International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-4 Number-2 Issue-15 June-2014

A Survey on Web Spam and Spam 2.0

Ashish Chandra[1], Mohammad Suaib[2]

Abstract

In current scenario web is huge, highly distributive, open in nature and changing rapidly. The open nature of web is the main reason for rapid growth but it has imposed a challenge to Information Retrieval. The one of the biggest challenge is spam. We focus here to have a study on different forms of the web spam and its new variant called spam 2.0, existing detection methods proposed by different researches and challenges that are still unanswered.

Keywords

Web spam, web 2.0, web spam 2.0, search engine, search engine ranking, social media spam.

1. Introduction

The term Spamming is defined by Gyongyi (2005)[1] as: any deliberate action solely to boost ranking of a web page is known as spamming. Web Spam adds cost to both the end users as well as web services providers. Spam websites are means of malware, scams, phishing and adult content. Search engines are designed to provide the web pages that are highly relevant for users' query so they are preferred entry point to billions of pages on the web. Experts say only first 3 to 5 links are clicked by users in a search result page. A webmaster has very limited options to increase relevancy with respect to search requests as well as get high rank in result page without falling into spam tactics. These limited options are under domain of Search Engine Optimization. In fact spamming and SEO have a very thin line between them. As estimated by Erdelyi (2011) [2] 10% of web sites and 20% of web pages are spam. Due to the wide spread of user generated content in web 2.0 sites, spam content is increasing at enormous rate and imposing new challenges for search engines.

Manuscript received June 08, 2014.
Ashish Chandra, Department of Computer Science and Engineering, Integral University, Lucknow, India.
Mohammad Suaib, Department of Computer Science and Engineering, Integral University, Lucknow, India.

2. Search Engine Spamming

According to Fetterly [3] "web pages that hold no actual information value but created to lure web searches to sites that would otherwise not visited".

2.1 Search Engine Result Ranking

Search engines rank pages to search result according to two main features. **Relevancy** of page with respect to query (Dynamic Ranking) and **Authoritativeness** of the page (Static Ranking).

Dynamic Ranking is calculated at query time. It depends on query, user's location, day, time, query history, location of page etc. **Static Ranking** involves query independent features of the page to avoid spam. It is pre-computed at index- time. Some among few hundred features are:

- Total length of the content,
- Frequency of keyword (most frequent words),
- Ratio of text to HTML,
- Number of images in the page,
- Compression ratio (=size of compressed visible text/size of uncompressed text).

Most of the features are under control of the page author, so usefulness of each feature is not disclosed by the search engines.

2.2 Content Spam

Search engine use information retrieval model based on content of page e.g. vector space model, BM25, statistical language model. Malicious web master analyse weakness of these models to exploit them. Locations of the spam contents are:

- Body, title, Meta tags of the page.
- Long URLs of page with keywords stuffing.
- unrelated anchor text of a link.

According to Gyongyi (2005)[1] content spam are:

- Repetition of terms to boost their TF (Term Frequency) value in TF.IDF weighting.
- Dumping unrelated terms or phases into the page to make the page at least partially relevant for multiple queries.
- Weaving of spam phrases into non spam content copied from other high quality pages.

- Stitching together non spam content to create new artificial content that might be attractive for the search engines.

2.2.1 Content Spam Detection

There are different methods such as:

- Document Classification Models
- Near Duplication
- Statistical Analysis Methods

Document Classification Models

It uses different features of content such as

- Number of words in title,
- Abnormal Language Model
- More Popular terms than non-spam pages.

Ntoulas (2006)[4] suggests a 3-gram language model. Piskorski (2008)[5] used Part of Speech tags to produce morphological class of each word (adjective, noun, verbs etc.) which leads to a POS-n-gram. Attenberg (2008)[6] suggested term distance feature which computes frequency of pair of words at certain distance. Zhou(2008)[7] calculated the maximum query specific score in page and found that scores close to maximum are more likely to be spam. Blei [8] suggested Latent Dirichlet Allocation (LDA) which was used by Biro [9] to detect spam by using multi-corpus LDA. Biro(2009)[10] suggested Linked LDA model where topics are propagated along links.
Nepotistic Links as described in [11][12] if source and target pages of links have entirely different content then they are likely to be spam pages. Benezur(2006)[11] says to reduce cost of comparison of two pages, compare only anchor text and words around it of source to the content of target page.

Near Duplication

Urvoy (2008)[13] suggested to check coding style similarity to detect machine generated pages. Fetterly (2005)[14] suggested text automatically generated by sketching of random phrases copied from other sources is near duplication. The largest cluster in the duplicate content are spam. To find such duplicate content and clusters, they applied shingling method based on fingerprint method.

Wu (2006)[15] gave concept of **Duplicate complete link**. If both anchor text and target URLs are copied to create link farms then source and target pages may be duplicate pages.

Statistical Analysis Methods

Fetterly (2004)[3] suggested statistical features like:

- Rapid changes in content.
- Duplicate in nature i.e. Low Word Count Variance in pages hosted on same server.
- URLs contains exceptional number of dots, dashes, digits and URLs with long length

Erdelyi (2011) [2] achieved superior classification results by using learning models LogitBoost and RandomForest and less computation hungry features. They tested 100,000 hosts from WebspamUK2007 and 190,000 hosts from DC2010 datasets and investigated the trade-off between feature generation and spam classification accuracy. It proved that more features improve performance but complex features like PageRank improve accuracy marginally.

2.2.2 Cloaking

Web server provides the crawler a page that is different than the webpage shown to the normal user. It may be helpful for search engines as they do not have to process GUI elements, scripts and ads on the page. **Semantic Cloaking** or malicious cloaking is used by malicious webmasters to deceive search engine by providing spam content to them [16].

2.2.3 Temporary (302) Redirection

A malicious webmaster creates a page u and submits it to the search engine for indexing. When crawler visits the page u, then it is redirected using 302 to a reputable page v. Search Engine considers u an v as canonical URLs and would pick arbitrarily any one of them in search result. When actual user visits the page, it is not redirected.

2.2.4 Redirection Spam

Similar to semantic cloaking, it redirects user to semantically different page in contrast to the cloaking where different copy of page is provided to the crawler. It is implemented by JavaScript which redirects the browser at loading of page. There are many such obfuscated codes available in JavaScript. Wang (2007)[17] suggested that the presence of such obfuscated code is a signal of spam. According to Matt Cutts [18] crawlers do not fully execute JavaScript so they could not detect redirections.

2.3 Link Based Ranking

Link based ranking uses the link structure to determine importance or popularity of a website.

PageRank
According to Page and Brin (1998) [19] PageRank is an estimate of the global importance (authority or reputation) score of a webpage on the web.
Richardson (2006)[20] believed that the contribution of PageRank to the final ranking of the page is small as compared to other factors.

HITS (Hyperlink Induced Topic Search)
HITS was proposed by Kleinberg (1999)[21] for ranking of pages. It starts with constructing a set of pages related to a particular topic by querying the search engine. Then it expands this set by retrieving incoming and outgoing links. It calculates two scores of each page: **Hub Score** representing outgoing links of the page to authority pages and **Authority Score** representing incoming links from authority pages.

2.3.1 Link Spam
There are two broad categories of link spam Outgoing link spam and Incoming link spam. **Outgoing Link Spam** can be easily manipulated by malicious webmasters because they have full control on their pages. They can create a large set of authoritative links on their page to target HITS algorithm by high hub score. Its example is directory cloning i.e. copying a large portion of a web directory like Yahoo directory. **Incoming Link Spam** is used to raise PageRank of target pages (boosted pages). Here spammers gets high authority pages containing links to the target page. For it they create honeypots, post links to user generated content, purchase expired domains, insert links in web directories etc.

2.3.1.1 Link Bombs (Google Bomb)
It is cooperative attempt to place a web page in search result list for a typically obscure search query.
According to Moulton (2007) [22] Google addressed this issue with an update in its ranking system.

2.3.1.2 Link Farm
Link Farm is a set of highly interconnected pages linked together with the sole purpose of boosting the search engine ranking of a subset of these pages known as boosted pages. The pages used in this strategy are called boosting pages or hijacked pages. Link Farms are used in two types of attacks:

- **Sibyl Attack** is where attacker is a single person which creates multiple identities.
- **Collusion Attack:** A group of attackers are agreed to mutually link their pages in a manner that is independent of the quality or relevance of each other's resources.

2.3.1.3 Link Farm Detection
There are two main detection approaches:

Detecting Dense Sub-graph
There is no efficient exact solution present, some of the approximate methods proposed by researchers are as follows.
Gibson (2005)[23] proposed a method based on hash sketches. The graph is represented by adjacency list. Groups of n nodes from each adjacency list are converted into hash sketches. After this an inverted index of sketches is created. This is a list of ordered sketches in which each sketch s is associated with a list of nodes in the graph, in whose out-link, sequence of nodes represented by s can be found. The posting lists of each sketch are scanned and groups of n nodes are sketched again to find dense sub-graph.
Wu (2005)[24] used following method to discover link spam. Find a candidate set of web pages whose in-link and out-link have a sufficient number of domains in common (It is a kind of sub-graph). This list of candidates is then expanded by finding pages with sufficient links to confirm spam.

Detecting Anomalies in Graph
Fetterly (2004)[3] examined in-degree and out-degree of a model representing pages. If values are found well beyond the expected Zipfian distribution then it correspond to presence of spam. Becchetti (2008)[25] investigated on link features extracted from a subset of nodes including degrees, average degree of neighbours, edge reciprocals etc. Benezar (2005) [26] investigated the distribution of PageRank score in the neighbourhood of the page. Da Costa-Corvalha (2006)[27] studied reciprocal link patterns and suggested anomalous link should be removed before estimating authority. Zhou (2008)[7] suggested spamicity approach. Ntoulas (2004)[28] found abnormal fast rate of link changes are signal of spam.

2.3.1.4 Trust Propagation
Zieglar and Lausen (2005)[29] suggested that there are two types of trust computation: **Local Trust inferences** are calculated from the perspective of a single node so each node in the network can have more than one trust values. **Global Trust inferences** are calculated from the perspective of whole network so each node has only one trust value.
VoteLinks suggested by Marks(2005)[30] is hyperlink with rev attributes in HTML anchor tag.

- rev="vote-for" tells to propagate trust
- rev="vote-against" tells to propagate distrust
- rev="vote-abstain" trust is unknown

Trust Rank
Page having high PageRank is more likely to be spam if it has no relationship with a set of trusted pages. TrustRank [31] uses a small set of trustworthy pages that are carefully select by human experts. Random walk with a restart to the seed set is executed for a small fixed set of iterations. Authors used the restart probability 0.15 and number of iterations 20.

2.3.1.5 Distrust Propagation

AntiTrustRank
AntitrustRank [32] is opposite of TrustRank. In-links are not under control of webmasters whereas out-links can be manipulated freely. Castillo (2007)[33] believed non spam pages do not generally link to spam pages. Antitrust Rank uses a random walk that follows links backward and restarts to known set of spam pages. Krishnan (2006)[34] suggested the badness of the page is its probability in the stationary state of the random walk..

SimRank
SimRank[35] is a generalization of the co-citation and can be used as spam classification feature. A page that has high link similarity (SimRank) to a spam page is likely to be spam. Wu (2007)[36] suggested that start with a given set of confirmed spam nodes and then walk randomly to find other nodes that might be involved in same spam activity.

2.3.1.6 Other Recent Approaches

WITCH
Web spam Identification through Content and Hyperlink[37] uses Graph Regularization method. It is machine learning method that uses hyperlink structure and page features. The hyperlink data is processed using graph regularization. It uses the approximate isolation of good pages that argue for asymmetric regularizer. Cheng(2011)[38] suggests to extract web spam URLs from SEO forums as spammers share on SEO forums the links of their websites to find partners for building global link farms.

Temporal Features for spam detection
Shen (2006)[39] found spammers create millions of machine generated pages and links. They may very quickly remove or regenerate pages as well as links if these are blacklisted. So linkage change is a feature that can be used to detect spam. It can be measured as In-link Growth Rate , In-link Death Rate, Out-link Growth Rate and Out-link Death Rate. Erdelyi (2011)[40] proposed graph similarity based temporal feature to detect link change of neighborhood hosts.

2.4 Search Engine Usage Data

Usage Information known as wisdom of crowd is very vital for search engines to optimize the user's search experience. Search engines collect these data as Query log, Browser log and Ad-click log.
Query Log stores keywords searched, link results clicked by user, sequence of user actions (query session), location of user, IP address of user etc.
Browser Log is obtained from users who opt-in to a system of tracking their activities. An example of such system is toolbar extension installed in user's browser. The sequence of actions of users being recorder are known as browsing trails.
Ad-click Log is the data of user activity with respect to ad-network pay per click ads such as query monetization (revenue generated by all ads that are displayed in respect to the query).
Search engine boosts ranking of the pages which are more clicked by users. For this query log is used. As per Craswell (2008)[41] for ranking purpose search engine also consider the position of the link in the result page. Liu (2008)[42] says Search engine also use browser log for ranking. e.g. BrowserRank.

2.4.1 Usage Spam

Spammers try to rank their pages in the search results. For this they try to manipulate usage data by artificially generating search and browsing actions.
According to Daswani (2007)[43] to be hidden from detection, they deploy scripts on many machines or in large botnets. Some examples are following.

Click Fraud [44] is the practice of manipulating pay per click advertising data by generating illegitimate events. Jansen (2007)[45] believes it is prevalent and potentially very harmful for the sponsored search business model.
Search Spam Spammers deploy scripts that automatically make searches of predefined queries and generate automatic clicks on the target pages.
Referrer Spam [46] is low impact spamming. Spammers create web crawlers that selectively visit to web pages but instead of leaving referrer field blank, they insert a link of their own target web page.

2.4.2 Usage Spam Detection

Buehrer (2008)[47] found that with a given sequence of queries associated with client IP address and cookie, we can classify the session into normal and automated with over 90% accuracy. The features for

the classification may be number of queries, entropy of key words in query, number of queries issued in a short period (<10 seconds) and the click through rate.
Bacarella (2004)[48] created a traffic graph for browsing trial where nodes of the graph represent web pages and edges represent visit from node u to v. Here the relative traffic of site v is the average in-traffic of site v is the average in traffic of in-neighbours of v. If the relative traffic of a site v is 90% or more then it is spam using deceptive means like popup, redirects etc.
Ntoulas (2006)[4] and castillo (2007) [33] suggested that a list of top popular queries submitted to search engines can be collected and if a page contains abnormally high fraction of these queries then it may be spam. According to Castillo (2008)[49][50] if a page attract traffic for many unrelated queries then it is likely to be spam. Chellapilla (2006)[51] provides methods to use query log to detect cloaking. They used 5000 most popular queries and 5000 most monetized queries. Then requested top 200 links for each query four times by providing various agent fields to initial request for a user(u).and a crawler(c) calculated cloaking score. They reported 0.75 and 0.985 precision for popular and monetized query.

2.5 User Generated Content (Web 2.0) Spam

Web 2.0 describe dynamic World Wide Web which allows users to actively contribute contents instead of passively viewing the contents. e.g. blogs, forums, social networking, video sharing sites, wikis etc.

Spam 2.0

Web spam in web 2.0 sites is known as spam 2.0 Hayati (2010) [52] observed that spam 2.0 spreads through legitimate websites such as government, universities, reputed companies or personal websites.
Spam 2.0 gets undeserved ranking for spam content and damages reputation of legitimate websites. As per Live spam zeitgeist[53] daily spam detection rate in 2014 is approximately double than that of in 2013.
When a spam 2.0 is posted, its content can be read by a large number of users. Many users may be seriously affected if it contains spyware, malware, phishing or fraud. Spammers create eye catching user profiles on social networking sites to do things. An awareness of online spam among users will be helpful to reduce impact of such malicious campaign. According to a survey by Potdar (2013)[54] 91.6% people had heard about online spam, 53% never noticed pages only filled with repeated keywords. 26.4% people said they have poor knowledge about spam. Only 45.4% had the idea that automatic software can be used to register spam accounts. 53.55% considered themselves as vulnerable to spam where 25.3% considered themselves not vulnerable at all. 33.1% users thought they were not likely to be spammed where 33.9% said they were highly likely.

2.5.1 Splogs

Also known as spam blogs or fake blogs, Splogs are used to promote junk /hijacked content or to generate link farms. Kolari (2006) [55] characterized splogosphere which was based on a blog collection and collection of pings and found that splogs generate more than 75% pings and have periodic patterns. Benevenuto (2008) [56] gave a behaviour based approach which suggests writing behaviour to detect spam. These behaviours include writing interval, writing structures, writing topic. Spammer tries to focus on same topic at a time with same size of content. Legitimate bloggers writes on different topics with different blog length. This approach is not language dependent so it can be applied to non-English blogs also. Lin (2008) [57] used temporal pattern to detect malicious blogs. Legitimate blogs are regular but not precise at time whereas machine generated blogs show machine like regularities. In Splogs, distribution of content into topic varies either rapidly or not at all. Splogs have small variations in links. Hayati (2012)[58] retrieved content from different systems, extracted keywords, meta content and perform classification to detect spam.

2.5.2 Forum Spam

Forum spam is used to increase linked based authority by posting new thread or replying to existing thread [59].
Potdar(2012) [60] collected following findings:

- Spammers prefer to use one time or short time email address.
- Email addresses of spammers normally have exceptional numbers of dots or random addition of letters and numbers.
- All emails of same spammer may be of same length.
- Insertion of dots in same Gmail id to exploit the inherent feature of Gmail.
- Repetition of post content with same body and same or different subject.
- Spam origin was Brazil 26%, Russia 26%, India 10%, China 10%, Argentina 11% etc.

Georgia (2007)[61] proposed detection method which searches for spam keywords, templates, attachments etc in Contents of blogs or forum post. Hayati (2010)[62] focused on spambot identification rather

than analyzing spam content in web 2.0 forums. They proposed action navigation as a new feature set and achieved 95.55% accuracy.

2.5.3 Comment Spam

Comment Spam is used to distribute promotional, fake, junk content. It may be attractive and contain link of target page. There are two main commercial comment spam protection plug-ins available:
Akismet (http://akismet.com) and
Mollom (http://mollom.com)
According to Akismet Live report 83% of posted comments were spam where this plugins was installed and according to Mollom 90% messages were spam.
No Follow Attribute [63] proposed by major blogging service providers Google, Yahoo and MSN in 2005 to combat comment spam. It says to crawlers that the link with rel="nofollow" attribute has unknown trust. They do not want to propagate authority to that link. As per our knowledge there is no study done on the impact of no follow attribute. Huang(2010)[64] proposed content based detection method to detect comment spam in blogs. They used SVM, Naïve Bayes and C4.5 to 4 features.

- Length of comment
- Similarity between comment and blog post
- Divergence between comment and blog post
- Ratio of propaganda and popular words.

The best result was achieved with accuracy 84% using C4.5 method.

2.5.4 Opinion & Review Spam

There are 3 types of spam reviews[65]. **False Review** is misleading or false judgment about a product. **Non Review** is advertisement of a product or brand. and **Brand Review** is like an advertisement of a brand instead of a product.
Nitin (2008)[66] used 36 features for classification in supervised machine learning Algorithm Logistic Regression. It is language independent approach so can be used in non-English contents. Jindal[67][68] used content specific approach and extracted 36 features from opinion /review content in review websites (e.g. Amazon). They trained the classifier to differentiate between spam and ham reviews using these features e.g. # of feedbacks, ratio of # of reviews written as first review. They achieved 98.7% accuracy in detecting brand review and non-review. Accuracy was not good in case of false reviews because false reviews are very hard to detect even by humans. Jindal (2010)[69] used behavioural features such as multiple rating of same product, multiple reviews of same product, deviation of rating by other reviews of same product. Caverlee(2008)[70] suggested social trust framework to compute trust in global manner in social network. It propagates trust score from user to friends. Sirivianos (2009)[71] suggested a mechanism to verify credential of a user from a third party.

2.5.5 Social Media Spam

Spam on sites which allow post comments, bookmark, tagging, images, video etc are social media spam. Normally these sites allow users to report abuse/spam. Some sites are Facebook, Twitter, Delicious, YouTube, Flicker etc. Social media spam can be tagging spam, video spam, voting spam.

2.5.5.1 Tagging Spam

Tags are descriptive text to annotate a resource. According to Yang (2011) [72] there is 3 levels of granularity in tagging spam detection User Level, Post Level, Tag Level

User Level Spam can be detected by these features:

- **TagSpam** measures the probability of user being spammer according to predefined tag vocabulary.
- **TagBlur** measures degree of un-relatedness among tags in a post.
- **DomFp** Spam pages generated by machine have same document structure. It is measure of structural similarity between pages. Here we extract the fingerprint of a page's DOM structure and strip away content of the page and left with HTML elements.
- **NumAds** Spammers often create pages displaying ads e.g. Google AdSense. We can count frequency of googlesyndication.com on page.
- **Plagiarism** Extract a random sequence of 10 words from page. Submit it to Yahoo Search API (http://developer.yahoo.com/download). The number of results returned is the measure of plagiarism.
- **ValidLink**=Number of valid links/Total links. If user is malicious, many of link may have blacklisted or removed.

Krause (2008)[76] suggested machine learning approach to detect user level spam detection. They proposed 25 characteristics of users in bookmarking. Some of the characteristics are number of digits in their name and mail Id, network location (IP), Tags used by them (black list of tags). They created a

classifier with accuracy of 0.93. they categorized these features in 4 categories.
(i) User Profile (ii) User Location (iii) User Integration (iv) Semantic features.

Post Level Spam Detection identifies individual spam from each post by any user. When a post is identified as spam all tags in this post are spam.
Yang (2011) [72] suggested post level spam detection approach which uses semantic similarity between web pages and its tags. They extracted keywords from webpage and compared with tags.

Tag Level Spam Detection allows spam tags detection which co-occurs with normal tags in a post.
Koutrika (2007)[73][74] provided a model for malicious and normal behaviour of users in tagging system to detect user level spam. Neubauer [75] studied abnormal tagging pattern and found that collection of spammers appear in large connected components.
Markines (2009) [77] used 6 features to detect spam in bookmark tagging.

- Probability of tags being used by Spammers.
- Dissimilarity of tags being used in the post.
- Likelihood of target page being machine generated
- Number of ads on the target page.
- Likelihood of plagiarism in target page
- Fraction of user's post still referencing to valid resources.

Clean Tweet, an extension of FireFox web browser, hides posts from accounts that are less than one day old or containing too many tags. It provides a user level spam detection.
Georgia (2007)[78] counted number of common tags among other users and assigned a resource, a relative ranking. Malicious and legitimate tags can be separated using this ranking. It is language independent content base approach.

2.5.5.2 Video Spam

Spammers can post irrelevant content either as video response or as a related video to the most popular video to gain popularity or promote their product.
Benevenuto (2008)[56] proposed spam detection in video sharing website YouTube. They examined attributes of objects, users and the social network connecting them. They found that the malicious users and spam objects had distributions that are different from legitimate users. They created content classifier that uses machine learning to classify video spam using metadata information. The results were not satisfactory as accepted by the authors themselves. The classifiers used SVM , the dataset size was 829.

2.5.5.3 Voting Spam

Bian (2008)[79] studied voting system of yahoo Answers and found that it is vulnerable to spammers.
Tran (2009)[80] described a voting method that analyses social network of users and gives less weight to votes from users who are not well connected to other users.

2.5.6 Wiki Spam

Spammers modify wiki pages and put back-links to their link farms. It is very difficult to detect wiki spam and there is no particular method found. We feel that presently the most effective way could be users' awareness to recover the old version of article from history.

3. Analytical Summary

Now we can summarize different approaches of researchers in following table.

Table 1: Summary

Approach	Detection Level
Language Model [4,5,7,9]	Content, Comment Spam[64]
Near Duplicate[12, 13, 14]	Content, Comment[64]
Trust Propagation	Link
Temporal features	Link[39,40], Splog[57]
Hybrid Approach [37]	Link, Content
Statistical features[3]	Link, Content
Obfuscated JavaScript[17]	Cloaking/ Redirection, Content Hiding
Query Log[51]	Cloaking/ Redirection
User Behavior	Splog[56], Forum Spam[60], Opinion[69], Social Spam [73,74,75,76,77,78]
Spam-Bot detection [62]	Forum Spam
Classification Features	Opinion[66,67], Video[56]
Social Trust [70,71]	Opinion/Review Spam
Semantic Similarity	Content, Social Spam[72]

Some key points that we observed in this survey are:

- Spammers create link farms and try to exploit high authoritative web 2.0 pages which are open to all. They create large number of spam links that impact performance of link mining algorithms. We can remove or down-weight these links while calculating authority.
- Almost all approaches work on the text, there is little work on images, video, audio.

- Careful features selection is very important for training of machine learning model. Many researchers proved in their experiment that complex global feature like PageRank contribute marginally.
- Most of the content based approaches work on monolingual models but this is not helpful in case of web 2.0 pages which may contain text contributed by people speaking different languages. By using temporal features we can improve the link based methods to detect spam in multilingual environment where content base analysis cannot help.

4. Conclusion

In this paper we first described what a search engine spam is and described about various types of web spam and the work done to combat spam. We also described new type of spam i.e. Spam 2.0 and challenges imposed by it to our huge Information Retrieval System i.e. our Search Engines. We noticed that normal users are not well aware about the spam and are highly vulnerable to it. We feel that spam is a socio-economic problem which can be dealt with to some extent if normal users are aware of it and there are such preventive measures that add extra cost to web spam generation. All the detection techniques we discussed are capable of detecting spam up to some extent but if we rely on single method, spammers will find the way to bypass it. So we have to use a combination of all possible techniques. We also observed that there is little work done in multimedia and multilingual content area; we hope researchers will address this challenge in near future.

References

[1] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy," in Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 39–47, May 2005.

[2] M. Erdelyi, A. Garzo, and A. A. Benczur. Web spam classification: a few features worth more. In Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality, WebQuality'11, Hyderabad, India, 2011.

[3] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages," in Proceedings of the Seventh Workshop on the Web and databases (WebDB), pp. 1–6, June 2004.

[4] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam Web pages through content analysis," in Proceedings of the 15th International Conference on World Wide Web (WWW), pp. 83–92, May 2006.

[5] J. Piskorski, M. Sydow, and D. Weiss, "Exploring linguistic features for Web spam detection: A preliminary study," in Proceedings of the Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 25–28, New York, NY, USA: ACM, 2008.

[6] J. Attenberg and T. Suel, "Cleaning search results using term distance features," Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 21–24, New York, NY, USA: ACM, 2008.

[7] B. Zhou, J. Pei, and Z. Tang, "A spamicity approach to Web spam detection" in Proceedings of the SIAM International Conference on Data Mining, April 2008.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[9] I. Bıro, J. Szabo, and A. A. Benczur, "Latent dirichlet allocation in Web spam filtering," 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 29–32, New York, NY, USA: ACM, 2008.

[10] I. Bıro, D. Siklosi, J. Szabo, and A. Benczur, "Linked latent dirichlet allocation in Web spam filtering," 5th International Workshop on Adversarial Information Retrieval on Web (AIRWeb), pp. 37–40, ACM Press, 2009.

[11] A. A. Benczur, I. Bıro, K. Csalogany, and M. Uher, "Detecting nepotistic links by language model disagreement," 15th International Conference on World Wide Web (WWW), pp. 939–940, ACM Press, 2006.

[12] J. Martinez-Romo and L. Araujo, "Web spam identification through language model analysis," 5th International Workshop on Adversarial Information Retrieval on Web (AIRWeb), pp. 21–28, ACM Press, 2009.

[13] T. Urvoy, E. Chauveau, P. Filoche, and T. Lavergne, "Tracking Web spam with HTML style similarities," ACM Transactions on the Web, vol. 2, no. 1, 2008.

[14] D. Fetterly, M. Manasse, and M. Najork, "Detecting phrase-level duplication on the World Wide Web," in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 170–177, New York, NY, USA: ACM, 2005.

[15] B. Wu and D. B. Davison, "Undue influence: Eliminating the impact of link plagiarism on Web

search rankings," in Proceedings of The 21st ACM Symposium on Applied Computing (SAC), pp. 1099–1104, April 2006.

[16] B. Wu and D. B. Davison, "Cloaking and redirection: A preliminary study," 1st International Workshop on Adversarial Information Retrieval on Web (AIRWeb), 2005.

[17] Y.-M. Wang, M. Ma, Y. Niu, and H. Chen, "Spam double-funnel: Connecting Web spammers with advertisers," in Proceedings of the 16th International Conference on World Wide Web (WWW), pp. 291–300, New York, NY, USA:ACM Press, 2007.

[18] E. Enge, "Matt cutts interviewed by eric enge," Article online at http://www.stonetemple.com/articles/interview-matt-cutts-012510.shtml, April 2010.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Techincal Report, Stanford University, 1998. Available from http://dbpubs.stanford.edu/pub/1999-66.

[20] M. Richardson, A. Prakash, and E. Brill, "Beyond PageRank: Machine learning for static ranking," 15th International Conference on World Wide Web (WWW), pp. 707–715, New York, NY, USA: ACM Press, May 2006.

[21] M. J. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol. 46, no. 5, pp. 604–632, 1999.

[22] R. Moulton and K. Carattini, "A quick word about Googlebombs," http://googlewebmastercentral.blogspot.com/2007/01/quick-word-about-googlebombs.html, January 2007.

[23] D. Gibson, R. Kumar, and A. Tomkins, "Discovering large dense subgraphs in massive graphs," 31st International Conference on Very Large Data Bases, pp. 721–732, VLDB Endowment, 2005.

[24] B. Wu and D. B. Davison, "Identifying link farm spam pages," in Special interest tracks and posters of the 14th International Conference on World Wide Web (WWW), pp. 820–829, New York, NY, USA: ACM Press, 2005.

[25] L. Becchetti, C. Castillo, D. Donato, R. Baeza-Yates, S. Leonardi, "Link analysis for Web spam detection," ACM Transactions on Web, vol. 2, no.1, pp. 1–42, February 2008.

[26] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher, "SpamRank: Fully automatic link spam detection," First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), May 2005.

[27] A. L. C. da Costa-Carvalho, P.-A. Chirita, E. S. de Moura, P. Calado, and W. Nejdl, "Site level noise removal for search engines," in Proceedings of the 15th International Conference on World Wide Web (WWW), pp. 73–82, New York, NY, USA: ACM Press, 2006.

[28] A. Ntoulas, J. Cho, and C. Olston, "What's new on the Web?: The evolution of the Web from a search engine perspective," 13th International Conference on World Wide Web, pp. 1–12, New York, NY, USA: ACM Press, 2004.

[29] C.-N. Ziegler and G. Lausen, "Propagation models for trust and distrust in social networks," Information Systems Frontiers, vol. 7, no. 4–5, pp. 337–358, December 2005.

[30] K. Marks and T. Celik, "Microformats: Vote links," Technical Report, Technorati, 2005. Online at http:// microformats.org/wiki/vote-links.

[31] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen, "Combating Web spam with TrustRank," 30th International Conference on Very Large Data Bases (VLDB), pp.576–587, Morgan Kaufmann, August 2004.

[32] M. Sobek, "PR0 — Google's PageRank 0 penalty," http://pr.efactory.de/e-pr0.shtml, 2002.

[33] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology," in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, July 2007.

[34] V. Krishnan, R. Raj, "Web spam detection with anti-TrustRank," 2nd International Workshop on Adversarial Information Retrieval on Web (AIRWeb), pp. 37–40, 2006.

[35] B. A, K. Csalogany, T. Sarlos", Link-based similarity search to fight Web spam," Second International Workshop on Adversarial Information Retrieval on Web 2006.

[36] B. Wu and K. Chellapilla, "Extracting link spam using biased random walks from spam seed sets," 3rd International Workshop on Adversarial Information Retrieval on Web, pp. 37–44,: ACM Press, 2007.

[37] J. Abernethy, O. Chapelle, and C. Castillo, "Graph regularization methods for web spam detection," Machine Learning Journal, vol. 81, no. 2, pp. 207–225, 2010.

[38] Z. Cheng, B. Gao, C. Sun, Y. Jiang, and T.-Y. Liu. Let web spammers expose themselves. In Proceedings of the fourth ACM International Conference on Web search and Data Mining, WSDM'11, Hong Kong, China, 2011.

[39] G. Shen, B. Gao, T.Y. Liu, G. Feng, S. Song, H. Li, "Detecting link spam using temporal information," IEEE International Conference on Data Mining, December 2006.

[40] M. Erdelyi and A. A. Benczur. Temporal analysis for web spam detection: An overview. 1st International Temporal Web Analytics Workshop in 20th International World Wide Web

Conference in Hyderabad, India. CEUR Workshop Proceedings, 2011.

[41] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in Proceedings of the First International Conference on Web Search and Data Mining (WSDM), pp. 87–94, New York, NY, USA: ACM, 2008.

[42] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li, "Browse-Rank: Letting web users vote for page importance," 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 451–458, ACM, 2008.

[43] N. Daswani and M. Stoppelman, "The anatomy of Clickbot.A," USENIX HOTBOTS Workshop, April 2007.

[44] L. A. Penenberg, "Click fraud threatens Web," Wired, October 2004.

[45] J. B. Jansen, "Click fraud," Computer, vol. 40, no. 7, pp. 85–86, 2007.

[46] K. Yusuke, W. Atsumu, K. Takashi, B. B. Bahadur, and T. Toyoo, "On a referrer spam blocking scheme using Bayesian filter," Joho Shori Gakkai Shinpojiumu Ronbunshu, vol. 1, no. 13, pp. 319–324, In Japanese, 2005.

[47] G. Buehrer, J. W. Stokes, and K. Chellapilla, "A large-scale study of automated Web search traffic," 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 1–8, ACM, 2008.

[48] V. Bacarella, F. Giannotti, M. Nanni, and D. Pedreschi, "Discovery of ads Web hosts through traffic data analysis," in Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), pp. 76–81, ACM, 2004.

[49] C. Castillo, C. Corsi, D. Donato, P. Ferragina, and A. Gionis, "Query log mining for detecting polysemy and spam," KDD Workshop on Web Mining and Web Usage Analysis (WEBKDD), Springer: LNCS, August 2008.

[50] C. Castillo, C. Corsi, D. Donato, P. Ferragina, and A. Gionis, "Query-log mining for detecting spam," 4th International Workshop on Adversarial Information Retrieval on the Web, ICPS: ACM Press, April 2008.

[51] K. Chellapilla, D. M. Chickering, "Improving cloaking detection using search query popularity and monetizability," 2nd International Workshop on Adversarial Information Retrieval on Web, pp. 17–24, August 2006.

[52] P. Hayati, V. Potdar, A. Talevski, N. Firoozeh, S. Sarenche, and E. A. Yeganeh, "Definition of spam 2.0: New spamming boom", In the IEEE International Conference on Digital Ecosystems and Technologies (DEST 2010), Dubai, UAE, pp. 580-584, 2010.

[53] Live-Spam-Zeitgeist: Some Stats, Akismet. [Accessed online by May 2014] http://akismet.com/about (2014).

[54] F. Ridzuan, V. Potdar, W. Hui, 2013. Awareness, Knowledge and Perception of Online Spam, JNIT: Journal of Next Generation Information Technology, Vol. 4, No. 3, pp. 9 ~ 22, 2013. DBLP.

[55] P. Kolari, A. Java, T. Finin, "Characterizing the Splogosphere," 3rd Annual Workshop on Weblogging Ecosystem, 2006.

[56] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross, "Identifying video spammers in online social networks," in Proceedings of the Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 45–52, New York, NY, USA: ACM, 2008. References 471.

[57] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and L. B. Tseng, "Detecting splogs via temporal dynamics using self-similarity analysis," ACM Transations on the Web, vol. 2, no. 1, pp. 1–35, 2008.

[58] P. Hayati, and V. Potdar, "Spam 2.0 State of the Art", International Journal of Digital Crime and Forensics (IJDCF), vol. 4, no. 1, pp.17-36, January-March 2012.

[59] Y. Niu, Y.-M. Wang, H. Chen, M. Ma, and F. Hsu, "A quantitative study of forum spamming using context-based analysis," 14th Annual Network and Distributed System Security Symposium (NDSS), pp. 79– 92, February 2007.

[60] V. Potdar, Y. Like, N. Firoozeh, D. Mukhopadhyay, F. Ridzuan, D. Tejani, "The changing nature of Spam 2.0," In the Proceedings of the CUBE International Information Technology Conference, Pune, India, pp. 826-831, 2012.

[61] Paul, H., Georgia, K., & Hector, G.-M. (2007). Fighting spam on social web sites: A survey of approaches and future challenges. IEEEInternet Computing, 11(6), 36–45. doi:10.1109/MIC.2007.125.

[62] P. Hayati, V. Potdar, K. Chai, and A. Talevski, "Web Spambot Detection Based on Web Navigation Behaviour," in 24th IEEE International Conference on Advanced Information Networking and Applications (AINA 2010), Perth, Western Australia, 2010.

[63] K. Marks and T. Celik, "Microformats: The rel=nofollow attribute," Techical Report, Technorati, 2005. Online at http://microformats.org/wiki/relnofollow.

[64] Huang, C., Jiang, Q., & Zhang, Y. (2010). Detecting comment spam through content analysis. In H. T. Shen, J. Pei, M. T. Özsu, L. Zou, J. Lu, T.-W. Ling, (Eds.), 2nd International Workshop on Web-based Contents Management

Technologies, Jiuzhaigou, China (LNCS 6185, pp. 222-233).
[65] N. Jindal and B. Liu, "Review spam detection," 16th International Conference on World Wide Web (WWW), pp. 1189–1190,ACM Press, 2007. References 477.
[66] J. Nitin and L. Bing, "Opinion spam and analysis," in Proceedings of international conference on Web search and web data mining Palo Alto, California, USA: ACM, 2008.
[67] N. Jindal and B. Liu, "Analyzing and detecting review spam," in Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), pp. 547– 552, 2007.
[68] N. Jindal and B. Liu, "Opinion spam and analysis," International Conference on Web Search and Data Mining (WSDM), pp. 219–230, ACM, 2008.
[69] Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In Proceedings of 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada (pp.939-948). New York, NY: ACM.
[70] J. Caverlee, L. Liu, S. Webb, "Socialtrust: Tamper-resilient trust establishment in online communities," in Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 104–114, 2008.
[71] M. Sirivianos, X. Yang, and K. Kim, "FaceTrust: Assessing the credibility of online personas via social networks," Technical Report, Duke University, 2009. http://www.cs.duke.edu/msirivia/publications/facetrust-tech-report.pdf.
[72] Yang, H. C., & Lee, C. H., Post-level spam detection for social bookmarking web sites. In Advances in Social Networks Analysis and Mining (ASONAM), 2011 international Confierence on (pp. 180-185). IEEE.
[73] G. Koutrika, A. F. Effendi, Z. Gy¨ongyi, P. Heymann, and H. Garcia-Molina, "Combating spam in tagging systems," in Proceedings of 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 57–64, New York, NY, USA: ACM Press, 2007.
[74] G. Koutrika, A. F. Effendi, Z. Gyongyi, P. Heymann, and H. Garcia-Molina, "Combating spam in tagging systems: An evaluation," ACM Transactions on the Web, vol. 2, no. 4, pp. 1–34, 2008.
[75] N. Neubauer, R. Wetzker, and K. Obermayer, "Tag spam creates large nongiant connected components," 5th International Workshop on Adversarial Information Retrieval on Web (AIRWeb), pp. 49–52, ACM Press, 2009.
[76] B. Krause, H. A. Schimitz, and G. Stumme, "The anti-social tagger - detecting spam in social bookmarking systems," Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), April 2008.
[77] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 41–48, New York, NY, USA: ACM, 2009.
[78] K. Georgia, E. Frans Adjie, Zolt, G. n, ngyi, H. Paul, and G.-M. Hector, "Combating spam in tagging systems," 3rd international workshop on Adversarial information retrieval on the web Banff, Alberta, Canada: ACM, 2007.
[79] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "A few bad votes too many?: Towards robust ranking in social media," 4th International Workshop on Adversarial Information Retrieval on the Web, pp. 53–60, ACM, 2008.
[80] N. Tran, B. Min, J. Li, L. Submaranian, "Sybil-resilient online content voting," 6th Symposium on Networked System Design and Implementation, 2009.

Ashish Chandra is currently working as a Project Manager at Kenbit Software Technologies. He has a fine blend in Software Development in .NET technologies. Ashish has mastered in Computer Science and Engineering from Integral University, Lucknow, India. He has a keen interest in Search Engines, Adversarial Information Retrieval and Artificial Neural Network.