국내 웹 분석을 통한 웹 스팸의 특성

최승진° 김성권
중앙대학교 컴퓨터공학부
sjchoi@alg.cse.cau.ac.kr , skkim@cau.ac.kr

Characterization of Web Spam through the Korean Web Analysis

Seung-Jin Choi° Sung-Kwon Kim
School of Computer Science and Engineering, Chung-Ang University

요 약

웹 스팸(Web Spam)은 스패머가 원하는 페이지를 검색 결과 상단에 올리는 기술이다. 이러한 웹 스팸에 의해 상위 랭크된 페이지는 사용자에게 올바른 정보를 전달해 주지 않는다. 해외에서는 웹 스팸의 심각성을 인식하고 이에 대한 연구 또한 활발히 진행되고 있다. 하지만 국내의 경우 아직 웹 스팸에 대한 연구가 미흡한 실정이다. 또한 해외에서 연구되고 있는 웹 스팸 탐지 기술들은 국내의 웹에 적용시키기 힘들다. 그래서 본 논문은 다양한 방식으로 국내 웹과 검색 사이트의 특성을 분석하고 해외와의 차이점에 대해 알아본다. 그리고 이 차이점을 통해 국내 웹에서 나타날 수 있는 웹 스팸과 앞으로의 연구 방향에 도움을 주고자 한다.

1. 서 론

웹이 대중화 되면서 사람들은 보다 많은 정보를 얻으려 한다. 하지만 많은 정보들 중에서 사용자가 원하는 정보를 정확히 찾아내는 것은 매우 힘들다. 그래서 정보의 검색이 용이한 검색 엔진(Search Engine)이 웹으로 접근하는 주된 입구가 되었다. 또한 검색 엔진의 결과에 랭킹을 부여함으로써 원하는 정보를 보다 쉽고 빠르게 찾을 수 있게 되었다. 유명한 구글(Google)의 페이지랭크(RageRank) [3] 알고리즘이 페이지들 간의 링크 관계를 분석하여 검색 결과에 랭킹을 부여하는 방법중의 하나이다. 구글뿐만이 아니라 다른 검색 엔진들도 검색 엔진 결과에 랭킹을 부여하고 있다. 오늘날 정보 검색과 검색 엔진은 뗄 수 없는 관계에 놓여있다.

하지만 검색 엔진의 긍정적인 발전뿐만이 아닌 이를 악용하는 방법 또한 발전하였다. 웹 스팸(Web Spam) 혹은 검색 엔진 스팸(Search Engine Spam)이라고 하는데, 스패머가 원하는 페이지나 사이트를 검색 엔진 결과의 상단에 랭크 시키는 행위를 말한다. 예를 들면 사용자가 임의의 검색어로 검색 시, 검색 결과 상단에 위치한 페이지를 선택 하였을 때 검색어와 관계없는 내용 또는 광고를 볼 수 있다. 이러한 웹 스팸으로 인해 크게 두 가지 문제점이 있다. 첫 번째, 필요 없는 정보로 인해 검색어 처리 과정에 많은 비용이 든다. 두 번째는 검색 엔진 결과의 신뢰성이 떨어질 수 있다.

그래서 최근부터 웹 스팸 탐지에 관하여 많은 연구가 진행 중이다 [10]. 해외의 경우 웹 스팸을 효과적으로 탐지하는 연구가 활발히 진행 중이다.

그러나 국내의 연구는 아직 걸음마 단계이다. 국내에서는 아직 웹 스팸이라는 단어가 생소하다. 또한 웹 규모가 양적인 발전을 거듭할 뿐 질적인 면은 제자리걸음이다. 예를 들어 게시판이나 블로그 등에서 많이 나타나고 있는 댓글 스팸(Comment Spam) [11]의 경우 사용자의 신고나 운영자의 모니터링 같은 수작업을 통해 관리되고 있는 실정이다. 또한 국내의 웹은 해외와는 다른 특성을 지니고 있다. 그래서 기존의 기법들을 쉽게 적용시키지 못하고, 적용되더라도 제 성능을 발휘하지 못한다.

본 논문은 국내의 웹을 다양한 방법으로 분석하였다. 웹 페이지와 사이트, 유명 검색 엔진, 웹 사이트 연결성과 구조, 검색 결과 등의 문항을 분석 후 특징을 찾아냈다. 그리고 그것을 해외의 웹과 비교하여 차이점을 도출하였다. 이 차이점을 바탕으로 국내의 웹만이 가지고 있는 특징을 찾아냈다. 그리고 특징을 통해 실제 나타날 수 있는 국내의 웹 스팸에 대해서 알아본다. 이와 같은 분석을 통해 앞으로 국내의 웹 스팸과 검색 엔진 연구에 도움을 주고자 한다.

논문의 구성은 다음과 같다. 2장 관련 연구에서는 기존의 웹 스팸과 해외의 연구 동향을 살펴본다. 3장에서는 다양한 방법을 통해 국내 웹 페이지와 검색 엔진을 분석하여 해외의 웹과의 차이점과 특징을 도출한다. 4장에서는 국내의 웹에서 발생할 수 있는 웹 스팸에 대해 논하고 마지막으로 5장에서 본 논문의 결론을 내리고 향후 연구 방향에 대해 논한다.

2. 관련 연구

2.1 웹 스팸의 정의와 분류

웹 최대의 백과사전 위키피디아(http://wikipedia.org)는 웹 스팸(Web spam)을 다음과 같이 정의하고 있다.

"검색 엔진에 의해 인덱싱된 자원의 정확도나 중요도를 부당한 방법으로 속이는 모든 기법들"

웹을 사용하는 사람들은 다양한 웹 스팸 기법들을 많이 접하게 된다. 그리고 사용자들은 많은 불편을 겪고 있다. 하지만 아직까지 웹 스팸에 대한 정의나 기법의 분류가 완벽히 이루어지지 않고 있다. 예를 들면 웹 스팸의 경우 검색 엔진 스팸(Search Engine Spam)이나 스팸덱싱(Spamdexing)이라는 이름으로도 많이 알려져 있다.

이러한 웹 스팸 기법들을 특징에 따라 분류하는 연구도 진행 중이다. 검색 엔진 랭킹을 올리는 Boosting technique와 스패머가 원하는 내용을 감추는 Hiding Technique로 나누는 방법이 있고 [1], 웹 페이지들 간의 링크 기반과 내용 기반으로 분류하는 방법 또한 있다 [2]. 이것 외에도 하나의 기법이 실제 쓰이는 용도에 따라 분류하는 방법도 존재한다. 웹 스팸 분류 작업의 목적은 많은 웹 스팸 기법들을 충분히 이해하고 그에 맞는 효과적인 탐지 기법을 찾는데 도움이 되고자 하는데 있다.

2.2 해외의 웹 스팸 연구 동향

해외의 웹 스팸 연구는 2005년부터 본격적으로 이루어지기 시작했다. 초기에는 주로 링크 기반 스팸이나 내용 기반 스팸을 탐지하는 기법들이 소개되었다. 현재에는 다양한 웹 스팸 기법들을 효과적으로 탐지하는 방법들이 소개되고 있다. 예를 들면 블로그 내에 존재하는 웹 스팸을 탐지하는 기법들 [12], 리다이렉션이나 클로킹을 탐지하는 기법들 [7] 등이 있다. 또한 대기업에서도 웹 스팸의 심각성을 인식하고 이 연구에 많은 지원을 보내고 있다 [13].

현재 진행되고 있는 연구들을 분류하면 2가지로 볼 수 있다. 첫 번째로 검색 엔진 개선이다. 기존의 유명한 검색 엔진 알고리즘인 페이지랭크(PageRank) [3]나 HITS [4]를 응용하여 개선하는 방법이다. 이 연구는 검색 엔진의 결과가 웹 스팸에 영향을 받지 않고 정확성과 신뢰성을 향상시키는 것이 목표이다. 또한 이 연구는 모든 웹 스팸 기법을 고려하는 연구이다. 현재 발표된 기법들은 TrustRank [8], SpamRank [9]등이 있다. 두 번째로 특정 기법과 환경에 맞는 탐지 기법 연구이다. 예를 들면 링크 기반 스팸의 경우 페이지들 간의 링크를 분석하여 스팸을 탐지 한다 [5]. 그리고 내용 기반 스팸의 경우 키워드를 검사하거나 [6] 정상적인 것과 비교를 통해 탐지한다 [14]. 또는 블로그와 같은 특정 환경에서 발생하는 웹 스팸 [12]에 대해서만 탐지하는 기법이 존재한다.

3. 웹 페이지와 검색 사이트 분석

본론에서는 다양한 방법을 이용하여 국내 웹과 검색 사이트의 특징을 살펴본다. 그리고 해외와의 비교를 통해 차이점을 살펴본다. 마지막으로 도출된 특징을 기반으로 국내의 웹에서 발생할 수 있는 웹 스팸과 그것의 특징에 대해서 논한다.

3.1 크롤러를 이용한 웹 분석

본 절에서는 국내 웹의 특징을 알아보기 위해 프로그램을 이용한 분석을 시도하였다.

사용한 프로그램은 웹 크롤러(Web Crawler)로, 이것을 통해 얻어진 자료들을 바탕으로 특성을 도출하였다. 웹 크롤러란 웹에 있는 내용들을 자동으로 가져오는 스크립트 혹은 프로그램을 말한다. 크롤러를 통해 이미지와 비디오 사용량, 도메인의 종류, 페이지의 Status code에 대해서 알아보았다.

본 논문에서 사용한 웹 크롤러는 Heritrix Crawler (http://crawler.archive.org/)이다. 크롤러 실행 환경과 실행 결과는 아래의 [표 1]과 같다.

[표 1] 크롤러 실행 환경과 실행 결과

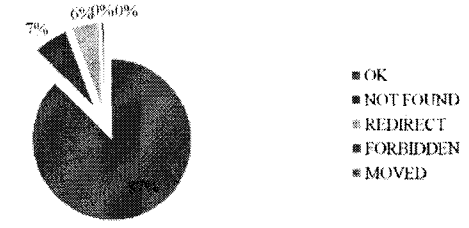
항목	내용
Total number of hosts crawled	1519
Total documents (files) crawled	73426
Total downloaded data size	1662110697 (1.5GB)
Time	3h 42m 30s
H/W	CPU Intel P4 3.2G
	Memory 256MB
S/W	Vmware 5.5.2
	Ubuntu 6.10
	Heritrix crawler 1.12.1

조사 대상은 2007년 7월 현재 상위 5개의 검색 사이트를 조사했다(http://www.top100.co.kr) 대상 검색 사이트는 네이버, 다음, 네이트, 야후코리아, 엠파스 이다. 첫 번째로 Status code를 알아보았다.

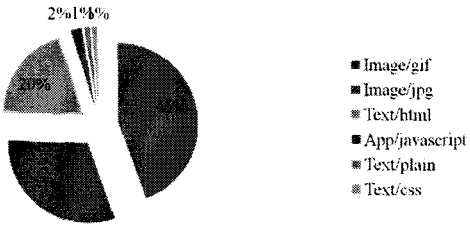
[그림 1]에서 중요한 점은 페이지의 리다이렉션이 6% 발생한다는 점이다. 해외와는 다르게 리다이렉션의 비중이 상대적으로 큰 편이다. 두 번째로 파일의 타입을 분류해 보았다.

[그림 2]에서 보듯 이미지 파일의 비중이 매우 높다. 그리고 단순 텍스트 파일의 비중이 적은 것을 볼 수 있다. 마지막으로 내부 도메인과 외부 도메인의 비율을 조사해 보았다.

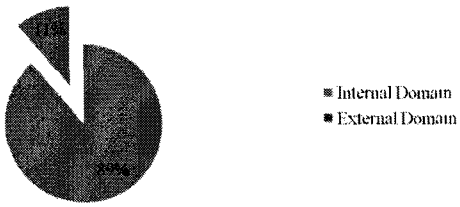
[그림 3]를 보면 조사 대상 사이트와 다른 도메인을 가진 외부 도메인의 비율이 10%정도이다. 대부분이 사이트 내부의 도메인을 사용하고 있는 페이지들이다. 크롤러를 이용한 웹 분석을 바탕으로 다음과 같은 결론을 얻었다.



[그림 1] Status Code 결과



[그림 2] File Type 결과



[그림 3] Domain 분류 결과

- 이미지 파일의 압도적인 양을 통해 국내 웹이 시각화 되고 있다는 것을 알 수 있다.
- 내부 도메인의 압도적인 비율은 사이트의 거대화를 의미한다.
- 외부 도메인의 매우 낮은 비율은 사이트 자체가 제작한 컨텐츠를 중심으로 운영되고 있음을 알려주고 있다.

3.2 검색 사이트 비교

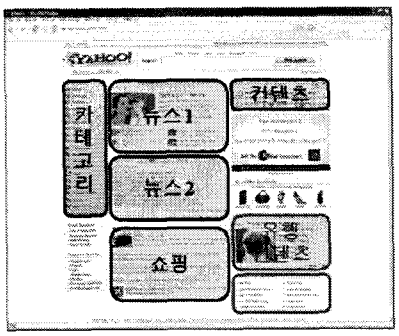
본 절에서는 실제 검색 사이트의 생김새와 특징, 그리고 검색 결과를 분석하였다. 분석한 대상은 국내와의 정확한 비교를 돕기 위해 한국과 미국의 야후 검색 사이트를 선정했다. 분석을 통해 국내만의 특징을 볼 수 있다.

비교 분석 대상은 앞에서 말한 것처럼 두 나라의 야후 사이트이다. 비교 문항은 사이트 첫 페이지 비교, 검색 결과를 보여주는 방식 비교이다.

우선 첫 번째 사이트 첫 페이지 비교이다. 웹 페이지에서는 다양한 내용들을 효과적으로 배치하는 것이 중요하다. 그리고 이용자의 성향과 필요를 정확히 판단하여 서비스 하는 것 또한 매우 중요한 작업 중의 하나이다. 양국 각각의 야후 사이트 주소를 입력 시 브라우저에 보이는 첫 페이지의 구조와 특징을 비교한다.

[그림 4]의 한국의 웹 페이지는 [그림 5]의 미국의 웹

[그림 4] 야후 한국의 첫 페이지 구성



[그림 5] 야후 미국의 첫 페이지 구성

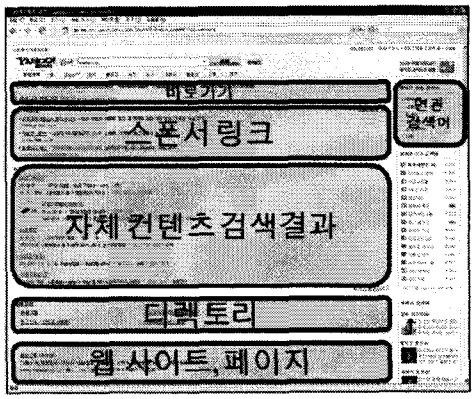
페이지와는 다르게 자사가 제작한 다양한 컨텐츠가 배치되어 있다. 또한 인기 검색어 파트가 있어 이용자들에게 시대의 유행이나 관심거리를 알려주고 있다. 이는 야후 뿐만이 아니라 국내에 있는 인기 검색 사이트들이 이러한 서비스를 제공하고 있다.

두 번째 비교 항목은 검색 결과를 보여주는 방식 비교이다. 이것 또한 앞의 비교와 마찬가지로 많은 차이를 보여주고 있다. 두 개의 사이트에서 검색어는 "Samsung"으로 입력 했다.

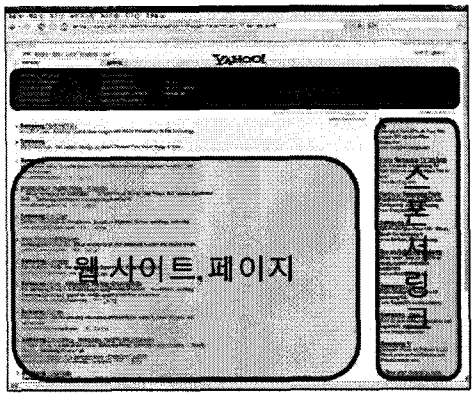
[그림 6]은 한국의 검색 결과를 보여주고 있다. [그림 7]의 야후 미국보다 다양한 검색 결과를 보여주고 있다. 두 사이트의 검색 결과의 공통점은 검색 결과 상단에 스폰서 링크가 배치되어 있는 것이다. 검색 사이트 사용자들은 검색 결과의 상단에 가장 큰 신뢰감을 가지고 있다 [15]. 이와 같은 배치는 순수한 검색 결과가 아닌 사이트의 이윤을 위한 인위적인 배치이다.

야후 한국을 포함한 국내 검색 사이트의 검색 결과의 특징은 3가지로 정리 해볼 수 있다. 우선 첫 번째, 자체 컨텐츠(블로그, 카페, 디렉토리) 검색 결과를 신뢰하여 상단에 위치시킨다. 두 번째, 해외에 비해 웹 페이지 검색에 소홀하다. [그림 6]에서 보듯 웹 페이지 검색은 제일 하단에 위치한다. 마지막으로 검색 결과에서 각각의 카테고리(바로가기, 스폰서링크, 자체 컨텐츠)는 고정된 노출 순서를 가지고 있다.

검색 사이트 첫 페이지와 검색 결과 보여주기 방식 비교를 통해 다음과 같은 결론을 도출하였다.



[그림 6] 야후 한국의 검색 결과



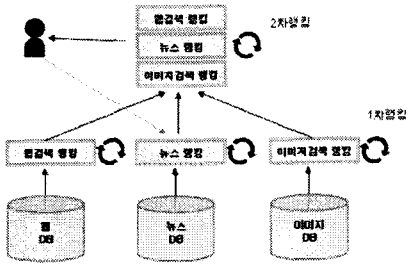
[그림 7] 야후 미국의 검색 결과

- 하나의 검색 사이트가 날씨, 뉴스, 카페, 블로그, 쇼핑, 동영상 등 다양한 컨텐츠를 서비스하고 있다.
- 블로그, 멀티미디어, 실시간 검색어를 통해 대중의 관심사나 유행을 보여주고 있다.
- 웹 페이지 검색보다 스폰서 링크, 자신의 사이트가 보유하고 있는 데이터베이스 검색 결과를 상단에 배치하여 인위적인 검색 결과를 보여준다.

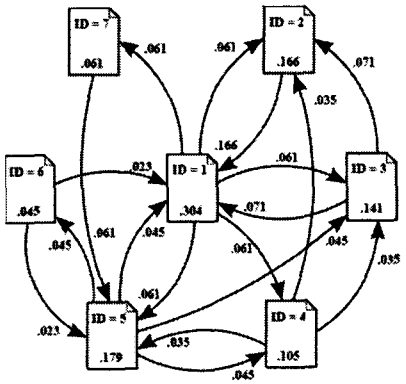
3.3 네이버 vs 구글

웹 스팸은 검색 엔진의 발달에 따른 부작용으로 생긴 것이다. 그래서 검색 엔진의 동작 방식과 웹 스팸 기술은 뗄 수 없는 관계이다. 앞의 3.2절과 본 절의 분석 목표는 검색 사이트의 특징을 도출하여 그 특징에 적합한 웹 스팸에 연구 방향에 대해 논하려 한다. 이 내용은 4장에서 다루겠다.

본 절은 3절의 마지막 분석으로 국내와 해외에서 가장 많이 사용되고 인기 있는 검색 사이트인 네이버와 구글을 비교한다. 비교 문항은 검색 알고리즘의 특징, 검색 결과 제공 방식이다



[그림 8] 네이버의 검색 랭킹 알고리즘
http://story.nhncorp.com/story.nhn?story_id=27



[그림 9] 구글의 검색 랭킹 알고리즘
http://www.rxiao.com/collection/doc/pagerank_cn/

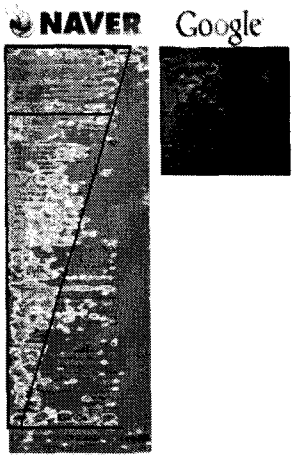
첫 번째로 각 사이트의 검색 알고리즘을 알아보겠다. 우선 네이버의 검색 알고리즘 동작 방식은 [그림 8]와 같다.

네이버는 멀티 랭킹 시스템(Multi-Ranking System)이라는 알고리즘을 사용하고 있다. 웹, 이미지, 뉴스 등의 콜렉션(Collection)에 각각 다른 랭킹 모델을 적용한다. 그 이유는 각각의 콜렉션마다 다른 특성을 지니고 있기 때문이다. 예를 들어 뉴스의 경우 최신성을 고려해야 하고 웹의 경우 정확도를 고려해야 한다. 각각의 콜렉션이 가지는 랭킹 모델을 적용 후 다시 통합하여 2차 랭킹을 계산한다. 이렇게 함으로써 이용자들에게 중복되지 않고 좀 더 정확한 검색 결과를 보여줄 수 있다.

네이버의 검색 엔진에서 가장 특이한 점은 웹 페이지의 검색 능력이 떨어지는 것이다. 또한 웹 페이지 검색에 그리 신경을 쓰지 않는다. 이는 웹 검색 보다 자신의 사이트가 보유하고 있는 데이터베이스 검색을 더 신뢰하고 있다고 볼 수 있다.

구글은 페이지랭크(PageRank) [3] 알고리즘을 이용하여 검색 결과를 보여준다.

[그림 9]에서 보면 자신에게 들어오는 링크의 개수와 들어오는 링크의 가중치를 기준으로 모든 웹 페이지들의 점수를 부여한다. 이 점수를 기준으로 검색 결과의 랭킹이 만들어진다. 구글 검색 엔진의 특징은 모든 작업이 컴퓨터의 계산에 의해 이루어진다. 사람의 힘이 들어가는 부분은 스폰서링크를 제외하고는 없다. 이는 사용자



[그림 10] 네이버와 구글의 Eye Tracking
www.joonj.com/wordpress/archives/308/feed/

들로 하여금 검색 결과에 대해 신뢰성을 줄 수 있다.

두 검색 엔진 알고리즘의 가장 큰 차이점은 웹 페이지 검색이다. 네이버는 웹의 내용보다 자신들이 만든 인공적인 데이터를 더 신뢰한다. 하지만 구글은 웹에 존재하는 수많은 웹 페이지들 중에서 사용자가 원하는 정보를 알려주려 한다. 사용자의 입장에서 네이버의 경우가 좀 더 정보를 쉽고 정확하게 찾을 수 있는 장점이 있다. 구글도 네이버처럼 다양한 검색 결과를 보여주기 위해 변화를 주려하고 있다 [16].

이용자들에게 검색 결과를 보여주는 방식은 두 가지가 있다. 구글처럼 하나의 내용을 보여주는 수평 검색(Horizontal Search)과 네이버처럼 다양한 내용을 한꺼번에 보여주는 수직 검색(Vertical Search)이 있다. 이 두 가지 방식을 비교하여 특징과 장단점을 살펴보겠다.

수평 검색은 검색어와 맞는 웹 문서를 보여주는 방식이다. 이때의 웹 문서는 웹 크롤러가 찾은 웹 문서 내에서 검색이 된다. 수직 검색은 다양하고 세부적인 정보들을 보여준다. 3.2절의 야후 한국의 검색 결과와 본 절의 네이버 검색 결과가 수직 검색의 예이다. 이 방식은 웹 문서 뿐만이 아니라 서비스 제공자가 만든 데이터베이스에서도 정보를 탐색하여 보여준다.

두 방식의 가장 큰 차이점은 사용자의 검색 결과에 대한 기대치이다. [그림 10]은 네이버와 구글의 검색 결과를 보는 사용자의 시선을 표시한 것이다. 네이버는 다양한 타입의 검색 결과를 보여준다. 그래서 사용자는 검색 결과 상단부터 하단까지 원하는 정보를 기대할 수 있다. 구글의 경우 모든 검색 결과를 줄 세우는 랭킹 기반이다. 그래서 검색 결과 상단과 하단의 정확도가 많은 차이를 가진다.

두 가지 비교를 하여 네이버와 구글의 차이점과 각각의 장단점에 대하여 살펴보았다. 두 사이트 모두 정보를 검색하는 사이트이지만 추구하는 바가 다르다는 것을 알 수 있었다. 분석을 통해 얻은 네이버와 구글의 특징은 [표 2]와 같다. 그리고 [표 2]의 네이버의 특징은 국내 대부분의 검색 사이트가 가지고 있는 특징이다.

[표 2] 네이버와 구글의 특징 비교

네이버	구글
컨텐츠 중심 검색	웹 중심 검색
인간의 수작업	극도의 컴퓨팅 정렬
대중을 위한 검색 포털	검색을 위한 검색 포털
수직 검색 결과	수평 검색 결과
인위적인 검색 결과	계산에 의한 검색 결과

4. 국내의 웹 스팸

앞서 3절에서 다양한 분석을 통해 여러 특징을 살펴볼 수 있었다. 특성들 중 웹 스팸과 관련된 것들은 다음과 같다.

- 사이트의 거대화
- 자체 컨텐츠 중심의 운영과 검색
- 대중의 관심과 유행을 포착

사이트의 거대화는 하나의 사이트가 다양한 서비스를 제공한다. 다양한 서비스를 제공함으로써 이용자들을 다른 사이트로의 이동 막는 효과가 있다. 그래서 점점 거대 사이트에 이용자들이 몰리게 된다. 스패머들은 거대 사이트에서 많은 이용자들을 대상으로 손쉽게 웹 스패밍을 할 수 있다.

그리고 두 번째 특징인 자체 컨텐츠 중심의 운영과 검색은 이용자에게 검색 시의 편의를 제공한다. 현재 인기 있는 컨텐츠들은 동영상, 이미지, 뉴스, 날씨, 블로그, 카페 등이다. 이 특징 또한 첫 번째 특징과 마찬가지로 스패머들에게 편의를 제공한다. 예를 들면 복잡한 웹 스팸 기법을 사용하지 않고 유명한 블로그나 카페에 들어가서 광고 댓글을 남길 수 있다. 그리고 검색 결과가 서비스 제공자가 만든 인위적인 것들이기 때문에 이용자들에게는 검색의 편의성을 제공하지만 내용의 신뢰성은 보장할 수 없다.

마지막으로, 인기 검색어 등을 통해 대중의 관심과 유행을 포착하는 것은 스패머들이 원하는 내용을 감추어 보여줄 수 있는 도구로 사용될 수 있다. 예를 들면, 스패머는 자신이 광고하고자 하는 페이지에 인기 검색어와 관련된 내용을 삽입하여 이용자들을 유혹할 수 있다. 이런 방식은 국내 검색 사이트에서 종종 볼 수 있는 방식이다. 인기 검색어 관련 동영상이나 이미지, 블로그에 광고 덧글을 작성하거나 광고 페이지로 리다이렉션을 시키는 것들이 있다.

검색 내용 측면을 보면 국내는 웹 페이지 검색에 소홀한 경향이 있다. 대신 서비스 제공자의 데이터베이스나 블로그, 카페 검색을 중요하게 생각한다. 이러한 차이점 때문에 해외에서 연구되고 있는 웹 스팸 탐지 기법들을 적용시키기 힘들다. 국내의 웹 스팸은 웹 페이지에서의 스패밍보다 블로그, 이미지, 동영상, 특정 컨텐츠에서 주로 발생하고 있다.

그래서 국내의 웹 스팸 연구 방향은 두 가지로 나누어 생각할 수 있다. 우선 첫 번째, 국내 검색 엔진을 개선하여 웹 스팸을 필터링하거나 랭킹 알고리즘을 개선하는 새로운 기법을 개발하는 것이다. 그리고 두 번째로 이미지, 비디오, 카페, 블로그 등의 각 컨텐츠에는 그것의 특징을 이용한 웹 스팸이 존재한다. 이것들을 각각의 특징을 이용하여 탐지하는 기법을 개발하는 것이다.

5. 결 론

검색 엔진의 발달에 따른 부작용으로 인해 웹 스팸이라는 새로운 기술이 발생했다. 웹 스팸은 검색 결과의 상단에 스패머가 원하는 페이지를 위치시키는 행동이나 기술이다. 이로 인한 웹 이용자들의 불편은 커져 가고 있다. 그래서 해외에서는 웹 스팸의 심각성을 인식하여 활발한 연구를 진행 중이다. 그러나 국내의 웹 스팸 연구는 아직 걸음마 단계이다. 또한 국내의 웹이 가지는 특성 때문에 해외의 기술 적용이 힘들다. 그래서 국내 웹과 검색 사이트를 다양한 방법을 통해 비교, 분석하여 특징을 도출하였다. 사이트의 거대화, 컨텐츠 중심의 운영과 검색, 인위적인 검색 결과, 이용자의 관심과 유행을 포착하려는 특징을 발견하였다. 각 특징마다 일어날 수 있는 웹 스팸에 대해 논하였고 실제 어떻게 발생되고 있는지 예를 들어 설명하였다. 그리고 국내의 웹 스팸 연구 방향을 두 가지로 나누어 보았다.

- 검색 엔진 단계에서 웹 스팸 필터링이나 웹 스팸에 영향 받지 않는 랭킹 알고리즘 개발
- 이미지, 비디오, 카페, 블로그 등의 컨텐츠에 나타나는 웹 스팸들을 탐지하는 기법 개발

본 논문은 블로그 내에 존재하는 웹 스팸을 효과적으로 검출하기 위한 프로젝트의 조사 부분을 정리, 보완하여 작성된 것이다. 그리고 본 논문을 통해 국내에 웹 스팸의 심각성과 연구의 필요성을 각인시키고자 한다. 그리고 실제 연구를 시작하려는 이들에게 도움을 줄 것이다.

참 고 문 헌

[1] Z. Gyongyi, H. Garcia-Molina. "Web Spam Taxonomy". 1st International Workshop on Adversarial Information Retrieval on the Web. May. 2005.

[2] Alan. Perkins. "The Classification of Search Engine Spam". http://www.the-indexing-company.org/whitepapers/spam-classification/. September. 2001.

[3] L. Page, S. Brin, R. Motwani and T. Winograd. "The PageRank Citation Ranking: Bringing order to the web". Technical report. Stanford University. 1998.

[4] JM. Kleinberg. "Authoritative Sources in a Hyperlinked Environment". Journal of the ACM, 1999.

[5] B. Wu, B. Davison. "Identifying Link Farm Spam Pages". International World Wide Web Conference, 2005

[6] A. Ntoulas, M. Najork, M. Manasse, D. Fetterly. "Detecting Spam Web Pages through Content Analysis". Proceedings of the 15th International conference on World Wide Web. 2005.

[7] B. Wu, B. Davison. "Cloaking and Redirection: A Preliminary Study". Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web. 2005.

[8] Z. Gyongyi, H. Garcia-Molina, J. pedersen. "Combating Web Spam with TrustRank". Proceedings of the 30th International Conference on Very Large Data Base, 2004.

[9] AA. Benczur, K. Csalogany, T. Sarlos, M. Uher. "SpamRank-Fully Automatic Link Spam Detection Work in Progress". Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web. 2005.

[10] 과학기술정보 글로벌동향브리핑 http://www.yeskisti.net/yesKISTI/Briefing/Trends/View.jsp?ct=TREND&clcd=J3&clk=J&lp=SI&cn=GTB2007050397

[11] G. Mishne, D.Carmel, R. Lempel. "Blocking Blog Spam with Language Model Disagreement". Proceedings of the First International Workshop on International Workshop on Adversarial Information Retrieval on the Web. 2005.

[12] P. Kolari, A. Java, T. Finin. "Characterizing the Splogosphere". Proceedings of the 3rd Annual workshop on World Wide Web. May. 2006.

[13] 과학기술정보 글로벌동향브리핑 http://www.yeskisti.net/yesKISTI/Briefing/Trends/View.jsp?ct=TREND&clcd=J5&clk=J&lp=SI&cn=GTB2006060150

[14] D. Fetterly, M. Manasse, M. Najork. "Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages". Proceedings of the 7th International Workshop on the Web and Databases. June. 2004.

[15] Eyetracking Research http://www.eyetools.com/inpage/research_google_eyetracking_heatmap.htm

[16] Google's Universal Search http://googleblog.blogspot.com/2007/05/universal-search-best-answer-is-still.html