제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

# 블로그 포스트 랭킹을 위한 액션 발생의 순서 활용 방안

황원석*, 도영주**, 김상욱*
*한양대학교 전자컴퓨터통신공학과
**매크로임팩트(주) 시스템소프트웨어연구소
e-mail: hws23@hanyang.ac.kr

# Using Action Occurrence Orders in Blog Post Ranking

Won-Seok Hwang*, Young-Joo Do**, Sang-Wook Kim*
*Dept of Electronics and Computer Engineering, Hanyang University
**MacroImpact Inc.

요 약

블로그 이용이 활성화됨에 따라 포스트 랭킹 알고리즘의 필요성이 증가하고 있다. 본 논문에서는 액션의 발생 순서를 이용하여 포스트에 랭킹을 부여하는 방법을 제안한다. 또한, 실제 블로그 데이터를 이용한 실험을 통하여 본 논문에서 제안하는 방법의 성능의 우수성을 보인다.

## 1. 서론

블로그란 블로거가 작성한 웹 상의 글인 포스트를 게재하는 일종의 개인 웹사이트이다. 블로거는 블로그의 포스트에 다양한 행위를 통하여 다양한 의견을 밝힐 수 있다. 블로거가 포스트에 행할 수 있는 모든 행위를 액션(action)이라 정의한다. 이러한 액션에는 작성, 덧글 쓰기, 스크랩 등이 있다. 블로그에는 이런 액션들의 정보와 함께 그들의 발생 시각이 저장되어 있다.

최근 블로그의 편의성으로 인해 블로그 이용이 증가하고 있다. 이에 따라 포스트의 양이 증가하고, 검색 결과로 다수의 포스트들이 나타날 확률이 높아진다. 따라서 다수의 포스트들 중 유저가 원할 것으로 생각되는 양질의 포스트를 상위 결과로 보여주는 랭킹 알고리즘이 필요하다.

이를 위해 PageRank[1]와 HITS[2]를 활용한 다양한 포스트 랭킹 알고리즘들이 제안되었다. 이들은 액션을 이용하여 포스트에 랭킹을 부여하였다. 하지만, 액션의 발생 순서를 랭킹의 정확도를 위해 활용한 알고리즘은 없었다.

본 논문에서는 정확한 포스트 랭킹을 위하여 액션의 발생 시각을 통해 액션의 순서를 파악하고, 이에 따른 가중치를 구한다. 이를 AuthHub에 적용하는 방안을 제안하고, 실험을 통해 기존 방법과 비교한다.

## 2. 관련 연구

본 장에서는 기존의 포스트 랭킹 알고리즘들에 대해 소개한다. Indegree는 포스트를 스크랩한 블로거의 수에 비례하도록 포스트에 랭킹을 부여하는 방법이다[4].

PostRank는 액션뿐만 아니라 그 액션을 부여한 블로거의 작성능력을 고려하여 포스트에 랭킹을 부여하는 방법이다[4]. 이때, 블로거의 작성능력은 양질의 포스트를 작성한 정도를 의미한다. 포스트의 랭킹은 포스트를 스크랩한 블로거들의 작성능력에 비례하도록 부여된다.

EigenRumor는 액션과 블로거의 작성능력, 평가능력을 동시에 고려하는 방법이다[3]. 블로거의 작성능력은 PostRank와 동일한 의미를 지닌다. 블로거의 평가능력은 양질의 포스트에 덧글을 단 정도를 의미한다. 포스트의 랭킹은 포스트를 작성한 블로거의 작성능력과 이를 스크랩한 블로거의 평가능력에 비례하도록 부여된다.

AuthHub는 액션과 블로거의 평가능력을 동시에 고려하는 방법이다[4]. EigenRumor와 달리 블로거의 평가능력은 양질의 포스트에 스크랩을 한 정도를 의미한다. 포스트의 랭킹은 포스트를 스크랩한 블로거의 평가 능력에 비례하도록 부여된다.

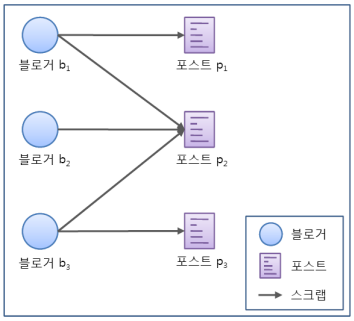
## 3. 제안하는 방안

본 논문에서 제안하는 방안은 AuthHub와 동일하게 스크랩과 블로그의 평가능력을 고려한다. 단, 블로거의 평가능력은 양질의 포스트를 스크랩한 정도뿐 만 아니라 그 스크랩이 동일한 포스트에서 일어난 다른 스크랩보다 순서상 빠른 정도 또한 고려한다. 이는 타 블로거보다 양질의 포스트를 빠르게 찾아 스크랩 하는 블로거가 더 평가를 잘할 것이라고 생각하기 때문이다.

이를 위해 먼저 블로그 환경을 그래프로 모델링해야 한다. 블로거와 포스트는 각각 블로거 노드와 포스트 노드로, 액션은 노드들 사이의 링크로 하는 블로거-포스트 그래프로 나타낸다. 포스트 노드에는 포스트의 질을 나타내는 권위점수를, 블로거 노드에는 블로거의 평가능력을 나타내는 허브점수를 부여하고, 블로거-포스트 그래프를 이용하여 이들 점수를 계산한다. (그림 1)은 블로거-포스트 그래프를 나타낸 예이다.

권위점수는 포스트 노드와 링크로 연결된 블로거 노드들의 허브점수의 합으로 계산된다. 이는 $\boldsymbol{a} = S\boldsymbol{h}$로 나타낼 수 있다. 전체 포스트의 수가 m, 블로거의 수가 n일 때, $\boldsymbol{a}$

는 모든 권위점수를 나타내는 m×1 벡터이고, $\boldsymbol{h}$는 모든 허브점수를 나타내는 n×1 벡터이다. $S$는 m×n 행렬로 각 인자 $s_{i,j}$는 포스트 노드 $p_i$와 블로거 노드 $b_j$ 사이에 링크가 있는 경우는 1, 없는 경우는 0의 값을 가진다.




(그림 1) 블로거-포스트 그래프

허브점수는 블로거 노드와 링크로 연결된 포스트 노드들의 권위점수의 가중치의 합으로 계산된다. 이때, 가중치는 동일한 포스트에서 스크랩이 일어난 순서에 의해 0에서 1사이의 값이 부여된다. 본 논문에서는 가중치를 결정하기 위한 방법으로 EDA(Equal Difference Attenuation)와 ERA(Equal Ratio Attenuation)를 제안한다. 이 가중치에 비례하여 허브점수가 받는 값의 크기가 결정된다.

EDA에서 가중치는 식 1에 의하여 계산된다. $order(p_i, b_j)$는 $b_j$가 $p_i$를 몇 번째로 스크랩하였는지에 대한 순서이다. $deg(p_i)$는 $p_i$의 차수(degree)를 의미한다. $min$은 사용자가 주는 0에서 1사이의 값으로, 에지에 부여될 가중치의 최소값을 의미한다. 이는 가중치가 일정 수치만큼씩 작아져서 0 이하가 되는 것을 막기 위한 값이다.

$$wd_{i,j} = 1 - \left(\frac{order(p_i, b_j)}{1 - deg(p_i)} \times min\right) \qquad \text{(식 1)}$$

$wd_{i,j}$를 허브점수의 계산에 이용하기 위해, $wd_{i,j}$를 $i$행과 $j$열의 인자로 하는 m×n 행렬 $WD$로 구성할 수 있다. 이를 통해 허브점수는 $\boldsymbol{h} = WD^T \boldsymbol{a}$로 계산된다.

ERA에서 포스트 노드 $p_i$와 블로거 노드 $b_j$ 사이의 링크의 가중치는 식 2에 의하여 계산된다. 이때, $\rho$는 감쇄하는 비율의 값으로 0에서 1사이로 주어진다.

$$wr_{i,j} = \rho^{order(p_i, b_j) - 1} \qquad \text{(식 2)}$$

$wr_{i,j}$를 허브점수의 계산에 이용하기 위해, $wr_{i,j}$를 WD와 동일한 방법으로 m×n 행렬 $WR$로 구성할 수 있다. 이를 통해 허브점수는 $\boldsymbol{h} = WR^T \boldsymbol{a}$로 계산된다.

EDA와 ERA의 권위점수와 허브점수는 각각 HITS와 동일하게 파워 메소드(power method)를 통해 계산한다[2]. 이 결과들을 $\text{AuthHub}_{EDA}$, $\text{AuthHub}_{ERA}$로 부른다.

## 4. 실험

실험에 사용한 데이터는 2006년 4월부터 수개월간 수집하여 익명으로 처리한 블로그 데이터이다. 알고리즘을 평가하기 위해 HITS[2]에서 사용한 질의 20개를 사용한다.

실험을 위해 각 포스트 랭킹 알고리즘의 결과 중 권위점수가 가장 높은 10개의 포스트를 질의마다 선택한 뒤, 포스트의 질을 11명의 평가자를 통해 '상', '하'로 평가하도록 한다. 각 평가자들이 평가한 결과의 최빈치(mode)를 해당 포스트의 질로 간주한다. 포스트 랭킹 알고리즘의 성능 척도로는 정밀도(precision)와 평균정밀도(average precision)을 이용한다. 이 평가 척도는 기존의 웹 문서 랭킹알고리즘을 평가하기 위해 주로 사용되는 방법으로, 좋은 알고리즘일수록 큰 값을 가진다. 본 실험에서는 각 질의 별로 측정된 정밀도와 평균정밀도의 평균값을 해당 알고리즘의 성능으로 간주한다.

본 실험에서는 AuthHub, $\text{AuthHub}_{ERA}$, $\text{AuthHub}_{EDA}$를 비교한다. EDA의 min값은 0.5, 0.75로, ERA의 ρ값은 0.8, 0.9, 0.95로 세팅한다.

<표 1> 포스트 랭킹 알고리즘의 정밀도와 평균정밀도

| 척도 \ 알고리즘 | AuthHub | $\text{AuthHub}_E$ DA (min=0.5) | $\text{AuthHub}_E$ DA (min=0.75) | $\text{AuthHub}_E$ RA (ρ=0.8) | $\text{AuthHub}_E$ RA (ρ=0.9) | $\text{AuthHub}_E$ RA (ρ=0.95) |
|---|---|---|---|---|---|---|
| 정밀도 | 0.863 | 0.858 | 0.863 | 0.874 | 0.868 | 0.868 |
| 평균 정밀도 | 0.876 | 0.876 | 0.888 | 0.894 | 0.895 | 0.900 |

<표 1>은 정밀도와 평균정밀도를 측정한 결과이다. 정밀도에서는 AuthHub에 비해 $\text{AuthHub}_{ERA}$가 높은 값을 보인 반면, $\text{AuthHub}_{EDA}$는 낮은 값을 보였다. 이는 $\text{AuthHub}_{ERA}$가 AuthHub에 비해 정확한 랭킹을 부여할 수 있음을 나타내는 것이다. 평균정밀도에서는 제안하는 모든 알고리즘이 높은 값을 보였다. 이는 액션의 발생 순서를 고려하는 방법이 AuthHub보다 양질의 포스트에 높은 랭킹을 부여한다는 것을 의미한다. 전체적으로 $\text{AuthHub}_{ERA}$가 가장 정확한 랭킹을 부여함을 확인할 수 있다.

## 5. 결론

본 논문에서는 액션의 발생 순서를 액션의 발생 시각을 통해 구하고, 이를 통해 정확한 블로거의 평가능력을 계산하기 위한 가중치를 제안하였다. 또한, 가중치를 기존의 AuthHub에 적용하여 더 정확한 랭킹을 부여하는 알고리즘인 AuthHubEDA와 AuthHubERA를 제안하였다. 제안하는 알고리즘들의 성능을 실제 블로그 데이터를 이용한 실험을 통해 보였다.

## 감사의 글

본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음. (NIPA-2010-(C1090-1011-0009))

## 참고문헌

[1] S. Brin and L. Page, "PageRank: Bringing Order to the Web," *Stanford Digital Libraries Technologies Project*, Working Paper 1999-0120, 1998.

[2] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," In *proc. of the 9th ACM-SIAM Symposium on Discrete Algorithm*, 1998.

[3] K. Fujimura, T. Inoue and M. Sugisaki, "The EigenRumor Algorithm for Ranking Blogs," In *proc. of the 14th Int'l WWW Conference*, 2005.

[4] W. Hwang, S. Kim, D. Bae and Y. Do, "Post Ranking Algorithms in Blog Environment," In *proc. of the 2nd Int'l FGCNS Conference*, 2008.