URL 리다이렉션 스팸 탐지 기법

백지현°, 김성권
중앙대학교 컴퓨터공학과
jhbaek@alg.cse.cau.ac.kr°, skkim@cau.ac.kr

Detecting Method for URL Redirection Spam

Jee-Hyun Baek°, Sung Kwon Kim
School of Computer Science & Engineering, Chung-Ang University, Seoul, Korea

요 약

인터넷의 급속한 성장은 사람들의 정보 습득 방식에 큰 변화를 주었다. 인터넷 이용자들은 과거와 비교도 할 수 없을 만큼의 많은 지식을 손쉽게 접할 수 있게 되었다. 하지만, 그로 인해 여러 가지 문제점들이 생겨나게 됐는데, 웹 스팸도 그 중 하나이다. 웹 스팸은 웹을 통한 불법적인 활동으로 이득을 보려는 활동을 통칭할 수 있다. 웹 스팸은 검색 엔진 결과 리스트의 순위를 올리기 위해 사용되는 것이 대부분이지만, 점점 검색 엔진 결과 리스트의 순위와 관련 없는 것들에서도 나타 생겨나고 있다. 웹 스팸은 종류도 다양할뿐더러, 아직까지 모든 웹 스팸을 예방할 확실한 방법이 제시되지 못하고 있다.

이 논문에서는 여러 웹 스팸 중 페이지-하이딩 스팸에 속하는 URL 리다이렉션에 대해 다루고자 한다. 다른 웹 스팸과 마찬가지로, 현재까지 자동적으로 URL 리다이렉션을 탐지하는 방법이 제시되지 못하고 있는 실정이다. 이 논문에서는 검색 엔진 결과 리스트의 순위를 사용하여 URL 리다이렉션의 탐지 기법을 제안하고자 한다.

1. 서 론

인터넷의 급속한 성장은 과거 사람들이 정보를 접하는 방식을 크게 변화시켰다. 과거의 사람들에게는 책이 정보를 습득하는 주된 수단이었지만, 현재의 사람들은 과거보다 훨씬 더 많은 정보를 인터넷을 통해 습득한다.

인터넷의 정보가 크게 늘어감에 따라 어떻게 인터넷에서 필요한 정보를 쉽게 찾을 수 있을 것인가가 중요한 문제가 되었다. 초기의 인터넷 이용자들을 그들이 원하는 정보에 접근하기 위해 브라우저의 '즐겨찾기'를 사용하였다. 하지만, 브라우저의 '즐겨찾기' 만으로 늘어나는 정보들을 관리하는 것은 결코 쉬운 일이 아니다. 이러한 이유 때문에 쉽게 정보를 찾을 수 있는 검색 엔진이 개발되게 되었다.

검색 엔진의 사용이 증가할수록, 검색 엔진은 인터넷 사용자들에게 점점 더 많은 영향력을 끼치게 되었다. 이는 상업적인 웹 사이트를 운영하는 사람들에게 특히 중요한 이슈였다. 상업적인 웹 사이트를 운영하는 사람들은 그들의 회사의 발전을 위해 많은 사람들에게 자신들의 웹 사이트를 보여야만 했다. 이로 인해, 상업적인 웹 사이트는 검색 엔진에 종속적으로 운영될 수밖에 없는 처지가 되었다.

많은 인터넷 이용자들은 검색 엔진이 만들어낸 전체 결과 리스트를 보지 않는다. 그들은 단지 결과 리스트의 상위 몇몇 곳만 살펴볼 뿐이다. 이 때문에, 상업적인 웹 사이트의 운영자들에게는 검색 엔진 리스트에서 좀 더 높은 순위에 오르는 것이 중요하게 되었다.

이러한 이유로 인해, 전문적으로 검색 엔진의 순위를 올리는 SEO들이 생겨나게 되었다. SEO(Search Engine Optimizer)들은 웹 사이트 개발자들이 잘 구조화되고, 주제에 맞게 키워드가 달린 컨텐츠를 개발할 수 있도록 도움으로서, 검색 엔진의 순위를 높여준다. 하지만, 모든 SEO들이 이런 합법적인 방법을 사용하는 것은 아니다. 불법적인 방법을 사용하는 SEO들은 웹 스팸(Web Spam)을 사용하여 검색 엔진의 순위를 높이려고 한다 [1].

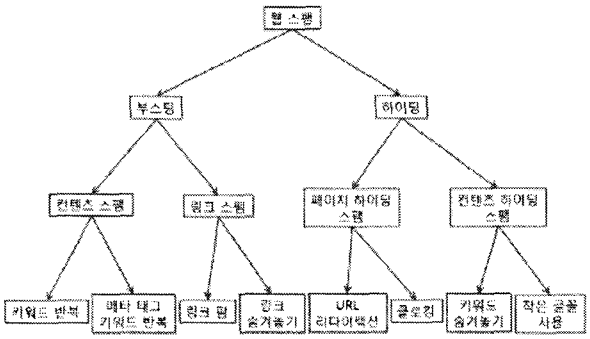
현재까지 다양한 종류의 웹 스팸이 알려져 있으며, 계속해서 새로운 웹 스팸이 개발되고 있다. 그 중 이 논문에서는 페이지-하이딩 스팸에 속하는 URL 리다이렉션 스팸을 소개하고, 그 방지 기법을 제안하고자 한다.

논문의 나머지는 다음과 같다. 2장은 관련 연구로, 웹 스팸에 대한 소개와 종류, URL 리다이렉션에 대한 소개를 담고 있다. 3장은 제안 알고리즘을 소개하며, 4장은 향후 연구 과제를 다룬다. 5장은 결론이다.

2. 관련 연구

2.1 웹 스팸(Web Spam)

웹 스팸은 스팸덱싱(Spamdexing)이나 검색 엔진 스팸(Search Engine Spam) 등으로도 불리며, 검색 엔진의 랭킹을 임의로 조작하려는 행위를 일컫는다. 하지만, 인



[그림 1] 웹 스팸 분류

터넷의 활용이 증가함에 따라 현재는 웹을 통한 다양한 불법적인 행위로 그 범위가 넓어지고 있다.

웹 스팸의 판단은 상당히 애매모호한 작업이다. 상당수의 경우, 웹 스팸인지 아닌지의 판단에는 개인적인 차이를 보인다. 예를 들어, 어떤 페이지는 형편없는 정보를 담고 있지만, 검색 엔진이 좋아하는 형식으로 구성되어 있기 때문에 웹 스팸으로 분류되지 않는다. 이에 반해, 다른 페이지는 아주 훌륭한 정보를 담고 있지만, 검색 엔진이 싫어하는 형식으로 제작되었기 때문에 웹 스팸으로 판단될 수 있다[2].

2.2 웹 스팸의 분류

웹 스팸은 [그림 1]과 같이 크게 부스팅(boosting) 영역과 하이딩(hiding) 영역으로 나뉠 수 있다. 부스팅 영역은 컨텐츠 스팸, 링크 스팸으로 구성되며, 하이딩 영역은 페이지-하이딩 스팸과 컨텐츠-하이딩 스팸으로 되어 있다.[3, 4, 5]

2.2.1 컨텐츠 스팸(content spam)

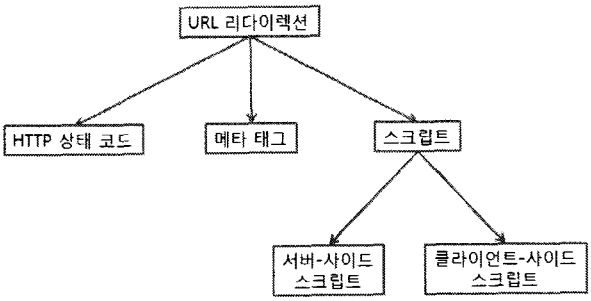
컨텐츠 스팸은 웹 문서 안에서 겉으로 드러나는 문서를 이용한 스팸이다. 웹 페이지의 연관도를 높이기 위해 같은 단어를 반복적으로 사용한 키워드 반복(Keyword Stuffing), 메타 태그 내에 반복적으로 키워드를 사용하는 메타 태그 키워드 반복(Meta Tag Stuffing) 등이 여기에 속한다.

2.2.2 링크 스팸(link spam)

링크 스팸은 구글(Google)의 페이지랭크와 같은 링크 기반의 순위 결정 알고리즘을 사용하는 검색 엔진에 적용되는 스팸이다. 이에 대표적인 방법이 링크 팜(Link Farm)이라 불리는 스팸 기법인데, 이것은 밀접하게 연결된 페이지들 간의 상호 참조에 의해 대상이 되는 웹 사이트의 순위를 임의로 상승시킨다.

2.2.3 페이지-하이딩 스팸(Page-Hiding Spam)

페이지-하이딩 스팸(Page-Hiding Spam)은 검색 엔진 결과 리스트를 통해 접근한 사용자에게 검색 엔진과는



[그림 2] URL 리다이렉션 분류

전혀 다른 페이지를 보여주는 스팸 기법이다. URL 리다이렉션(URL Redirection)이나 클로킹(Cloaking)은 가장 일반적인 페이지-하이딩 스팸들 중의 하나이다.

2.2.4 컨텐츠-하이딩 스팸(Content-Hiding Spam)

컨텐츠-하이딩 스팸은 사용자로부터 웹 문서 내의 컨텐츠를 보이지 않게 숨겨 이익을 취하려하는 스팸 기법이다. 여기에는 웹 문서의 글자의 색을 배경색과 같게 하여, 사람이 쉽게 볼 수 없게 만드는 키워드 숨겨놓기 기법(Invisible Text), 글자를 눈에 쉽게 띄지 않는 작은 글꼴을 사용하여 보여주는 작은 글꼴 사용 기법(Tiny Text) 등이 속한다.

2.3 URL 리다이렉션 스팸

앞서 2.1에서 언급한 것처럼, URL 리다이렉션 스팸은 페이지-하이딩 스팸 중의 하나이다. URL 리다이렉션 스팸은 웹 크롤러에게 인덱싱된 페이지를 웹 브라우저를 통해 볼 때, 전에 전혀 다른 페이지로 자동으로 옮겨가도록 만든 웹 스팸이다. 웹 페이지의 리다이렉션은 페이지가 로딩된 시점에 발생하거나 타이머에 의해서 혹은, 마우스 이동과 같은 사용자 이벤트에 의해 발생될 수 있다.[6]

[그림 2]에서처럼 오늘날의 웹 브라우저에서는 3가지 방식으로 URL 리다이렉션이 발생한다. 그것들 각각은 HTTP 상태 코드에 따른 리다이렉션, 메타 태그에 따른 리다이렉션, 그리고 스크립트에 따른 리다이렉션이다.

2.3.1 HTTP 상태 코드(HTTP Status Code)에 의한 리다이렉션

HTTP 상태 코드는 서버가 요구 메시지를 수신하여 처리한 결과를 알려주는 3자리 정수로 된 처리 결과 번호이다. 각 정수 값의 첫 자리 숫자는 상태 유형을 나타내는데, 아래와 같이 리다이렉션은 3으로 시작하는 상태 코드 값들이다.[7]

① 300: Multiple Choices
② 301: Moved Permanently
③ 302: Found, Redirect
④ 303: See Other, Redirect Method
⑤ 307: Temporary Redirect

URL 리다이렉션의 목적지는 HTTP 응답에서 Location 헤더에 해당하는 부분이다.

```
HTTP/1.1 301 Moved Permanently
Date: Fri, 31 Aug 2007 05:13:03 GMT
Server: Apache/2.0.46 (Red Hat)
Location: http://www.qsl.net/ds2dma/
Vary: Accept-Encoding
Content-Length: 314
Content-Type: text/html; charset=iso-8859-1
```

HTTP 상태 코드에 의한 리다이렉션은 실제 페이지를 다운로드 받기 전에 이루어지며, 가장 효율적인 리다이렉션 방법이라 할 수 있다.

2.3.2 메타 태그(META Refresh Tag)에 의한 리다이렉션

다음과 같은 메타 태그를 사용하는 웹 페이지는 사용자를 다른 페이지로 옮길 수 있다.

```
<meta http-equiv = "refresh" content = "0; url =
http://www.destination.com/">
```

리다이렉션을 위한 메타 태그는 http-equiv와 content라는 2개의 속성을 갖는다. http-equiv는 refresh로 설정되는데, 이는 이 페이지가 리다이렉션됨을 나타낸다. content는 대개 세미콜론으로 구분된 2개의 값을 갖는데, 첫 번째 부분은 리다이렉션이 발생하기까지의 대기시간을, 두 번째 부분은 이동하게 되는 대상 페이지를 나타낸다.

메타 태그에 의한 리다이렉션은 페이지의 일부분만을 받은 상황에서 발생하는데, 대개는 헤더 부분이 해당된다.

2.3.3 스크립트(Script)에 의한 리다이렉션

스크립트에 의한 리다이렉션은 서버-사이드 스크립트(Server-Side Script) 리다이렉션과 클라이언트-사이드 스크립트(Server-Side Script) 리다이렉션이 있다. 스크립트에 의한 리다이렉션은 여러 이벤트 상황 하에서 실행될 수 있기 때문에, 다양한 방법으로 활용이 가능하다.

```
Document: onload, onunload, onchange, onsubmit
          onreset, onblur, onfocus
Keyboard: onkeydown, onkeypress, onkeyup
Mouse:    onclick, ondblclick, onmousemove,
          onmousedown, onmouseover, onmouseout,
          onmouseup
```

앞서 2가지 방법의 리다이렉션과 달리 이 방법은 전체 페이지가 다운로드된 뒤에 실행하게 된다.

2.4 리다이렉션의 사용

모든 리다이렉션이 불법적으로 동작하는 건 아니다. 몇몇 합법적인 리다이렉션이 존재하는데, 그 중 대표적인 것이 트래픽을 분산시키기 위해 사용하는 경우이다. 방문객이 많은 대부분의 웹 사이트는 하나의 도메인에 여러 IP가 속해 있다. 인터넷 이용자가 이 사이트에 접근하게 되면, 웹 서버는 적절한 IP로 트래픽을 분산시킨다.

리다이렉션은 도메인 포워딩을 위해 사용되거나 대문 페이지를 구성하기도 한다. 긴 이름을 가지는 웹 사이트들은 짧은 이름을 가지는 웹 사이트에 비해 상업적으로 불리하다. 인터넷 이용자들은 긴 이름보다 짧은 이름을 가지는 웹 사이트를 더 선호한다. 그렇기 때문에, 짧은 이름으로의 도메인 포워딩은 긴 이름으로 인한 약점을 제거시킬 수 있다. 또, 오타로 웹 사이트에 접근하지 못하는 경우를 방지할 수 있다. 예를 들어, 구글의 정식 도메인은 http://www.google.com 이지만, http://www.goooogle.com 의 잘못된 입력으로도 구글 웹 사이트에 접근할 수 있다. 대문 페이지의 경우, 영화 홍보 사이트처럼 그들의 웹 사이트에 접근하는 인터넷 이용자들의 흥미를 유발시킬 수 있다.

하지만, 이 두 가지 리다이렉션의 사용은 불법적으로 사용될 가능성이 아주 높다. 현재로선, 검색 엔진들은 명확하게 불법적인 리다이렉션인지 아닌지를 구별하지 못한다. 그래서 모든 리다이렉션 페이지에 대한 페널티를 부과하고 있는 실정이다.

2.5 웹 스팸 탐지 기법

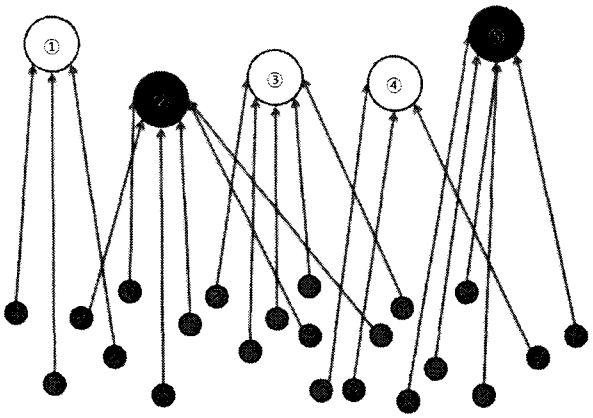
통계적 방법이나 기계 학습을 통한 다양한 웹 스팸 탐지 기법들이 소개되었지만, 어느 것도 모든 웹 스팸에 대해 확실한 성공을 보장하지는 못한다. 이는 시간이 흐를수록 다양한 웹 스팸이 소개되고 있기 때문이다.

이런 상황 속에서, 웹 스팸은 스팸 메일이나 컴퓨터 바이러스와 같은 다른 일반적인 불법적인 문제들보다 덜 심각해 보인다. 하지만, 가까운 미래에는 이러한 것들보다 더욱 문제의 심각성이 커지게 될 것으로 예상되고 있다.[2]

3. 제안 알고리즘

이 논문에서 제안하는 탐지 기법의 모티브는 구글의 순위 결정 알고리즘인 페이지랭크 알고리즘의 모티브와 유사하다. 또한, 검색 엔진 결과 리스트의 순위를 부가적으로 사용한다.

[그림 3]과 같이 URL 리다이렉션을 사용해서 사이트의 순위를 높이려는 SEO들은 다수의 페이지에서 리다이렉션을 사용할 것이다. 이는, 좀 더 많이 검색 엔진에 노출시키기 위해서이다. 그에 반해, 일반적인 개인 웹 페이



[그림 3] URL 리다이렉션 페이지들 간의 관계

지 관리자들은 단지 1~2개 정도만의 리다이렉션을 사용하는데 그칠 뿐이다. 이는 리다이렉션 스팸을 탐지하는데 큰 도움이 된다. [그림 3]에서 ①, ③, ④번은 합법적인 리다이렉션인 경우를, ②, ⑤번은 리다이렉션 스팸을 나타낸다.

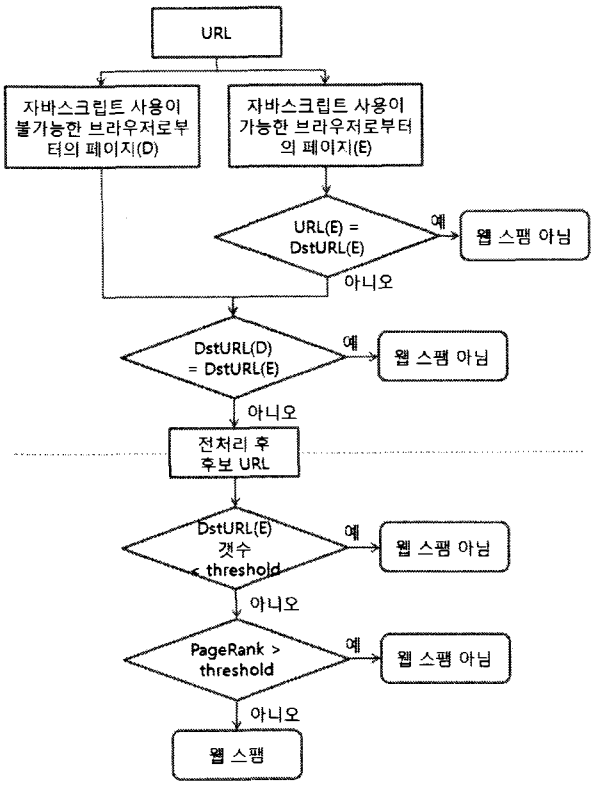
하지만, [그림 3]의 ③번처럼 리다이렉션이 많다고 모두 웹 스팸으로 판단할 수 없다. 합법적인 사이트들도 다수의 리다이렉션을 가지는 경우가 있기 때문이다. 이를 해결하기 위해 검색 엔진의 순위를 사용한다. 이 논문에서는 구글의 페이지랭크 알고리즘[8]을 사용하여 순위를 매긴다.

하나의 페이지로 이동하는 다수의 리다이렉션 페이지가 있을 때, 이 페이지의 순위가 낮다면 이것은 다른 페이지보다 웹 스팸일 확률이 높다. 목적지 페이지(DstURL)를 조사하여 순위가 낮은 페이지를 찾고, 이 페이지로 이동하는 원시 페이지를 조사한다. 이러한 과정을 거쳐 선택된 페이지들을 URL 리다이렉션 스팸으로 판단한다.

3.1 알고리즘

이 제안 기법은 크게 2단계로 구성되어 있다. 첫 번째는 전처리 단계로, 실제 리다이렉션이 발생하는 웹 사이트를 추출하는 것이다.

[그림 4]에서처럼 2가지의 페이지를 초기 데이터로 취한다. 이 페이지들은 동일한 URL의 페이지이지만, 하나는 자바스크립트가 동작하는 브라우저에서 다운로드받고, 다른 하나는 자바스크립트가 동작하지 않는 브라우저에서 다운로드받는다. 우선 전자에 해당하는 페이지의 실제 페이지와 목적지 페이지가 동일할 경우 리다이렉션이 발생하지 않은 것으로 판단하여 후보군에서 제외시킨다. 그리고 남은 페이지와 자바스크립트가 동작하지 않는 브라우저에서 다운로드된 페이지를 비교한 후, 동일한 페이지일 경우 후보군에서 제외시킨다. 이와 같은 과정을 거쳐 남겨진 것들을 리다이렉션 후보군으로 설정한다. 여기에는 불법적인 리다이렉션도 존재하지만, 합법적인 리다이렉션도 다수 존재할 수 한다.



[그림 4] URL 리다이렉션 스팸 탐지 알고리즘

다음 단계에서는 이 후보군들을 대상으로 다운로드받은 페이지들을 조사한다. 먼저, 임계값(threshold) 이상되는 목적지 페이지를 찾는다. 이 때, 임계값 이하인 목적지 페이지로 이동하는 페이지는 스팸이 아니라도 판단한다. 다음으로 페이지랭크 값을 비교한다. 합법적인 페이지도 다수의 리다이렉션을 가질 수 있기 때문에, 임계값 이상의 페이지랭크 값을 가지면 웹 스팸이 아니라고 판단한다.

이와 같은 과정을 거쳐 최종적으로 남게 되는 페이지를 웹 스팸으로 판정한다.

4. 향후 연구 과제

현재까지 URL 리다이렉션 스팸 탐지에 대한 연구는 거의 없었다. 기존의 방법은 전문가의 수작업에 의한 탐지가 전부였다. 그렇기 때문에, URL 리다이렉션 스팸의 탐지는 비용이 많이 들고, 매우 시간 소비적이었다.

현재의 한국의 인터넷은 세계의 인터넷과 많은 차이를 보인다. 아직까지 국내에서는 URL 리다이렉션에 의해 심각한 문제점들을 찾기가 쉽지 않다. 그렇다 하더라도, 결코 간과해서는 안 될 문제이다.

이 논문에서 제안한 알고리즘은 기존의 수동적인이었던 리다이렉션 스팸 탐지 방법에 대한 자동화된 방법이

다. 향후 이 알고리즘에 대해 국내 인터넷과 해외 인터넷을 대상으로 실험을 할 예정이다.

5. 결 론

과거 인터넷은 몇몇 사람들의 전유물로 보였다. 하지만, 시간이 흘러 인터넷이 널리 퍼지게 되었고, 매우 많은 정보를 인터넷에서 구하는 세상이 되었다. 이젠 인터넷이 없는 세상을 상상조차 할 수 없다. 하지만, 인터넷의 발달은 비단 우리에게 유익한 면만 보여주지는 않는다. 인터넷으로 인해 여러 가지 문제점들이 나타나게 되었다. 비록, 그러한 기술들이 불법을 위해 개발되어지지 않았지만 말이다.

웹 스팸도 그 중 하나이다. 현재로서는 스팸 메일이나 컴퓨터 바이러스에 비해 심각성이 알려지지 않았지만, 가까운 미래에는 웹 스팸에 대한 문제는 다른 문제들을 능가할 것이다. 좀 더 많은 연구가 진행되어야 할 것이다.

5. 참고문헌

[1] Dennis Fetterly, Mark Manasse, Marc Najork, "Spam, Damn Spam, and Statistics", Seventh International Workshop on the Web and Databases (WebDB 2004), 2004

[2] Kumar Chellapilla, David Maxwell Chickering, "Improving Cloaking Detection Using Search Query Popularity and Monetizability", 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2006), 2006

[3] Zoltan Gyongyi, Hector Garcia-Molina, "Web Spam Taxonomy", 30th International Conference on Very Large Data Bases (VLDB 2004), 2004

[4] Baoning Wu and Brian D. Davision, "Detecting Semantic Cloaking on the Web", 15th International World Wide Web Conference (WWW 2006), 2006

[5] Wikipedia, "Spamdexing", Online at http://en.wikipedia.org/wiki/Spamdexing

[6] Kumar Chellapilla and Alexey Maykov, "A Taxonomy of Javascript Redirection Spam", 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007), 2007

[7] R. Fielding, J. Getttys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T.Berners-Lee, "RFC 2616: Hypertext Transfer Protocol -- HTTP/1.1", Online at ftp://ftp.isi.edu/in-notes/rfc2616.pdf

[8] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford Digital Libraries Technologies Project, 1998

[9] A. Perkins, "White Paper: The Classification of Search Engine Spam", Online at http://www.silverdisc.co.uk/articles/spam-classification/2001

[10] Baoning Wu and Brian D. Davision, "Cloaking and Redirection: A Preliminary Study", 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2003), 2003