다차원적 랭킹을 통한 웹 스팸 탐지

정연성[1] 황지영[2‡]
[1]성균관대학교 통계학과
[2]성균관대학교 컴퓨터공학과
{jys3548, jjwhang}@skku.edu

Multidimensional Ranking for Web Spam Detection

Yeonsung Jung[1] Joyce Jiyoung Whang[2‡]
[1]Department of Statistics, Sungkyunkwan University
[2]Department of Computer Science and Engineering, Sungkyunkwan University

요 약

웹 스팸(web spam)이란 웹상에서 특정 목적을 가지고 웹 페이지의 랭크 점수(rank score)를 인위적으로 높게 만들어 해당 페이지가 높은 빈도로 노출되도록 하는 것을 의미한다. 이러한 스팸 페이지는 사용자들의 진성(genuine) 추천 결과에 위배되는 검색 결과를 만들어내므로 검색엔진의 성능을 저하시키는 주요 요인으로 작용한다. 일반적으로 널리 사용되고 있는 링크 기반 스팸 탐지(link-based spam detection)는 링크 그래프 구조를 이용하여 웹 스팸을 제어하는 기법이다. 기존의 링크 기반 스팸 탐지 알고리즘들은 페이지 혹은 호스트 중 한 가지를 선택하여 페이지 단위 혹은 호스트 단위에서의 연결성을 고려한 알고리즘들이 대부분이었다. 본 논문에서는 페이지 단위와 호스트 단위를 모두 고려하여 스팸의 특성을 파악하고, 더 나아가서 호스트들의 클러스터링 구조를 활용하여 보다 거시적인 관점에서 스팸의 성질을 분석하였다. 이러한 분석을 종합하여, 결과적으로 다차원적 랭킹을 통한 웹 스팸 탐지 알고리즘의 설계가 가능하다는 것을 보여주었다.

1. 서 론

검색엔진이란, 사용자가 검색어를 입력하면 검색어와 연관된 페이지 중 검색엔진의 자체 랭킹 알고리즘(ranking algorithm)을 기반으로 높은 랭크 점수를 획득한 페이지들을 사용자에게 제공해주는 시스템이다. 높은 랭크 점수를 획득하여 검색결과 상단에 위치하는 것은 사용자 유입, 광고 등의 측면에서 많은 경제적·비경제적 이익을 가져다준다. 이러한 이익을 얻기 위해 인위적으로 랭크 점수를 높게 만드는 악의적인 유저(user)를 스패머(spammer)라하고, 이러한 목적으로 만들어진 페이지를 웹 스팸(web spam)이라고 한다.

검색엔진 상에서 좋은 점수(혹은 높은 랭크)를 얻기 위해 스패머들은 주로 용어 스패밍(term spamming), 링크 스패밍(link spamming) 등의 기법들을 사용한다. 용어 스패밍은 페이지의 제목, 메타 데이터(meta data), 페이지 내용 등에 이용자들이 많이 검색하는 단어들을 삽입하여 랭크 점수를 높이는 기법이다. 링크 스패밍은 다수의 신뢰도가 높은 페이지와의 링크 수를 인위적으로 늘리거나, 스팸 페이지끼리의 내부적인 순환을 만드는 행위(link farm)를 통해서 랭크 점수를 높이는 기법이다 [1]. 최근에는 검색엔진의 크롤러를 피하기 위해서, 크롤러와 사용자에게 서로 다른 내용(content)을 보여주는 은신(hiding) 기법도 사용한다.

이러한 다양한 스팸들을 탐지하기 위한 기법으로는 크게 내용기반 스팸탐지(content-based spam detection)와 링크기반 스팸탐지(link-based spam detection)으로 나눌 수 있다. 내용기반 스팸탐지는 페이지 당 단어 수, 사용된 언어, 앵커 텍스트(anchor text)의 양, 질의(query)의 패턴 등을 이용해서 탐지한다. 링크기반 스팸탐지는 각 페이지를 노드(node)로, 페이지간의 하이퍼링크(hyper link)를 간선(edge)으로 하는 그래프 모델에서, 스팸 페이지들의 링크 구조의 특정한 패턴을 찾아내어 탐지한다. 이 기법은 내용기반 스팸탐지보다 더 경제적이기 때문에 많은 검색엔진에서 주로 사용하는 기법이다. 하지만 대부분의 기존 알고리즘들은 페이지 레벨(page level) 혹은 호스트레벨(host level) 중 한 가지의 차원을 선택하여 랭킹(ranking)을 하였다.

본 논문에서는 페이지 레벨, 호스트 레벨, 더 나아가서 호스트의 클러스터 레벨에서의 스팸의 특성을 분석하고, 이러한 레벨별 분석 결과를 모두 종합하여 스팸 탐지에 활용한 다차원적 랭킹 알고리즘을 통한 스팸 탐지 프레임워크를 제안한다.

2. 관련 연구

기존의 대표적인 링크기반 탐지기법에는 PageRank[3]와 TrustRank[2]가 있다. PageRank는 영향력 있는 페이

‡ 교신저자 (Corresponding author)
이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2016R1D1A1B03934766, NRF-2010-0020210). 또한, 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(2015-0-00914). 본 연구에서 사용된 모든 데이터는 네이버(주)에서 제공받았음.

지들에게 많은 링크를 받는 페이지일수록 신뢰도가 높다는 아이디어에 기반을 두어, 평가하고자 하는 페이지의 들어오는 링크(inlink)의 질과 양을 이용하여 랭크 점수를 측정한다. TrustRank는 신뢰도가 높은 페이지일수록 또 다른 신뢰도가 높은 페이지에게 링크를 많이 한다는 아이디어에 기반을 두어, PageRank가 들어오는 링크(inlink)에 집중한 것과는 반대로 나가는 링크(outlink)에 초점을 맞춘 알고리즘이다. 이외에도 비신뢰점수(distrust rank score)를 전파하여 랭크를 측정하는 Anti-TrustRank[4], 신뢰점수와 비신뢰점수를 동시에 전파하여 두 점수를 선형 결합하여 랭크를 측정하는 trust and distrust propagation 등이 있다.

3. 다차원적 랭킹 알고리즘

3.1 TrustRank

TrustRank는 먼저 PageRank를 역으로 적용하여 나가는 링크가 많은 상위 페이지들을 씨앗(seed) 노드로 선택한다. 각 씨앗 노드들의 정상·비정상 여부를 가린 뒤, 정상과 비정상 노드들에게 차별적으로 신뢰 점수를 전파하여 최종적인 각 노드의 랭크 점수를 통해 스팸 여부를 판별하는 알고리즘이다. 본 논문에서는 일반적으로 널리 쓰이고 있는 TrustRank 알고리즘을 다뤘다.

3.2 페이지 레벨

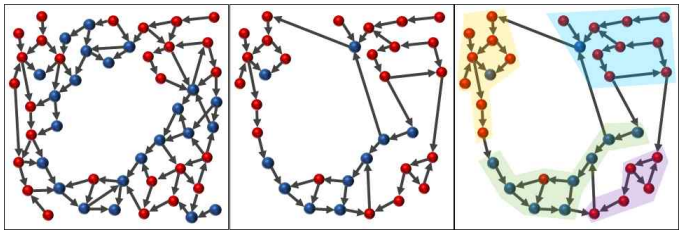
페이지는 웹 사이트를 구성하는 하나의 문서이다. 문서와 문서는 하이퍼링크를 이용하여 비순차적인 이동이 가능하다. 각각의 페이지가 노드가 되고, 각 페이지 간의 하이퍼링크가 방향성이 있는 간선이 되어 하나의 그래프 구조를 이룬다. 각 노드는 정상 노드와 스팸 노드로 구분된다. <그림1-(a)>와 같이, 노드의 수가 다른 레벨에 비해 상대적으로 많기 때문에 랭킹을 하는데 더 많은 시간이 소요 된다 (그림에서 붉은 색 노드는 스팸 페이지를 의미한다). 또한 스팸 페이지들은 스팸 페이지들끼리 다수의 링크를 주고받는 특징이 있다.

3.3 호스트 레벨

호스트란 인터넷을 통해 통신이 가능한 기기를 의미한다. 이러한 의미에서 각 호스트는 네트워크 구조에서 하나의 노드에 해당한다. 하나의 호스트가 여러 개의 페이지를 서비스 할 수 있으므로, 호스트 레벨에서의 분석은 페이지 레벨보다 적은 수의 노드가 존재하기에 랭킹을 하는데 더 적은 시간이 든다. 또한 하나의 호스트가 다수의 스팸 페이지를 서비스하는 특징이 존재한다. <그림 1-(b)>는 호스트 레벨에 대한 예시를 보여주며 붉은 색 노드는 스팸 호스트를 의미한다.

3.4 호스트 클러스터 레벨

그래프 상에서의 클러스터란 그래프 구조상에서 비슷한 특징을 공유하는 노드들 간의 응집력 있는 그룹을 의미한다. 호스트 노드들을 그래프 클러스터링 알고리즘[5]을 사용하여 클러스터링하게 되면, 호스트 레벨보다 한 단계 상위 레벨로 전체 네트워크의 구조를 보다 거시적인 관점에서 분석할 수 있게 된다. <그림1-(c)>는 호스트들을 클러스터링한 결과를 보여주며, 이러한 클러스터 레벨에서 스팸의 분포는 특정 클러스터에 집중되는 흥미로운 결과를 발견하였다.



<그림1-(a) 페이지> <그림 1-(b) 호스트> <그림 1-(c) 클러스터>

4. 실험 결과

4.1 실험 데이터

실제 검색 엔진 상에서의 웹 그래프 데이터를 제공받아서 사용하였다. 기본적인 데이터 요약 정보는 <표 1>과 같다.

<표 1 데이터 요약>

노드 수	간선 수	연결 요소 수	GCC 크기
1,009,982	5,328,876	630	1,006,090

*GCC (giant connected component)

4.2 노드 분류 (node classification)

<표 2>는 전체 웹 그래프에서 그래프 샘플링을 통해 일부 그래프를 추출하여, 각 노드들을 정상인 노드와 스팸인 노드로 분류한 결과다. 각 호스트 당 하나 이상의 스팸 페이지가 존재하면 스팸 호스트라고 정의하였다.

<표 2-(a) 페이지 노드 분류>

페이지 레벨	
정상	스팸
1,000,629 (99.07%)	9,353 (0.93%)

<표 2-(b) 호스트 노드 분류>

호스트 레벨	
정상	스팸
429,603 (99.95%)	209 (0.05%)

4.3 간선 분류 (edge classification)

4.3.1 페이지 레벨

<표 3-(a)>는 실험 데이터에서 페이지 레벨의 간선을 분류한 표이다. 정상노드와 스팸노드 간의 연결성의 비율을 분석하기 위해, 랜덤 모델을 생성하여 비교하였다. <표 3-(b)>는 페이지 간에 간선이 랜덤하게 생성되었을 경우의 기댓값을 나타낸다. 실제 간선의 비율과 랜덤 모델에서의 간선의 비율을 비교했을 때, 스팸노드에서 정상노드로 향하는 간선이 두드러지게 많다는 것을 알 수 있다.

<표 3-(a) 페이지 간선 분류>

	정상	스팸
정상	4,998,545 (93.80%)	9 (0.00%)
스팸	330,307 (6.20%)	15 (0.00%)

<표 3-(b) 기댓값>

	정상	스팸
정상	5,233,313 (98.21%)	48,649 (0.91%)
스팸	46,490 (0.87%)	423 (0.01%)

4.3.2 호스트 레벨

<표 4-(a)>는 4.3.1과 같이, 실험데이터의 호스트 레벨의 간선을 분류한 표이고, <표 4-(b)>는 랜덤 모델 상에서의 간선의 비율을 나타낸다. 페이지 레벨의 분석과는 달리 호스트 레벨에서는 스팸 호스트에서 정상 호스트로 향하는 간선의 비율이 기댓값보다 낮게 나타났다. 따라서 페이지 레벨과 호스트 레벨에서의 구조가 다른 양상을 보인다는 것을 유추하였으며, 본 논문에서 제시하는 다차원적 랭킹의 필요성을 뒷받침 해준다.

<표 4-(a) 호스트 간선 분류>

	정상	스팸
정상	1,006,374 (96.67%)	38 (0.00%)
스팸	34,662 (3.33%)	13 (0.00%)

<표 4-(b) 기댓값>

	정상	스팸
정상	713,795 (68.56%)	42,781 (4.11%)
스팸	269,641 (25.90%)	14,869 (1.43%)

4.4 군집화 (clustering)

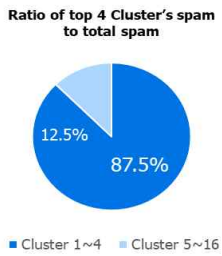
호스트 레벨의 노드들을 Graclus[5]를 이용하여 64개로 군집화 하였다. 각 클러스터의 크기를 고려하여, 클러스터 내의 스팸 개수에 대한 기댓값을 계산할 수 있다. 실제 스팸의 수와 기댓값을 비교하여 기댓값보다 많은 스팸의 수를 보유한 클러스터를 추려내니, 64개 중에 4개의 클러스터가 해당하였다. 그 결과는 <표 5>에 나타나 있으며, <그림 2>에서 볼 수 있듯이 전체 스팸 호스트 중 약 88%가 상위 4개의 군집에 속해 있다는 것을 알 수 있다.

4.5 스팸 탐지 성능평가

페이지 레벨과 호스트 레벨에서 TrustRank를 바탕으로 스팸 탐지를 실행한 결과, Recall 값이 각각 99.92% 99.84%로 거의 100%에 근사하게 나오는 것을 확인할 수 있었으며, 이 두 레벨을 결합한 효과를 확인하기 위해서는 스팸 비율이 조금 더 높은 그래프 샘플링 결과가 필요한 것으로 생각된다. 또한, 호스트 클러스터 레벨에서 64개의 클러스터 단위로 TrustRank를 구현한 결과, 4.4에서 제시했던 높은 비율의 스팸을 보유한 상위 4개의 클러스터들이 다른 클러스터보다 현저히 낮은 점수를 받는 것을 확인하였다. 따라서 클러스터 레벨의 분석결과를 반영하여 스팸 탐지 성능을 더욱 향상 시킬 수 있을 것으로 기대된다.

<표 5 군집화 요약 (64개)>

	군집 크기	스팸 수	기대 스팸 수
1	11,795	84	6
2	11,808	66	6
3	11,563	20	6
4	11,902	13	6



<그림 2 군집화 그래프 >

5. 결론 및 향후 연구

본 논문에서는 웹 그래프 상에서의 페이지 단위, 호스트 단위, 그리고 클러스터 단위에서의 웹 스팸의 특성을 분석하였으며, 각 레벨별로 스팸에 대한 두드러진 특징을 발견하였다. 이러한 분석을 통해 그래프의 구조를 다각적으로 고려한 다차원적 랭킹을 통한 스팸 탐지 프레임워크를 제시하였다. 향후 보다 구체화된 랭킹 메커니즘을 개발하고, 다양한 실제 데이터에 적용함으로써 보다 정밀한 성능 평가를 실행할 계획이다.

참 고 문 헌

[1] B. Wu and B. D. Davison, “Identifying link farm spam pages,” *WWW* poster, 2005.

[2] Z. Gyongyi et al., “Combating web spam with TrustRank,” *VLDB*, 2004.

[3] L. Page et al., “The PageRank citation ranking: Bringing order to the web,” Tech. Report, 1999.

[4] V. Krishnan and R. Raj, “Web spam detection with anti-trust rank,“ *AIRWeb*, Vol. 6, 2006.

[5] I. Dhillon et al., “Weighted Graph Cuts without Eigenvectors: A Multilevel Approach,” *PAMI*, 2007.