시맨틱 링크를 사용한 페이지순위 알고리즘 개선

(Improvement of PageRank Algorithm using Semantic Link)

전 희 국 † 임 동 혁 †† 김 형 주 †††
(Hee-Gook Jun) (Dong-Hyuk Im) (Hyoung-Joo Kim)

요 약 페이지랭크는 웹정보 검색의 중요도를 평가하는 대표적 방법이다. 그러나 페이지랭크가 가진 중요도 판단의 특성상 의미 없는 문서지만 인링크 개수가 많아 중요한 문서로 인식될 수 있는 가능성이 존재한다. 기존 방법들이 링크를 차등 평가하는 대안을 제시했으나 하이퍼링크 기반 웹 구조의 성격상 링크의 중요성을 직접 평가할 수 없는 문제가 있다. 이러한 문제를 해결하기 위해 본 논문에서는 시맨틱 링크를 사용하여 웹 구조를 의미를 가진 링크 기반의 웹 구조로 변경해 링크의 중요성을 직접 판단하도록 중요도 계산 방법을 개선했다. 실험결과 제안한 방법이 상위 순위에 보다 많은 관련 문서를 제시해 더 높은 적합도를 가지는 것을 보였다.

키워드 : 온톨로지, 페이지랭크, 시맨틱웹, RDFa

Abstract PageRank is a representative method to evaluate the importance of web pages for web information retrieval. However, it is possible for meaningless pages that have many inlinks to be recognized as important pages because of characteristics of PageRank. Existing papers that provide methods to stratify weight of links still have a problem these methods cannot evaluate the weight of links directly in hyperlink based web structure. We propose a new approach that changes the hyperlink based web structure to the semantic link based web structure to evaluate the weight of links directly by using semantic links. Experiment shows that our approach performs better than the existing PageRank in terms of evaluating the importance of web pages.

Key words : Ontology, PageRank, Semantic Web, RDFa

1. 서 론

페이지랭크[1]는 웹정보 검색을 위해 문서의 중요도를 평가하는 대표적인 방법이다. 페이지랭크는 문서 간 링크에 기반해 중요도를 계산하며, 이전 키워드 매치에 기반한 방법보다 더 나은 검색 결과를 보인다. 그러나 페이지랭크는 인링크를 많이 가지고 있으면 실제의 내용에 상관 없이 중요한 문서로 인식될 가능성이 높아지는 구조적 취약점이 있다.

이러한 문제를 개선하기 위해 문서간 링크를 차등 평가해 중요도를 계산하는 방법들이 제안되어 왔다. 그러나 월드와이드웹의 하이퍼링크는 단순한 연결 기능만 가지고 있으므로 차등 평가를 위한 링크의 중요도는 직접 판단할 수 없는 문제가 있다. 이러한 이유 때문에 지금까지의 개선 방법들은 링크 구조를 분석하거나[2], 저자 신뢰도 네트워크를 구축[3]하는 등 추가 정보를 이용해 링크의 중요도를 추정하는 방법을 사용했다.

본 논문에서는 직접 링크의 의미를 판단할 수 있도록 시맨틱 링크를 활용한 새로운 페이지랭크 방법을 제시한다. 기존 연구들이 하이퍼링크 기반 웹 구조는 유지한 채 페이지랭크를 개선해온 반면, 본 논문에서는 웹에서

· 본 연구는 BK-21 정보기술 사업단의 연구결과로 수행되었음
· 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 20120009186)

† 비 회 원 : 서울대학교 컴퓨터공학부
hgjun@idb.snu.ac.kr
(Corresponding author임)
†† 비 회 원 : 서울대학교 컴퓨터연구소 객원연구원
dhim@idb.snu.ac.kr
††† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2012년 7월 24일
심사완료 : 2012년 9월 8일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제18권 제12호(2012.12)

RDFa[5]를 활용해 시맨틱 정보를 추출해 내어 의미를 가진 링크(Semantically labeled link) 기반의 웹 구조 영역으로 문제를 가져와 해결한다. 그 결과 링크의 의미를 직접 평가해 문서의 중요도를 계산하는 것이 가능해진다. 또한 중요도를 단순히 링크 개수로 평가하는 것에서 더 나아가 문서가 얼마나 중요한 정보를 가지고 있는가로 평가하는 기준을 제시할 수 있게 되었다.

논문의 구성은 다음과 같다. 2장에서는 페이지랭크를 개선한 기존 방법들을 살펴보고 3장에서는 제안하는 시맨틱 링크 기반 페이지랭크 알고리즘에 대해 설명한다. 4장에서는 실험 방법 및 결과를 제시하고 5장에서 결론 및 향후 연구에 대해 언급한다.

2. 관련연구

2.1 PageRank

페이지랭크[1]는 웹문서의 중요도를 계산하는 알고리즘이다. 한 문서의 페이지랭크 값은 자신을 가리키는 링크를 가진 문서들의 페이지랭크 값의 합이다. 반대로 자신은 다른 문서로 향하는 아웃링크의 개수만큼 균등하게 페이지랭크 값을 나누어 분배한다. 그림 1을 예로 들면, 문서 A는 자신의 페이지랭크값 20을 B, C에 각각 10씩 전달한다. 문서 D는 B, C에 각각 30씩 전달한다. 따라서 문서 B의 페이지랭크값은 40이 된다. 식 (1)은 링크 고립문제를 다루기 위한 상수 d가 추가된 페이지랭크 수식이다.

그러나 링크 기반의 중요도 계산 방법으로 인해 의미 없는 문서지만 인링크가 많아 높은 페이지랭크 값을 가질 수 있는 문제가 발생할 수 있다.

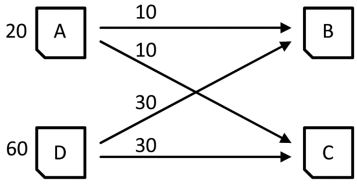

그림 1 PageRank의 예

$$PR(r_i) = d\sum_{j \to i} \frac{1}{N_j} \cdot PR(r_j) + (1-d) \qquad (1)$$

2.2 Weighted PageRank

Weighted PageRank[2]는 식 (2)처럼 문서의 인링크 비율과 아웃링크 비율을 사용해 링크에 가중치를 계산한다. 예를 들면, 그림 2에서 B의 A에 대한 인링크 가중치는 A와 연결된 B와 C의 인링크 총합 3으로 B의 인링크 개수 2를 나눈 값이다.

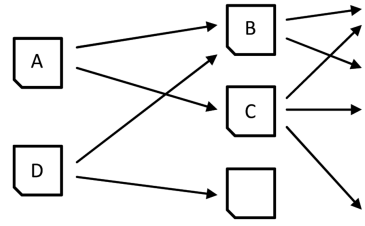

그림 2 Weighted PageRank의 예

$$W_{in}(A, B) = I_B / (I_B + I_C) = 2/3$$

아웃링크 비율도 마찬가지 방법으로 계산한다.

$$W_{out}(A, B) = O_B / (O_B + O_C) = 2/5$$

그러나 이 알고리즘도 역시 링크의 의미보다는 개수에 기반을 두고 있다는 한계가 있다.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p}, \quad W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p} \qquad (2)$$

2.3 PageRank with Author Trust Network

[3]은 각 문서를 작성한 저자들의 신뢰도 네트워크를 별도로 구축한다. 이를 통해 저자들이 작성한 문서를 평가할 수 있음과 동시에 저자들 간의 신뢰도를 사용하여 페이지랭크 값을 개선할 수 있다.

그러나 저자 신뢰도 네트워크의 관리비용이 추가로 발생하며, 문서의 가치에 대한 평가는 시간이 지남에 따라 진부화(Obsolete)될 수 있다는 문제점이 있다.

2.4 Weighted Page Content Rank

[4]는 Weighted PageRank[3]에 웹 컨텐츠 마이닝 기법을 도입해 페이지랭크 알고리즘을 개선했다. 링크 가중치를 계산할 뿐만 아니라, 주어진 검색어에 문서가 얼마나 관련 있는지를 컨텐츠 마이닝 기법으로 측정한다. 하지만 두 문서 사이의 링크에 대한 의미를 측정하는 방법은 여전히 링크 개수에 기반을 두고 있으며, 문서의 관련성을 위해 컨텐츠 마이닝 과정을 거쳐야 한다는 추가 비용이 발생한다.

3. 시맨틱 링크 기반 페이지랭크

본 절에서 시맨틱 링크에 기반한 개선된 페이지랭크 알고리즘을 설명한다. 이 방법은 단순히 인링크의 개수가 아닌 실제 문서가 얼마나 중요한 정보를 많이 가지고 있는지로 문서의 중요도를 평가하는 새로운 기준을 제안한다.

시스템은 크게 4가지로 분류된다(그림 3). 첫 번째 단계는 웹문서를 돌아다니며 시맨틱 정보를 추출한다. 두 번째 단계에서 추출한 정보들을 병합해 RDF 그래프를 구축한다. 세 번째 단계는 구축한 RDF 그래프에서 자원(Resource)들의 랭크값을 계산한다. 마지막으로 네 번째 단계에서 자원들의 랭크값을 활용하여 웹문서의 페이지랭크 값을 구한다.

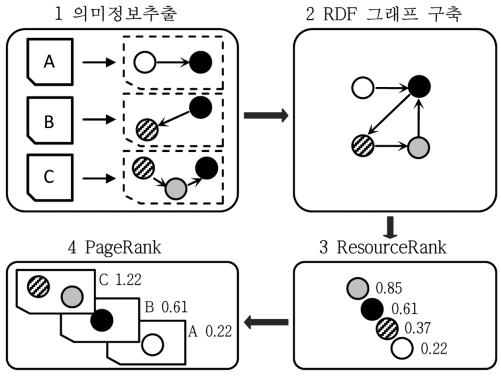

그림 3 시맨틱 링크 기반 페이지랭크 시스템

3.1 의미 정보 추출

본 논문에서는 웹문서가 RDFa[5]를 이용해 Semantic Annotation이 추가된 상태라고 가정[1]을 한다. RDFa는 일반 XHTML 페이지에 RDF정보를 정의할 수 있도록 하는 기술이다. RDF[6]는 표현의 대상이 되는 자원(Resource)간의 시맨틱 관계 정보를 그림 4처럼 "주어부(Subject)-서술부(Predicate)-목적부(Object)" 형태의 트리플(Triple)로 표현하므로, 이를 이용해 의미 있는 단위인 자원을 가지고 문서의 중요도를 측정할 수 있게 되고, 트리플 구조에 의해 의미 있는 서술부를 가지고 링크의 가중치를 측정할 수 있다.

그림 4 RDF 트리플의 예

우선 첫 단계로 웹문서를 크롤링하면서 RDF 파싱을 진행한다. 그림 5는 RDFa가 선언된 HTML문서에서 RDF 파싱을 하는 예를 보여준다.

3.2 RDF 그래프 구축

첫 번째 단계에서 모든 웹문서의 RDF 파싱이 끝나면, 두 번째 단계로 만들어진 RDF 트리플들을 모두 병합한다.

RDF 자원은 URI라는 유일한 값을 가지고 있으므로, 서로 다른 트리플에 URI가 동일한 자원이 각각 있다면 그 두 자원은 동일한 것으로 보고 RDF 트리플들을 합쳐나갈 수 있다(그림 6).

1) RDFa는 2004년 W3C에 의해 만들어져 2008년 표준으로 채택되었다. RDFa가 적용된 XHTML문서는 브라우저에 보일 수 있을 뿐 아니라 RDF 파싱이 가능하다. 더 나아가 [7,8]과 같이 웹문서에 자동으로 RDFa를 작성해주는 방법들이 제안되고 있다. RDFa는 시맨틱 웹으로의 발전을 가속화할 수 있는 기술로서 추후 활용도는 더욱 높아질 것으로 보이며, 이에 근거하여 본 논문의 연구도 RDFa가 적용된 웹문서를 이용해 연구 및 실험을 하였다.

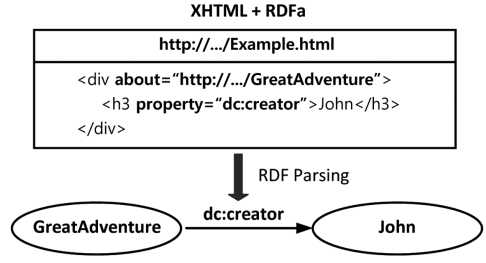

그림 5 RDFa가 선언된 HTML 문서의 RDF 파싱

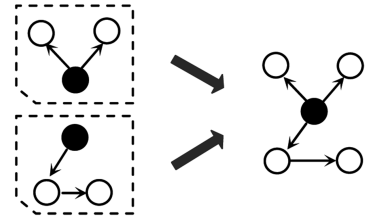

그림 6 RDF 트리플 병합 과정

3.3 ResourceRank

두 번째 단계에서 만들어진 RDF 그래프를 사용해 자원의 랭크(ResourceRank)를 계산한다. 계산 방법은 페이지랭크 알고리즘의 기본 원리와 비슷하나 자원 간의 링크는 의미가 있는 Predicate으로 연결되어 있으므로 링크의 가중치를 직접 계산할 수 있다. 링크의 가중치 계산 방법은 직접적(Manually) 혹은 반자동적(Semi-automatically)인 방식이 있다. 반자동적인 방식은 TF- IDF를 이용한 방법[9] 등이 있다. 본 논문에서는 직접적으로 링크의 가중치를 계산했다. 이렇게 계산된 가중치를 이용해 자원의 랭크를 계산하는 방법은 식 (3)과 같다.

$$RR(r_i) = d \sum_{j \in outlink(i)} \frac{RR(r_j) \cdot weight(r_j, p)}{\sum_{j \in outlink(i)} weight(r_j, p)} + (1-d) \quad (3)$$

3.4 PageRank

세 번째 단계에서 구한 자원들의 랭크값을 원래 자원이 추출된 웹문서로 보내 웹문서의 페이지랭크값을 최종 계산한다. 계산 방법은 식 (4)와 같으며 웹문서의 PageRank는 웹문서가 가지고 있는 모든 자원의 ResourceRank값의 합이다.

$$PageRank(p_i) = \sum_{r \in p_i} RR(r) \quad (4)$$

4. 실험 결과

4.1 실험 데이터

실험을 위해 400개의 웹문서가 포함된 가상의 웹 환경을 구축하였다. 각 웹문서 안에는 440개의 자원(Re-

source)에 대한 RDF 트리플이 선언되어 있다.

웹문서의 내용은 위키피디아 데이터의 일부를 발췌하여 작성하였다. 페이지랭크와 시맨틱 링크 기반 페이지랭크의 수행결과를 보다 직관적으로 비교할 수 있도록 각 웹문서는 “영국문학”과 관련한 주제로만 작성하는 것으로 실험에 제약을 두었다.

4.2 실험 결과

본 절에서는 구축한 웹 환경에 대해 페이지랭크와 시맨틱 링크 기반 페이지랭크를 수행한 결과를 비교한다. 검색어는 “이상한 나라의 앨리스”(이하 “앨리스”)를 사용하였다.

그림 7은 결과 문서 개수 별 관련문서의 개수를 비교한 그래프이다. x축은 결과로 제시하는 문서의 개수, y축은 그 중 실제로 검색어와 관련된 문서의 개수이다. 결과 문서 개수가 많을수록 페이지랭크와 시맨틱 링크 기반 페이지랭크가 제시하는 관련문서 개수는 비슷해지나 결과 문서 개수가 적을 때는 시맨틱 링크 기반 페이지랭크가 제시하는 관련문서 개수가 상대적으로 많다. 즉 시맨틱 링크 기반 페이지랭크 방법을 사용하면 사용자가 원하는 결과를 상위 문서로 보다 많이 제시할 수 있음을 알 수 있다.

그림 8은 페이지랭크와 시맨틱 링크 기반 페이지랭크의 적합도[2]를 비교한 결과이다. x축은 결과로 제시하는 문서의 개수, y축은 적합도의 로그값이다. 적합도는 관련 문서들이 결과 문서의 상위 순위로 제시될수록 값이 커진다. 한가지 주목할 점은 결과 문서 개수가 200개인 경우 두 방법 모두 동일한 개수인 107개의 관련문서를 제시했지만(그림 7), 적합도는 시맨틱 링크 기반 페이지랭크가 더 높은 값(그림 8)을 가지고 있다. 두 방법이 제시한 관련문서의 개수는 같으나 시맨틱 링크 기반 페이지랭크가 관련문서들을 보다 상위 순위에 나오도록 제시해 주었기 때문이다.

표 1은 문서에 대한 중요도 판단 방법의 차이를 볼 수 있는 실험결과이다. 페이지랭크에서 “앨리스”의 출판사인 Macmillan과 미국의 랭킹결과는 33위, 5위로 미국의 중요도가 더 높다. 그러나 시맨틱 링크 기반 페이지랭크에서는 15위, 16위로 Macmilllan의 중요도가 올라간다.

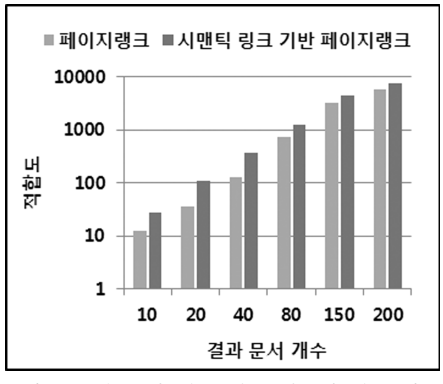

그림 7 결과 문서 개수 별 관련문서 개수 비교

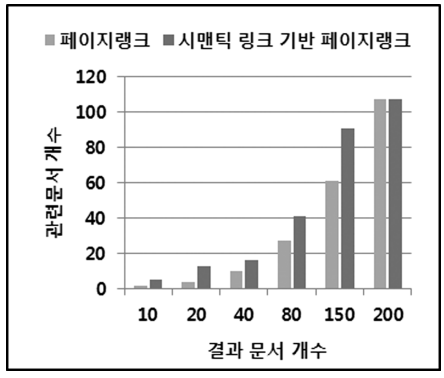

그림 8 결과 문서 개수 별 적합도 비교

위 결과를 자세히 살펴보기 위해 표 2로 출판사와 미국 문서가 포함한 자원에 대해 비교하였다. 시맨틱 링크 기반 페이지랭크의 세 번째 단계인 자원의 랭킹(ResourceRank)에 의한 순위를 살펴보면 다음과 같다. Macmillan 웹문서에는 두 자원 “macillan”과 “publish company”가 있으며 각각의 자원 랭킹 값은 0.659, 0.468이다. 반면 웹문서 미국 안의 자원 “unitedState”의 자원 랭킹 값은 1.107이다. 개별 자원의 자원 랭킹 값으론 “unitedState”가 “macmillan”과 “publish company”보다 높지만, 웹문서는 자신 안의 모든 자원의 자원 랭킹 값의 총합으로 계산되므로(식 (4)) 최종적으로 계산된 미국의 시맨틱 링크 기반 페이지랭크 값 1.107보다 Macmillan의 값 1.127이 더 커져 순위가 바뀌게 된 것을 관찰 할 수 있다.

이 결과는 시맨틱 링크 기반 페이지랭크가 전통적 방법의 페이지랭크보다 검색어에 대해 더욱 관련 있는 문서를 중요한 문서로 판단함을 보인다. 앞서 언급한 Macmillan은 “앨리스”의 출판사이므로 많은 다양한 일반적 개념들과 연결되는 미국보다 검색어 “앨리스”에 밀접한 관련이 있는 개념이다. 그러므로 기존 페이지랭크와 반대로 Macmillan을 미국보다 상위 랭크로 계산한 시맨틱 링크 기반 페이지랭크가 더 의미 있는 중요도 판단 결과를 제시했음을 알 수 있다.

표 1 출판사와 미국 문서에 대한 순위 결과 비교

문서	페이지랭크	시맨틱 링크 기반 페이지랭크
Macmillan	33위	15위
미국	5위	16위

표 2 출판사와 미국 문서 내 자원에 대한 비교

문서	문서 내 자원	자원랭크	시맨틱 링크 기반 페이지랭크
Macmillan	*macillan*	0.659	1.127
	publish company	0.468	
미국	*unitedState*	1.107	1.107

5. 결론 및 향후연구

온톨로지 언어인 RDF는 의미 있는 연결관계를 가진 트리플로 자원을 정의할 수 있다. 본 논문에서는 이러한 특성을 활용해 페이지랭크를 개선하였다.

시맨틱 링크 기반 페이지랭크는 웹문서가 얼마나 중요한 자원을 많이 가지고 있냐에 따라 중요도를 결정하므로 기존 페이지랭크에 비해 상대적으로 의미있는 랭킹 결과를 도출할 수 있다. 또한 기존 페이지랭크는 링크 기반의 중요도 계산으로 인해 상위 순위의 웹문서임에도 불구하고 사용자가 원하는 내용이 없을 수 있는 상황이 발생할 수 있지만, 시맨틱 링크 기반 페이지랭크는 자원에 대한 랭킹을 기반으로 웹문서의 랭킹을 결정하므로 상위 순위에 있는 웹문서는 의미 있는 자원을 가진 내용을 사용자에게 제공하는 것을 보장해준다.

향후 연구로는 의미를 가진 연결의 가중치를 계산하는 방법을 독자적으로 개발할 계획이다. 그리고 웹문서에 자동으로 RDFa 주석을 다는 기술을 사용해 RDFa가 정의되지 않은 웹문서에도 시맨틱 링크 기반 페이지랭크를 사용할 수 있도록 기능을 개선하고자 한다.

참 고 문 헌

[1] S. Brin, L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, vol.30, no.1-7, pp.107-117, 1998.

[2] W. Xing, A. Ghorbani, "Weighted PageRank Algorithm," In *proceedings of the 2nd Annual Conference on Communication Networks & Services Research*, pp.305-314, 2004.

[3] K. Stein, C. Hess, "Information Retrieval in Trust-Enhanced Document Networks," *Lecture Notes in Computer Science*, vol.4289, pp.65-81, 2006.

[4] P. Sharma, D. Tyagi, P. Bhadana, "Weighted Page Content Rank for Ordering Web Search Result," International Journal of Engineering Science and Technology, vol.2, no.12, pp.7301-7310, 2010.

[5] RDFa Taskforce, RDFa-W3C Semantic Web Deployment Wiki, http://www.w3.org/2006/07/SWD/wiki/RDFa.html

[6] RDF Working Group, Resource Description Framework, W3C, http://www.w3.org/RDF

[7] R. De Virgilio, F. Frasincar, W. Hop, S. Lachner, "A Reverse Engineering Approach for Automatic Annotation of Web Pages," *Multimedia Tools and Applications*, Published online, 2011.

[8] M. Duma, "RDFa Editor for Ontological Annotation," In *proceedings of the Student Research Workshop associated with RANLP*, pp.54-59, 2011.

[9] N. Toupikov, J. Umbrich, R. Delbru, M. Hausenblas, G. Tummarello, "DING! Dataset Ranking Using Formal Descriptions," In *Proceedings of the WWW 2009 Workshop on Linked Data on the Web*, 2009.

전 희 국

2004년 동국대학교 컴퓨터공학과 학사 2007년 고려대학교 전자컴퓨터공학과 석사. 2011년~현재 서울대학교 컴퓨터공학부 박사과정 재학 중. 관심분야는 데이터베이스, 시맨틱 웹, 온톨로지, 빅데이터

임 동 혁

2003년 고려대학교 컴퓨터교육과 학사 2005년 서울대학교 컴퓨터공학부 석사 2011년 서울대학교 컴퓨터공학부 박사 2012년 서울대학교 치학연구소 선임연구원. 2012년~현재 서울대학교 컴퓨터공학부 박사후과정. 관심분야는 데이터베이스, 시맨틱 웹, 온톨로지, 빅데이터

김 형 주

정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 39 권 제 4 호 참조