12]는 웹페이지를 상호강화관계를 갖는 Hub와 Authority로 구분함으로써 중요한 자원을 보다 효과적으로 발견할 수 있는 방법을 제안하였다. 그러나 웹의 구조가 점점 더 복잡해짐에 따라 웹페이지를 Hub와 Authority로 구분할 경우 네트워크에 Hub와 Authority가 혼재되어 Authority를 효과적으로 검색하지 못하는 문제가 발생하게 되었다. 이에 따라, 본 연구에서는 Hub-Authority 구조가 갖는 몇 가지 문제점을 제시한 후에, 이들의 관계를 조정하는 Super Hub의 도입을 제안한다. 즉, 웹페

이지를 콘텐츠 중심의 Authority와 이를 연결하는 Hub, 그리고 상위의 내비게이션 기능을 제공하는 Super Hub로 구분함으로써, Authority와 Hub를 보다 효과적으로 검색하는 SHA(Superhub-Hub-Authority) 모델을 제시한다. 본 연구는 웹사이트의 웹페이지를 3가지로 분류하여 계층화하기 때문에 콘텐츠의 구조(contents structure) 혹은 사이트 맵(site map)을 자동적으로 구성하는 애플리케이션의 개발에도 적용될 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구의 기반이 되는 Kleinberg의 알고리즘에 대해 기술하며, 제3장에서는 본 연구에서 제시하는 상호강화모델의 확장에 대해 기술한다. 제4장에서는 SHA 알고리즘을 이용한 웹사이트의 구조화 분석에 대해 기술하며, 제5장에서는 본 연구의 결론 및 향후 연구방향에 대해 기술한다.

2. 상호강화모델에 대한 고찰

2.1 HITS의 개념

Kleinberg[11]가 제안한 HITS 알고리즘은 웹페이지들이 주어진 주제에 대해 갖는 관련정도를 평가하여 우선순위를 부여한다. 주제와 관련 있는 웹페이지들을 중심으로 base set을 구성한 후, Hub와 Authority간의 상호강화관계에 따라 각 웹페이지의 Hub 가중치 및 Authority 가중치를 계산한다. 여기서, 상호강화관계란 좋은 Hub는 좋은 Authority들을 연결하며 좋은 Authority는 좋은 Hub들로부터 연결됨을 의미하는 것으로서, 서로를 강화시키는 반복 알고리즘을 통해 가중치를 갱신함으로써 각 페이지의 Hub 및 Authority를 계산한다.

여기서 base set을 구성하는 각 웹페이지 i는 0보다 크거나 같은 Hub 가중치와 Authority 가중치를 가지며, 이 가중치의 값이 클수록 좋은 Hub와 좋은 Authority로 평가된다. 반복 수행의 과정에서 Hub 및 Authority가 서로에게 영향을 미치는 정도를 조정하기 위해 그 제곱합의 값이 1이 되도록 다음과 같이 정규화한다.

$$\sum_p h(i)^2 = 1, \quad \sum_p a(i)^2 = 1$$

웹페이지 i의 Hub 가중치를 나타내는 $h(i)$는 이 웹페이지로부터 연결되는 모든 웹페이지 j가 갖는 Authority 가중치들의 합으로 계산되며, 웹페이지 i의 Authority 가중치를 나타내는 $a(i)$의 값은 이 웹페이지로 연결되는 모든 웹페이지 j가 갖는 Hub 가중치들의 합으로 계산된다.

$$h(i) = \sum_{i \mapsto j} a(j), \quad a(i) = \sum_{j \mapsto i} h(j)$$

이 과정은 행렬계산으로 이루어지는데 $h(i)$와 $a(i)$의 값이 일정한 값에 수렴할 때까지 반복적으로 수행된다. 즉, base set을 구성하는 n개의 웹페이지로 구성되는 인접행렬 B는 0 혹은 1의 값을 갖는 $n \times n$의 행렬로 표현될 수 있는데, 웹페이지 i로부터 j로 연결되는 하이퍼링크가 존재하면 B[i, j] 값이 1이 되며, 그렇지 않을 경우에는 0이 된다. 여기서, Hub와 Authority의 가중치를 $h = \langle h_1, h_2, \cdots, h_n \rangle$과 $a = \langle a_1, a_2, \cdots, a_n \rangle$의 벡터로 표현하면, 앞의 계산식은 $h = B \cdot a$와 $a = B^T \cdot h$로 표현될 수 있다. 이 계산식을 풀어 쓰면 다음과 같다.

$$h = B \cdot a = BB^T h = (BB^T)^2 h = \cdots = (BB^T)^k h$$
$$a = B^T \cdot h = B^T B a = (B^t B)^2 a = \cdots = (B^T B)^k a$$

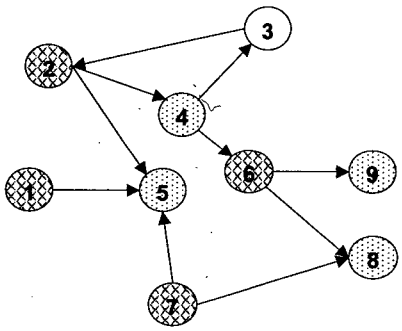
따라서 Hub와 Authority값은 각각 BB^T와 $B^T B$의 Principal Eigen 벡터로서 행렬의 초기값에 관계없이 일정한 값으로 수렴한다. 부가하여, Hub와 Authority의 가중치는 앞의 계산식을 반복적으로 수행하거나 혹은 표준 Eigen 분석패키지를 이용하여 계산될 수 있는데, Kleinberg[11]에 따르면 약 20회의 반복 수행으로도 충분히 수렴된다.

2.2 HITS의 문제점

Kleinberg[11]가 제시한 Hub-Authority 구조가 웹의 연결특성을 잘 설명하고는 있으나, 웹의 규모

가 점점 더 커지고 웹의 구조가 복잡화됨에 따라 이 구조로 설명하기 어려운 문제들이 나타나고 있다. 본 절에서는 몇 가지 예를 통해 HITS 알고리즘이 도출하는 불합리한 결과에 대해 기술한다.

[그림 1]은 Kleinberg[11]가 root set과 base set의 개념을 설명하기 위해 제시한 네트워크의 예로서, 이 네트워크에 대해 Kleinberg의 알고리즘에 따라 Hub와 Authority를 구분하면, 웹페이지 1, 2, 6, 7은 Hub, 웹페이지 4, 5, 8, 9는 Authority가 되며 웹페이지 3은 어떤 속성도 갖지 않는다. 여기서 웹페이지 2는 Authority인 웹페이지 5를 연결하는 Hub의 기능하며, 웹페이지 2를 연결하는 Hub가 존재하지 않기 때문에 Authority값은 갖지 않는다. 따라서 Hub의 역할을 하는 웹페이지 2로부터 연결되는 웹페이지 4는 Authority가 된다.

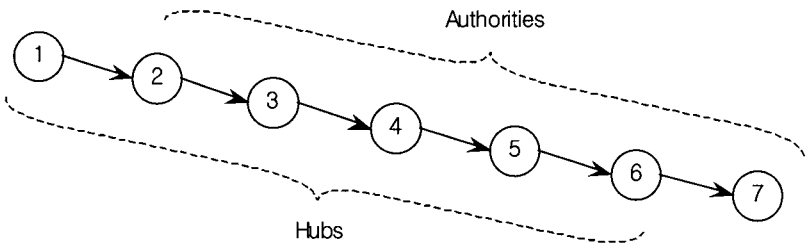


[그림 1] Hub-Authority 연결구조의 예

웹페이지 6은 Authority인 웹페이지 8과 9를 연결하기 때문에 Hub가 된다. 여기서, 웹페이지 4와 6의 관계를 살펴보면 Authority인 웹페이지 4가 Hub인 웹페이지 6을 가리키는 문제가 발생함을 알 수 있다. 이 네트워크를 직관적으로 평가할 경우 웹페이지 5, 8, 9는 네트워크의 단말노드로서 Authority가 되고, 웹페이지 1, 6, 7은 이들을 직접 연결하는 Hub가 되며, 웹페이지 2, 3, 4는 Hub를 통해 Authority로 연결하는 상위의 내비게이션을 위한 웹페이지인 것으로 판단할 수 있다.

다음으로 [그림 2]는 7개의 웹페이지가 순차적으로 연결된 구조를 갖는 네트워크의 예로서, HITS에 따르면 웹페이지 1~6은 모두 동일한 Hub 값을 가지며, 웹페이지 2~7은 모두 동일한 Authority 값을 갖는다. Kleinberg[11]은 일반적으로 Hub는 다른 Hub에 연결되지 않으며, Authority는 다른 Authority에 연결되지 않음을 가정하였으나, 이 네트워크에 HITS 알고리즘을 적용한 결과는 Hub간의 연결 혹은 Authority간의 연결로 나타나고 있다. 여기서 웹페이지 2~6은 Hub와 Authority의 특성을 모두 갖고 있는 것은 사실이지만, 개발자의 관점에서 보면 웹페이지 1에 가까울수록 Hub의 특성이 강하고 웹페이지 7에 가까울수록 Authority의 특성이 강한 것으로 판단될 수 있다.

위에서 소개한 두 가지 예는 복잡한 네트워크 구조에 Hub-Authority 모델을 그대로 적용했을 때 바람직하지 못한 결과를 도출할 수 있음을 잘 보여주고 있다. 다음 장에서는 웹페이지의 기능을 Super Hub, Hub, 그리고 Authority 등의 세 가지로 분류하고, 이들 간의 상호강화 및 약화 관계에 따라 각 웹페이지의 속성을 결정하는 SHA 알고리즘의 개발에 대해 기술한다.



[그림 2] 순차적 연결구조에서 HITS의 결과

3. 상호강화모델의 확장

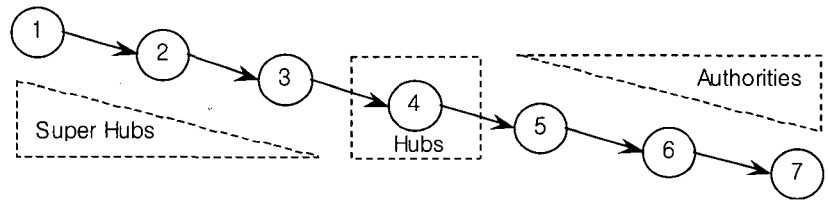
3.1 연결구조의 관계 분석

Google 검색엔진에서 사용하고 있는 PageRank는 모든 웹페이지를 동일한 가중치로 평가하였으나, Kleinberg의 HITS는 각 웹페이지를 Hub와 Authority로 구분함으로써 검색하고자 하는 콘텐츠를 담고 있는 웹페이지와 이들을 연결하는 Hub를 발견하는데 많은 기여를 했다. 그러나 2장에서 기술한 바와 같이, HITS는 웹페이지를 Hub와 Authority의 두 가지로 구분하고, 이들이 서로를 강화시키기 때문에 네트워크에서 Hub와 Authority가 혼재되는 문제가 발생한다. 이에 따라, 본 연구에서는 점점 더 복잡해져 가는 웹의 연결구조를 설명할 수 있도록 Super Hub의 개념을 도입함으로써 웹의 연결구조를 계층화하여 분석하는 방법을 제안한다.

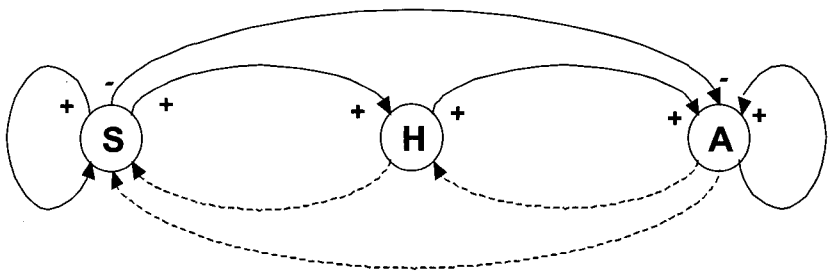
본 연구에서 제안하는 Super Hub는 기능적으로는 기존의 Hub와 동일한 기능을 수행하지만, 구조적으로는 Super Hub는 Hub와 연결되며 Hub는 Authority와 연결되는 계층적 구조로 모형화 한 것이다. 이는 시장에서 도매상-소매상-소비자로 연결되는 구조에 비유될 수 있다. Super Hub와 Hub의 구분은 도매상이 소비자에게 직접적으로 상품을 판매하지 않는 것과 마찬가지로, Super Hub는 광범위한 내비게이션을 지원하기 위한 웹페이지이므로 콘텐츠를 담고 있는 Authority로 직접 연결되지 않지만, Hub는 Super Hub와 Authority를 매개의 역할을 한다.

웹은 사용자의 관점에서 볼 때에는 웹페이지들이 상호간에 연결되어 있는 네트워크 구조를 갖지만, 개발자의 관점에서는 계층구조에 따라 개발한다. 계층의 상부는 주로 전체적인 내비게이션을 관리하는 메뉴 및 인덱스 페이지들로 구성되며, 계층의 하부로 갈수록 직접적인 정보 콘텐츠를 담고 있는 웹페이지들로 구성한다. 또한 계층의 중간에는 상부와 하부를 연결하는 Hub들이 놓이게 된다. 본 연구에서는 [그림 3]에서 보는 바와 같이, 수직적인 연결관계에서 좌측(top)의 웹페이지들은 Super Hub의 속성을 가지며, 우측(bottom)의 웹페이지들은 Authority의 속성을 갖고, 중간(middle)의 웹페이지들은 Authority로 직접 연결하는 Hub의 속성을 갖도록 알고리즘을 개발하였다.

[그림 4]는 하이퍼링크로 연결되어 있는 웹페이지



[그림 3] 확장된 상호강화모델의 기본 개념



[그림 4] 상호강화모델의 확장

간에 각 속성이 영향을 미치는 방향과 영향의 정도를 그림으로 나타낸 것으로서, 실선으로 표시된 부분은 영향을 미치지만 점선으로 표시된 부분은 영향을 미치지 않음을 나타낸다. 예를 들어, 웹페이지 p가 웹페이지 q로 연결되어 p→q의 연결관계를 갖고 있다면, 웹페이지 q의 A(uthority)는 웹페이지 p의 H(ub)에 정(+)의 영향을 받으며, 웹페이지 p의 H(ub)는 웹페이지 q의 A(uthority)에 정(+)의 영향을 받는다. 이는 Kleinberg의 HITS에서 사용하고 있는 개념과 동일한 것으로서, 좋은 Hub에 연결된 웹페이지는 높은 Authority 값을 가지며, 좋은 Authority들을 연결하는 웹페이지들은 높은 Hub 값을 갖게 됨을 알 수 있다. 또한, 웹페이지 p가 Super Hub라면 웹페이지 q는 Authority일 가능성이 낮으며, 반대로 웹페이지 p가 Authority라면 웹페이지 q는 Super Hub일 가능성이 낮기 때문에 이들 간에는 음(-)의 영향을 주는 것으로 분석된다. 이에 부가하여 본 연구에서는 컴퓨터 시뮬레이션을 통해 속성간의 다양한 영향관계를 발견하였다.

- p가 Super Hub의 속성을 갖는다면, q는 Authority보다는 Hub일 가능성이 높다.
- p가 Hub의 속성을 갖는다면, q는 Authority일 가능성이 높다.
- p가 Authority의 속성을 갖는다면, q도 Authority일 가능성이 높다.
- q가 Super Hub의 속성을 갖는다면, p또 Super Hub일 가능성이 높다.
- q가 Hub의 속성을 갖는다면, p는 Super Hub일 가능성이 높다.
- q가 Authority의 속성을 갖는다면, p는 Super Hub보다는 Hub일 가능성이 높다.

3.2 알고리즘 구성

Base Set을 구성하는 모든 웹페이지들은 0보다 크거나 같은 가중치를 가지며, 이 가중치의 값이 클수록 좋은 Super Hub, Hub, 혹은 Authority로 평가된다. 각 가중치는 서로에게 영향을 미치는 정도를 조정하기 위해 다음과 같이 정규화 한다. 참고로, Kleinberg[11]가 제안한 HITS 알고리즘에서는 제곱의 합이 1이 되도록 정규화하였는데, 일반적으로 제곱합은 음수를 양수화하거나 혹은 가중치의 영향을 증폭하기 위해 사용한다. 그러나 뒤에서 기술하는 바와 같이 $Max(0, X)$ 함수를 사용하기 때문에 가중치가 음수가 되지 않을 뿐만 아니라, 시뮬레이션을 실시한 결과 제곱합이 단순합에 대해 나은 결과를 도출하지 않는 것으로 나타났기 때문에, 본 연구에서는 단순합을 사용하였다.

$$\sum_p s(i)=1,\ \sum_p h(i)=1,\ \sum_p a(i)=1$$

앞에서 기술한 연결구조의 영향관계에 따라 각 웹페이지의 Super Hub, Hub, 그리고 Authority의 가중치를 다음과 같이, 현재의 각 가중치로부터 순환적으로 계산한다. 이 때, 가중치는 비음수(non negative)의 속성을 가지며, PageRank에서 사용한 damping factor(d)를 도입하여 다음과 같이 계산한다.

$$s(i)' = d+(1-d)\times Max(0, \sum_{i\mapsto j}(s(j)+h(j) - w\cdot a(j))/ID(j))$$

$$h(i)' = d+(1-d)\times Max(0, \sum_{i\mapsto j} a(j)/ID(j) + \sum_{k\mapsto i} s(k)/OD(k))$$

$$a(i)' = d+(1-d)\times Max(0, \sum_{k\mapsto i}(a(k)+h(k) - w\cdot s(k))/OD(k))$$

여기서, $ID(j)$는 웹페이지 j로 연결되는 링크를 갖고 있는 웹페이지의 개수(input degree)를 의미하며, $OD(k)$는 웹페이지 k에서 다른 웹페이지로 연결되는 웹페이지의 개수(output degree)를 의미한다. HITS는 연결되어 있는 모든 웹페이지의 가중치를 합산하는 방법을 사용하였으나, SHA는 가중평균에 따라 가중치를 계산한다. 또한, w는 Super Hub와 Authority간의 영향정도를 나타내는 가중치로서 Super Hub↔Hub 혹은 Hub↔Authority 간의 상호강화관계와 달리, 상호간에 음(-)의 영향을 줌으로

써 웹페이지의 속성을 보다 명확히 구별하는 역할을 한다. 본 알고리즘에서는 다음과 같이 로그(log) 함수를 사용함으로써 반복횟수(t)가 증가할수록 그 영향정도가 감소하도록 하였다.

$$w = \frac{1}{1+\log(t)}$$

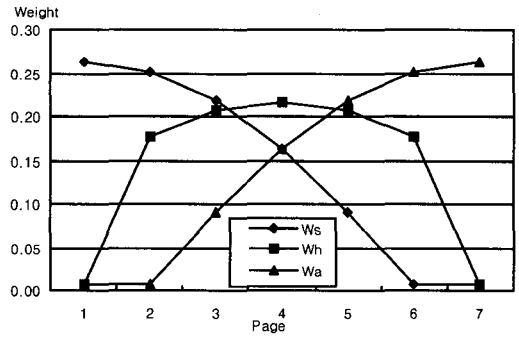
본 연구에서 개발한 SHA 알고리즘을 의사코드로 표현하면 [그림 5]와 같다. 여기서, ε는 각각의 가중치가 갱신된 정도를 나타내는 값으로써 임의로 설정한 δ보다 작아지면 계산을 종료한다. 또한, d는 PageRank에서 제시한 damping factor로서 본 연구에서는 임의로 0.01로 설정하였다.

```
Procedure SHA(Vertex V, Edge E) {
Let ID(j) is the input degree of j∈V
Let OD(k) is the output degree of k∈V
Set δ=0.001, d=0.01, t=1
∀i∈V, s(i)=h(i)=a(i)=1.0
Do
  w=1/(1+log(t++))
  ∀i∈V, s(i)' = d+(1-d)×Max(0, Σ(i,j)∈E (s(j)
                + h(j) - w · a(j))/ID(j))
        h(i)' = d+(1-d)×Max(0, Σ(i,j)∈E a(j)
                /ID(j) + Σ(k,i)∈E s(k)/OD(k))
        a(i)' = d+(1-d)×Max(0, Σ(k,i)∈E (a(k)
                + h(k) - w · s(k))/OD(k))
      Normalize s(i)', h(i)', and a(i)',
      obtaining s(i), h(i), and a(i), respectively.
  ε = Σi∈V |a(i)' - a(i)| + |s(i)' - s(i)|
      + |h(i)' - h(i)| / |V|
While (ε > δ)
Return (S, H, A)
}
```

[그림 5] 확장된 상호작용모델 알고리즘

3.3 알고리즘의 적용

본 연구에서 개발한 SHA 알고리즘의 성과를 평가하기 위해 2장에서 소개한 순차적인 연결구조를 갖는 네트워크에서 Super Hub, Hub, 그리고 Authority의 가중치를 계산하면 [그림 6]과 같다. 여기서, 웹페이지 1과 7은 각각 완전한 Super Hub 및 Authority의 속성을 갖고 있으나, 중간에 위치한 웹페이지들은 우측으로 갈수록 Authority의 특성이 높고, 좌측으로 갈수록 Super Hub의 특성을 많이 갖고 있음을 알 수 있다. 또한, 중간에 위치한 웹페이지들은 Hub의 속성을 모두 갖고 있으나 균형을 이루고 있으며, 가운데 위치한 웹페이지 4가 Hub의 역할을 함을 알 수 있다.



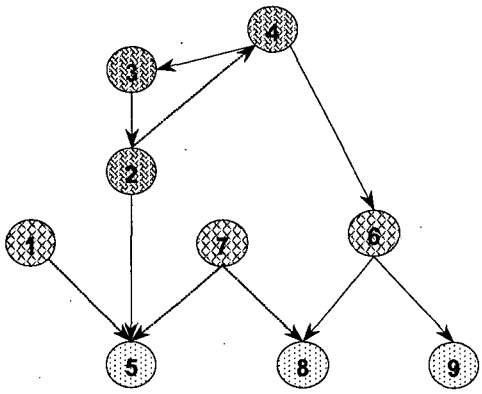
[그림 6] SHA을 이용한 순차구조 웹의 가중치 계산 결과

다음으로 Kleinberg가 Hub-Authority의 관계를 설명하기 위해 사용했던 네트워크에 SHA 알고리즘을 적용하면 <표 1>과 같다. 여기서 페이지 4, 2, 3은 Super Hub, 페이지 6, 7, 1은 Hub, 그리고 페이지 5, 8, 9는 Authority인 것으로 나타났으며, 페이지 2는 Super Hub인 3, 4의 영향을 받아서 Super Hub로 결정되었으나 Hub 혹은 Authority의 속성도 상당부분 갖고 있음을 알 수 있다.

〈표 1〉 SHA를 이용한 웹 연결분석의 결과

Web Page	Ws	Wh	Wa	Class
4	0.3494	0.1803	0.1464	S
2	0.2985	0.2527	0.1786	S
3	0.2958	0.1764	0.1060	S
6	0.0055	0.1752	0.1060	H
7	0.0162	0.0709	0.0072	H
1	0.0175	0.0379	0.0072	H
5	0.0055	0.0898	0.2044	A
8	0.0055	0.0102	0.1355	A
9	0.0055	0.0062	0.1082	A

이러한 결과로부터 Super Hub↔Hub, Hub↔Authority로 연결되는 링크에 우선순위를 부여한 후, 동일한 우선순위를 갖는 경우 가중치가 높은 링크를 선택할 경우 [그림 7]에서 보는 바와 같은 계층구조를 얻을 수 있다. 이는 관심을 갖고 있는 사이트 혹은 영역에 대한 콘텐츠 구조로서, 본 연구에서 제시한 SHA 알고리즘은 콘텐츠 구조 혹은 사이트 맵의 작성에도 응용될 수 있을 것으로 판단된다.



[그림 7] SHA를 이용하여 도출한 콘텐츠 구조도

4. SHA를 이용한 구조화 분석

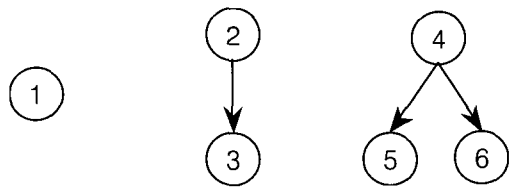
본 연구에서 제시한 SHA 알고리즘은 웹페이지를 Super Hub, Hub, 그리고 Authority로 구분하기 때문에, 이를 이용하면 웹사이트가 계층적으로 구조화되어 있는 정도를 평가하는 것이 가능하다. 즉, Super Hub는 Hub로 연결되며, Hub는 Authority로 연결되어 있는 정도를 분석함으로서 가능하게 되는데, 본 연구에서는 Hub가 Super Hub와 Authority를 얼마나 잘 연결하는가에 따라 구조화 정도를 평가하는 척도를 제시한다.

$$P(H)=\frac{P(H)_{in}+P(H)_{out}}{2}$$

여기서, $P(H)_{in}$은 Hub로 오는 링크중에서 Super Hub로부터 오는 링크의 비율을 나타내며, $P(H)_{out}$은 Hub에서 나가는 링크중에서 Authority로 나가는 링크의 비율을 나타낸다. 따라서, Hub로 들어오는 링크가 Super Hub로부터 오며, Hub에서 나가는 링크가 Authority로 가는 비중이 높을수록 네트워크 구조화 정도는 높아진다.

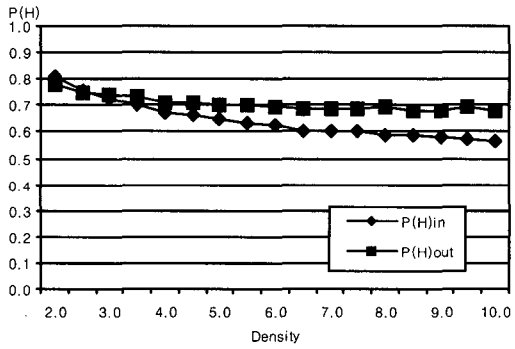
본 연구에서는 구조화 분석에 대한 예를 제시하기 위해 컴퓨터 시뮬레이션을 실시했다. 먼저 컴퓨터 시뮬레이션 환경 및 데이터 세트의 구성에 대해 기술하면 다음과 같다. 본 연구에서 제안한 알고리즘은 2.66GHz CPU를 갖는 윈도우즈 2000서버 운영체제에서 객체지향언어인 자바(Java)로 구현되었으며, 데이터관리는 오라클 데이터베이스를 이용하였다. 본 연구에서는 실제 웹에 존재하는 페이지를 이용하지 않고, 웹을 모형화하여 각 페이지와 페이지간의 연결관계를 컴퓨터 시뮬레이션을 통해 자동으로 생성하여 데이터베이스에 저장한 후, 이를 이용하여 알고리즘의 성과를 측정하였다.

웹을 모형화한 데이터의 생성 방법은 다음과 같다. 컴퓨터의 처리속도를 고려하여 웹사이트를 구성하는 웹페이지를 1,000개로 한정하였으며, 웹페이지간을 연결하는 링크의 개수는 웹페이지의 2배인 2,000개로부터 10,000개까지 500개씩 순차적으로 증가시키면서 각 조건에서 30개씩의 데이터 세트를 생성하여 총 510개의 데이터 세트에 대해 시뮬레이션을 실시하였다. 네트워크는 다음과 같이 구성하였다. 첫째로, 모든 웹페이지가 [그림 8]에서 제시하는 네트워크를 구성하는 기본요소의 한 가지 형태를 갖도록 구성한 후, 둘째로 Input Degree를 갖지 않는 모든 웹페이지를 임의의 웹페이지에 연결하였다. 끝으로, 나머지 링크는 임의로 두 개씩의 웹페이지를 선택하여 연결함으로써 가상적인 웹사이트의 구성을 완료하였다.



[그림 8] 웹페이지를 연결하는 기본요소의 예

[그림 9]는 위에서 생성한 가상의 웹사이트에 SHA 알고리즘을 적용했을 때, 구조화 척도의 변화를 그래프로 나타낸 것이다. 참고로, 밀도(density)는 웹페이지가 갖는 평균 링크의 수로서 전체 링크의 개수를 웹페이지의 개수로 나누어서 계산하며, damping factor를 나타내는 d는 0.01, 반복종료 조건을 나타내는 δ는 0.001로 설정하였다. 시뮬레이션 결과, 이때의 반복횟수는 평균 10.8회로서, Super Hub↔Hub간의 관계보다 Hub↔Authority간의 관계가 약간 더 잘 구성되었으며, 밀도가 높아질수록 구조화 정도는 약간씩 감소하는 것으로 나타났다.



[그림 9] 가상 웹사이트에서 구조화 정도의 변화

5. 결 론

웹은 비구조화된 정보들이 하이퍼링크를 통해 자발적으로 연결된 구조를 갖고 있기 때문에, 정보를 검색하거나 혹은 자사의 정보를 효과적으로 조직하여 정보를 제공하는 것이 용이한 일이 아니다. 이에 따라, 각각의 웹페이지가 갖고 있는 콘텐츠뿐만 아니라, 웹페이지간의 연결구조를 파악함으로써 정보를 보다 효율적/효과적으로 활용하기 위한 연구가 활발히 이루어져 왔다.

웹의 연결구조를 이용하는 대표적인 두 가지 접근법은 Google 검색엔진의 검색방법에 적용된 Page-Rank 알고리즘과 웹페이지간의 상호강화관계를 통해 웹페이지를 Hub와 Authority로 구분한 HITS 알고리즘으로서, 본 연구에서는 HITS 알고리즘의 문제점을 분석한 후 이를 확장한 SHA 알고리즘을 제시하였다. 본 연구에서 제시한 알고리즘은 기존의 Hub 및 Authority에 Super Hub의 개념을 추가함으로써, 웹페이지간의 관계를 보다 잘 설명할 수 있음을 몇 가지 예를 통해 제시하였다. 또한, 연결구조분석을 위한 구조화 척도의 개발 및 컴퓨터 시뮬레이션에 대해서도 기술하였다.

본 연구의 결과는 웹페이지간의 관계를 보다 명확히 파악함으로써 정보검색에 활용될 경우, 웹으로부터 사용자가 보다 많은 관심을 갖고 있는 정보를 우선적으로 제시할 수 있을 것이다. SHA의 알고리즘을 특정 웹사이트에 적용할 경우 자동적으로 각 페이지들을 Super Hub, Hub, 그리고 Authority로 계층화시킴으로써, 웹 콘텐츠의 구조 혹은 사이트 맵을 자동으로 구성하는 애플리케이션에 적용될 수 있을 것으로 생각한다. 끝으로, 본 연구의 한계점으로는 현실에서 운영되고 있는 웹을 이용하는 대신 임의로 데이터 세트를 생성하여 시뮬레이션을 실시했다는 것과 알고리즘의 성과를 객관적으로 측정할 수 없었다는 것이다. 이에 따라 향후의 연구에서는 성과측정 방법에 대해 보다 심도 깊은 연구를 수행할 뿐만 아니라, 현재 운영되고 있는 주요 정보검색엔진을 메타검색엔진으로 활용함으로써 SHA 알고리즘을 실제 웹에 적용하는 방안을 연구하고자 한다.

참 고 문 헌

[1] 김성진, 이상호, 방지환, "페이지랭크 알고리즘 적용을 위한 구현기술," 「정보처리학회논문지 D」, 제9권, 제5호(2002), pp.745-754.

[2] Andrew, Y. Ng, Alice X. Zheng and Michael I. Jordan, "Stable Algorithm for Link Analysis," *Proceedings on ACM SIGIR '01*, September 2001, New Orleans, Louisiana, USA, pp.258-266.

[3] Brin, S. and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search

Engine," *Proceedings on WWW7*, Brisbane, Australia, April, 1998.

[4] Carriere, J. and R. Kazman, "Webquery : Searching and Visualizing the Web Through Connectivity," *In Proceedings of the 6th International Conference on WWW*, 1997.

[5] Devanshu Dhyani, Wee Keong Ng and Sourav S. Bhowmick, "A Survey of Web Metrics," *ACM Computing Surveys*, Vol.34, No.4(2002), pp.469-503.

[6] Gibson, D., Jon Kleinberg and Prabhakar Raghavan, "Inferring Web Communities from Link Topology," *Proceedings on ACM HyperText*, 1998, Pittsburgh PA, USA, pp.225-234.

[7] Gui-Rong Xue, Hua-Jun-Zeng, Zheng Chen, Wei-Ying Ma, Hong-Jiang Zhang and Chao-Jun Lu, "Implicit Link Analysis for Small Web Search," *Proceedings on ACM SIGIR'03*, Jul.~Aug. 2003, Toronto, Canada, pp.56-63.

[8] Hung-Yu Kao, Shian-Hua Lin, Jan-Ming Ho and Ming-Syan Chen, "Entropy-Based Link Analysis for Mining Web Informative Structures," *Proceedings on ACM CIKM' 02*, Nov. 2002, McLean, Virginia, USA. pp. 574-581.

[9] Ke Wang and Ming-Yen Thomas Su, "Item Selection By Hub-Authority Profit Ranking," *Proceedings on ACM SIGKDD '02*, Edmonton, Alberta, Canada, pp.652-657.

[10] Kessler, M., "Bibliographic Coupling Between Scientific Papers," *Am. Doc.*, Vol.14 (1963), pp.10-25.

[11] Kleinberg, J., "Authoritative Sources in a Hyperlinked Environment," *In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithm*, 1998, pp.668-677.

[12] Kleinberg, J., "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, Vol.46, No.5(1999), pp.604-632.

[13] Kleinberg, J., "Hubs, Authorities, and Communities," *ACM Computing Surveys*, Vol.31 (1999), pp.1-3.

[14] Kleinberg, J. and Andrew Tomkins, "Applications of Linear Algebra in Information Retrieval and Hypertext," 1997, pp.185-193.

[15] Lempel, R. and Aya Soffer, "PicASHOW: Pictorial Authority Search by Hyperlinks On the Web," *ACM Transactions on Information Systems*, Vol.20, No.1(2002), pp.1-24.

[16] Lempel, R. and S. Moran, "SALSA : The Stochastic Approach for Link-Structure Analysis," *ACM Transactions on Information Systems*, Vol.19, No.2(2001), pp.131-160.

[17] Mao Lin Huang, Peter Eades and Wei Lai, "On-line Visualization and Navigation of the Global Web Structure," *International Journal of Software Engineering*, Vol.13, No.1(2003), pp.27-52.

[18] Norodin, A., Gareth O. Roberts, Jeffery S. Rosenthal, and Panayotis Tsaparas, "Finding Authorities and Hubs From Link Structures on the World Wide Web," *Proceedings on ACM WWW10*, May 2001, Hong Kong, pp.415-429.

[19] Osareh, O., "Bibliometrics, Citation Analysis and Co-Citation Analysis : A Review of Literature I," *Libri*, Vol.46(1996), pp.149-158

[20] Small, H., "Co-citation In the Scientific Literature : A New Measure of The Relationship Between Two Documents," *J. Am. Soc. Inf. Sci.*, Vol.24(1973), pp.265-269.

[21] Wen-Syan Li, Necip Fazil Ayan, Okan Kolak and Quoc Vu, "Constructing Multi-

Granular and Topic-Focused Web Sited maps," *Proceedings on ACM WWW10*, May 1-5, 2001, Hong Kong, pp.343-354.

[22] Zheng Chen, Shengping Liu, Liu Wenyin, Geguang Pu, and Wei-Ying Ma, "Building a Web Thesaurus from Web Link Structure," *Proceedings on ACM SIGIR'03*, July ~Aug. 2003, Toronto, Canada, pp.48-55.