International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference & workshop on Advanced Computing 2013 (ICWAC 2013) – www.ijais.org

Spam Detection Techniques: Issues and Challenges

Amrita Mathur
M.E, computer
Sardar Patel Institute of
Technology, Andheri Mumbai

Prachi Gharpure, Ph.D
Principal,
Sardar Patel Institute of
Technology, Andheri Mumbai

ABSTRACT
In recent years social websites have become important components of the web. With their success, however has come a growing influx of spam. If left unchecked, spam threatens to undermine resource sharing, interactivity and openness. This article focuses on various countermeasures that are based on detection, demotion and prevention. Although various countermeasures have been proposed for e-mail and web spam but still there are challenges .This paper surveys various anti-spam strategies for social networking. In literature we have studied that many anti- spam strategies have been discovered but still there are some open challenges to these techniques. Some of them are highlighted in this article.

General Terms
Spam detection , security.

Keywords
Social networking; spam; anti-spam strategies

1.INTRODUCTION
A social network is the set of human beings or their digital representations that refer to the registered users who are linked by relationships extracted from the data about their activities, common communication or direct links gathered in the internet based system. The increasing popularity of the World Wide Web and the internet caused a increasing number of types of services available on computer network. People who use these services have created a new kind of virtual societies usually called online social network or web based social network or computer supported social network or virtual communities. The main concern of web surfers is to get relevant data based on their keyword or query. The spammers insert frequent words into the document and made their document worthwhile to read. Resulting which web surfer ends up with unknowingly reading the entire content which is not relevant to their query. Also internet is becoming a common media ofcommunication whether through e-mails or social websites. So fighting with these spams is becoming a big challenge for service providers.

There are six main forms of spam, and they have different effects on Internet users like: (1) E-mail spam , (2) comment spam, (3) Instant Messenger Spam, (4) Unsolicited text message, (5) social networking spam.[1]

To reduce or eliminate spams various anti-spam methods have been proposed in state-of-the art research. [2] Heymann et al classified anti-spam strategies into three categories: (1) prevention based, (2) detection based, (3) Demotion based. There are various anti spam strategies as content based, link based, graph analysis , clocking scheme but in spite of having various anti-spam strategies there are various open challenges to these anti-spam strategies which need to be addressed. Some of these are highlighted in the paper.

2. ANTI-SPAM STRATEGIES

2.1 Prevention based
This approach aims at making it difficult for spam content to contribute to social tagging system by restricting certain access types through interfaces (such as CAPTCHA which stands for "completely automated public Turing test to tell computers and humans apart") or through usage limits (such as tagging quota e.g Flickr introduced a limit of 75 tags per photo).[1]

2.2 Detection based

These approaches identify likely spams either manually or automatically by making use of machine learning (such as text classification) or statistical analysis (such as link analysis) and then deleting the spam content or visibly marking hidden to the user. For these methods, we can treat the corpus as set of objects with associated attributes. In e-mail spam, the messages are objects and the headers are attributes. In web spam , the web pages are objects and attributes might be inlinks, outlinks, page content and various external meta data.[1]

2.3 Demotion based

This approach reduces the prominence of content likely to be spam. For instance rank based methods produce ordering of a system's content, tags or users based on trust score.

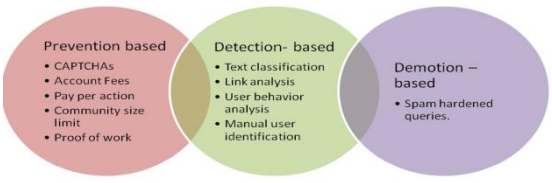

Figure 1: Anti-spam strategies [1]

3 NEED FOR SPAM DETECTION

Spam detection is becoming a big challenge for the service providers because of their following negative effect:[2]

- Spam deteriorates the quality of search result and deprive legitimate websites of revenue.
- Spam have economic impact since a high ranking provides large free advertising and so an increase in web traffic volume.
- It weakens the trust of a user in search engine provider which is a especially tangible issue due to zero cost of switching from one search provider to another.
- Spam websites are means of malware and adult content dissemination and fishing attack.
- Spam forces a search engine company to waste a significant amount of computation and storage resource.
- One important challenge in tagging is to identify the most appropriate tags for the given content and to eliminate the spam tag.

4 SPAM DETECTION TECHNIQUES

Algorithms for spam detection can be categorized into following 4 groups:

4.1 Content based:

Techniques which analyze content features such as word count or language models and content duplication.Fetterly et al proposed that web spam pages can be identified through statistical analysis. Spam pages exhibit some anomalous properties as :(1) URL of spam pages have exceptional number of dots, dashes. digits and length, (2) Most spam pages that resides on the same host have very low word count variance, (3) Content of spam pages changes very rapidly. T. urvoy et al introduced features based on HTML page structure to detect script generated spam pages[5]. In this preprocessing is made by removing all the content and considering only layout of the page. They applied finger printing technique with subsequent clustering to find groups of structurally near spam pages [5]. G. Mishne et al proposed a line of work on language modeling for spam detection. They proposed an approach of spam detection in blogs bycomparing the language models for blog comments and pages linked with these comments. They use KL divergence as a measure of discrepancy[7]. In other work by M. sydow linguistic features were analyzed for web spam detection by considering lexical validity and content diversity, syntactical diversity and entropy , usage of active and passive voices and various other NLP features.[9]

4.2 Link based

Another group analyzes link based information such as neighbor graph connectivity. Based on identification of suspicious nodes and links and their subsequent down weighting. Extracting link based features for each node and use various machine learning algorithm to detect spam. Graph regularization techniques for spam detection. In this link information is used to compute global importance scores for all the pages on the web. The link from page pi to page pj shows trust of page pi to page pj. Algorithm follows repeated improvement principle i.e the true score is computed as convergence point of an iterative Updating process.[5]

Algorithms belonging to this category represent pages as feature vectors and perform standard classification or clustering analysis. studies link-based features to perform website categorization based on their functionality. Their assumption is that sites sharing similar structural pattern, such as average page level or number of outlinks per leaf page, share similar roles on web. For e.g web directories mostly consist of pages with high ratio of outlinks to inlinks, form a tree like structure and the number of outlinks increases with the depth of pages while spam site have specific topology aimed to optimize Page Rank boost and demonstrate high content duplication. Overall, each website is represented as a vector of 16 connectivity features and a clustering is performed using cosine as a similarity measure.[4]

4.3 Algorithms that exploits click stream

data and user behavior data, query popularity data or information and HTTP session information. Since click spam aims to push “malicious noise” into a query log with the intention to corrupt data, used for the ranking function construction, most of the counter methods study the ways to make learning algorithms robust to this noise. Other anti-click-fraud methods are driven by the analysis of the economic factors underlying the spammers ecosystem. Interesting idea to prevent click spam is proposed . The author suggests using personalized ranking functions, as being more robust, to prevent click fraud manipulation.[3]

4.4 Sematic based spam detection

To overcome the drawback of content based spam detection , semantic based detection method is used where instead of content semantic of the web site is analyzed.

5. CHALLENGES IN SPAM DETECTION TECHNIQUES

In literature we have studied that many anti- spam strategies have been discovered but still there are some open challenges to these techniques. Some of them are highlighted below:

- In trust modeling system user’s trust tends to vary over time according to the user’s experience and involvement of social networks. Only a few approaches deals with the dynamics of trust by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling in real world application.
- Most of the existing approaches based on text information assuming monolingual environment.
- However social network services are used by people from various countries, so various languages simultaneously appears in tags and comment. In such cases some text information may be regarded as wrong or considered as spam due to language spam. Therefore incorporating multilinguism in trust modelingwould solve this problem.
- It is observed that interaction across social network become popular. For e.g. users can use their Facebook accounts to log in some other social network services. Thus future challenge is to investigate how trust model across domains can be effectively connected and shared.
- Trust modeling most of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis while audio and visual content features of multimedia content can also provide useful information about the relevance of the content and

content tag relation. In future challenge could be to combine multimedia content analysis with the conventional tag processing and user profile analysis.

6. CONCLUSION

From the above study we have studied various spam detection techniques and explored the open challenges spam detection techniques which has to be addressed and left as open challenge for research.

7. REFERENCES

[1] ZoltanGyongi; Hector Garcia-Molina.,"Web Spam Taxonomy ",.First International workshop on Adversarial Information Retrieval on the We (at the 14th International World Wide Web Conference) chiba, Japan, 2005..

[2] Paul Heymann; Georgia Koutrika.," Fighting Spam on Social Website", 2007; IEEE INTERNET COMPUTING; ISSN 1089-7801/07

[3] Nikita Sprin, Jiawei Han, "Survey on web spam detection: principles and algorithms", Proceedings of ACM SIGKDD Explorations newsletter, vol. 13 issue 2, December 2011.

[4] Vijay Krishnan, Rashmi Raj "Web Spam Detection with Anti Trust Rank". - Airweb.cse.lehigh.edu/2006/proceedings.pdf

[5] Jacob Abernethy, Olivier Chapelle, Carlos Castillo., "Graph regularization methods for web spam detaction", springer.com , MachLearn(2010)81:207-225, DOI 10.1007/s10994-010-5171-1, 25th march 2010

[6] Katarzyna Musial, Prezemyslaw, "Social Networks on the Internet", World Wide Web, DOI10.1007/s11280-011-0155-z., 27th December 2011.

[7] Saeed Abu-Nimeh., Thomas M.Chen, "proliferation and detection of Blog Spam", published by IEEE computer and reliability society,1540-7993, september2010.

[8] Zi Chu, Steven Gianvecchio, Haining Wang., "Detecting Automation of Twitter Accounts : Are You A Human, Bot or Cyborg?", IEEE transactions on dependable and secure computing, DOI 10.1109/TDSC.2012.75, pp No 1545-5971/12

[9] Jacob Piskorsike et.al "Exploring Linguistic Features for Web Spam Detection":AIRWeb '08 , Procedings of the 4th international workshop on Adversarial information retrival on the web,pages25-28, ACM, DOI 10.1145/1451983.1451990 ISBN: 978-1-60558-159-0

[10] Win, B,Goel, & Davison, BD Topical Trust Rank using topicality to combat Web Spam.WWW2006.