사용자 행동과 사회적 관계 기반의 블로그 랭크 모델

황재선*, 김장원**
*고려대학교 소프트웨어공학과
**고려대학교 컴퓨터·전파통신공학과
e-mail : neovis@korea.ac.kr, ikaros1223@korea.ac.kr

A Model for Blog Rank based on User Behavior and Social Relationship

Jae-Seon Hwang*, Jangwon Kim**
*Dept. of Software Engineering, Korea University
**Dept. of Computer Science and Radio Communications Engineering, Korea University

요 약

블로그는 누구나 쉽게 이용할 수 있는 도구이며, 블로그를 통한 콘텐츠의 생산과 소비는 빠른 속도로 증가하고 있다. 이런 블로그의 글은 단순히 정보를 전달하는 웹 페이지 이상의 사회적 관계를 포함하고 있다. 하지만 지금까지 웹 페이지 및 블로그에 대한 검색은 이러한 사회적 관계를 고려하지 않고 있다. 따라서 본 논문에서는 사용자 행동과 사회적 관계에 기반한 블로그 랭크 모델을 제안한다. 이를 기반으로 국내의 서로 다른 서비스에서 제공한 블로그 랭킹을 새롭게 제안한 블로그 모델과 비교하였고, 이를 통해 제안하는 블로그 모델의 타당성을 제시하였다.

1. 서론

웹 2.0 의 선풍적인 인기와 함께 개인의 일상부터 전문적인 지식까지 다양한 영역의 정보를 담고 있는 블로그(Blog)[1]는 일반인들에게 아주 일상적인 도구가 되었다. 블로그는 기술적 진입장벽 없이 누구나 쉽게 다룰 수 있는 도구이며, 웹을 통한 콘텐츠의 생산과 소비의 중요한 도구로서 그 가치가 점점 증대되고 있다[3].

또한 블로그는 인터넷 사용자들의 콘텐츠 소비 행태도 바꾸고 있다. 지금까지 특정 웹 페이지의 정보를 얻고자 하는 경우에는 직접 해당 웹 페이지를 방문하여야만 가능했다. 하지만 대부분의 블로그는 XML 기반의 RSS[4]를 지원하고 있기 때문에 블로그에서 작성되는 글에 대해 사용자들이 해당 블로그를 직접 방문하지 않고, RSS 구독하는 것만으로도 정보의 습득이 가능하게 되었다. 이런 새로운 소비 행태는 정보 습득의 효율성을 높일 뿐만 아니라 콘텐츠 소비의 양도 증가하고 있다[2]. 이처럼 블로그는 콘텐츠 생산과 소비의 중요한 도구이며, 블로그 콘텐츠의 양이 폭발적으로 늘어나고 있기 때문에 블로그에 대한 검색과 평가의 중요성도 계속 증가하고 있다.

웹 검색에 활용되는 PageRank[5] 및 HITS[6]와 같은 웹 페이지에 대한 랭크 분석 연구는 이미 상용화되어 널리 사용되고 있는 기술이다. 이들 연구에서처럼 하이퍼링크를 중심으로 블로그 콘텐츠를 평가한다면 블로그가 가지고 있는 폭발적인 콘텐츠 생산력과 사용자들간의 커뮤니케이션 및 상호작용과 같은 블로그 고유의 특성들을 반영하지 못하는 단점이 생긴다. 또한 아주 작은 수의 블로그 글만이 참조를 가지고 있어 기존 페이지 링크 방식만으로는 품질 저하가 일어나는 문제점도 발생한다[7, 9]. 영향력을 가진 블로거(Blogger)가 자신의 블로그에 새로운 글을 올리는 경우를 생각해보면 페이지 랭크와는 무관하게 짧은 시간 안에 많은 사람들에 의해 소비되고, 영향력을 발휘하며, 중요한 콘텐츠로 인식된다. 이처럼 지금까지의 검색을 통한 콘텐츠 소비 방식과 블로그를 통한 콘텐츠 소비 방식이 다르다는 것을 알 수 있으며, 전통적인 웹 페이지 랭크 방식은 블로그의 특성을 완벽하게 대응할 수도 없다[10].

이러한 문제를 해결하기 위해 BlogRank, EigenRumor 알고리즘이 제안되었고, 페이지 랭크를 결정하기 위해 새로운 요소들이 정의되었다. BlogRank 의 경우 블로그가 가진 문서들간의 연결의 특징을 분석하여 이들을 구조화하였으며, EigenRumor 의 경우 문서들간의 참조 링크뿐만 아니라 사이버 커뮤니티 구성원의 영향도와 해당 페이지의 평판을 통해 블로그 랭크를 구현하고자 하였다. 하지만 이들 또한 앞서 이야기한 페이지와 무관하게 짧은 시간 안에 많은 사람들에 의해 소비되는 블로그 콘텐츠에 대한 중요도와 블로그의 영향력은 판별할 수 없는 문제점이 발생한다.

본 논문의 구성은 다음과 같다. 2 절에서는 본 연구에 필요한 기반 기술과 관련 연구에 대해 설명한다. 3 절에서는 블로그 랭크를 위한 팩터들을 기술하며, 4 절에서는 블로그를 위한 랭크 모델을 제안한다. 5 절에서는 실제 블로그 데이터를 이용하여 새롭게 제안

한 블로그 랭크 모델에 대한 실험을 통해 모델 검증을 하며, 마지막으로 6 절에서는 결론 및 향후 연구를 제시한다.

2. 기반 기술 및 관련 연구

블로그는 일반 사용자도 손쉽게 콘텐츠를 생산할 수 있는 도구로서 일반 게시판의 게시물과는 달리 하나의 블로그 글은 변하지 않는 고유 주소(Permanent Link)를 가지며, XML 기반의 RSS 를 통해 블로그 글을 공개하고, 구독할 수 있다. 또한 블로거와 블로거, 블로거와 독자들의 코멘트(comment)와 트랙백(trackback)을 통해 서로의 의사를 교환하고 있으며, 이를 통해 블로그는 느슨하지만 소셜 네트워크의 특징을 가지게 된다[11]. 특히 트랙백은 코멘트나 기존의 하이퍼링크와는 달리 트랙백을 받은 블로그의 글 하단에 트랙백을 보낸 블로그 글이 노출되는 메커니즘을 가지고 있으며, 블로그를 통한 능동적인 커뮤니케이션의 수단이다[12]. 뿐만 아니라 트랙백이 하이퍼링크 형태로 서로 다른 블로그를 연결하기 때문에 블로그 랭크에 대한 연구에 있어 중요한 팩터로 활용하고 있다.

BlogRank 알고리즘은 블로그의 랭킹을 구하기 위해 블로그의 글인 포스트(Post)와 블로그 자체가 가지고 있는 속성들을 문서의 링크 그래프 특징에 반영하여 랭크 알고리즘을 제안하였다. BlogRank 는 PageRank 를 확장한 것으로, 블로그 문서가 가지는 링크들의 특성을 반영하여 랭크 알고리즘을 제안하는데 그쳤기 때문에 블로그 랭킹 및 블로그 콘텐츠 소비의 중요한 요소인 블로그 자체의 영향력, 블로거와 사용자와의 상호작용 등은 반영되지 못하였다.

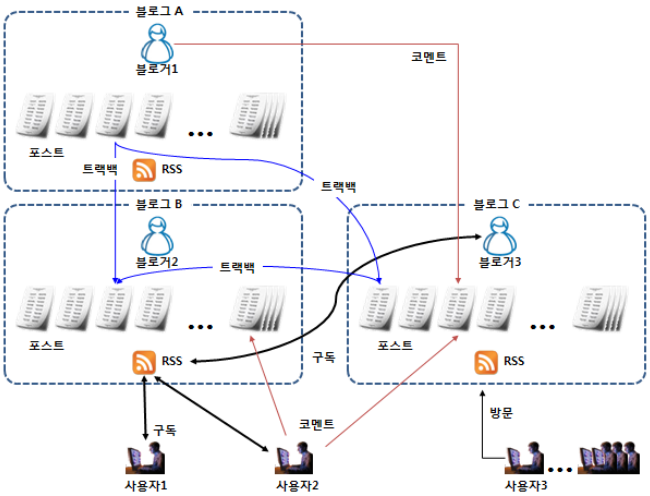
EigenRumor 알고리즘은 블로그 랭크를 위해 블로그에 대한 권위(Authority) 값과 허브(Hub) 값 이외에도 블로그 글에 대한 평판(Reputation) 값을 새롭게 제안하였다. 권위 값은 얼마나 의미 있는 글을 많이 생성하였는지를 나타내고, 허브 값은 이러한 의미 있는 글을 얼마나 평가하였는지에 따라서 계산된다. 마지막으로 글에 대한 평판 값은 글의 인기도를 반영하는 지표로서 허브 값이 높은 구성원들로부터 얼마나 많은 평가를 받았는가에 대한 계산 결과이다. EigenRumor 알고리즘은 평판 값을 통해 블로그의 특성을 반영하여 랭크를 구하였으나, 평판 값을 구하는데 트랙백에 대한 연결만 포함하고 있다는 단점이 있다. 실제 블로그 랭크를 위한 고려사항에는 트랙백 이외에도 블로그에 대한 구독 수, 방문수, 포스트 수, 코멘트 수 등 다양한 팩터들도 고려해야 한다.

블로그에 포함된 사용자 행동, 사회적 관계에 의미를 두고, 블로그를 소셜 네트워크로서의 하나로 정의한 연구도 진행되었다. 본 연구에서는 블로그를 통한 관계가 다른 커뮤니케이션 미디어나 오프라인 미팅의 관계로 확장되는 사회적 관계에 주목하여, 블로그 포스트가 포함하는 링크, 트랙백, 코멘트에 대한 네트워크 구조를 분석하였다. 블로그 중심의 사용자 행동과 사회적 관계에 초점을 두었지만 블로그 랭킹을 위한 랭크 알고리즘이 없다.

따라서 본 논문에서는 기존 PageRank 에서 중요한 팩터가 되는 하이퍼링크뿐만 아니라 블로그에 대한 평가에 사용자 행동과 사회적 관계에 기반한 팩터들을 제시하고, 이에 따른 새로운 블로그 랭크 모델을 제안한다.

3. 블로그 랭크를 위한 팩터

블로그 콘텐츠 소비에는 사용자의 검색에 따른 정보 취득을 목적으로 한 일회성의 콘텐츠 소비와 특정 블로그에 대한 관심 또는 관계에 기반한 적극적인 콘텐츠 소비의 형태인 RSS 구독이 있다. 또한 트랙백 및 코멘트를 통해 상호작용을 할 수 있으며, 이런 관계 또한 콘텐츠 소비에 영향을 미치고 있다.



(그림 1) 블로그 콘텐츠의 소비, 링크 및 사회적 관계

사용자 행동과 사회적 관계 기반의 블로그 랭크 모델을 위해 고려할 팩터에는 다음과 같은 항목들이 있다.

분류	영향력	상호작용	
		평가	주목도
팩터	구독 수 방문수 포스트 수 갱신 주기	아웃바운드 트랙백 아웃바운드 코멘트	인바운드 트랙백 인바운드 코멘트

(표 1) 블로그 랭크를 위한 팩터

영향력(influence)은 블로그 랭크 평가의 가장 중요한 요소로, 사용자들에 의해 소비되는 콘텐츠의 양으로 평가한다. 영향력이 높다는 것은 해당 블로그의 콘텐츠가 사용자로부터 높은 인기를 얻고 있으며, 실제 많은 소비가 일어나는 것을 의미한다. 영향력을 결정하는 팩터는 다음과 같다.

- **구독 수** : 블로그를 RSS 를 통해 구독하는 행위는 블로그 콘텐츠에 대한 일회성의 소비가 아닌 지속적인 관심을 보여주는 팩터로서, 구독 수는 블로그 영향력을 결정하는 가장 중요한 요소가 된다.
- **방문수** : 방문수가 많다는 것은 해당 블로그 콘

텐츠에 대한 소비가 많다는 것을 의미한다. 참조가 많아 PageRank 가 높은 블로그 뿐 만 아니라 새롭게 생성된 블로그 또는 새로운 글이라고 할지라도 많은 사람들로부터 관심을 받을 수 있는 블로그의 특징이 반영된 팩터이다. 단, 글 하나로 이슈가 집중되어 잠시 동안 주목을 끄는 것이 아니라 블로그 전체의 영향력을 반영하기 위해 방문수는 특정 기간 동안의 평균 값을 사용한다.

- **포스트 수** : 블로그 랭크가 중요한 이유는 블로그를 통한 콘텐츠의 생산과 소비에 있다. 이런 관점에서 블로그 포스트가 많다는 것은 콘텐츠 소비에 있어 사용자들에게 미치는 영향력이 높다.
- **갱신 주기** : 포스트 수와 마찬가지로 지속적인 업데이트를 보여주는 블로그가 그렇지 못한 블로그에 비해 콘텐츠의 생산과 소비에 더욱 많은 영향을 미치게 된다. 블로그 글에 대한 갱신 주기 또한 블로그의 영향력을 위한 팩터로 활용된다.

블로그는 단순한 정보 전달이 아닌 다른 사용자들과의 관계가 존재하고 있으며, 사용자들과의 상호작용(interaction) 또한 블로그 랭크를 구성하는 중요한 요소가 된다. 상호작용은 크게 다른 블로그에 대한 평가(evaluation)와 자기 블로그에 대한 주목도(attention)로 구분한다. 평가와 주목도에 해당하는 팩터는 다음과 같다.

- **아웃바운드 트랙백(outbound trackback)** : 다른 블로그 글에 대한 능동적인 커뮤니케이션을 나타내는 팩터로 아웃바운드 코멘트보다는 더욱 높은 가중치를 가진다. 특정 콘텐츠와의 관계를 맺는 것으로서 다른 블로그에 대한 평가 수단으로 취급한다.
- **아웃바운드 코멘트(outbound comment)** : 다른 블로그 글에 대한 커뮤니케이션으로 트랙백보다는 보다 쉬운 커뮤니케이션 수단이다. 코멘트를 많이 남기는 것은 다른 블로그와의 사회적 관계 확장을 활발히 하는 것을 의미한다.
- **인바운드 트랙백(inbound trackback)** : PageRank 의 하이퍼링크처럼 다른 블로그에서 해당 블로그를 참조하는 것이다. 주목도를 위한 중요한 팩터가 된다. 향후 PageRank 처럼 트랙백을 보낸 블로그의 영향력을 가중치로 활용한다면 보다 정확한 값을 구할 수 있다.
- **인바운드 코멘트(inbound comment)** : 자신의 블로그 글에 다른 사용자들이 남긴 코멘트를 의미하는 것으로 크게 익명과 블로그 주소를 밝힌 코멘트로 구분할 수 있다. 트랙백보다는 손쉬운 커뮤니케이션이지만 많은 사람들이 코멘트를 남긴다는 것은 그 만큼 주목을 받고 있는 것을 의미한다.

4. 블로그 랭크 모델

지금까지 소개한 다양한 팩터들을 기반한 사용자 행동과 사회적 관계 기반의 블로그 랭크 모델(kBlogRank)을 다음과 같이 정의하였다.

$$B(i) = \beta(\alpha \cdot I(i) + (1-\alpha)E(i)) + (1-\beta)A(i)$$

$B(i)$는 블로그 i에 대한 kBlogRank 이며, I는 영향력, E는 평가, A는 주목도를 나타낸다. α와 β는 댐핑(damping) 팩터로 0 과 1 사이의 값을 가지며, 본 논문에서는 각각 0.9, 0.8 로 정의하였다.

$$I(i) = w_R \frac{R(i)}{\sum_{j=1}^{n} R(j)} + w_P \frac{P(i)}{\sum_{j=1}^{n} P(j)} + w_C \frac{C(i)}{\sum_{j=1}^{n} C(j)} + w_F \frac{F(i)}{\sum_{j=1}^{n} F(j)}$$

$I(i)$는 블로그 i에 대한 영향력을 나타내며, R은 구독자수, P는 방문수, C는 전체 포스트 수, F는 갱신주기를 나타낸다. W_R, W_P, W_C, W_F는 각 항목에 대한 가중치로 40, 40, 15, 5 로 정의한다.

$$E(i) = w_{OT} \frac{OT(i)}{\sum_{j=1}^{n} OT(j)} + w_{OC} \frac{OC(i)}{\sum_{j=1}^{n} OC(j)}$$

$E(i)$는 블로그 i에 대한 평가를 나타내며, OT는 아웃바운드 트랙백, OC는 아웃바운드 코멘트 수치이다. W_{OT}와 W_{OC}는 각 항목에 대한 가중치로 80, 20 으로 정의한다.

$$A(i) = w_{IT} \frac{IT(i)}{\sum_{j=1}^{n} IT(j)} + w_{IC} \frac{IC(i)}{\sum_{j=1}^{n} IC(j)}$$

$A(i)$는 블로그 i에 대한 주목도를 나타내며, IT는 인바운드 트랙백, IC는 인바운드 코멘트 수치이다. W_{IT}와 W_{IC}는 각 항목에 대한 가중치로 80, 20 으로 정의한다.

5. 실험

국내의 경우 독자적인 모델을 이용하여 블로그 랭킹을 보여주는 서비스가 있다. 하지만 대표적인 서비스인 야후코리아[16]와 블로그코리아[17]의 블로그 랭킹 상위 30 위를 비교해보면 단지 2 개 블로그만 겹치며, 동일한 블로그에 대한 랭킹의 차이가 20 만 이상 발생하는 모습도 볼 수 있다.

블로그 URL	블로그코리아랭킹	야후코리아랭킹
http://blog.naver.com/xezenan	1위	38,063위
http://savenature.tistory.com/	2위	34위
http://blog.naver.com/yang456	3위	137위
http://conodont.egloos.com/	4위	192위
http://photohistory.tistory.com/	5위	2위
http://bizworld.tistory.com/	6위	388위
http://fiancee.tistory.com/	7위	2,955위
http://blog.naver.com/funkstyle	8위	2,250위
http://blog.naver.com/cyongjoon	9위	6,486위
http://manualfocus.tistory.com/	10위	209위
http://www.designlog.org	22위	14위
http://how2learn.tistory.com/	30위	224,515위

(표 2) 서로 다른 블로그 랭킹 서비스 비교

그래서 본 논문에서는 특정 블로그 랭킹을 표본으로 삼지 않고, 여러 블로그 서비스에서 선정한 우수 블로그를 이용하여 실험한다[13, 14, 15].

사용자 행동과 사회적 관계 기반의 블로그 랭크 모델(kBlogRank)에 대한 검증을 위해 서로 다른 서비스 업체에서 선발된 우수블로그 20 개와 랜덤방식으로 추출한 일반블로그 20 개에 대한 kBlogRank 와 블로그 랭킹을 구해보았다. 또한 kBlogRank 에서 방문수는 최근 1 개월 동안의 1 일 평균 값을 채택하여 일시적인 이슈에 따른 방문수 증가의 영향을 최소화하였고, 갱신 주기는 최근 1 개월 동안의 포스트 수를 사용하였다. ID 는 블로그 주소를 대신한 고유 숫자이며, 이를 통해 개별 블로그를 구분할 수 있다. 우수 블로그 항목은 우수 블로그로 선정된 경우에는 O 로 그렇지 않은 경우에는 X 로 표기하였다. 40 개 블로그에 대한 kBlogRank 는 다음과 같다.

ID	랭킹	kBlogRank	우수블로그	ID	랭킹	kBlogRank	우수블로그
6	1	8.361	O	5	21	1.659	O
10	2	6.394	O	11	22	1.650	O
15	3	5.848	O	19	23	1.430	O
9	4	5.277	O	39	24	1.290	X
8	5	5.141	O	7	25	0.926	O
17	6	4.729	O	27	26	0.912	X
1	7	4.688	O	38	27	0.810	X
12	8	4.421	O	31	28	0.723	X
3	9	4.075	O	26	29	0.707	X
18	10	3.626	O	37	30	0.705	X
21	11	3.131	X	36	31	0.621	X
13	12	2.878	O	25	32	0.617	X
14	13	2.687	O	32	33	0.606	X
4	14	2.685	O	35	34	0.528	X
29	15	2.642	X	24	35	0.517	X
28	16	2.435	X	33	36	0.460	X
20	17	2.300	O	40	37	0.333	X
16	18	1.992	O	23	38	0.325	X
2	19	1.966	O	34	39	0.166	X
30	20	1.691	X	22	40	0.127	X

(표 3) kBlogRank 및 블로그 랭킹

위 결과에서 보듯이 여러 서비스에서 추출한 우수 블로그들은 상대적으로 높은 kBlogRank 지수를 가지는 것을 확인할 수 있다. 상위 20 위권 이내에 우수블로그로 선정되지 않은 블로그들의 특징을 살펴보면 블로그 ID 21 번의 경우 월간 방문수가 40 만 명으로, 실험을 위한 모집단에서 월간 방문수 랭킹으로만 5 위에 해당하는 수치를 보여준다. 블로그 ID 29 번의 경우 비록 우수블로그로 선정되지 못했지만 등록되어 있는 전체 포스트 수가 2300 개, 최근 1 개월 동안 올라온 글이 80 개로 모집단 중 가장 높은 수치를 보여주었으며, 인바운드 코멘트 또한 10 위 정도의 수치를 보여주어 이와 같은 결과가 나왔다.

6. 결론

지금까지 블로그를 단순한 웹 페이지로 보고, PageRank 또는 이와 유사한 알고리즘만으로 그 가치를 평가하는 것은 블로그가 가진 하이퍼링크 이상의 사회적 관계, 사용자들의 콘텐츠 소비 행태를 반영하지 못하는 결과를 가져왔다. 사용자 행동과 사회적 관계 기반의 블로그 랭크 모델은 이런 기계적 계산이 아닌 다양한 팩터들이 추가적으로 고려하여 모델을 제안하였으며, 실험을 통해 제안 모델이 블로그 환경에 적합하다는 것을 확인할 수 있었다. 다만, 본 논문에서 사용한 모집단은 그 규모가 작았고, 단순히 우수블로그 여부를 두고 비교하였기 때문에 향후 연구에서는 대규모의 블로그에 대한 블로그 랭크 모델을 검증하고, 블로그 랭크 모델에 필요한 각각의 팩터들을 자동으로 추출하는 방법에 대한 연구가 진행되어야 할 것이다.

참고문헌

[1] Wikipedia, “Blog,” http://en.wikipedia.org/wiki/Blog

[2] Technorati, “State of Blogosphere 2007,” http://www.sifry.com/alerts/archives/000493.html

[3] Universal McCann, “Power To The People – Wave3 Report,” 2008

[4] RSS(RDF Site Summary), http://web.resource.org/rss/1.0

[5] S. Brin and L. Page, “The Anatomy of a Large-scale Hypertextual Web Search Engine,” In Proceedings of 7th International World Wide Web Conference, 1998

[6] J. M. Kleinberg, “Authoritative sources in hyperlinked environment,” Journal of the ACM, Vol. 46, No. 5, 1999

[7] K. Fujimura, T. Inoue and M. Sugisaki, “The EigenRumor Algorithm for Ranking Blogs,”WWW 2005, 2005

[8] K. Fujimura and N. Tanimoto, “The EigenRumor Algorithm for Calculating Contributions in Cyberspace Communities,” Trusting Agents, LNAI 3577, pp. 59-74, 2005

[9] J. Shen, Y. Zhu, H. Zhang, C. Chen, R. Sun and F. Xu, “A Content-based Algorithm for Blog Ranking,” 2008 International Conference on Internet Computing in Science and Engineering

[10] A. Kritikopoulos, M. Sideri and I. Varlamis, “BLOGRANK: RANKING WEBLOGS BASED ON CONNECTIVITY AND SIMILARITY FEATURES,” ACM International Conference Proceeding Series; Vol. 198, 2006

[11] N. Ali-Hasan and L. Adamic, "Expressing Social Relationships on the Blog through Links and Comments," ICWSM’2007 Boulder, 2007

[12] SixApart, “TrackBack Technical Specification,” http://www.sixapart.com/pronet/docs/trackback_spec 2002

[13] Tistory, “2008 TOP 100 Award,” http://www.tistory.com/supporters/

[14] Egloos, “Egloos TOP 100,” http://top100.egloos.com/5705

[15] Allblog, “Allblog Award 2008,” http://award.allblog.net/

[16] Yahoo! Korea, http://kr.blog.search.yahoo.com

[17] BlogKorea, http://www.blogkorea.net/