International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 467 - 470

Survey on Web Spam Detection using Link and Content Based Features

Mr.Rahul C. Patil
Assistant Professor, Dept of Computer Engineering
JESITMR, Nasik
Maharashtra, India
e-mail: rahulpatil0830@gmail.com

Ms.Vishakha R. Bhadane
Assistant Professor, Dept of Computer Engineering
JESITMR, Nasik
Maharashtra,India
e-mail: bhadanevishakha@gmail.com

***Abstract*—** Web spam is one of the recent problems of search engines because it powerfully reduced the quality of the Web page. Web spam has an economic impact because spammers provide a large free advertising data or sites on the search engines and so an increase in the web traffic volume. In this paper we Survey on efficient spam detection techniques based on a classifier that combines new link based features with language models. Link Based features are related to qualitative data extracted from the web pages and also to the qualitative properties of the page links. Spam technique applies LM approach to different sources of information from a web page that belongs to the context of a link in order to provide high quality indicators of web spam. Specifically Detection technique applied the Kullback Leibler divergence on different combinations of these sources of information in order to characterize the relationship between two linked pages.

Keywords - *Language Model, Qualified Link Analysis, Information Retrieval, Web Spam Detection.*

*****

I. INTRODUCTION

Spam is the misuse of electronic messaging systems to send unsolicited bulk messages continuously. Search engine is the dominant method for finding data or content. From the last decade search engines have been the necessary tool for information retrieval. Many People get spam sites when they look for legitimate content or data. Web spam is one of the recent problems of search engines because it powerfully reduced the quality of the result. Web spam has an economic impact because spammers provide a large free advertising data or sites on the search engines and so an increase in the web traffic volume. For this reason it is essential to build an anti spam techniques to get over this problems. Spamming remains financially viable because advertisers have no operating costs beyond the management of their mailing lists and it is difficult to grasp senders accountable for their mass mailings. While the most used form of spam is e-mail spam and web spam. People also get the web search engine spam, blog spam, wiki spam, online classified advertising spam, mobile phone messaging spam, internet spam, social networking spam and file sharing network spam. In internet spam people get the illegal data that they did not ask for and they do not want [1].

This data can be advertisements for products both legitimate and illegal. spam data contains the spiteful code that causes damage to user's computer. While in e-mail spam user get the unsolicited bulk mail. In general terms Link spam, content spam and cloaking these are the three types of web spam. In Link spam illegal or useless links present between the pages which they have no value. In type of web spam it consist of creation of link structure to take advantage of link based ranking algorithm such as page rank which gives a higher ranking to a website the more other highly ranked websites link to it. In content spam illegal data can be present on the internet for advertisements. While In cloaking it is the process of sending different content to a search engine than to a regular visitor of web sites .In type of web spam, content presented to the search engine spider is different to the browser of the user.Here SVM classifier used for the content & link data.For cloaking technique, cost sensitive tree is used.In this cost sensitive tree content & topology informaton combined [2].

In web spam detection technique it take different values for spam and non spam pages.These values are used to implement a classifier which can be able to detect spam pages.Technique build new features to characterize web spam pages while most of using content and link based features to detect spam.To improve the web spam detection technique new qualitative features can be group in two sets. In first set, a group of link based features which check the reliability of links. In second set, a group of content based features extracted with the help of a Language model (LM) approach. Here Detection technique implement an automatic classifier that combines both types of features and they also improve the result of each type separately and those obtained by other proposals [3].

Generally links in non spam pages are described by the corresponding anchor text and its context. Here technique used a number of features which evaluated these differences with the spam pages. Some features are based on the behavior of standard search engines applied to queries composed of pieces of information that pages provide for their links. In general, information associated to a link is given in the URL, anchor text & other context of the link. Here technique introduced other features related to the links which can be in the form of broken links in the page and the presence of links pointing to spam pages [4].

In web spam detection some other sets of features are considered which can be used to evaluate the coherence between a page and the pages it point to. In detection technique it measure the coherence between the degrees of relationship is expected between the information associated to a link in the studied page and the content of page.LM approach is used for this measurement.Language models are probabilistic method which have been developed to calculate linguistic features hidden in the form of texts, words or word sequences in a language.In Language model approaches a set of features are used to extended with new features related to sources of information associated to the URL,achor text,the title page or the meta tags as well as the page pointed to content of the page.It make a language model from every source of data and calculate the Kulback Leibler (KL) divergence between their LM's.Method build a spam classifier which defines a set of features which are the coherence between the different combinations of the considered sources [5].

II. RELATED WORK

Luca becchetti et al., have Proposed Link based Characterization and Detection of Web Spam.In this method they performed a statistical analysis of a large collection of web pages, focusing on Spam Detection.They study several metrics such as degree corelations, number of neighbors and rank propagation through links, trust rank and others to build several automatic Web Spam classifiers.Their work presents a study of the performance of each of these classifiers alone, as well as their combined performance. They have used truncated page rank and probabilistic estimation of the number of neighbors to build an automatic classifier for link spam using several link based features [2].

Chapelle et al. have Proposed a web spam identification through content & hyperlinks methodology.In this method web spam can significantlly detect the quality of search engine relults.Two pages linked by a hyperlink should be topically related even through this this were a weak contextual relation.They had analyzed different sources of information of a web page that belongs to the context of a link & they have applied kullback leibler divergence on them for characterizing the relationship between two linked pages.In this method they present an efficient spam detection technique based on a hybrid clustering that combines k means & SVM and then classified by using C 5.0 with qualified link based features & language model [1].

Benczur et al. have Proposed a Detecting Nepostistic links by language model disagreement method.In this method they proposed several qualitative features to improve web spam detection Technique.They are based on the set of features.This features checks reliability of links & a group of content based features extracted with the help of Language model.Finally They construct an Automatic Classifier that combines both three type of features.In this method they Increase the spam detection rate [3].

Benczur et al. have Proposed a Spam rank fully automatic link spam detection method. In this method spammers intend to increase the page rank of certain spam pages by creating a large number of links pointing to them. They proposed a novel method based on the concept of personalized page rank that detects pages with a undeserved high page rank value without the need of any kind of white or blacklists or other means of human intervention. They assume that spammed pages have a biased distribution of pages that contribute to the undeserved high page rank values [4].

Carlos Castillo et al. have Propsed Web Spam Detection using the web topology Method.In this method they study impact nepotistic link in a web graph whch is in terms of page rank.They Proved bound on the page rank increase that depends both on the reset Probability of the random walk & on the original page rank of the collusion set [7].

Gilad Mishne et al. have Proposed Blocking Blog Spam with Language Model Disagreement Method.They present an approach for detecting link spam common in blog comments by comparing the language models used in the blog post, the comment and pages linked by the comments. They presented an approach for classifying blog comment spam by exploiting the difference between the language used in a blog post and the language used in the comments to that post.Method works by estimating language models for each of these components and comparing these models using well known methods [8].

Hector Garcia Molina et al. have Proposed Web Spam Taxonomy Method. Here they proposed web spamming refers to actions intended to mislead search engines in to ranking some pages higher than they deserve. Their work presents a comprehensive taxonomy of current spamming techniques which they believe can help in developing appropriate counter measures. In this method they presented a variety of commonly used web spamming techniques and organized them in to taxonomy [9].

A. Ntoulas et al. have Proposed a Detecting spam web pages through content analysis.Here they Proposed a methodology using Qualified link analysis.They study on the divergences between the linked pages.In this method they used the C4.5 Classifier algorithm [11].

B.Wu et al. have Proposed Propogating trust & distruct to demote web spam method.In this method they proposed several alternative methods to propogate trust on the web.Here they experiments on a real web dataset.They show that these methods can greatly decrease the number of top portion of the trust ranking.Here method show that combining trust & distruct values can demote more spam sites than the sole use of trust values [12].

F.Javier Ortega et al. have Proposed Propagating Content Based Information through a Web Graph to Detect Spam. Here Spam Web pages have become a problem for information Retrieval systems due to the negative effects that this phenomenon can cause in their results. In this work they tackle the problem of detecting these pages with a propagation algorithm that taking as input a web graph chooses a set of spam likelihood over the rest of the netwok.Thus they take advantage and their spam. Their intuition consist to giving a high reputation to those pages related to relevant ones and giving high spam likelihood of the pages and propagate this information. Their graph based algorithm computes two scores for each node in the graph [13].

L. Araujo et al. have Proposed Web Spam Detection based on Qualified link analysis & language models.Here they present an efficient spam detection method based on a classifier that combines new link based features with language model.They apply an LM approach to different sources of information from web pages.Here they used the Kullback Lieblesr Divergence on different combinations of features [14].

III. METHODOLOGY

In web spam detection method Qualified link Features, Language model based method and Meta tags used for the building more accurate classifier.

A. Qualified Link Analysis

In Qualified link analysis nepotistic links can be find out. These nepotistic links are present for reasons other than merit. Here parameters of page links can be studied.Parametrs of page links can be find out such as testing if they are broken or measuring the difference between internal and external links or between outgoing and incoming lnks.others refer to the anchor text whether it is just a URL, a number, a punctuation mark or even just an empty chain [4].

Other parameters are related to the different aspects of the coherence between a link and a pointed page and between the pages containing the link.For study the parameters we have developed an information system.By using information retrieval system it provide us with a quality factor from every page which is represented by a set of features by a set of features about its links. In this analysis it analyzes the Web links, broken links, Incoming Outgoing links, External Internal links and anchor text topology .firsly feature analyzed the web links.Information retreival system analyzes the links in a page and extracts several features from that page.The technique not only offers information about the number of links whose pointed page can be recovered using information from the link and the page that contains it,but also data about every link.Technique based on classical information retrieval techniques and natural language processing and it mainly consists of two stages [5].

In web spam detection method Qualified link Features, extract relevant information on a link.Here they used the anchor text as the main source of information to recover a link.Method construct an complex queries and request to a search engine.The Original query is composed of the terms extracted from the anchor text and this query is expanded using the terms extracted from the other sources of information considered.The all expanded queries are submitted to the selected search engine and the top ranked documents are retrieved.In this method it consider that a link has been recovered if the page pointed by the link is in the set of pages retrievd with some of the queries [6].

B. Language Model

Language model method based on the distribution analysis which uses the KL divergence.KL divergence is used to compute the divergence between the probability distributions of terms of two particular documents considered.KL divergence is applied to measure the difference between two text units of the source and target pages. KL divergence characterizes the relationship between two linked web pages according to different values of divergence. For calculating divergence source of information from the source page is used.In web spam detection technique anchor text, URL terms and Internal & External Links sources of information considered [7].

In detection technique three sources of information consider from the source page.Anchor text is first source of informtaion consider here.The main function of anchor text is when a page links to another this page has only a way to convince a user to visit this link that is by showing relevant and summarizes information of the target page.Due to this function a great divergence between this piece of text and the linked page shows a clear evidence of spam.In this LM based features Surrounding Anchor text is second Source of information.In this Source of information Sometimes anchor text provide small or no descriptive value.Due to this Text surrounding a link can provide contextual information about the pointed page.In this LM based features they used several words around the anchor text to extend it.Here URL terms is last feature of the LM based features.URL information available of a link in it.A URL is made up of a protocol, a domain, a path, and a file.These component provide more information from the target page [8].

C. Meta Tags

Meta tags provide structured data about a web page.Meta tags are used in Search engine optimization.From the long time, Meta tags are the target of the spammers.The less and less pages are considered by the search engines.In spam detection technique, it considered the attributes “Description” and “Keywords” from the meta tags to build a virtual document with their terms.They decided to use these data to calculate its divergences with other sources of information from the source page.The source page is in the form of anchor text, surrounding anchor text and URL terms.Most of the sources of information used these meta tags to measure the divergence between two web pages.Many source of information are considered here.Combination of sources of information is one of most important source of information.In this source of information it creates an virtual documents which provide more information.Hre two types of new sources of information combined.first in between the anchor text and URL terms.while in second combination it considered the Surrounding anchor text and URL terms.In this method other source of information also used for spam detection.Content page, Title and meta tags are the another source of information [9].

IV. FEATURE EXTRACTION

A. Content Based Feature

For building more accurate classifier, content based features are used .In detection method content based features are combined with the link based features. Due to this, Detection technique obtained a good quality of spam detection method. Anchor text, surrounding anchor text URL terms and Meta tags these are the main component of the Content based features. .The main function of anchor text is when a page links to another this page has only a way to convince a user to visit this link that is by showing relevant and summarizes information of the target page. Due to this function a great divergence between this piece of text and the linked page shows a clear evidence of spam. Meta tags provide structured data about a web page. These Meta tags are used in Search engine optimization. From the long time, Meta tags are the target of the spammers. The less and less pages are considered by the search engines. In this spam detection technique, it considered the attributes “Description” and “Keywords” from the Meta tags to build a virtual document with their terms [10].

B. Link Based Feature

In Link based features, it contains the extraction of web links. Detection technique considers the Incoming Outgoing, External Internal & Broken Links. In spamming spam pages link to non spam pages but non spam pages do not link to spam pages. Detection method analyzed the how many sites point to the Incoming links. In web spam detection technique it takes an External & Internal features. It represents the rate of these two types of links for spam and non spam pages. Broken links is another component of the link based features. Broken links are the common problem for spam and non spam pages. Link based features are more important for the web spam detection technique. It combined with Content based feature, LM based feature & QL based feature to get the more efficient spam detection technique [11].

COCLUSION

In this paper we Survey on a spam detection technique based on classifier that combines new link based features with language model based.Web spam detection technique is based on the analysis of QLs and a study of the divergence between linked pages. In Web spam detection it can analyzed the relationship between a page and those that point to it. Language model method based on the distribution analysis which uses the KL divergence.KL divergence is used to compute the divergence between the probability distributions of terms of two particular documents considered.

REFERENCES

[1] J. Abernethy, O. Chapelle and A. C. Castillo,"Web Spam Identification Through Content and Hyperlinks". *In International workshop on adversarial Information retrieval on the web*, China, pp. 41-44 , 2008.

[2] L. Becchetti, C. Dastillo, D .Donato, S. Leonardi and R.Baeza, " Link Based Characterization and Detection of Web Spam", *In Second Workshop On Adversarial Information Retrieval On the Web*,Seattle, pp. 1-8, 2006.

[3] A. A. Benczur,I.Brio,K. Csalogany and M. Uher,"Detecting Nepotistic Links by Language Model Disagreement", *In Proceedings of 15th International Conference World Wide Web*, New York, pp. 939-940, 2006.

[4] A. Beczur, K. Csalogany, T. Sarlos and M. Uher , "Spam Rank Fully Automatic Link Spam Detection". *In the 1st International Workshop On adversarial Information Retrieval On the Web*, Japan, pp. 25-38, 2005.

[5] K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment", *In Proceedings of 21st Annual International Conference On Research and Development In Information Retrieval*, New York, pp. 104-111, 1998.

[6] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini and S. Vinga , " A Reference Collection for Web Spam", *IGR forum*, Vol 40 , 2006.

[7] C. Catillo, D. Donato, A. Gionis, V. Murdock and F. Silverstri, "Web Spam Dtection Using the Web Topology", *In Proceedings of 30th Annual International Conference on Research and Development In Information Retrieval*, New York, pp. 423-430 , 2007.

[8] G. Msihne, D. Carmel and R. Lempel, "Blocking Blog Spam with Language Model Analysis", *In the proceedings of 1st Annual International Workshop On Adversarial Information Retrieval On the Web*, Japan,pp. 1-6, 2005.

[9] Z. Gyongyi and H. G. Molina , " Web Spam Taxonomy",*In the Proceedings of the 1st International Workshop On Adversarial Information Retrieval On the Web*, pp. 222-231, 2005.

[10] R. Jin, A. Hauptmann and C. Zhai, " Title Language Model for Information Retrieval", *In the Proceedings of the 25th Annual International Conference on Research and Development In Information Retrieval*, New York, pp. 42-48,2002.

[11] A. Ntoulas, M. Najork, M. Manasse and D. Fetterly, " Detecting Spam Web Pages Through Content Analysis", *In the Proceedings of the 15th Annual International Conference*,New York,pp. 83-92, 2006.

[12] B. Wu, V. Goel and Brian , " Propogating Trust & Distruct to Demote Web Spam", *In the proceedings of the 1st International Workshop*, Bethlehem, pp. 42-57. 2005.

[13] A. Javier Ortega, J. A. Troyano, F. L. Cruz and C. G. Vallejo, " Propogating Content Based Information Through a Web Graph to Detect Web Spam"' *In International Journal of Innovative Computing*, Vol 8, 2012.

[14] Lourdes Araujo and Juan Martinez Romo, " New Classification Features Based on Qualified Link Analysis and Language Models", *In IEEE Transactions on Information Forensics and Security*, pp. 581-590, 2010.