Veb spamlar və onlarla mübarizə metodları

Xəyyam Nurəliyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
xeyyam777@gmail.com

Xülasə — **Veb spamlar axtarışın nəticəsində yüksək mövqe qazanmaq məqsədilə informasiya-axtarış sistemlərini aldatmaq və onun nəticələrini manipulyasiya etməyə çalışan veb-saytlar və ya səhifələrdir. Bu gün saytların internetdə irəliləməsi (daha geniş tanıdılması) məqsədilə veb-spamların tətbiqinə axtarış sistemləri tərəfindən qadağa qoyulmasına baxmayaraq, praktiki olaraq, bu texnologiyadan çox istifadə olunur. Veb-spamların istifadə edilməsi qısa zamanda veb-saytın tanınmasına və ona müraciətlərin sayının artmasına kifayət qədər əhəmiyyətli təsir göstərə bilər.**

Açar sözlər— web spam detection, content spam, link spam, cloaking, collusion, Pagerank, TrustRank, classification, clustering, web search, user behavior.

I GİRİŞ

Veb spam anlayışı ilk dəfə 1996-cı ildə meydana gəlmişdir. Axtarış sistemləri və istifadəçilər üçün əsas problemə çevrilmişdir. Axtarış sistemləri arasında getdikcə artan rəqabət daha təmiz internet anlayışını ortaya çıxardı və veb spamların mənfi təsiri araşdırılmağa başlandı. Veb spamlar axtarış sistemlərinin axtarış nəticələrinə və saytların qiymətləndirmə dərəcələrinə çox təsir göstərir. Hazırda saytlar müəyyən gəlir əldə etmək məqsədilə veb spamlar yaradırlar. Aştarış sistemləri interneti zibilləmiş veb saytlardan təmizləmək üçün yeni metodlar işləyib hazırlayırlar. Son vaxtlarda axtarış sistemləri zərərli informasiyalar almağa başladılar. İlkin olaraq veb spamlar axtarış sistemlərinin qiymətləndirmə nəticələrini pozurdu, veb saytların qanuni yolla qazanc əldə etməsinə imkan vermirdi, buda istifadəçilərin axtarış sistemlərinə inamını zəiflədirdi.

İstifadəçilər veb saytlarını axtarış sistemlərinin sıralamasında ön sıralara çıxarmaq üçün veb spamlar yaradırlar. Araşdırmalar göstərir ki, 85% istifadəçilər ilk 3-5 bağlantıya ziyarət edir və nəticədə veb sayt sahibləri bu nəticəni yaxşılaşdırmaq üçün veb spam metodlarına baş vurur [1].

Bu aşağdakı şəkildə olur:

Bir səhifəyə əsas kontentin yerinə aid olmayan kontent əlavə etmək.

Həddindən çox və haqsız bağlantı yaratmaq, gizlətmə, klik etmə saxtakarlığı və istənməyən etiketlər.

Araşdırmalar göstərir ki, veb saytların 6-22% veb spam olur. [1]

Dillərə görə veb spam faizi aşağdakı kimidir.

İngilis-13,8%,

Yapon-9%,

Alman-22%,

Fransız-25%

Domen adlarına görə veb spam aşağıdakı kimidir.

.biz - 70%,

.com - 20%

Veb saytların getdikcə inkişafı nəticəsində çoxlu sayda veb spamlar ortaya çıxdı və hazırda çoxlu sayda veb spamlar mövcuddur.

Hazırda veb spamlarla mübarizədə ən böyük problem metodların tez-tez yenilənməsi və hər bir dilin özünə məxsus xüsusiyyətləridir.

II VEB SPAMLARIN NÖVLƏRİ

A. Kontent spamlar

Kontent spamlar ilk və ən çox yayılan və məşhur olanıdır. Kontent spamlar dedikdə mahiyyət etibarı ilə saytın və ya veb-səhifənin əsas mətninə açar sözlərin və ifadələrin əlavə edilməsi, həddən artıq doldurulması, məzmunun qarışdırılması başa düşülür.

Kontent spamlarının 5 alt tipi var.

• Title Spamming.

• Body Spamming.

• Meta-Tags Spamming.

DOI: 10.25045/NCInfoSec.2018.56

• Anchor Text Spamming.

• URL Spamming.

A. Link spamları

Link spamları veb-spamların daha geniş yayılmış növüdür. Link spamlarının mahiyyəti müxtəlif internet ehtiyatlarında istinadların istifadəsinin süni manipulyasiya edilməsindən ibarətdir. Burada spam rolunda həm daxili, həm də xarici linklər çıxış edə bilər.[5]

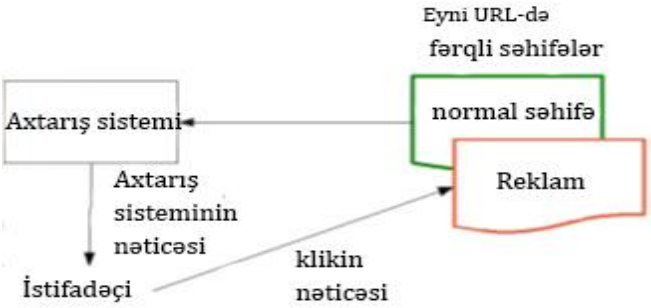
Link spamlarının iki kateqoriyası vardır.

Outgoing- çıxan linklər

İncoming- daxil olan linklər

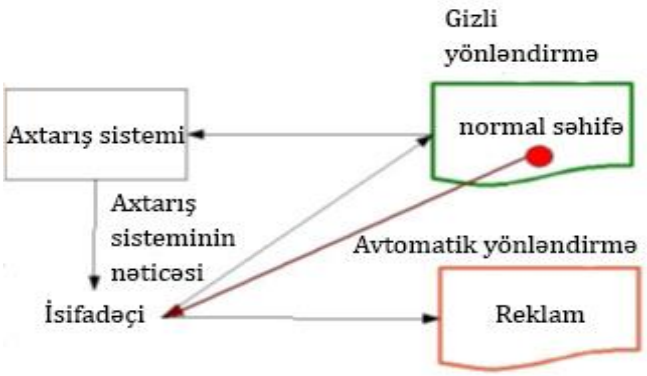
B. Cloaking and Redirection (gizləmə və yönləndirmə spamları)

Gizləmə - veb səhifələrə gələn ziyarətçilərin təbii yoxsa bot olmasının təyin edilməsindən sonra fərqli səhifələrə yönləndirmə edərək axtarış sistemlərinə xüsusi yaradılmış səhifələr göstərərək axtarış sistemlərinin botlarını aldadır.



1) Cloaking spamin işləmə sxemi [3]

Redicrection - həm istifadəçilər, həm də axtarış sistemləri tərəfindən göndərilən sorğuya cavab olaraq gələn URL-in javascript vasitəsi ilə başqa bir URL-ə yönləndirilməsi başa düşülür.



2) Redicretion spamlarının işləmə sxemi [3]

III VEB SPAM METODLARI

İnternet dedikdə ağlımıza gələn ilk saytlardan biri olan Google axtarış sistemidir. Google dünyada ən çox istifadə olunan axtarış sistemidir. Google axtarış sistemi veb saytların ranqlaşdırılmasında istifadə olunan bir çox alqoritmi işləmişdir. Bunlara aşağdakı alqoritmləri göstərmək olar [4]:

-Google Panda

-Google Penguin

-Google Sandbox

1. Page Rank alqoritmi

Google-un veb səhifələrə verdiyi ranqlaşdırma qiyməti şkalasıdır. Bu qiymət ümumi olaraq xüsusi bir məzmuna bağlantı və keçid verən səhifənin yüksək ranqlaşdırma dərəcəsinə bağlıdır. PageRank qiyməti saytınızın Google axtarış mühərrikində internet istifadəçiləri tərəfindən axtarış aparıldığında səhifəin ön sıralarda olmasını təmin edir. T1...Tn, A səhifəsinə istinad edən səhifələr olaraq qəbul etsək. d parametrini 0 ilə 1 arasında dəyişən və ümumilikdə 0,85 qəbul edilən bir ədəddirdd. C (A) isə A səhifəsindən çıxan linklərin sayıdır. Bu halda PR (A), A səhifəsinin PageRank qiymətini ifadə edir.

PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) [2]

2.TrustRank alqoritmi

Stanford universiteti və Yahoo tərəfindən təyin edilən ve spama qarşı mübarizədə istifadə edilən bir link analiz texnologiyasıdır. Spam olan veb səhifələrin yarım avtomatik bir şəkildə təyin olunmasıdır. Hazırda bir çox veb səhifə gəlir əldə etmək məqsədilə axtarış sistemlərində nəticələrini artırmaq üçün spam texnologiyalarından istifadə edir. Hər səhifənin sayt

meniceri tərəfindan araşdırılması çox çətin olduğundan, bəzi etibarlı səhifələr qeyd edilir və ondan alınan link çıxışlarındakı saytlara xallar verilir.

TrustRank ilə editorlar tərəfindən birinci dərəcədə etibarlı saytlar qeyd edilir. Bu saytlara kök sayt deyilir və bu saytlardan link çıxışları ilə digər saytlara etibarlılıq xalı verilir. [2] [6]

NƏTİCƏ

Bu işdə veb spamların növləri və onlarla mübarizə üçün alqoritmlər araşdırılmışdır. Veb spamların istifadəçilərə və axtarış sistemlərinə necə təsir etdiyi, axtarış sistemlərinin nəticələrinə təsiri qısa olaraq araşdırlmışdır.

Minnətdarlıq: Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir **– Qrant № EİF-BGM-4-RFTF-1/2017-21/8/1.**

ƏDƏBİYYAT

1. Nikita Spirin, Jiawei Han, “Survey on Web Spam Detection: Principles and Algorithms,” Department of Computer Science University of Illinois at Urbana-Champaign, December 2011
2. Gökhan Eğri, “Arama Motoru OptimizasyonuTeknikleri,” T.C. İstanbul Kültür Üniversitesi Fen Bilimleri Enstitüsü, Ocak 2013
3. L. Becchetti , P. Boldi , C. Castillo , D. Donato , A. Gionis , S. Leonardi , V.Murdock , M. Santini , F. Silvestri , S. Vigna , “Web Spam Detection, Yahoo! Research Barcelona – Catalunya, Spain 2. Universit`a di Roma “La Sapienza” – Rome, Italy 3. Yahoo! Research Santiago – Chile 4. ISTI-CNR –Pisa,Italy 5. Universit`a degli Studi di Milano – Milan, Italy, 2007, pp. 7-8.
4. https://moz.com/blog/google-algorithm-cheat-sheet-panda-penguin-hummingbird
5. Marc Najork, ‘Web Spam Detection’, Microsoft Research, Mountain View, CA, USA, 2009
6. Vijay Krishnan, Rashmi Raj, ‘Web Spam Detection with AntiTrust Rank, Stanford University Stanford, CA 4305, 2006

WEB SPAMMING AND FIGHTING WEB SPAM METHODS

Xayyam Nuraliyev
Institute of Information Technology of ANAS,
Baku, Azerbaijan
xeyyam777@gmail.com

Abstract—Web spammers are websites or pages that seek to deceive information-searching systems and manipulate its results to gain a high ranking position. Today, although the use of web spammers is banned by search engines for the promotion of websites (wider popularity), it is practically used more than this technology. The use of web spammers can have a considerable impact on the popularity of the website and the number of applications it receives soon.

Keywords — web spam detection, content spam, link spam, cloaking, collusion, Pagerank, TrustRank, classification, clustering, web search, user behavior.