HITS 에 기반한 포스트 랭킹 알고리즘: 개선 방안 및 성능 평가

황원석*, 도영주**, 김상욱*
*한양대학교 전자컴퓨터통신공학과
**매크로임팩트㈜ 시스템 소프트웨어 연구소
e-mail : hws23@agape.hanyang.ac.kr

Post Ranking Algorithms Based on HITS: Improvement and Performance Evaluation

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 방법의 필요성이 증가하고 있다. 기존 연구에서는 HITS 에 기반을 둔 포스트 랭킹 방법인 BAITS 와 BAITS 를 수정한 포스트 랭킹 방법들을 제안한 바 있다. 본 논문에서는 이러한 포스트 랭킹 방법들을 다양한 척도를 이용한 비교 실험을 통해 비교하여 가장 정확도가 높은 랭킹 방법을 판별하고자 한다.

1. 서론

블로그(blog)는 그 소유자인 블로거(blogger)가 작성한 포스트(post)들이 아카이브(archieve) 형태로 구성되어 있는 일종의 개인 웹 사이트이다. 블로거는 자신의 블로그에 포스트를 작성하거나 타 블로거가 작성한 포스트들에 블로그 액션(blog action)을 통해 다양한 행동을 할 수 있다 [2].

블로그의 편리한 기능들로 인해 다양한 내용을 담은 다수의 포스트들이 지속적으로 생성되었고, 이는 포스트 검색에서 풍부함의 문제 [1]를 가져왔다. 이 문제를 해결하기 위하여 검색에서 포스트에 랭킹을 부여하는 랭킹 방법이 필요하다. 포스트 랭킹 방법은 검색 사용자가 원할 것이라 생각되는 품질이 좋은 포스트에 높은 랭킹을 부여하여 사람들이 자신이 원하는 포스트를 쉽게 찾을 수 있도록 한다.

참고 문헌 [2]에서는 블로그 액션들 중 스크랩(scrap)이나 포스트엮기(trackback)를 이용하여 포스트에 랭킹을 부여하고자 하였다. 스크랩은 한 블로거가 다른 블로거의 포스트를 자신의 블로그로 복사하는 것이고, 포스트엮기는 한 블로거가 자신의 포스트의 하단에 다른 블로거의 포스트의 주소를 남기는 블로그 액션이다. 이 두 가지 블로그 액션은 블로거가 포스트의 품질에 만족하였다는 것으로 해석할 수 있으므로 이를 통해 포스트의 품질을 파악할 수 있다. 참고 문헌 [2]에서는 포스트 랭킹 방법들을 소개하고, 다양한 실험을 통해 포스트 랭킹 방법들의 정확도를 비교하여 HITS [1]를 변형한 방법인 BAITS 가 정확함을 보였다.

또한, 참고 문헌 [2]에서는 BAITS 의 계산 과정에서의 문제점을 지적하고, 이를 해결하기 위하여 블로거 점수를 제한하는 랭킹 방법인 *BloggerAVG* 와 *BloggerAtK* 를 제안하였다. 참고문헌[2]에서 BloggerAVG 는 BAITS 보다 정확한 것으로, BloggerAtK 는 부정확한 것으로 평가되었다. 그러나 참고 문헌 [2]에서 언급한 바와 같이, 두 가지의 척도만을 이용한 비교 실험의 결과만으로는 블로거 점수를 제한하는 랭킹 방법들이 BAITS 의 문제를 해결하고 더 정확한 랭킹 결과를 도출하는지 판단하기 어렵다.

본 논문에서는 추가적인 실험들을 통해 블로거 점수를 제한하는 랭킹 방법들이 BAITS 의 정확도를 향상시킬 수 있는지를 분석하고자 한다. 이를 위하여 각 포스트 랭킹 방법의 정확도를 *intersection* [3], *weighted intersection* [3], *Kendall distance* [4]를 통해 측정한다.

2. BAITS와 블로거 점수 제한하는 BAITS

BAITS 는 블로그 액션을 잘 부여하는 블로거들에 의해 추천을 많이 받은 포스트를 품질이 높은 포스트로 정의한다. 블로그 액션을 잘 부여하는 블로거란 좋은 품질의 포스트를 많이 찾아 스크랩하거나 엮인글로 엮은 블로거를 의미하고, 블로거가 액션을 잘 부여하는 정도는 블로거 점수로 나타낸다. 포스트의 품질은 포스트 점수로 표현된다.

BAITS 는 블로거가 의도적으로 다수의 포스트를 스크랩하거나 엮인글로 엮어서 높은 블로거 점수가 되도록 조작하기 쉽다는 문제점이 있다. 또한 포스트의 품질에 상관없이 다수의 포스트를 무작위로 스크랩하거나 엮인글로 엮는 블로거들에게 높은 블로거 점수를 부여한다는 문제점이 있다. 이 문제점들로 인하여 블로거가 블로그 액션을 잘 부여하지 못함에도 불구하고 높은 점수를 부여 받아 잘못된 랭킹을 포스트에 부여할 수 있다. 이는 품질이 좋지 않은 포스트에 높은 랭크를 부여하게 만들어 랭킹의 정확도를 감소시키는 요인이 될 수 있다.

이러한 BAITS 의 단점을 보완하기 위해 블로거 점수 계산에 제한을 두는 알고리즘들이 BloggerAVG 와 BloggerAtK

이다 [2]. 이 알고리즘들과 같이 블로거 점수의 계산을 변화시키면 포스트 점수 또한 변화하게 되고, 최종적으로 다른 랭킹 결과를 얻을 수 있다. 이는 블로거 점수를 제한하는 알고리즘들이 HITS 와 동일하게 포스트 점수와 블로거 점수를 반복하여 계산하기 때문이다.

BloggerAVG 는 블로거가 다수의 포스트를 스크랩하거나 엮인글로 엮었어도 그 포스트들이 좋은 품질이 아니라면 블로거 점수가 높게 부여되지 않도록 하는 방안이다. 이를 위해 블로거 점수는 그 블로거 노드와 에지로 연결된 포스트들의 점수의 평균으로 부여된다.

BloggerAtK 는 품질 낮은 포스트를 스크랩하거나 엮인글로 엮더라도 높은 품질의 포스트를 일정 수 이상 스크랩하거나 엮인글로 엮었다면, 블로그 점수를 감점시키지 말아야 한다고 생각하는 방법이다. 이를 위해 블로거가 스크랩하거나 엮인글로 엮은 포스트들 중 가장 높은 포스트 점수를 가진 것들만을 블로거 점수의 계산에 이용한다.

3. 실험

실험을 위해 국내 블로그 서비스 중 하나인 네이버 블로그에서 수개월간 수집하여 익명으로 처리한 데이터를 사용하였다. 포스트 랭킹 방법의 정확도는 *intersection* [3], *weighted intersection* [3], *Kendall distance* [4] 을 통해 나타냈다. Intersection 은 두 상위 k 개의 포스트들 중 일치하는 포스트의 수를 나타내는 값이다. Weighted intersection 은 상위 1 개에서 k 개까지 각 위치에서 intersection 을 계산하고, 이를 평균하여 계산한 값이다. Kendall distance 는 두 순열을 구성하는 요소들의 순서를 비교하는 척도이다. 실험에서 Kendall distance 의 파라미터 p 는 0.5 로 두고 실험하였다. 랭킹 방법이 정확할수록 intersection, weighted intersection 의 값은 높게, Kendall distance 는 낮게 계산된다.

본 실험에 이용한 모든 척도들은 두 순열을 비교하는 척도로써 이 중 하나는 11 명의 사용자들의 평가를 통해, 나머지는 랭킹 방법을 통해 생성하였다. 실험에서 BloggerAtK 의 파라미터 K 는 블로거 노드들의 차수의 평균과 차수의 중앙값을 각각 이용하였고, 각각을 *BloggerAtK(AVG)*, *BloggerAtK(MED)*로 표시하였다.

표 2 와 3 은 각각 intersection 과 weighted intersection 의 결과이다. 이 결과에서는 블로거 점수를 제한하는 랭킹 방법들의 성능이 전반적으로 우수한 것으로 나타났다. 특히 BloggerAtK(AVG)가 가장 정확한 것으로 평가되었다는 점에서 참고 문헌 [2]의 결과와 차이를 보인다. 또한, BloggerAtK(MED)도 BAITS 보다 더 정확한 것으로 평가되었다. 표 4 는 Kendall distance 의 결과를 나타낸다. 여기서도 BloggerAtK(AVG)가 가장 정확한 것으로 평가되었으나, 전반적으로 블로거 점수를 제한하는 랭킹 방법들이 BAITS 보다 정확한 것으로 평가되지 못하였다. 전체적인 결과를 보았을 때, 블로거 점수를 제한하는 랭킹 방법들은 BAITS 의 정확도를 향상시킨다고 할 수 있다.

<표 2> 포스트 랭킹 방법들의 Intersection

# Posts / Method	Top 10	Top 20	Top 30
BAITS	3.45	8.40	13.50
BloggerAVG	2.60	8.85	14.10
BloggerAtK(AVG)	**4.05**	**10.65**	**15.00**
BloggerAtK(MED)	3.05	9.45	14.50

<표 3> 포스트 랭킹 방법들의 Weighted Intersection

# Posts / Method	Top 10	Top 20	Top 30
BAITS	**1.505**	3.770	6.248
BloggerAVG	1.140	3.563	6.375
BloggerAtK(AVG)	1.475	**4.673**	**7.515**
BloggerAtK(MED)	1.170	3.813	6.663

<표 4> 포스트 랭킹 방법들의 Kendall Distance

# Posts / Method	Top 10	Top 20	Top 30
BAITS	0.483	0.449	0.453
BloggerAVG	**0.476**	0.465	0.455
BloggerAtK(AVG)	0.497	**0.421**	**0.402**
BloggerAtK(MED)	0.500	0.463	0.477

4. 결론

기존의 포스트 랭킹 방법들 중 BAITS 는 가장 우수한 성능을 보였으나, 블로거 점수 계산에 문제점이 있었다. 이를 해결하기 위하여 블로거 점수를 제한하는 랭킹 방법들이 제안되었으나, 그 효과에 대해 충분한 실험이 이루어지지 못하였다. 본 논문에서는 다양한 실험을 통하여 블로거 점수를 제한하는 랭킹 방법들의 정확도를 평가하였다. 그 결과 블로거 점수를 제한하는 랭킹 방법들이 BAITS 의 정확도를 향상시킴을 보였다.

감사의 글

본 연구는 2010 년도 정부(교육과학기술부)의 재원으로 한국연구재단 (No.2008-0061006) 및 지식경제부 및 정보통신산업진흥원의 'IT 융합 고급인력과정 지원사업' (NIPA-2011-C6150-1101-0001)과 NHN(주)의 지원을 받았습니다. 그러나, 본 논문에서 제시된 의견이나 결론, 또는 권고 등은 온전히 저자(들)의 것이며, 반드시 지원회사의 입장을 대변하는 것은 아닙니다.

참고문헌

[1] J. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” In *Proc. of the 9th ACM-SIAM Symp. on Discrete Algorithms*, pp.668–677, 1998.

[2] 황원석, 도영주, 김상욱, “스크랩 기능을 지원하는 블로그 공간에서 포스트 랭킹 방안: 알고리즘 및 성능 평가,” *정보처리학회논문지*, 2011. (accepted to appear).

[3] A. Borodin, R. Gareth, S. Jeffrey, and T. Panayiotis, “Link analysis ranking: Algorithms, theory, and experiments,” ACM Transactions on Internet Technology, 5(1):231–297, 2005.

[4] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing top k lists,” In ACM SIAM Symposium on Discrete Algorithms, pages 28-36, 2003.