Title

Classification of Malicious Web Pages through a J48 Decision Tree, aNaïve Bayes, a RBF Network and a Random Forest Classifier forWebSpam Detection

Journal
International Journal of u- and e- Service, Science and Technology, 10(4)

Author
Alam Kazmi, Syed Hasnain

Publication Date
2017-04-28

Peer reviewed

Classification of Malicious Web Pages through a J48 Decision Tree, a Naïve Bayes, a RBF Network and a Random Forest Classifier...

Article · April 2017

DOI: 10.14257/ijunesst.2017.10.4.05

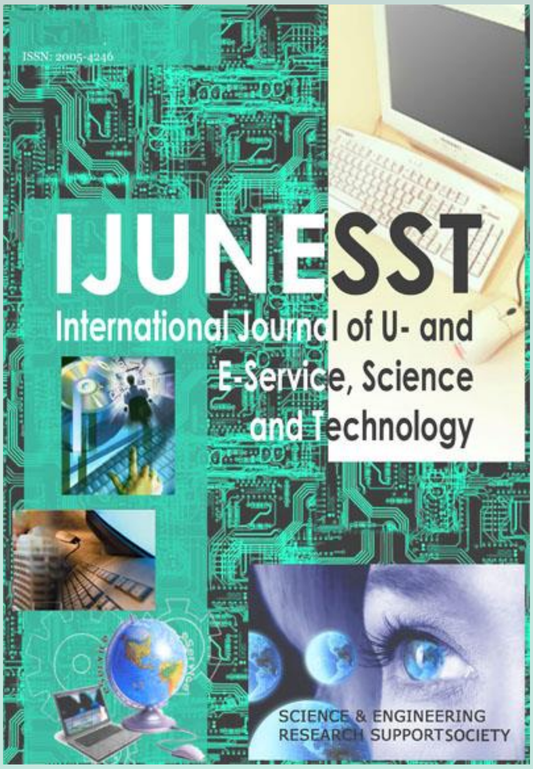
Muhammad Iqbal
Bahria University Karachi Campus

Malik Muneeb Abid
International Islamic University, Islamabad

Syed Hasnain Alam Kazmi
Southwest Jiaotong University

Classification of Malicious Web Pages through a J48 Decision Tree, a Naïve Bayes, a RBF Network and a Random Forest Classifier for WebSpam Detection

Muhammad Iqbal, Malik Muneeb Abid, Usman Waheed and Syed Hasnain Alam Kazmi


ISSN: 2005-4246
IJUNESST
International Journal of U- and E-Service, Science and Technology
SCIENCE & ENGINEERING RESEARCH SUPPORT SOCIETY

ISSN 2005-4246
Volume 10, Number 4, 2017

International Journal of u- and e- Service, Science and Technology
Vol. 10, No. 4 (2017), pp.51-72
http://dx.doi.org/10.14257/ijunesst.2017.10.4.05

SERSC
Science & Engineering Research Support soCiety

International Journal of u- and e- Service, Science and Technology
Vol.10, No.4 (2017), pp.51-72
http://dx.doi.org/10.14257/ijunesst.2017.10.4.05

Classification of Malicious Web Pages through a J48 Decision Tree, a Naïve Bayes, a RBF Network and a Random Forest Classifier for WebSpam Detection

Muhammad Iqbal[1, 3], Malik Muneeb Abid[2,*], Usman Waheed[3] and Syed Hasnain Alam Kazmi[4]

[1]*School of Information Sciences and Technology, Southwest Jiaotong University, Sichuan, Chengdu, PR China*
[2]*Department of Civil Engineering, Mirpur University of Science and Technology, AJK, Pakistan*
[3]*Department of Computer Science, Bahria University, Karachi, Pakistan.*
[4]*Department of Management Sciences, Muhammad Ali Jinnah University, Karachi, Pakistan*
**Correspondence: muneeb.ce@must.edu.pk; Tel.: +0092-3015552846*

Abstract

Web spam is a negative practice carried out by spammers to produce fake search engines results for improving rank position of their Web pages. It is available on arena of World Wide Web (WWW) in different forms and lacks a consistent definition. The search engines are struggling to eliminate spam pages through machine learning (ML) detectors. Mostly, search engines measure the quality of websites by using different factors (signals) such as, number of visitors, body text, anchor text, back link and forward link etc. information and, and spammers try to induce these signals into their desired pages to subvert ranking function of search engines. This study compares the detection efficiency of different ML classifiers trained and tested on WebSpam UK2007 data set. The results of our study show that random forest has achieve higher score than other well-known classifiers.

Keywords: *Web Spam, Spam Classification, Supervise Machine Learning*

1. Introduction

The internet is a global system of TCP/IP based networks and its applications can be seen across the world; such as, business, education, science, teleconference, telemedicine, video-on-demand and online gaming *etc.* [1]. The modern era demands everyone to know the benefits of this communication infrastructure and get maximum opportunities from this offering. The World Wide Web ("WWW" or simply the "Web") is a way of accessing information from internet. Web is an information-sharing model based on graph theory that is built on top of the internet. In short, the Web has given an opportunity to different domains to access wider global audience. Websites, Online Social Networking (OSN) [3], Blogs and forums *etc.* are the major tools to approach target people.

The size of web, which is now believed to be a largest repository ever built, can be described by different factors; such as penetration rate among users, the size of indexable web *etc.* Everyday a fraction of this large repository is crawled and index by different search engines including Google. Authors reported that Google crawl and index more than 45 billion pages on January 2015 [2]. DMR [41], a digital marketing company revealed that Google processes an average of 2.3 million users' queries every second. Furthermore, the growth of www has increased substantially over the last two decades because of technological boosts in communication infrastructure, the miniaturization of

ISSN: 2005-4246 IJUNESST

electronic devices, and the web have led to the growth of data at an astounding rate. The internet is a global system of TCP/IP based networks and its applications can be seen across the world; such as, business, education, science, teleconference, telemedicine, video-on-demand and online gaming *etc.* [1]. The modern era demands everyone to know the benefits of this communication infrastructure and get maximum opportunities from this offering. The World Wide Web ("WWW" or simply the "Web") is a way of accessing information from internet. Web is an information-sharing model based on graph theory that is built on top of the internet. In short, the Web has given an opportunity to different domains to access wider global audience. Websites, Online Social Networking (OSN) [3], Blogs and forums *etc.* are the major tools to approach target people.

The size of web, which is now believed to be a largest repository ever built, can be described by different factors; such as penetration rate among users, the size of indexable web *etc.* Everyday a fraction of this large repository is crawled and index by different search engines including Google. Authors reported that Google crawl and index more than 45 billion pages on January 2015[2]. DMR [41], a digital marketing company revealed that Google processes an average of 2.3 million users' queries every second. Furthermore, the growth of www has increased substantially over the last two decades because of technological boosts in communication infrastructure, the miniaturization of electronic devices, and the web have led to the growth of data at an astounding rate. This makes www as major avenue of people for their social & business lives

The search engines plays a pivotal role for people to retrieve useful contents from large data space of Web, and generally website owners are only interested in good ranked results in the first several pages of Search Engine Result Page (SERP).A vast majority of internet users rely on search engines to retrieve their desired information. Simply the search engines are the keys to find specific information on the vast expanse of the www, but most of the time users receive undesired results against their queries. Search engines create an index of search database and match this index with user generated search queries.

Mostly, website owner's burning desire is to present their contents on top of SERP. The SERP listing result is an outcome of search engine ranking function execution and keyword query submitted by the user. The visibility of specific website in SERP is heavily dependent on website traffic.

Search engines carry out following three important tasks to manage web contents on www (see Figure 1):

1. Web spider crawls the web to discover the location of web pages.
2. Index or record the web links (hyperlinks) for fast traversing of users queries.
3. Assign ranking score to web pages in the database, which ultimately reflects on SERP.

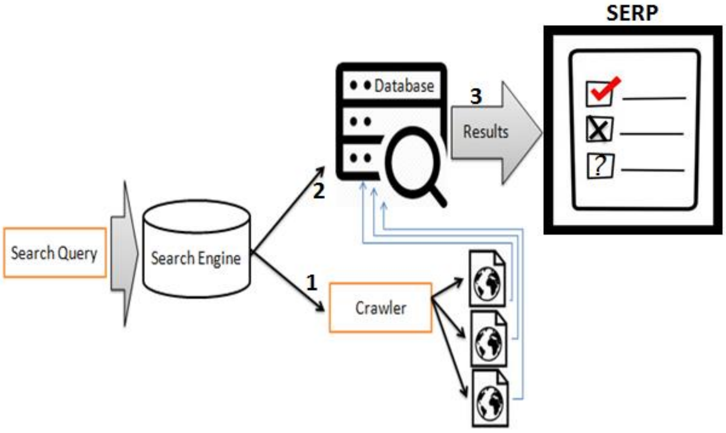


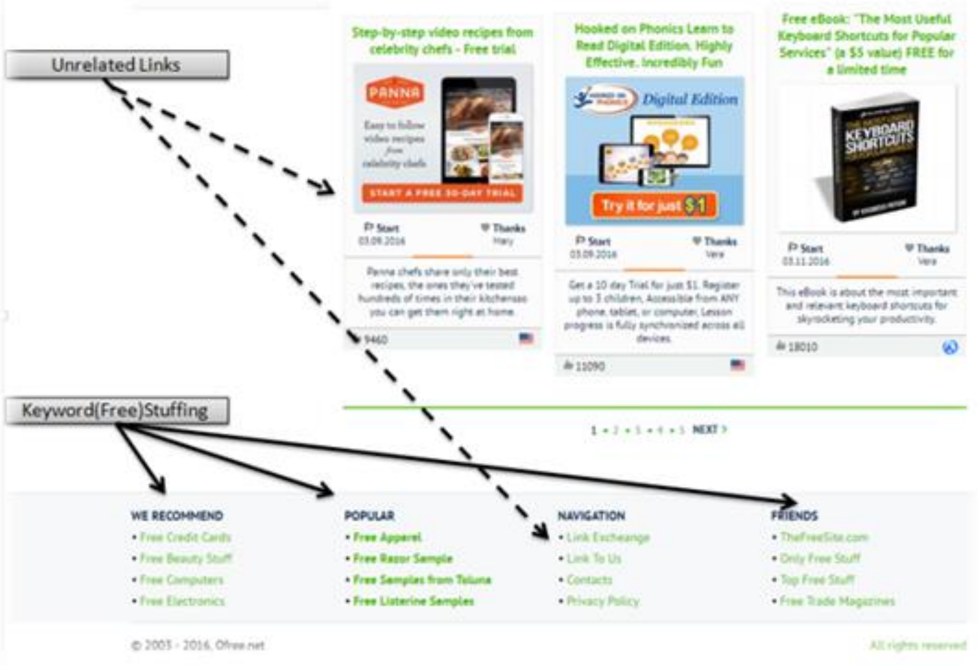
Figure 1. Query to Search Engine Return Model

SERP mostly contains two types of results *i.e*. organic and inorganic *i.e*. paid. Search engine ranking algorithms produces listing of top rank web pages in chronically order in SERP.

Top rank result page means the probability increases to bring more internet traffic to website and in return the site owner will earn more profit. Mostly, these economic incentives triggered the website administrators to subvert the ranking results of search engines through unethical Search Engine Optimization (SEO) techniques [28]. This kind of negative efforts of web site developers to manipulate their web pages to attract a large number of internet users is known as web spam or "spamdexing". Figure 2 is an example of web spam, where page is suffered with keyword stuffing and unrelated links.

SERP can be thought as "first impression opportunity" for website owners to display their contents in front of very large Web audience and this also compelling website creators to use spamming techniques. Pen *et al*. [5] reported in their research that mostly end-users pay attention to top rank result pages of SEs. In another study, which was conducted by Wang *et al*. [27], they reported that 56.6% of internet users only pay attention to the first two pages of SERP.

The increasing use and role of search engines (*e.g*. Google, Yahoo!, Bing, Baidu *etc*.) in information retrieval has made companies and web site developers concern about the ranking of their web sites. Hence, it is extremely important for search engines to filter out spam pages to keep their indices clean and only hold quality web pages information. Currently developing an effective spam detection solution is a challenging task for research community along with search engine companies and other stake holders of Web.



Step-by-step video recipes from celebrity chefs - Free trial

Panna chefs share only their best recipes, the ones they've tested hundreds of times in their kitchensso you can get them right at home

Hooked on Phonics Learn to Read Digital Edition. Highly Effective. Incredibly Fun

Get a 10 day Trial for just $1. Register up to 3 children. Accessible from ANY phone, tablet, or computer. Lesson progress is fully synchronized across all devices.

Free eBook: "The Most Useful Keyboard Shortcuts for Popular Services" (a $5 value) FREE for a limited time

This eBook is about the most important and relevant keyboard shortcuts for skyrocketing your productivity.

WE RECOMMEND
- Free Credit Cards
- Free Beauty Stuff
- Free Computers
- Free Electronics

POPULAR
- Free Apperel
- Free Razor Sample
- Free Samples from Toluna
- Free Listerine Samples

NAVIGATION
- Link Excheange
- Link To Us
- Contacts
- Privacy Policy

FRIENDS
- TheFreeSite.com
- Only Free Stuff
- Top Free Stuff
- Free Trade Magazines

© 2003 - 2016, 0free.net   All rights reserved

Figure 2. An Example of Web Spam Page

The objective of the study is to analyze the efficiency of different ML approaches in detecting web spam pages. We have adopted several link, content and obvious features to distinguished web spam from non-spam context. We also believe that such features could be common for the WEBSPAM–UK2006 and WEBSPAM–UK2007 data sets. To evaluate this hypothesis we created initially web spam detector by using WEBSPAM–UK2007 dataset. The corpus was retrieved from the Laboratory of Web Algorithmics, "Universit degli Studi di Milano", with the support of the DELIS EU - FET research project. We train and test our classifier though k-fold (k=10) Cross Validation (CV) scheme. Figure 3 shows the visual procedure of cross validation.

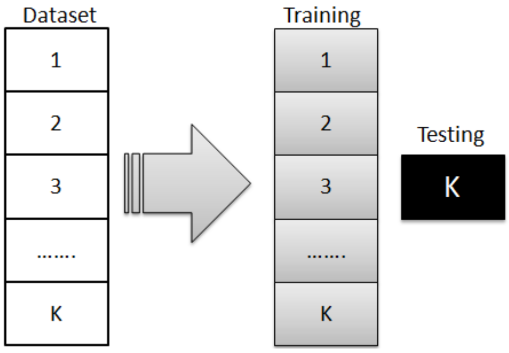


Figure 3. 10-cross Validation Scheme

The aim of selecting CV scheme is not to estimate parameters but to estimate the generalization performance and to bring stability in our learning model.

Low bias and low variances are good attributes for selection of estimation method. In k fold cross validation scheme the dataset is divided into k subsets of (approximately) equal size. The training is performed on k-1 times, each time leaving out one of the subsets from training and that omitted subset will be used for testing classifier accuracy. The main contribution of our work can be summed up as: (i) Using the different ML algorithms in spam detection and comparing and analyzing results and (ii) We applied Chi square test to collect most authoritative features to correctly classify web pages.

The organization of the paper is as follows. Section 2 discusses the background & related work. Section 3 provides the details of widely used ML algorithms. Section 4 describes the performance evaluation and parameters settings of our detection model. Section 5 evaluates the proposed approach and Section 6 finally concludes the paper and presents the future work.

The search engines plays a pivotal role for people to retrieve useful contents from large data space of Web, and generally website owners are only interested in good ranked results in the first several pages of Search Engine Result Page (SERP).A vast majority of internet users rely on search engines to retrieve their desired information. Simply the search engines are the keys to find specific information on the vast expanse of the www, but most of the time users receive undesired results against their queries. Search engines create an index of search database and match this index with user generated search queries.

Mostly, website owner’s burning desire is to present their contents on top of SERP. The SERP listing result is an outcome of search engine ranking function execution and keyword query submitted by the user. The visibility of specific website in SERP is heavily dependent on website traffic.

Search engines carry out following three important tasks to manage web contents on www (see Figure 1):

4. Web spider crawls the web to discover the location of web pages.
5. Index or record the web links (hyperlinks) for fast traversing of users queries.
6. Assign ranking score to web pages in the database, which ultimately reflects on SERP.

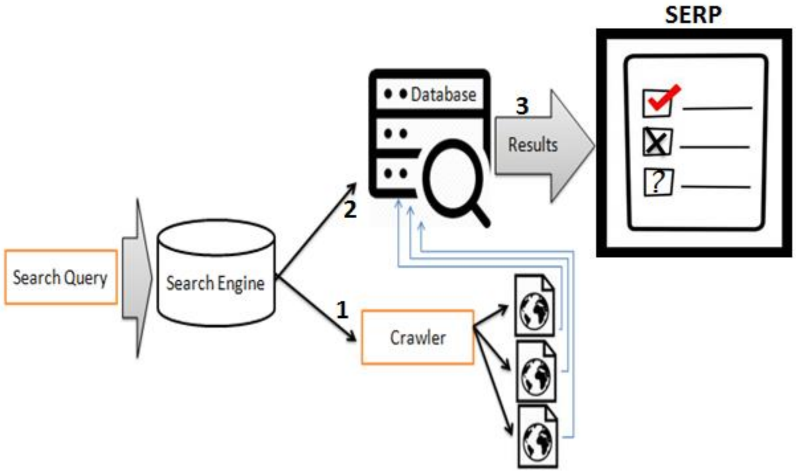


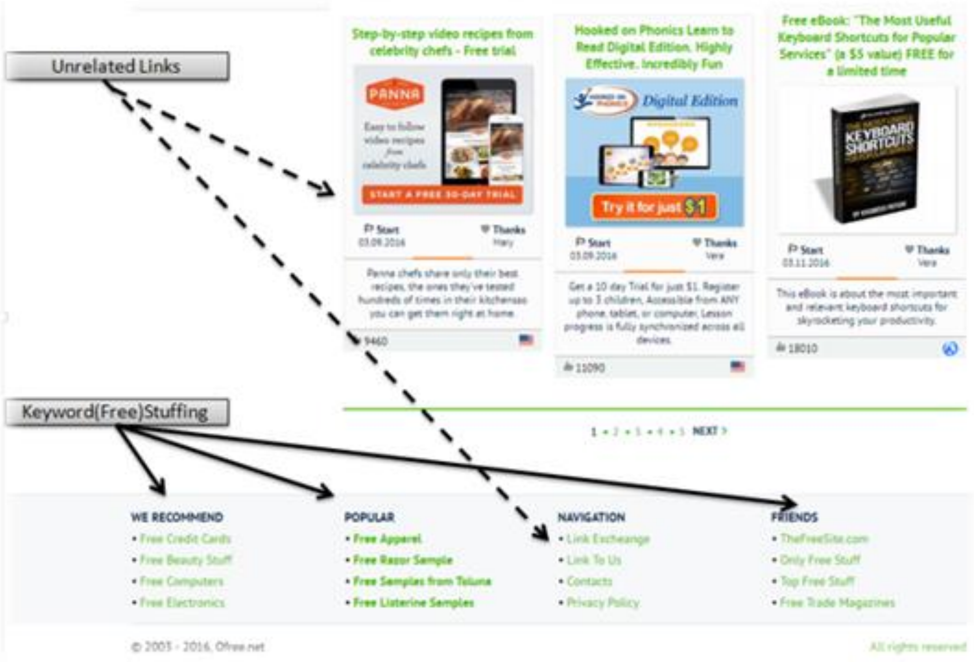
Figure 1. Query to Search Engine Return Model

SERP mostly contains two types of results *i.e*. organic and inorganic *i.e*. paid. Search engine ranking algorithms produces listing of top rank web pages in chronically order in SERP.

Top rank result page means the probability increases to bring more internet traffic to website and in return the site owner will earn more profit. Mostly, these economic incentives triggered the website administrators to subvert the ranking results of search engines through unethical Search Engine Optimization (SEO) techniques [28]. This kind of negative efforts of web site developers to manipulate their web pages to attract a large number of internet users is known as web spam or "spamdexing". Figure 2 is an example of web spam, where page is suffered with keyword stuffing and unrelated links.

SERP can be thought as "first impression opportunity" for website owners to display their contents in front of very large Web audience and this also compelling website creators to use spamming techniques. Pen *et al*. [5] reported in their research that mostly end-users pay attention to top rank result pages of SEs. In another study, which was conducted by Wang *et al*. [27], they reported that 56.6% of internet users only pay attention to the first two pages of SERP.

The increasing use and role of search engines (*e.g*. Google, Yahoo!, Bing, Baidu *etc*.) in information retrieval has made companies and web site developers concern about the ranking of their web sites. Hence, it is extremely important for search engines to filter out spam pages to keep their indices clean and only hold quality web pages information. Currently developing an effective spam detection solution is a challenging task for research community along with search engine companies and other stake holders of Web.



Step-by-step video recipes from celebrity chefs - Free trial

PANNA
Easy to follow video recipes from celebrity chefs
START A FREE 30-DAY TRIAL

Start 03.09.2016 · Thanks Mary

Panna chefs share only their best recipes, the ones they've tested hundreds of times in their kitchens so you can get them right at home.

9460

Hooked on Phonics Learn to Read Digital Edition. Highly Effective. Incredibly Fun

Digital Edition
Try it for just $1

Start 03.09.2016 · Thanks Vera

Get a 10 day Trial for just $1. Register up to 3 children. Accessible from ANY phone, tablet, or computer. Lesson progress is fully synchronized across all devices.

11090

Free eBook: "The Most Useful Keyboard Shortcuts for Popular Services" (a $5 value) FREE for a limited time

KEYBOARD SHORTCUTS

Start 03.11.2016 · Thanks Vera

This eBook is about the most important and relevant keyboard shortcuts for skyrocketing your productivity.

18010

1 • 2 • 3 • 4 • 5 NEXT >

WE RECOMMEND
- Free Credit Cards
- Free Beauty Stuff
- Free Computers
- Free Electronics

POPULAR
- Free Apparel
- Free Razor Sample
- Free Samples from Toluna
- Free Listerine Samples

NAVIGATION
- Link Excheange
- Link To Us
- Contacts
- Privacy Policy

FRIENDS
- TheFreeSite.com
- Only Free Stuff
- Top Free Stuff
- Free Trade Magazines

© 2003 - 2016. 0free.net · All rights reserved

Figure 2. An Example of Web Spam Page

The objective of the study is to analyze the efficiency of different ML approaches in detecting web spam pages. We have adopted several link, content and obvious features to distinguished web spam from non-spam context. We also believe that such features could be common for the WEBSPAM–UK2006 and WEBSPAM–UK2007 data sets. To evaluate this hypothesis we created initially web spam detector by using WEBSPAM–UK2007 dataset. The corpus was retrieved from the Laboratory of Web Algorithmics, “Universit degli Studi di Milano”, with the support of the DELIS EU - FET research project. We train and test our classifier though k-fold (k=10) Cross Validation (CV) scheme. Figure 3 shows the visual procedure of cross validation.

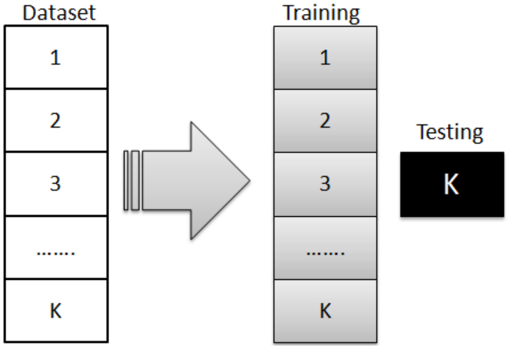


Figure 3. 10-Cross Validation Scheme

The aim of selecting CV scheme is not to estimate parameters but to estimate the generalization performance and to bring stability in our learning model.

Low bias and low variances are good attributes for selection of estimation method. In k fold cross validation scheme the dataset is divided into k subsets of (approximately) equal size. The training is performed on k-1 times, each time leaving out one of the subsets from training and that omitted subset will be used for testing classifier accuracy. The main contribution of our work can be summed up as: (i) Using the different ML algorithms in spam detection and comparing and analyzing results and (ii) We applied Chi square test to collect most authoritative features to correctly classify web pages.

The organization of the paper is as follows. Section 2 discusses the background & related work. Section 3 provides the details of widely used ML algorithms. Section 4 describes the performance evaluation and parameters settings of our detection model. Section 5 evaluates the proposed approach and Section 6 finally concludes the paper and presents the future work.

2. Background and Related Work in Web Spam Filtering

Web spam has shown its existence since the start of www and has been growing with fast pace due to economic incentives gained by spammers on this exponentially growing platform. However, the discussion about web spam in research community is quite recent. In the beginning, this harmful phenomenon was only limited to e-mail but with the passage of time and development of web technologies, spammers began to use this concept on different domains for making illegal profit.

The importance of spamdexing and quality of results against user’s queries to the search engines was discussed by Henzinger *et al*. [7]. Gyongyi and Garcia-Molina [17] suggested taxonomy of Web Spam pages and proposed TrustRank algorithm. Algorithm first selects a certain number of good seeds (pages) for experts’ manual evaluation and then propagates to other page by linking to them. The philosophy of TrustRank is that good pages seldom link to bad pages. Wu *et al*. [8] presented topical TrustRank, which improves TrustRank method by employing topical information. Wu and Davison [46] introduced the parent penalty algorithm to identify link farm spam pages by propagating negative values.

James Caverlee *et al*. [44-45] have discussed the importance of source-centric link analysis, such as source size, the presence of self-links and they developed a novel credibility-based Web ranking technique with the name CredibleRank, which integrates credibility information directly into the quality assessment of each page on the Web.

Most of the research in this domain focuses on some of the main types of web Spam *i.e*. Content, Cloaking, Click and Link Spam [42-46].Content spamming is believed to be the first web spam technique which was used to subvert the ranking of search engines. It is favorite spamming method for spammers because of the fact that most search engines apply the information retrieval models based on a page content to rank web pages, such as a vector space model [9], BM25 [10], or statistical language models [11].Hence, spammers analyze the weaknesses of these models and exploit. However, Ntoulas *et al*. [35] reported in their experimental work by using decision tree (DT) that 82-86% of spam pages of these characters can be detected by ML classifier. Moreover, the authors introduce various new content-based features for web spam detection. Castillo *et al*. [53] applied combination of link-based features and content-based features and obtained 88.4% of spam detection rate with 6.3% false positive using DT.

Nowadays, spammers are also targeting link based ranking algorithms, such as PageRank [15] and HITS[16] to subvert ranking of search engines because of the fact that they also consider web link structure information along with content based relevance metrics to measure ranking score of web page. A web page that participates in a link farm mostly has a high in-degree, but little relationship with the rest of the web (graph). Spammers can also receive web traffic in terms of http links from ham pages by buying advertising, or through buying expired domains which were used previously for non-spam sites. Link-based ranking algorithms, such as PageRank and HITS are the main targets for spammers. These algorithms can be considered as computing applications of various Markov chain processes over Web pages to rank their scores. Becchetti *et al*. [13] studied several link-based metrics, such as rank propagation for hyperlinks and probabilistic counting to improve the Web spam detection methods. Different researchers [42, 43] use the WEBSPAM–UK2006 and WEBSPAM–UK2007 data sets to classify web spam pages. We have also utilized WEBSPAM–UK2007 dataset to evaluate the performance of our algorithm.

2.1 Size of Problem & Existing Solutions

Web spamming is a widespread problem and it continuous to get worse. Search engines that fall victim to web spamming can end up losing a large pool of users. As a matter of fact we know that it is nearly impossible for human experts to manually classify ham and spam pages, so it has become a challenging task for researchers to improve and introduce new ways in currently available web spam detection algorithms. Presently, a large faction of web site creators employed aggressive black hat SEO tactics to achieve top position in search results. This speculative phenomenon is not made by anonymous people, but unfortunately a number of organizations acquire services of highly skilled personnel. Prieto *et al*. [29] discussed two important reasons that why spammers are interested to get involved in spamming activities :(i) Increase the ranking score in order to receive a top position and raise their income, (ii) Damage the business of competitor companies.

Ghiam and Nemaney [31] reported three prominent reasons to detect and control the web spam traffic from web: (i) web spam pages are destructive for both search engines and the victim's machines, (ii) mostly web spam pages waste visitors' precious time and this may cause adverse effects on search engine results, and finally (iii) web spam pages misuse significant resources of search engines.

Wang *et al*. [3] reported that one seventh web pages of English websites were identified as spam pages. Statistics reveal that due to presence of spam pages companies earning suffered more than US$100 billion globally [50]. It has been reported that a single spamming bot-net was approximately $2M per day [51]. In short, web spamming drops the workers' productivity directly or indirectly and cost implications of this phenomena show a very gloomy picture too.

Web spamming exists on web in various forms (see Figure 4) and lacks a consistent definition; but one attribute is common among spammers that they strive to earn top ranking of their websites so that they can attract large number of free advertisements. Web spam detection mechanism can be regarded as a binary classification problem, where a Machine Learning (ML) classifier is utilized to predict ham or spam web pages [30]. The adaptive learning capability to learn underlying patterns makes ML algorithms correct solution to detect spam pages.

At present, there are two approaches to combat web spam: (i) Web spam detection, where spam pages eliminated from search engine index, and (ii) spam demotion, to punish web-spam pages by demoting them in the search result ranking. A number of ML algorithms (decision-tree based classifiers *e.g.*, C4.5, SVM-based classifiers, Bayesian classifiers *etc.*) are being used to marked spam or ham traffic. In this paper, we have analyses the performance of different machine learning algorithms for Web spam detection. Decision Tree (C4.5) [4] and support vector machine (SVM) [52] are two widely used approaches among the adversarial information retrieval community. However, there are so many ML algorithms are being employed to handle web spam problem but we have not covered all.

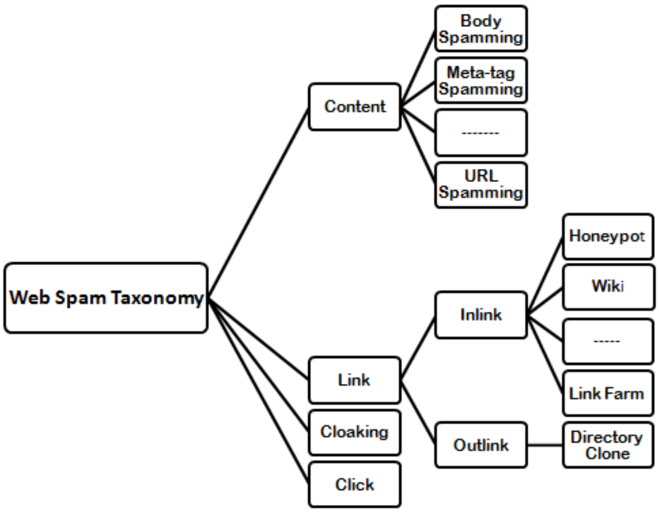


Figure 4. Spam Existence on Web in Different Forms

2.2. Tools and Heuristics Applied in Spam Attacks

In simplest form of web spamming *i.e.* content web spamming, a page is dwelt with irrelevant contents to improve its popularity. Content spamming targets the different data fields of web page; for example body, title, meta tag, anchor text or URL[17]. Spammer also utilizes dumping of unrelated terms and phrase stitching methods to achieve their objectives.

In link spamming the spammers manipulate the link structures of the web sites, by employing different techniques, such as creating link farms [14]. Link based spamming can be grouped into outgoing and incoming links. Outgoing link structure manipulation refers to add numerous outgoing links to popular pages, while the incoming link structure utilizes the link farm concept. A link farm is a densely connected set of pages and all are pointing to a single intended (target) page. The purpose of creating link farm is to betray link-based ranking algorithms. Since many search engines take into account the number of incoming links in ranking pages as an important parameter, ultimately the rank of the target page is likely to increase, and appear earlier in query result page. A schematic diagram showing the normal and link farm structure is indicated in Figure 5, where Figure 5a depicting normal link structure of web and 5b is an example of link farm topology. Drastic increase of in-links in page 5 can be observed from Figure 5b and with passage of time this structure gets denser

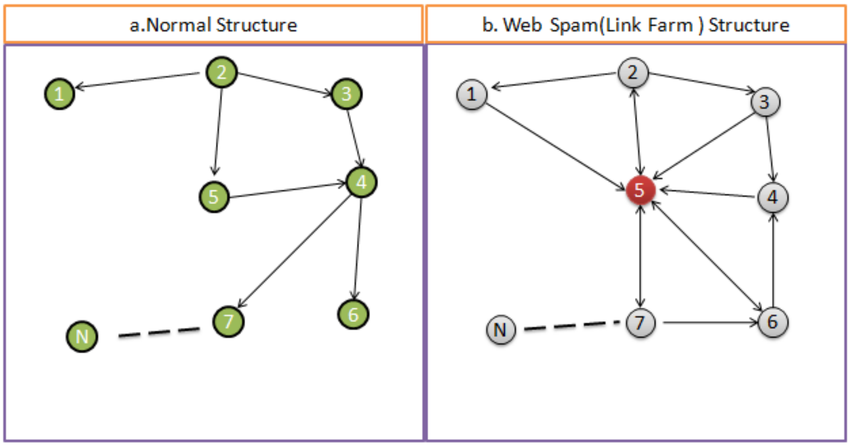


Figure 5. Schematic Depiction of Link Farm

Cloaking or hiding based spam delivers different content to search engines and end users receive entirely different contents from spammers [6]. Finally, click spam refers to the method of submitting queries to SE that retrieve target result pages and then ‘‘click’’ on these pages in order to simulate user interest in their content [30].

Most of these spamming methods like: Click, Cloaking, link farming, and keyword stuffing [17] are being succeeded in lots of cases to betray the ranking algorithms adopted by different search engines. The success of spamming techniques to betray a search engine yields non-relevant results to the query, and this hurts the reputation of search engine. This also frustrates the users and in many cases majority switches to another search engines.

Mostly the spammers tune the ranking function of search engines to receive good position in SERP by utilizing their excellent web engineering skills. For example, spammers mislead search engines by applying content spamming through forging of TFIDF score in their web sites [12]. Indeed, it is very difficult to distinguish between “ethical” and “un-ethical” SEO services because of a large gray area exists between black-hat (un-ethical) and white-hat (ethical) [13].

3. Methodology

In this section, we have discussed the famous ranking functions of search engines. This section further presents the four famous machine learning algorithms which are described and evaluated in this study. These are decision tree (DT), Naïve Bayes (NB), Random Forest (RF), and RBF Network (RBFN).

3.1 Famous Ranking Algorithms to Rank Web Pages

We now present the three most famous ranking algorithms and establishes how spammers attempt to deceive these three algorithms to obtain the best possible rank for the spammed Web pages in the SERP.

The Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF stands for term frequency-inverse document frequency, and is a numerical statistic method, often used in information retrieval and text mining. Through this technique we can evaluate the importance of word in a document or in a collection of documents (corpus).Different studies [18-19] shows that TF-IDF is an efficient and simple algorithm for matching text words in a query to documents that are relevant to that query.

Typically, the TF-IDF weight is composed by two terms, *i.e*. TF and IDF.

Equation (1) shows the formula to calculate importance of word.

$$\mathrm{tf}-\mathrm{idf}(t,d,D)=\mathrm{tf}(t,d)*\mathrm{idf}(t,D) \quad (1)$$

Where tf(t,d) calculate the frequency of term t appear in document d, idf(t,D) measures the importance of term *t*.

Spammers try to increase the tf-idf scores in their desired content-based spam Web pages. For example Spammers use many repeated and unrelated words in tags of an HTML such as: the <body> tag, Anchor text, URL, Headers (<h1> … <h6> tags), <meta> tags, and the Web page <title>, with many repeated and unrelated words in order to obtain a higher TF-IDF score [20].

Hyperlink-Induced Topic Search (HITS) Algorithm

Hyperlink-Induced Topic Search (HITS) algorithm, is a long-familiar method to find the Hubs and Authoritative Web pages, and is introduced by Jon Kleinberg in 1999, as a link analysis algorithm. It is aimed before the PageRank algorithm used for ranking Web pages [21]. HITS split the Web pages into two main types *i.e*. hubs and authorities. A hub page is one that contains a large number of links to web pages containing information about specific topic. An authoritative page actually store the information about the topic [22]. A web page in HITS calculate two values for each Web page: the first value is for the authority which represents the score of the content-based Web page, and the second value is for the hub, which estimates the score of its links to other Web pages [28].

∀p, we compute *A*(*p*) to be: equation (2).

$$A(p) = \sum_{i=1}^{n} H(i) \tag{2}$$

Where A(p) is the Authority for *p* Web page; *n* is the total number of Web pages that are linked to *p*; and the H(i) is the hub value for the Web page that points to *p*.

Below equation (7) expresses the Hub Update Rule:

∀p, we compute *H*(*p*) to be :equation (3)

$$H(p) = \sum_{i=1}^{n} A(i) \tag{3}$$

Where *H*(*p*) is the Hub for *p* Web page; *n* is the total number of Web pages *p* connected to; and the *A*(*i*) is the Authority values for page .

The Web page is considered to be as a good hub if it points to many good authoritative, and the Web page is assorted as a good authority if it is referred to by many good hubs. The hub values can be spammed through the use of link farms by adding the spam outgoing links to the reputable Web pages. So in this fashion spammers attempt to increase the hub values, and attract several incoming links from the spammed hubs to point to the target spam Web pages [22].

PageRank Algorithm

PageRank is a link analysis algorithm developed in 1998 by Google's founders (Larry Page and Sergey Brin) to create a new kind of search engine as a part of their postgraduate research project. This famed algorithm was first applied to rank the importance of web pages on the Web [23] and since then, PageRank has become popular in wide range of applications in a variety of domains within computer science such as distributed networks, data mining, Web algorithms, and distributed computing [24].PageRank gives each page a numeric score that determines the popularity of that page.

The http-link from web source page P to page Q is known as the forward-link of page P, and the back-link of page Q. The forward-link from page P to page Q presents a vote to page Q. Generally, a higher number of http-links to page Q results in a higher PR score of page Q.

The overall score of a page *p* is determined by the importance (PageRank scores) of pages which have out links to that page *p* [25].

According to Michal *et al*. [26] now PageRank has been frequently used for citation analysis but now it also been applied on the publication citation network.

It is important to note that algorithm does not rank the whole website, rather it evaluates each page individually. The generic formula which appears in the literature for calculating PageRank score for a page p_k is shown in the equation (4).

$$PR(p_k) = \frac{1-d}{N} + d \times \sum_i \frac{PR(T_i)}{\|C(T_i)\|} \tag{4}$$

where *PR* refers to the PageRank score; $C(T_i)$ is the number of forward links from $C(T_i)$ to page p_k:p_k∈P={p1,p2,p2…p_n};*N* is the total number of http pages on the web; $PR(T_i)$ is the PageRank of page T_i; *d* is the damping factor.

$$\forall PR(p) \Rightarrow d \in (0,1) \tag{5}$$

Generally, d is set to 0.85. At the beginning of an algorithm execution, PR of each page is set to1/N. According to equation (8), the PR of each web page can be calculated by using a simple iterative function and the PR of each page will converged. Linear algebra is applied on PageRank calculations. For example, if page p_i have outgoing links target to page p_j, the probability of surfing from p_i to p_{jis} computed as[37]:

$$p_{i,j} = \frac{1}{N(p_i)} \tag{6}$$

The PageRank algorithm is using transition matrix M to run the iteration process to achieve equilibrium value. The transition matrix A is generated as follows:

$$A=\begin{pmatrix} p_{1,1} & \cdots & p_{n,1} \\ \vdots & \ddots & \vdots \\ p_{1,n} & \cdots & p_{n,n} \end{pmatrix} \qquad (7)$$

A Web page with an eminent PageRank score will appear at the top of the list of SEPR as a answer to a particular query. Despite this achievement for those search engines that use PageRank as a ranking method, spammers and malicious Web administrators use some of PageRank algorithm weaknesses to boost the rank of their Web pages illegally by using techniques that violate the SEO tips, in order to gain more visits from Web surfers to their Website. As we know PageRank is based on the link structure of the Web, it is therefore useful to understand how addition or deletion of hyperlinks influences its score.

3.2 Famous Machine Learning Algorithms

This section presents the four famous machine learning algorithms which are described and evaluated in this study. These are decision tree (DT), Naïve Bayes (NB), Random Forest (RF), and RBF Network (RBFN).

A ML classifier is actually a mapping of the input vector space onto a set of classes and we have applied different features in terms of set to map this activity. The detection mechanism of spam Web pages can be regarded as a binary classification problem, where a ML classifier is applied to predict whether a given website is spam or ham. Therefore, in order to enhance the understanding web spamming, this paper presenting the concept of spamdexing in the form of mathematical equation.

Mathematically spamdexing of specific website W_s can be defined as:

$$WebSpam\,(W_s) = \forall W_s P \in W_s (\textstyle\sum_{i=1}^{N} CS(W_s) + LS(W_s) + CL(W_s) + CK(W_s)) \qquad (8)$$

Where

$$W_s = \begin{cases} 1 & When\ page\ is\ effected\ with\ S\ \therefore S \rightarrow\in\ content\ , Link, Cloaking\ and\ Click\ \ spam \\ 0 & when\ page\ is\ normal \end{cases}$$

Table 1. Equation Description

W_s	Website
W_sP	Pages in particular website
CS	Content Spam
LS	Link Spam
CK	Cloaking
CL	Click

We have applied CS and LS features in our training and testing models.

Detection of Spam pages by feeding different features into classifying algorithm $(\mathrm{ML}_{\mathrm{ca}})$ is calculated as:

$$S(W_s) \rightarrow \mathrm{ML}_{\mathrm{ca}}(\mathrm{C}_{\mathrm{spam\,|ham}}\ |\textstyle\sum_{i=1}^{n} f_n\ ) \qquad (9)$$

Consider the problem of classifying website by their content based features, into spam and ham website. Probability that the i^{th} feature of a given website occurs in a feature from class C can be written as:

$$P((W_{s_i} \in CS|C) \qquad (10)$$

If website is stemmed with all features *i.e.* link, content and obvious (W_{s_i}) then probability:

$$P((W_s|C) = \textstyle\prod_i P(W_{s_i}|C) \qquad (11)$$

The research community of ML has developed a large number of classification algorithms to address binary or multiclass problems, such as DT based classifiers, SVM-based classifiers, Bayesian classifiers *etc*. The performance and computational analysis of machine learning algorithms can be evaluated by using different statistical metrics, for example accuracy, F-measure *etc*.

Some of the existing ML classifiers for Web Spam detection are precisely discussed below:

DT (C4.5) - DT (a.k.a statistical classifier)

DT (C4.5) - DT (a.k.a statistical classifier)[56] is an inductive inference tools and mostly applied on classification problem domain.C4.5 is an extension of ID3 algorithm developed by Quinlan to address number of issues, such as; dealing with over fitting problem of data, handling continuous attributes, Improving computational efficiency *etc*. Whereas, J48 is an implementation of C4.5 in Java programming language. The theory of information entropy is applied on selected features to predict the target class. In DT, the nodes of the tree refer to attributes, the possible decision is represented by the branch of the tree and leaves are the target classes. The tree is generated by algorithm in top-down fashion. The information gain (IG) ratio IG Ratio (A, S) of an attribute A relative to the sample set S is defined as:

$$\mathrm{IGRatio}(A, S) = \frac{\mathrm{Gain}(A,S)}{\mathrm{SplitInfo}(A,S)} \qquad (12)$$

Where

$$\mathrm{Gain}(A, S) = \mathrm{Ent}(S) - \textstyle\sum_{a\in A}^{k} \frac{|S_a|}{|S|}\,\mathrm{Ent}(S_a) \rightarrow \forall_{S_a\in\, S}$$

$$\mathrm{SplitInfo}(A, S) = -\sum_{a\in A}^{k} \frac{|S_a|}{|S|}\log_2 \frac{|S_a|}{|S|}$$

The split information value refers to the potential information created by splitting the training dataset Sainto k partition and corresponding to k outcomes on attribute A.

Naive Bayes (NB)

The NB [57] classifier is a classification algorithm based on Bayes theorem with strong independent (naïve) assumptions between features. In order to understand NB, consider a set of training examples, where each example is made up from *i* discrete-valued attributes and a class from a finite set C. The NB classifier can probabilistically make a prediction about the class of an unknown example using the available training example to calculate the most probable outcome. The most predictable class C_{NB} of an unknown example with the conjunction $a_1, a_2, a_3. . ., a_i$ is given by [38]:

$$C_{NB} = arg \max_{c\in C} p\,(c|\ a_1, a_2, a_3, . . ., a_i) \qquad (13)$$

Random Forest (RF)

A RF[58] is an ensemble DT which will predict output value by constructing multiple decision trees on various sub-samples (subset) of the datasets and predict the class that appear most often of the decision trees. In the field of ML ,the utilization of RF has becomes a vital choice for classifying objects due to its prediction accuracy and robustness against noise [40].RF algorithm is using two parameters *i.e*. the number of variables in the random subset at each node and the number of trees in the forest.

The working mechanism of RF is start with production of random vector R_K and disseminated to all trees. Training dataset and RK vector are used in creation of each tree [47] and subsequently produces tree structure classifier T_S.

Where $T_S \rightarrow \{g(x, R_K), k = 1, . . ., N\}$ and given as input vector x.

The generalization error in RF is given by [48]:

$$\mathrm{GE} *= G_{x,y}(\mathrm{mf}(X, Y) < 0) \qquad (14)$$

where *x* &*y*(subscripts) are random vectors that indicate the probability is over the X, Y space and mf is the margin function which measures the extent to which the average number of votes at random vectors for the right output exceeds the average vote for any other output. Margin function is calculated as:

$$\mathrm{mf(X,Y)} = av_k I_f(g_k(X) = Y) - max_{j \neq Y} av_k I_f\,(g_k(X) = j)) \tag{15}$$

Where I_f is the indicator function [48]. In this study, random forest consisted of 100 trees with different feature subset selection.

RBF Network (RBFN)

RBF Network (radial nets) is an example of artificial neural network that employ radial basis function as activation function. Function approximation, time series prediction and classification of data are some well-known applications of Radial basis function networks.

Different radial basis function *e.g.* linear, thin plate spline, cubic, Gaussian *etc.* are available which can be used at hidden units. The most common is the Gaussian function and can be defined as [49]:

$$G_f\,(X) = \exp(- \frac{\|x-\mu_f\|^2}{2\sigma_f^2} \tag{16}$$

Where μ_f is the vector deciding the center of basis function *G, σ_f*is the width parameter and *X* is the dimensional input vector.

4. Data Set and Experimental Performance Evaluation Metrics

This section is divided into three subsections. We begin with dataset information and processing steps to select appropriate features for our experimental work. We then describe the evaluation metrics, which we have used in our study. Finally, experimental results are presented.

4.1. Dataset

We have used publically available WepSpam-uk-2007 dataset for our experimental work, which is a collection of 105,896,555 web pages from 114,529 hosts in the .uk domain and is created by Yahoo!. The percentage of Spam data is 6%. A team of volunteers have manually labeled (spam/non-spam/undecided) 6,479 pages only.

The dataset comprises of 4 sub datasets *i.e.* content based features, link based features, transformed linked based features and obvious/direct features. We have applied content and link features to train and test the model. The experimental work is carried out in the following fashion. (i) Pre-processing of data and Feature selection to reduce problem size, (ii) By applying χ2 score on labeled data, we have selected 139 features from larger feature pool as a concise representation.

Each host is presented as a 140-dimensional vector, which includes features and an associated class label.Table-2 illustrates the distribution of feature vectors in our study. Feature "O" refers to the obvious features, *i.e.* Number of Pages, Length of Hostname. Feature "C" is a list of 96 content features; for example, number of words in the page, number of words in the title, average word length, fraction of anchor text and visible text *etc.* Most of these features are extracted from the work of Ntoulas *et al.* [35]. Castillo *et al.* [36] discusses the information about significance of different features.

Featureset "L" is set of 41 link based features. These features are worked out on the home-page (hp) and the main page (mp), where mp refers to the page with maximum PageRank score in each host. The list comprises of in-degree, out-degree, and edge-reciprocity *etc.* features.

Table 2. Distribution of Feature Vector (Obvious +Content +Link)

Notation	Feature Set	Distribution
O	Obvious	2
C	Content	96
L	Link	41

The ratio of spam and ham is 1 to 17(see Table 3).

Table 3. Dataset with 139 Attributes (Obvious +Content +Link)

Class Label	Percentage in Dataset	Distribution
Spam	5%	208
Nonspam	95%	3641

4.2. Evaluation Mechanism

We have applied 10 cross validation technique in our experimental work. The working process of k-fold cross validation is combination of five steps :(i)The available dataset (D) is divided into k subsets of about equal size that consists in building k data subsets. (ii) D is split into k mutually exclusive parts, D1, D2...Dk.(iii)The instance is trained on D/D_i and tested against D_i.(iv) This action is repeated k times with different i value *i.e.* {i= 1, 2... k}.(v) Finally the performance is judged as the mean of the total number of tests.

In order to test and compare performances of well-known classifiers, we have used WEKA toolkit, a tool for automatic learning and data mining, which includes different types of classifiers and different algorithms for each classifier. In order to estimate the accuracy of different ML algorithms, we use a famous accuracy measure in the context of Information Retrieval *i.e.* F-1 Score or F-measure.

We have divided all instances into two classes *i.e.* spam and ham. Spam is our predicted positive class. We have employed the confusion matrix (see Table 4) to calculate sensitivity (True positive rate or recall) and specificity (True negative rate) measures.

Table 4. Confusion Matrix

	Positive(Spam) positive	Negative(Ham)
True(Spam)	TP	FP
False(Ham)	FN	TN

Where TP score presents the number of positive instances that are correctly sorted as positive, FP score shows the number of negative instances that are falsely sorted as positive, FN score shows the number of positive instances that are falsely sorted as negative, TN score presents the number of negative examples that are correctly sorted as negative. The TP and FP metrics are very useful, especially for imbalanced class problem, like ours. In WebSpam classification problem, where our core objective is to eliminate WebSpam pages; however, we also need to reduce the WebSpam pages that were mistakenly classified as WebSpam (False Positive). Simply, surfing of only WebSpam page is considered as “worse” than dealing with few spam pages in SERP. In short, TP score provides useful information than FP score.

The evaluations metrics (P_{rc},R_{ec},F_{me}) are defined as following:

Precision (P_{rc})refers the percentage of truly positive instances in those classified as positive by the classifier:

$$P_{rc} = \frac{TP}{TP+FP} \tag{17}$$

Recall (R_{ec}) refers the percentage of correctly classified positive instances out of all positive instances:

$$R_{ec} = \frac{TP}{TP+FN} \tag{18}$$

P_{rc} andR_{ec} are related to the FP and TP scores, where F-measure (F_{me}) is a weighted average between Precision and Recall:

$$F_{me} = \frac{2*(P_{rc})*(R_{ec})}{P_{rc}+R_{ec}} \tag{19}$$

4.3. Experimental Results and Evaluation of Classifiers

A machine with Intel Core 2 Duo Processor, 2.93 GHz, with 4 GB memory and running Windows 7 Ultimate has been used to run the algorithms. The dataset which we have used in our experimental work is highly imbalance (see Table 1) so spam detection is highly unbalanced classification issue. In addition, to avoid cost-sensitive effect on classifier *i.e.* classifying a ham web page into spam is much worse than classifying a spam page into ham class. Therefore, we need to measure algorithmic performance through different metrics (Precision, Recall, and F-measure) rather than rely on a single evaluation index. For summarized result presentation, we have opted F-measure score as a evaluation metric because it combines the precision and recall values to compute F_{me}.

The results of experimental work are reported in Table 5, 6 and 7. Two experimental phases *i.e.* detailed and summarized were performed for the evaluation of results. The results of Table 5 and 6 are from first experimental phase, where the comparative analysis of four different ML classifiers are being tested. The evaluation metrics which we have applied in first phase are accuracy, error, Model building time, True Positive Rate, False Positive Rate, Precision, Recall and Fmeasure. The highlighted bold values in result tables depicts the higher score. Table 5 is presenting results of obvious, content and link features. Table 6 illustrates the performance of classifiers with combination of obvious, content and link features with the same evaluation metrics.

Table 7 exhibits the summarized results in terms of Fmeasure score achieved by classifiers. Results indicates that RF and DT algorithms can distinguish the spam and non-spam pages more concisely through the use obvious features, and obtain better F-measure value. The highlighted bold F-measure results refer as the highest F_{me} result for the particular feature set. From Table 5 and 6, it can be observed that RF has outperformed other ML classifiers in terms of other evaluation metrics, such as CCI, ICI, PA *etc.* It can also be observed that RF model which was trained with the combination of obvious and content features has produced more accurate results. By evaluating results in Table5-7, it can be found that NB has produced poorest results, while RBF has generated steady results.

Table 5. Experiment Results with Obvious, Content and Link Features

Evaluation Metrics	2-Obvious Features Classification Algorithms				96-Content Features Classification Algorithms				41-Link Features Classification Algorithms			
	NB	DT	RBF	RF	NB	DT	RBF	RF	NB	DT	RBF	RF
CCI	3579	3642	3272	3642	541	3626	3639	3669	3560	3633	3642	3642
ICI	271	208	578	208	3309	224	211	181	290	217	208	208
PA	92.9	94.5	84.9	94.5	14.0	94.1	94.5	95.2	92.4	94.3	94.5	94.5
TTBM	0.04	0.11	0.3	1.31	0.21	1.45	3.69	1.24	0.09	0.55	2.31	1.03
TP	0.93	0.946	0.85	0.946	0.141	0.942	0.945	0.953	0.925	0.944	0.946	0.946
FP	0.911	0.946	0.856	0.946	0.117	0.701	0.937	0.733	0.906	0.946	0.946	0.914
Rec.	0.93	0.946	0.85	0.946	0.141	0.942	0.945	0.953	0.925	0.944	0.946	0.946
F_{me}	0.914	**0.92**	0.872	**0.92**	0.17	0.935	0.92	**0.941**	0.912	0.919	0.92	**0.923**
P_{re}	0.901	0.895	0.897	0.895	0.91	0.93	0.911	0.944	0.901	0.895	0.895	0.923

CCI=Correctly Classified Instances, ICI=Incorrectly Classified Instances, PA=Prediction Accuracy, TTBM=Time taken to build the Model, TP= True Positive , FP= False Positive, Rec.= Recall, F_{me}= Fmeasure , P_{re}= Precision

Table 6. Experiment Results with Combination of Obvious, Content and Link Features

Evaluation Metrics	98-Obvious+Content Features Classification Algorithms				43-Obvious +Link Features Classification Algorithms				137-Content+Link Features Classification Algorithms			
	NB	DT	RBF	RF	NB	DT	RBF	RF	NB	DT	RBF	RF
CCI	542	3621	3638	3681	3553	3638	3642	3643	939	3613	3642	3678

ICI	3308	229	212	169	297	212	208	207	2911	237	208	172
PA	14.0	94.0	94.4	95.6	92.2	94.4	94.5	94.6	24.3	93.8	94.5	95.5
TTBM	0.3	2.34	9.08	13.77	0.21	0.67	4.86	10.52	0.27	2.84	13.81	12.19
TP	0.141	0.941	0.945	0.956	0.923	0.945	0.946	0.946	0.244	0.938	0.946	0.955
FP	0.117	0.724	0.942	0.719	0.906	0.937	0.946	0.919	0.147	0.706	0.946	0.76
Rec.	0.141	0.941	0.945	0.956	0.923	0.945	0.946	0.946	0.244	0.938	0.946	0.955
F_{me}	0.17	0.933	0.92	**0.945**	0.911	0.92	0.92	**0.923**	0.329	0.932	0.92	**0.942**
P_{re}	0.91	0.928	0.904	0.951	0.9	0.909	0.895	0.926	0.921	0.928	0.895	0.953
CCI=Correctly Classified Instances, ICI=Incorrectly Classified Instances, PA=Prediction Accuracy, TTBM=Time taken to build the Model, TP= True Positive , FP= False Positive, Rec.= Recall, F_{me}= Fmeasure , P_{re}= Precision												

Table 7. F-measure Results on OCL Features from Different ML Classifiers

Classifier	O	C	L	O+C	O+L	C+L
NB	0.914	0.17	0.912	0.17	0.911	0.329
DT	**0.92**	0.935	0.919	0.933	0.92	0.932
RBF	0.872	0.92	0.92	0.92	0.92	0.92
RF	**0.92**	**0.941**	**0.923**	**0.945**	**0.923**	**0.942**

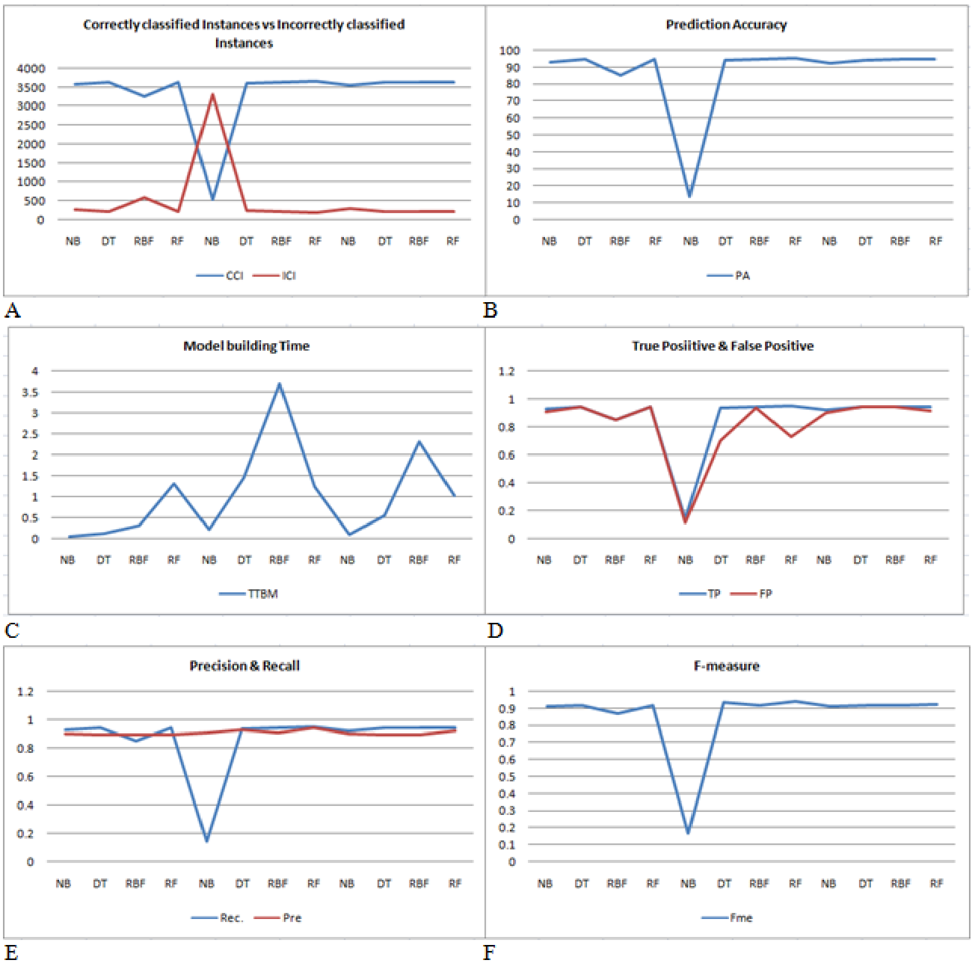


Figure 6. Phase 1Performance Assessment of each Classifier with OCL Feature Set

WebSpam is a complex and dynamic phenomena, where spammers continuously changes their tricks after knowing the important ranking signals to subvert search engine ranking. Hence, we believe that the performance of a classifier is dependent on the different feature set, and therefore experimental work is carried out on different features with different combinations. After classifier training and testing phase, different evaluation metrics were recorded. It can the seen in Figure 6 which is a product of Table 5 that RF, RBF and DT has achieved prominent score in terms of CCI and ICI in all feature set. However, NB produces poor results, especially when tested on content feature set. RF has achieved a prominent PA score against other ML classifiers.

The graph depicts that RF and RBF algorithms were computationally suffers in time space than DT and NB. NB performed well than all other classifiers. RF and DT achieves prominent scores *i.e.* more than 94% in terms of TP in all feature set, while NB produces weak results with content feature set and RBF performance declines with obvious feature. It can be observed from the graph that RBF holds larger FP score than rest algorithms, it shows its weakness, while NB brings very low FP score with content features.

Recall results of RF, DT, NB and RBF are almost replica of figure part A; however RF has outperformed with content and link feature set than RBF, NB and DT. With respect to the F-measure, the best ML classifier is RF that wins on two out of three feature set.

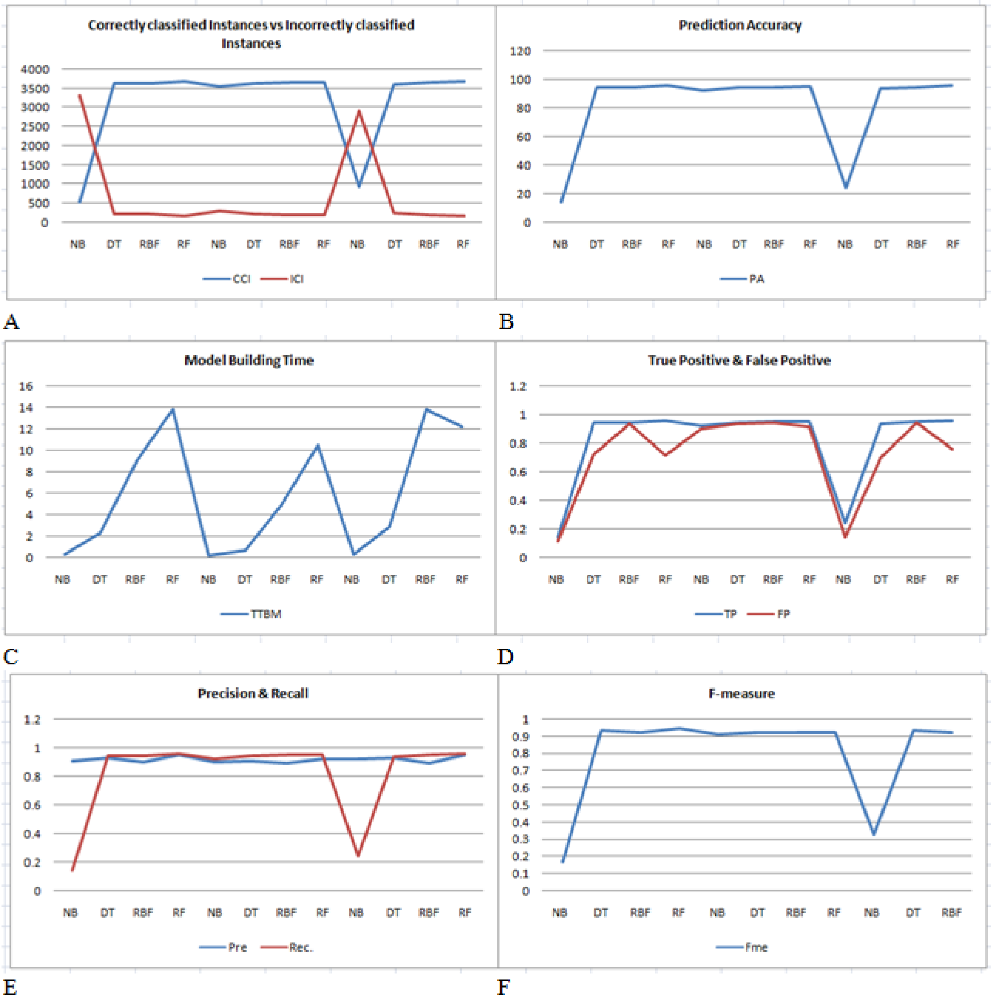


Figure 7. Phase I Performance Assessment of each Classifier with Combined Feature Set

Figure 7 is an extension of results with combination of feature set and generated from Table 6. Figure illustrates RF, RBF and DT has achieved eminent score in terms of CCI and ICI. However, NB again not able to produce significant result. It can be seen from figure that RF, RBF and DT has best predictive accuracy From graph, it can be understood that RF and RBF algorithms took quite some time to build the model, whereas DT and NB are useful in time critical applications. Figure shows the chart between TP and FP score. RF, RBF and DT has achieved prominent score in terms of TP in all feature set, while NB produces weak results with (O+C, O+L) feature set. It can also be observed from the graph that RBF has shown better results to achieve good TP score than rest algorithms. NB brings very low FP score with O+C and O+L feature set.

Recall results of RF, DT, NB and RBF are almost replica of figure part A; however RF has outperformed with O+C and C+L feature set than RBF, NB and DT. RF proved itself the best classifier in terms of F-measure score, whereas NB performed worst.

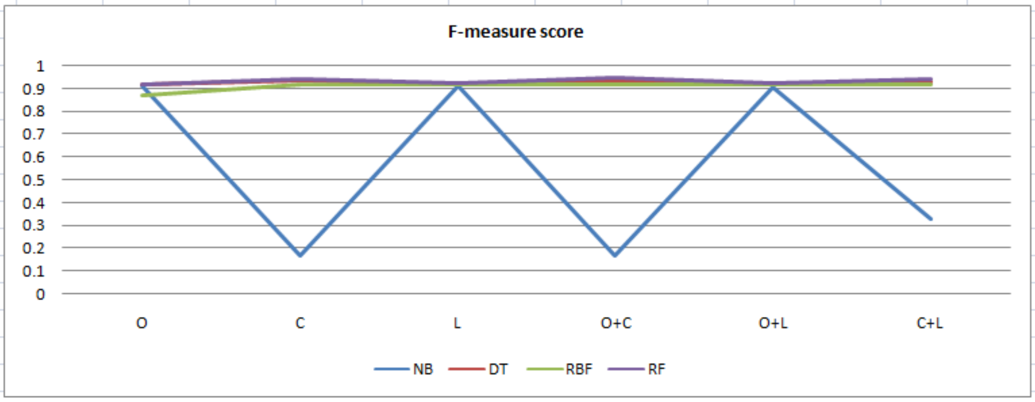


Figure 8. Performance Comparison of 4 Classifiers with Standalone and Combined Feature Set

From Figure 8, it can be seen that RF learning classifier has achieved significant f-measure score with all three types of features in two different input fashion *i.e.* standalone and with combination. Naive Byes fail to reach on prominent score, while the DT and RBF results are steady and performs well with content features.

5. Conclusions

This paper addresses the problem of web spam page detection. In this paper, we presented standalone and combined feature set experimental comparison on well-known ML classifiers to test their performance. Empirical studies show that RF learning model can be good choice to catch spam pages, thereby obtaining prominent results. RF has achieved highest F_{me} score and proved to be powerful learning algorithm than most other famous data mining tools. In addition, experiment result showed that DT and RBF classifiers are stable with different feature set. We also demonstrated the use of mathematical equations to understand spamdexing issue.

This work mainly focused on to evaluate the efficiency of the machine learning classifiers used for Web spam classification. In the future, we plan to introduce new features which could help for Web spam detection problem.

Acknowledgments

This work is made possible through the help and support from my Professor Dr. Zhu Yan. Additionally, We sincerely thank to our parents, family, and friends, who provide the advice and support. The product of this research paper would not be possible without all of them. Moreover, the authors thank the anonymous reviewers for their valuable comments.

Author Contributions

All authors contributed to conceive the paper, analyze the data, writing and polishing the manuscript.

Conflicts of Interest

Declare conflicts of interest or state “The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funding sponsors in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state “The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results”.

References

[1] Ioannis N., Nikolaos A., Zoe V., Dimitris V. Assessment of the gap and (non-)Internet users evolution based on population biology dynamics. Telecommunications Policy, 2015, Volume 39, Issue 1, 14–37.

[2] Antal V. B., Toine B., Maurice K. Estimating search engine index size variability: a 9-year longitudinal study”, Scientometrics, ISSN 0138-9130.

[3] Wang Y.M., Ma M., Niu Y., Chen H. Spamdouble-funnel: Connecting web spammers with advertisers. In: Proceedings of the 16thinternational conference on World Wide Web, ACM, 2007, 291–300.

[4] Farookh K. H., Elizabeth C.A Survey in Traditional Information Retrieval Models. 2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008),pp 397-402,Phitsanulok, Thailand: Institute of Electrical and Electronics Engineers (IEEE).

[5] Pan, B., Hembrooke, H. A., Joachims, T., Lorigo, L., Gay, G. K., & Granka, L. A. In Google we trust: users' decisions on rank, position, and relevance. Journal of Computer-Mediated Communication, 2007, 12, 801e823

[6] Zoltan G., Hector G., Jan P. Combating Web Spam with TrustRank. Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

[7] Henzinger, M.R. Motwani, R. Silverstein C. Challenges in web search engines SIGIR Forum, 36, 2002, 11–22.

[8] Baoning W., Vinay G., Brian D. D. Topical TrustRank: Using Topicality to Combat Web Spam, In WWW ’06: Proc. of the 15th international conference on World Wide Web, ACM, New York, NY, USA.

[9] Salton, G. Wong, A.and Yang C. S. A vector space model for automatic indexing. Commun. ACM, Vol.18, Nov. 1975.

[10] Robertson, S. Zaragoza, H. and Taylor. M. Simple bm25 extension to multiple weighted fields. In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management,CIKM’04, Washington, D.C., 2004.

[11] Zhai. C. Statistical Language Models for Information Retrieval. Now Publishers Inc., Hanover, MA, 2008.

[12] Nikita S., Jiawei H. Survey on Web Spam Detection: Principles and Algorithms. SIGKDD Explorations Volume 13, Issue 2,pp 50-64.

[13] Luca B., Carlos C., Debora D., Stefano L., and Ricardo B. Web Spam Detection: link-based and content-based techniques”, Available [online] http://chato.cl/papers/becchetti_2008_link_spam_techniques.pdf

[14] Ricardo B., Carlos C., and Vicente L. Pagerank increase under different collusion topologies. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[15] Page, L., Brin, S., Motwani, R., Winograd, T. The Pagerank Citation Ranking: Bringing Order to the Web ,Tech,rep, Stanford University, 1999.

[16] Kleinberg J. M. Authoritative Sources in a Hyperlinked Environment ,J. ACM, 46 (5), 1999, 604–632.

[17] Gyongyi, Z., and Garcia. Molina Web Spam Taxonomy. Hector, 2004. Technical Report. Stanford Available: http://airweb.cse.lehigh.edu/2005/gyongyi.pdf

[18] Ugo E. , Sabrina S., Fernando M. ,Giuseppe C. Approximate TF–IDF based on topic extraction from massive message stream using the GPU. Information Sciences Volume 292, 20 January 2015, Pages 143–161.

[19] Bruno T., Sasa M., Dzenana D. KNN with TF-IDF based Framework for Text Categorization. Procedia Engineering Volume 69, 2014, Pages 1356–1364.

[20] Mohammed N. A., Izzat M. A., Heider A. W. Evaluation of Spam Impact on Arabic Websites Popularity. Journal of King Saud University - Computer and Information Sciences Volume 27, Issue 2, April 2015, Pages 222–229.

[21] Chris H. Q. Ding, H. Z., Xiaofeng H., Parry H. and Horst D. S. Link Analysis: Hubs and Authorities on the World Wide Web. SIAM Review, Vol. 46, No. 2 , 2004, 256-268.

[22] Paul, A. C., Daniel O. and Wolfgang N. Finding Related Pages Using the Link Structure of the WWW. Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence, 2004, 632-635.

[23] Brin, S., Page, L. The anatomy of a large-scale hypertextual web search engine. Proc. of Seventh International World-Wide Web Conference (WWW), 1998, 107–117.

[24] Atish D. S., Anisur R. M., Gopal P., Eli U. Fast distributed PageRank computation. Theoretical Computer Science Volume 561, Part B, 2015, 113–121.

[25] Kang, F., Liu, X., Liu. A personalized ranking approach via incorporating users' click link information into PageRank algorithm. In: International conference on energy systems and electrical power, 2011, Vol. 13, pp. 275–284.

[26] Michal N. , Karel J. , Dalibor F., Martin D. PageRank variants in the evaluation of citation networks. Journal of Informetrics, 2014, Volume 8, Issue 3, , Pages 683–692.

[27] Hongwei W., Yuankai L., Kaiqiang G. Countering Web Spam of Link-based Ranking Based on Link Analysis. Procedia Engineering 23, 2011, 310 – 315.

[28] Xinyue L. , You W., Shaoping Z., Hongfei L. Combating Web spam through trust–distrust propagation with confidence. Pattern Recognition Letters 34, 2013, 1462–1469.

[29] Víctor M. P., Manuel Á., Fidel C. SAAD, a content based Web Spam Analyzer and Detector. The Journal of Systems and Software 86, 2013, 2906–2918.

[30] Marc N. Web Spam Detection, in Encyclopedia of Database Systems, Springer Verlag, September 2009.

[31] Ghiam, S., & Nemaney, P. A survey on web spam detection methods:taxonomy. International Journal of Computing Research Repository (CoRR). 2012. arXiv:1210.3131v1 [cs.IR]

[32] Chu, Z., Gianvecchio, S., Koehl, A., Wang, H., & Jajodia, S. Blog or block:Detecting blog bots through behavioral biometrics.Computer Networks, 2013, 57(3),634–646

[33] Muhammad N. M., Watheq E. M. , Fayez G. A spam rejection scheme during SMTP sessions based on layer-3 e-mail classification", Journal of Network and Computer Applications 32, 2009, 236– 257.

[34] Atefeh H., Mohammad A. T., Naomie S., Zahra H. Detection of review spam: A survey. Expert Systems with Applications 42, 2015, 3634–3642

[35] Alexandros N., Marc N., Mark M., and Dennis F. Detecting spam web pages through content analysis. In Proceedings of the World Wide Web conference, 2006, 83–92, Edinburgh, Scotland.

[36] Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., Vigna, S.: A reference collection for web spam detection. ACM SIGIR Forum 40(2), 2006, 11–24.

[37] Li L., Letian S., Shiping C. , Ming L. , Jun Z. K-PRSCAN: AclusteringmethodbasedonPageRank", Neurocomputing175, 2016, 65–80

[38] Damrongrit S. Classification of complete blood count and hemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassemia screening. Biomedical Signal Processing and Control 7, 2012, 202–212.

[39] Watkins, A., Timmis, J., Boggess, L. Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. Genetic Programming and Evolvable Machines, Kluwer Academic Publishers, 2004, 173—181.

[40] Zaklouta, F., Stanciulescu, B., Real-time traffic-sign recognition using tree classifiers, IEEE Trans. Intell. Transp. Syst. 13 (4) , 2012, 1507–1514.

[41] By the numbers:100 amazing Google statistics and Facts: Available at http://expandedramblings.com/index.php/by-the-numbers-a-gigantic-list-of-google-stats-and-facts/

[42] Miklós E. ,András G., András A. B. Web Spam Classification: a Few Features Worth More. Web Quality '11, March 28, 2011, Hyderabad, India.

[43] Luca B., Carlos C., Debora D., Stefano L., and Ricardo B. Web Spam Detection: link-based and content-based techniques. In Proceedings of the final workshop, Barcelona, Spain, European integrated Project DELIS, 2009.

[44] James C., Ling L., William B. R. Link-Based Ranking of the Web with Source Centric Collaboration. 2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing, Atlanta, 2006: 1-9.

[45] James C., Steve W., Ling L., William B. R. A Parameterized Approach to Spam-Resilient Link Analysis of the Web. In Parallel and Distributed Systems, 2009, 20(10):1422-1436.

[46] Wu, B., Davison, B.D. Identifying link farm spam pages. In: WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, New York, NY, USA, 2005, 820–829.

[47] Zerina M., Abdulhamit S. Congestive heart failure detection using random forest classifier. computer methods and programs in biomedicine 1 3 0 , 2 0 1 6 , 54–64.

[48] Breiman, L.Random forests, Mach. Learn. 45 (October (1)) ,2001, 5–32.

[49] Philip R. , Bogdan M. W. Efficient incremental construction of RBF networks using quasi-gradient method. Neurocomputing 150, 2015, 349–356.

[50] Bauer J.M., Eeten M, Wu Y. Itu study on the financial aspects of network security: Malware and spam. 2008.

[51] Kanich C., Weaver N., McCoy D., Halvorson T., Kreibich C., Levchenko K., Paxson V., Voelker G. M., Savage S Show me the money:Characterizing spam-advertised revenue. 2011, In: USENIX Security Symposium.

[52] Cortes C., Vapnik V.. Support-vector networks. Machine Learning, 1995, 20(3):273–297.

[53] Castillo C., Donato D., Gionis A., Murdock V., Silvestri F. Know your neighbors: web spam detection using the web topology. 2007,DOI 10.1145/1277741.1277814.

[54] Piskorski J., Sydow M. , Weiss D. Exploring linguistic features for web spam detection: a preliminary study. DOI 10.1145/1451983.1451990.

[55] Dai N., Davison B. D., Qi X. Looking into the past to better classify web spam. 2009, DOI 10.1145/1531914.1531916.

[56] Quinlan J.R. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. 1993.

[57] Russell S., Norvig P. Artificial intelligence: a modern approach, 1995.

[58] Breiman L. Random forests. Machine learning 45(1):2001, 5–32.