

AN IMPROVED FRAMEWORK FOR CONTENT
AND LINK-BASED WEB SPAM DETECTION: A
COMBINED APPROACH



PTTA UTHM
PERPUSTAKAAN TUNKU TUN AMINAH
ASIM SHAHZAD

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

STATUS CONFIRMATION FOR THESIS
DOCTOR OF PHILOSOPHYAN IMPROVED FRAMEWORK FOR CONTENT AND LINK-BASED WEB
SPAM DETECTION: A COMBINED APPROACH

ACADEMIC SESSION: 2020/2021

I, **Asim Shahzad**, agree to allow Thesis to be kept at the Library under the following terms:

1. This Thesis is the property of the Universiti Tun Hussein Onn Malaysia.
2. The library has the right to make copies for educational purposes only.
3. The library is allowed to make copies of this Thesis for educational exchange between higher educational institutions.
4. The library is allowed to make available full text access of the digital copy via the internet by Universiti Tun Hussein Onn Malaysia in downloadable format provided that the Thesis is not subject to an embargo. Should an embargo be in place, the digital copy will only be made available as set out above once the embargo has expired.
5. ** Please Mark (v)

CONFIDENTIAL

(Contains information of high security or of great importance to Malaysia as STIPULATED under the OFFICIAL SECRET ACT 1972) *Title and Abstract only*

RESTRICTED

(Contains restricted information as determined by the organization/institution where research was conducted)-
Title, Abstract and Introduction only

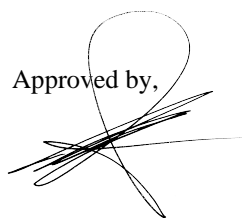
EMBARGO

_____ until _____
(date) (date)

FREE ACCESS



(WRITER'S SIGNATURE)



(SUPERVISOR'S SIGNATURE)

Permanent Address:

HAJIABAD NEAR
RHC SHINKIARI
PAKISTANPROF. DR. NAZRI BIN MOHD NAWI
Head Of Soft Computing & Data Mining Centre (SMC)
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn MalaysiaDate : 05/18/2021Date: 5/18/2021

NOTE: ** If this Thesis is classified as CONFIDENTIAL or RESTRICTED, please attach the letter from the relevant authority/organization stating reasons and duration for such classification.

This thesis has been examined on -----

and is sufficient in fulfilling the scope and quality for the purpose of awarding the Degree of Doctor of Philosophy.

Chairperson:

Faculty of Computer Science and Information Technology

Universiti Tun Hussein Onn Malaysia

Examiners:

Faculty of Computing

Universiti -----

Faculty of Computer Science and Information Technology

Universiti Tun Hussein Onn Malaysia



PTTA UTHM
PERPUSTAKAAN TUNKU TUN AMINAH

AN IMPROVED FRAMEWORK FOR CONTENT AND LINK-BASED WEB
SPAM DETECTION: A COMBINED APPROACH

ASIM SHAHZAD

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy of Information Technology



Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

MAY 2021

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

Student :
Asim Shahzad

Date : MAY 2021

Supervisor :
Prof. Dr. Nazri Mohd. Nawi

Date : MAY 2021



PTTA UTHM
PERPUSTAKAAN TUNKU TUN AMINAH

*With Love and Thanks,
To my beloved parents & wife for their love, encouragement, endless support &
sacrifices*



PTTA UTHM
PERPUSTAKAAN TUNKU TUN AMINAH

ACKNOWLEDGEMENT

My first and foremost thanks go to One and Only Allah ﷻ and his last Messenger Mohammad ﷺ. This Project would have been left out as a mere dream if my Creator's spiritual help and His Messenger were not there for me. I would particularly like to express my sincere gratitude to my supervisor, Prof. Dr. Nazri Mohd Nawi and his continuous support, technical guidance, and assistance in finishing this research for his patience, enthusiasm, motivation, and immense knowledge. His advice helped me all the time in study and writing this thesis. I could not have imagined a better research supervisor for my Ph.D. journey. Besides my supervisor, I would also like to thank my co-supervisor, Associate Prof. Dr. Hairulnizam Mahdin, for his support, guidance, valuable comments, suggestions, and words of encouragement during my research. Furthermore, my profound appreciation for the prestigious Universiti Tun Hussein Onn Malaysia (UTHM) for giving me an opportunity to study here and accomplish my dreams of becoming a researcher. In addition, I am very grateful to the ORICC of UTHM for supporting this research. With boundless love and appreciation, I would also like to extend my heartfelt gratitude and appreciation to my parents, beautiful wife, and adorable kids for motivating me and supporting me to continue my research journey and make this research a reality.

Finally, yet importantly, I am also thankful to my siblings for believing in me and supporting me in all my endeavors. In the end, I am grateful to all my friends who have encouraged me and helped me in this research.



ABSTRACT

In the modern digital era, the Web has been utilized for searching information by using different search engines (SE) as a tool. However, web spammers misuse the web for financial benefits by ranking the irrelevant and spam web pages higher than relevant pages in the search engine's results pages (SERPs) by using web spamming techniques. Furthermore, those top-ranked unrelated web pages contain insufficient or inappropriate information for the user. In addition, web spamming techniques dramatically affect the quality of the search engine. Researchers introduced several web spam detection techniques such as content-based features, link-based features, label propagation, label refinement, click-based web spamming detection, and real-time web spam detection. However, identifying all spam pages on the Web with high accuracy is still remains unsolved. This work proposes a content-based web spam detection framework, link-based web spam detection framework, and a combined approach to identify both types of web spams with high accuracy that can detect the newly evolved link pyramid. The content-based web spam detection framework uses three proposed and two improved content-based algorithms for web spam detection. The link-based web spam detection framework initially exposed the relationship network behind the link spamming and then used the paid-links database algorithm, spam signals algorithm, and improved link farms algorithm for link-based web spam identification. Finally, the combination of both content and link-based frameworks enhance the accuracy of web spam detection. The proposed combined approach's performance has been evaluated and compared with the J48 classifier, C4.5 decision tree classifier, SVM classifier, and heuristic combined approach. Some experiments were conducted to obtain the threshold values using the proposed collection architecture on well-known datasets WEB SPAM-UK2006 and WEB SPAM-UK2007. The results show that the proposed methods outperform other methods with 82.1% precision and an F-measure of 80.6% to illustrate the proposed framework's effectiveness and applicability.



ABSTRAK

Dalam era digital moden, jaringan telah digunakan secara meluas sebagai alat untuk mencari maklumat melalui enjin carian yang berbeza (*SE*). Walau bagaimanapun, penghantar spam telah menyalahgunakan jaringan untuk mengaut keuntungan dengan menyenaraikan laman jaringan yang tidak relevan dan menghasilkan lebih banyak spam daripada hasil carian halaman yang berkaitan di dalam halaman hasil enjin carian (*SERPs*) dengan menggunakan teknik jaringan web. Di samping itu, teknik jaringan web secara dramatik boleh menjejaskan kualiti hasil enjin carian dan memberikan kesan buruk kepada ekonomi kerana data pengiklanan percuma dihasilkan dengan banyak dan diindeks pada enjin carian sekali gus meningkatkan jumlah lalu lintas jaringan pada tapak jaringan yang disasarkan. Penyelidik berusaha memperkenalkan beberapa teknik pengesanan spam seperti pengujian kandungan, ciri berasaskan pautan, penyebaran label, penambahbaikan label, pengesanan jaringan web berasaskan klik, pengesanan jaringan web masa nyata dan sebagainya. Walau bagaimanapun, mengenalpasti semua halaman spam di jaringan dengan ketepatan yang tinggi masih menjadi isu utama dan tidak dapat diselesaikan hingga kini. Kajian ini mencadangkan gabungan rangka kerja pengesanan jaringan web berasaskan kandungan dan berasaskan pautan bagi mengenalpasti kedua-dua jenis spam jaringan dengan ketepatan yang tinggi dan dapat mengesan piramid pautan yang baru berkembang. Rangka pengesanan jaringan web yang dicadangkan menggunakan tiga kaedah dan dua algoritma berasaskan kandungan sementara rangka pengesanan jaringan web berasaskan pautan mendedahkan rangkaian hubungan di belakang pautan jaringan web dengan menggunakan algoritma pangkalan data pautan, algoritma isyarat spam, dan algoritma ladang pautan yang lebih baik untuk mengenalpasti jaringan berasaskan pautan. Prestasi pendekatan gabungan yang dicadangkan telah dinilai dan dibandingkan dengan teknik J48, pepohon keputusan C4.5, teknik SVM dan pendekatan gabungan heuristik. Keputusan eksperimen yang dijalankan terhadap dataset yang terkenal seperti WEB SPAM-UK2006 dan WEB SPAM-UK2007 menunjukkan menunjukkan bahawa kaedah yang dicadangkan lebih baik daripada kaedah lain dengan ketepatan 82.1% dan F-ukuran 80.6%.



TABLE OF CONTENTS

TITLE	i
DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	v
ABSTRAK	vi
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF SYMBOLS AND ABBREVIATIONS	xiii
LIST OF PUBLICATIONS	xv
CHAPTER 1 INTRODUCTION	1
1.1 Background of the Research	1
1.2 Problem Statement	5
1.3 Aims of the Research	6
1.4 Objectives of the Research	6
1.5 Scope of the Research	7
1.6 Significance of the Research	7
1.7 Outline of the Thesis	8
CHAPTER 2 LITERATURE REVIEW	9
2.1 Introduction	9
2.2 Search Engine Optimization and Web Page Spamming	9
2.3 Algorithms for Content-Based Web Spam Detection	12
2.4 Algorithms for Link-Based Web Spam Detection	16
2.4.1 Algorithms Based on Label Propagation	17
2.4.2 Algorithms Based on Link-based Features	22
2.4.3 Algorithms Based on Graph Regularization	23



2.4.4	Algorithms Based on Label Refinement	25
2.4.5	Algorithms Based on Link Pruning	26
2.5	Non-Traditional Techniques for Web Spam Detection	27
2.5.1	HTTP Analysis and Real-Time Detection	27
2.5.2	Web Spam Detection Using Evidence Theory	29
2.5.3	Unsupervised Spam Detection Algorithms	30
2.5.4	Click-Spam Detection Algorithms	31
2.5.5	Semantic-Based Spam Detection Algorithms	32
2.5.6	User's Browsing Behavior-Based Algorithms	33
2.6	Combined Techniques for Web Spam Detection	34
2.7	Standard Threshold Values	37
2.8	Data Collection Architecture	41
2.9	Research Gap Analysis on Combined Approach	42
2.10	Chapter Summary	45

CHAPTER 3 RESEARCH METHODOLOGY 46

3.1	Introduction	46
3.2	Data Collection	48
3.3	Revealing Relationship Behind Spam Network	48
3.4	The Proposed Improved Data Collection Architecture	51
3.5	Analysis of Groups, Forums, SEO Service Providers	54
3.6	Data Pre-Processing for Web Spam Framework	55
3.6.1	Paid Links Database	56
3.6.2	Spam Signals Identification	56
3.7	The Proposed Stopwords Algorithm	61
3.8	The Improved Keywords Algorithm	64
3.9	The Proposed Spam Keywords Database Algorithm	68
3.10	The Improved POS Algorithm	70
3.11	The Proposed Algorithm for Copied Content	74
3.12	The Proposed Content-based Framework	77
3.13	The Proposed Paid-Link Detection Algorithm	79
3.14	The Proposed Spam Signals Algorithm	80
3.15	The Improved Link Farm Detection Algorithm	82
3.16	The Proposed Link-based Framework	87



3.17	A Combined Approach for Web Spam Detection	88
3.18	Chapter Summary	89
CHAPTER 4	RESULTS AND DISCUSSIONS	90
4.1	Introduction	90
4.2	Experimental Results	91
4.2.1	The Proposed Content Framework Results	92
4.2.2	The Proposed Link-based Framework Results	93
4.2.3	The Proposed Combined Approach Results	94
4.3	Comparison with The Existing Techniques	95
4.3.1	Comparison with Smart-BT Technique	95
4.3.2	Comparison with CFS-PSO	96
4.3.3	Comparison with a Novel Framework WSF2	96
4.3.4	Comparison with Roul's Combined approach	97
4.3.5	Comparison with SVM Classifier	98
4.3.6	Comparison with C4.5 Decision Tree	99
4.3.7	Comparison with J48 Classifier	99
4.3.8	Comparison with Decision Tree Classifier	100
4.4	Chapter Summary	101
CHAPTER 5	CONCLUSIONS AND FUTURE WORKS	102
5.1	Introduction	102
5.2	Summary of Research Findings	103
5.3	Research Contributions	104
5.4	Future Works	105



LIST OF TABLES

2.1	The comparison of link-based detection algorithms	17
2.2	The existing combined techniques	36
2.3	The percentage of pso usage in documents	40
3.1	Existing and improved data collection architecture	53
3.3	Percentage of different spam signals	59
3.4	Spam score based on spam signals on website	60
3.5	Enhancement in stopword algorithm	61
3.6	Sample output files of custom script	62
3.7	Enhancement in keywords algorithm	65
3.8	Keywords frequency calculation on w_{p_i}	65
3.9	Spam keywords ratio calculation on w_{p_i}	68
3.10	Part of speech identification and tagging on w_{p_i}	71
4.1:	Experimental results for content-based framework	92
4.2:	Experimental results for link-based framework	93
4.3:	Experimental results for combined approach	94
4.4:	The proposed framework vs. smart-bt technique	95
4.5:	The proposed framework vs. cfs-pso technique	96
4.6:	The proposed framework vs. novel framework	97
4.7:	The proposed framework vs. Roul's approach	98
4.8:	The proposed framework vs. svm classifier	98
4.9:	The proposed framework vs. c4.5 decision tree	99
4.10	The proposed framework vs. j48 classifier	100
4.11	The proposed framework vs. decision tree	100
4.12	The proposed framework vs. all other techniques	101



LIST OF FIGURES

2.1	Categorization of web spamming techniques	11
2.2	Categorization of anti-web spamming techniques	12
2.3	Example of machine-generated spam page	38
2.4	Web page using keywords stuffing technique	38
2.5	Web page using spam keywords	39
3. 1	Research methodology	47
3. 2	An improved data collection architecture	51
3.6	Flowchart for the proposed stopwords algorithm	63
3.7	Pseudocode for the proposed stopwords algorithm	64
3.8	Flowchart for the improved keywords algorithm	66
3.9	Pseudocode for the improved keywords algorithm	67
3.10	Example of a web page with spam keywords	68
3.11	Flowchart for the proposed spam keywords algorithm	69
3.12	Pseudocode for the proposed spam keywords algo	70
3.13	Flowchart for the improved part of speech algorithm	73
3.14	Pseudocode for the improved part of the speech algo	74
3.15	Flowchart for the proposed copied content algorithm	75
3.16	Example of finding the publication date	76
3.17	Pseudocode for the proposed copied content algo	77
3.18	Flowchart of the proposed content-based framework	78
3.19	Flowchart for the proposed paid link database algo	79
3.20	Pseudocode for the proposed paid link database algo	80
3.21	Flowchart for the proposed spam signals algorithm	81
3.22	Pseudocode for the proposed spam signals algorithm	82
3.23	Web graph representing the websites and links	84
3.24	Flowchart for the improved link farm algorithm	85



3.25 Pseudocode for the improved link farm algorithm	86
3.26 The proposed link-based framework	87



PTTA UTHM
PERPUSTAKAAN TUNKU TUN AMINAH

LIST OF SYMBOLS AND ABBREVIATIONS

SE	-	Search Engine
TF-IDF	-	Term Frequency-Inverse Document Frequency
PR	-	PageRank
p	-	Page
SEO	-	Search Engine Optimization
TR	-	TrustRank
HITS	-	Hyperlink-Induced Topic Search
SAAD	-	Spam Analyzer and Detector
SERP	-	Search Engine Results Pages
TF	-	Term Frequency
IDF	-	Inverse Document Frequency
q	-	Query
t	-	Term
PP	-	Primitive Polynomial
SER	-	Search Engine Result
PPR	-	Personalize Page Rank
SR	-	SpamRank
TR	-	TrustRank
BP	-	BadPath
HR	-	Harmonic Rank
ATR	-	Anti TrustRank
σ	-	Random Jump Probability
W _{pi}	-	Web Page
RSW	-	Ratio of Stopwords
SW	-	Stopwords
KW _i	-	Keyword
T _{kw}	-	Total Number of Keywords



KWf	-	Keyword Frequency
KWp	-	Keyword Phrase
Nwp	-	Number of Word in Phrase
SKW	-	Spam Keywords
SKWR	-	Spam Keywords Ratio
T_{SKW}	-	Total Number of Spam Keywords
g_f	-	Grammatical Form
POS	-	Part of Speech
x	-	The Number of Occurrences of g_f
y	-	Total Number of Words Present in Wpi
API	-	Application Program Interface
PBM	-	Private Blog Network
G	-	Graph
V	-	Vertex
E	-	Edge
od(Wpi)	-	The out-degree (od) of a Web Page Wpi
id(Wpi)	-	The in-degree (id) of a Web Page Wpi
A_{ij}	-	Element of Matrix A
Wpj	-	Web Page
Wpn	-	Any Web Page
PC	-	Pseudo-Code
LF	-	Link Farm
DB	-	Database
THV	-	Threshold Value
DCA	-	Data Collection Architecture
TDL	-	Top Level Domain



LIST OF PUBLICATIONS

Journals:

1. **Shahzad, A.**, Mahdin, H., & Nawawi, N. M. (2020). An Improved Framework for Content-based Spamdexing Detection. *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 409–420.
2. **Shahzad, A.**, Jacob, D. W., Nawawi, N. M., Mahdin, H., Saputri, M. E. (2020). The new trend for search engine optimization, tools and techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, vol 18, no. 3, pp. 1568-1583
3. **Shahzad, A.**, Nawawi, N. M., Sutoyo, E., Naeem, M., Ullah, A., Naqeeb, S., & Aamir, M. (2018). Search Engine Optimization Techniques for Malaysian University Websites: A Comparative Analysis on Google and Bing Search Engine. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4), 1262-1269.
4. **Shahzad, A.**, Nawawi, N. M., Hamid, N. A., Khan, S. N., Aamir, M., Ullah, A., & Abdullah, S. (2017). The Impact of Search Engine Optimization on The Visibility of Research Paper and Citations. *JOIV: International Journal on Informatics Visualization*, 1(4-2), 195-198.
5. Ullah, A., Nawawi, N. M., Sutoyo, E., **Shahzad, A.**, Khan, S. N., & Aamir, M. (2018). Search Engine Optimization Algorithms for Page Ranking: Comparative Study. *International Journal of Integrated Engineering*, 10(6), 19-25.



CHAPTER 1

INTRODUCTION

1.1 Background of the Research

Web page spamming is an intentional act intended to trigger illegally favorable importance or relevance for some page, considering the web page's real significance (Metaxas and Pruksachatkun, 2017). Web page spamming is a well-known challenge to search engines because it massively degrades the quality of the search engine's results, and it annoys the user for getting inappropriate information (Jakubiček et al., 2020). As the World Wide Web (WWW) is growing with an unprecedented ratio, the size of available textual data has become huge to any end-user. A recent survey by an organization, "WorldWideWebSize" shows that the Web consists of 5.26 billion web pages (Kunder, 2021). Every day, thousands of web pages are being added to the Web Corpus, and many are either duplicated or spam web pages (Ardi and Heidemann, 2019). Web spammers are using several creative spamming techniques for dragging internet users to their websites in order to take different benefits from them. The goal behind creating spam web pages is to cheat the search engine in such a way that it presents spam pages to the web users, which are entirely non-beneficial and irrelevant to them (Practices, 2015). The ultimate target of any spammers is to improve their spam web page's rank in search engine results. Besides that, there is a substantial economic impact of web spamming. A website with a higher page rank can qualify for free advertisements and large web traffic volume. During the past couple of decades, researchers from industry and academia are working hard to develop advanced web spam detection methods. Still, web spamming methods are evolving, and spammers



PTTA UTM
PERPUSTAKAAN TUNJUNG
TUNJUNG AMINAH

are introducing new spamming methods every day (MCA & Prakash, 2016). Web spam detection research becomes an arms race to challenge an opponent who consistently introduces the latest advanced techniques. Currently, more efficient and advanced search engines are required, which can provide promising results according to the user's search query.

Web page spamming has several adverse effects on both end-user and search engines. Web page spamming is not only wasting time and processing resources but also destroying the storage space. As search engines need to index and store many web pages, so more space is required for storage. In 2017 the total financial losses worldwide caused by spam were considered 450 billion dollars (Admin, 2018). In 2016, the economic loss from spam in the United States surpassed 1.3 billion dollars. In 2009, spam's global financial losses were considered 130 billion dollars (Rao & Reiley, 2012). Moreover, based on the user's search query, a search engine needs to search the large corpus of web pages. Therefore, more time is required for searching and retrieving the relevant data. This decreases the search engine's effectiveness and reduces the end user's trust in the search engine (Li *et al.*, 2018).

Enhancement in anti-web spamming techniques is required to overcome the web spamming attacks. Generally, there are three different types of web spamming, (a) content-based web spamming, (b) link-based web spamming, and (c) cloaking. In cloaking, spammers offer content to the end-users entirely different from the content submitted to the search engine spiders (Guo & Guan, 2018). However, content and link-based web spamming are the most common types of web spamming. Furthermore, any web spammer technique for changing a search engine's logical view over the web page contents is known as content-based spamming (Shahzad *et al.*, 2020). And it is a widespread type of web spamming (Spirin and Han, 2012). As most of the search engines are using information retrieval models (IRM) such as BM25 (Robertson *et al.*, 2004), vector space model (Salton *et al.*, 1975), statistical language model (Zhai, 2008), and probabilistic model. These models are applied to the content of web pages for ranking the web page. Therefore, content web spamming is very popular among web spammers. They are utilizing these models' vulnerabilities for manipulating the content of the spam web page (El-Mawass & Alaboodi, 2017). For example, they might use the famous keywords on the spam web page many times to increase the keywords frequencies. Or copy the legitimate website's content, produce the machine-



generated content for spam web pages, and add all the words from a dictionary to a spam web page. Then, change the text color of dictionary words like background color so the user cannot see these words on the spam web page and only visible to search engine spiders.

The other widespread type of web spamming is link-based web spam. Petkova (2019) explained link-based spamming as if the links among several pages are present only for web spamming and is known as link-based web spamming. In link-based web spamming, spammers establish the link structure to get recognition from link-based algorithms. For instance, the PageRank algorithm will assign a higher rank to a web page if other highly ranked web pages point to the web page with backlinks. Web spammers are using link spamming techniques for various reasons, ranging from malware propagation to money-related activities. Due to the fast growth and volatile nature of the internet, spammers are introducing more sophisticated techniques to generate more revenue (Z. Li *et al.*, 2012; Stone-Gross *et al.*, 2011; Zarras *et al.*, 2014). Unfortunately, these practices have several negative impacts on both search engines and the user's experience. For instance, users get annoyed when they cannot find what they are looking for, and their systems face possibly high-security threats due to malicious content on these spam web pages.

Researchers developed several new anti-web spamming techniques to overcome this issue (Ledford, 2015). However, web spammers keep a close eye on anti-web spam techniques, and they constantly improve their spamming practices to avoid detection. Therefore, a couple of years old designed methods for web page spam detection might not detect spam pages with high accuracy after few years. Due to the continuously changing nature of web page spamming techniques and introducing new spamming methods every day, tweaking is also required in existing web page spam detection techniques. The techniques designed a few years ago cannot detect today's advanced spamming practices. Every search engine is changing and tweaking its algorithms every year, and Google changes its algorithm several times every year (Moz, 2020). Search engines already provided the guidelines to website owners in achieving excellent PageRank. However, spammers are still using Facebook search engine optimization (SEO) groups, SEO forums, paid link services, subreddits, and SEO service providers to achieve high PageRank using illegal SEO techniques (Sharma *et al.*, 2019).



Based on the literature review, it is noticed that a lot of work has been done to detect a single type of web page spamming techniques, and most of the researchers are focusing on improving the methods that can detect only one kind of web page spamming. The drawback of using these techniques is that they can only identify a single type of web page spam technique. For example, a technique designed for a specific kind of content-based spam page detection can only classify those web pages as spam that are practicing this content-based spamming technique. It will not identify other types of content and link-based web page spamming techniques.

However, very little work has been done using a combined approach. Egele *et al.* (2011) introduced a new technique and used a j48 decision tree classifier in their approach to differentiate the legitimate web pages from the spam web pages. They could detect one spam page out of five spam web pages by decreasing the false positive to zero. Fdez-Glez *et al.*, (2016) proposed a WSF2: a novel framework for filtering web spam, particularly fit for identifying spam content on web pages. Their framework allows the effortless combination of several filtering techniques. They combined the content and link-based features for spam web page detection and applied their proposed framework on publicly available WEBSpAM-UK datasets and achieved 79.8 percent accuracy. Roul *et al.* (2016) proposed a combined approach of content and link-based techniques for web spam detection. They used part of speech (POS) ratio and term density to detect content-based spam and explored the Personalized Page Rank to classify web pages as non-spam or spam. They used the WEBSpAM-UK2006 dataset to compare their experimental results with the existing techniques, and an excellent F-measure of 75.2 percent shows the effectiveness of their approach. Singh and Singh (2018) proposed a combined technique for web page spam detection using content and link-based features. They combined the characteristics of the Particle Swarm Optimization strategy and Correlation Based Feature Selection Technique. The authors assessed the performance of CFS-PSO on the WEBSpAM-UK dataset with five different classifiers Naive Bayes, AdaBoost, SVM, J48, and MLP classifier. Their experimental results show an 88 percent reduction in original features and achieved the maximum F-measure of 78.4 percent for the Naive Bayes Classifier. Asdaghi and Soleimani, (2019) proposed an effective technique for spam page identification. Their technique combined the content and link-based features and then applied the features reduction method to reduce the features. They used the standard WEBSpAM-UK



datasets with the Naïve Bayes classifier and achieved 71.8 percent accuracy with 63.6 percent recall.

Moreover, almost every researcher in the field uses the standard datasets provided by the Laboratory of Web Algorithmic to conduct their experiments. Conducting the experiments with thirteen years old dataset has some limitations, such as whether it is fit for the rapid development of today's web page spam detection techniques (J. Liu et al., 2020). Therefore, new datasets are required for the development of today's advanced web spam detection techniques as well as dataset collection where an improved data collection architecture is needed. In the first place, Zarras *et al.*, (2015) used the idea of collecting data using the non-traditional method by proposing the concept of data collection architecture for collecting web spam data from SEO forums. Zarras's idea can be enhanced to propose an improved data collection architecture for obtaining the new spam web page dataset.

Therefore, this research focuses on improved data collection architecture and a combined approach for content and link-based web spam detection. The benefit of improved data collection architecture is it will provide a new dataset that practices the current spamming techniques, and this dataset can be used to propose the new techniques for spam web page detection. Similarly, the combined approach's benefit is that it can identify web pages practicing either content or link-based web spamming techniques, which ultimately improves web spam detection accuracy. The existing conventional web spam detection methods cannot identify the newly evolved spams, such as link pyramids. Therefore, there is a need for an improved combined approach.

1.2 Problem Statement

From the prior studies on web spamming, it was realized that search engines are continuously ranking the malicious websites high on their results pages, which are usually related to pornography, pirated software, mortgage consolidation, and general consumer search terms. Users are getting directed to websites by search engines, which may be malicious in some sense. Most of the researchers focus on content or link-based techniques for web spam detection ignored the power of combined web spam detection techniques, only a few researchers worked on combined techniques. The recently combined techniques cannot achieve excellent web spam detection accuracy

and remain unsolved (Roul *et al.*, 2016),(Singh and Singh, 2018). Furthermore, the existing combined techniques cannot efficiently detect the newly invented spamming techniques such as link pyramids. Due to several adverse effects, web spam detection is becoming a considerable challenge for all search engines. Therefore, an improved combined approach is required to detect the existing spam pages with high accuracy and detect newly evolved techniques in both content and link-based web spamming.

Moreover, from the literature review, it was identified that there are only two standard datasets (WEB SPAM-UK2006 and WEB SPAM-UK2007) available for the researchers to conduct their experiments. The datasets were developed in 2006 and 2007, as web spamming is continuously evolving, and web spammers introduced several new web spamming techniques. Therefore, these standard datasets cannot be used to propose advanced techniques that can identify existing and newly evolved web spam (Liu *et al.*, 2020). Due to continuous evolution in web spamming techniques, there is a need for an improved data collection architecture to build a new dataset. The new dataset can be used to develop advanced and better web spam detection techniques.

1.3 Aims of the Research

This study aims to improve web spam detection accuracy by introducing an improved combined technique and providing the roadmap for further research in the area. This research advances by introducing improved techniques for content and link-based web spam detection. Moreover, this study pursues a data collection architecture to reveal the secret relationship behind the spam network, identify the spam signals on spam web pages, and collect the new dataset that is practicing the newly evolved web spamming techniques. The proposed improved combined approach improves web spam detection accuracy and detects the spam web pages that participate in link farms and link pyramids creation with high accuracy.

1.4 Objectives of the Research

This research includes the following three objectives:

1. To propose an improved data collection architecture that collects the updated spam data and reveals the secret relationships and techniques behind the spam network.
2. To propose a combined approach using the content and link-based approach that detects both content and link-based spam web pages accurately and can help identify the newly evolved web spamming technique link pyramid.
3. To evaluate and compare the proposed combined approach's performance in (b) with other existing techniques in terms of accuracy.

1.5 Scope of the Research

This study focuses on the content and link-based web spam detection algorithms and techniques to solve the problem of identifying several types of existing and newly evolved web spamming techniques. The proposed framework will identify the English language spam web pages. The study also focuses on the data collection architecture to obtain a new dataset for conducting the experiments. A database is used to store and preprocess the dataset. Moreover, the content and link-based techniques are combined to improve web spam detection accuracy. The accuracy of the proposed combined technique is verified and compared with other benchmark techniques.

1.6 Significance of the Research

This research contributes to the field of web spam detection in the following directions.

1. The proposed improved data collection architecture helped in collecting the updated spam data and revealed the secret relationships and techniques behind the spam network
2. The proposed improved framework for content and link-based web spam detection used a combined approach for detecting the content and link-based spam web pages with more accuracy than existing benchmark techniques.
3. The proposed improved combined approach helped in the identification of the newly evolved web spamming technique link pyramid.

1.7 Outline of the Thesis

We subdivided the thesis into five chapters, including the Introduction and Conclusion chapters. The outline of each chapter is as below.

In addition to presenting an outline of the thesis, Chapter 1 includes an overview of the background studies, the research's scope, aims of the research, objectives, and importance of the research undertaken.

Chapter 2 outlines the prior studies made on web spam detection with a detailed overview of the use of search engine optimization methods for spamming, and algorithms developed to detect these spamming methods are reviewed. After an in-depth review, problems and improvements in previously developed algorithms are highlighted, and the necessity for further improvements is intimated. The chapter examines the content-based algorithms, linked-based algorithms, combined techniques for web spam detection, and some non-traditional algorithms for web spam detection. Moreover, Chapter 2 outlines the prior studies made on data collection architecture. Finally, Chapter 2 ends while discussing the research gap analysis on the combined approach.

Moreover, Chapter 3 presents data pre-processing and data collection architecture used to reveal the secret relationships and techniques used in link-spam networks. Based on Chapters 2, Chapter 3 presents proposed content and link-based frameworks for web spam detection, for instance, the keywords density method, stopwords technique, spam keywords database, POS and Unicheck for identification of content-based spam, and spam link database, spam signals, and link farm algorithm for detection of newly evolved link pyramid spamming.

Finally, this chapter introduces the proposed combined approach for web spam detection with more accuracy. The proposed techniques, such as keywords density, stopwords density, POS ratio test, paid-links database, link farm algorithm, and all other techniques, are implemented and tested for accuracy.

Chapter 4 ends with the comparison of our proposed combined technique's results with other conventional methods. In chapter 5, we summarized the research, and this chapter also discussed the future works at the end.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter starts by describing the search engine optimization techniques and their types in detail. Furthermore, the chapter discussed the web spam taxonomy and the different techniques used for web spamming. In the same section, different spamming techniques used by web spammers are divided into subcategories based on their types. Researchers worked on various web spam detection and prevention algorithms and techniques for content and link-based web spamming, while very few worked on combined approaches. Their proposed algorithms and techniques for web spam detection are discussed thoroughly. In the same section, the web spam detection algorithms are divided into four subcategories based on their types. And all subcategories are discussed in detail. Then further down in the sections, there is a discussion on the importance of improved web spam detection techniques. Finally, the chapter is concluded with details on research gap analysis on the combined approach for web spam detection.

2.2 Search Engine Optimization and Web Page Spamming

Search engine optimization (SEO) improves web traffic to a given web page by improving the web page's visibility in the search engine (SE) results. SEO professionals enhance web page relevancy by improving the web page's content and

ensuring that the web page can be indexed accurately (Zhang & Cabage, 2017). There are two types of SEO, on-page and off-page SEO. On-page optimization is also known as on-site optimization, where the web pages are optimized to get a higher rank on search engine results pages (SERPs) to get the more relevant traffic in search engines (SEs). On-page optimization deals with a web page's contents that affect the search engine rankings (Carragher & Palmer, 2017).

Off-page optimization is also used for improving the rank of a website on search engine results pages. It is a common perception that only link building is off-page optimization, but it is not always true. There are many other techniques for off-page optimization, for instance, Social Media Marketing and Social bookmarking. However, these techniques mainly focus on developing the incoming links to a web page (Adams *et al.*, 2017) by using different link creation methods. Such as by submitting the web page's URL to search engines, social media, listing directories, and other link building repositories (Carragher & Palmer, 2017). SEO experts are using three different types of techniques on on-page and off-page search engine optimization. The first technique is white hat SEO, which is the proper SEO method, which fulfills the search engine's guidelines (Google, 2017), and it remains for a long time. The second technique is gray hat SEO, the usage of grey hat techniques is technically legal for increasing the web page rankings, but it is ethically doubtful, and one day, it could become a black hat technique (Regalado *et al.*, 2015). And the third technique is black hat SEO, which is not a proper method for search engine optimization. It is used to get extraordinary SE rankings unethically by abusing the SE guidelines.

Another name for these techniques is spamdexing or web poisoning (Giomelakis & Veglis, 2016). Spamdexing is a mixture of two different words, 'spam' and 'indexing'. Spamdexing is also known as search engine spamming, which produces intentional manipulation of SE ranking's results to get more web traffic on an undeserving web page to obtain a higher ranking in all vital search engines (Castillo *et al.*, 2006). This technique is volatile and aggressive, and search engines discouraged the technique of approaching the search engine optimization ranking process. Black hat techniques are mainly designed for search engines and not for real users (Carragher & Palmer, 2017). Web spammers are using several web spamming techniques for indexing their websites on top in SERPs. Based on the working mechanism, web



spamming techniques can be categorized into different categories. Figure 2.1 depicts the categorization of web spamming techniques, where it mainly categorized into two categories, on-page web spamming, and off-page web spamming and then these main categories are sub categorized in several categories.

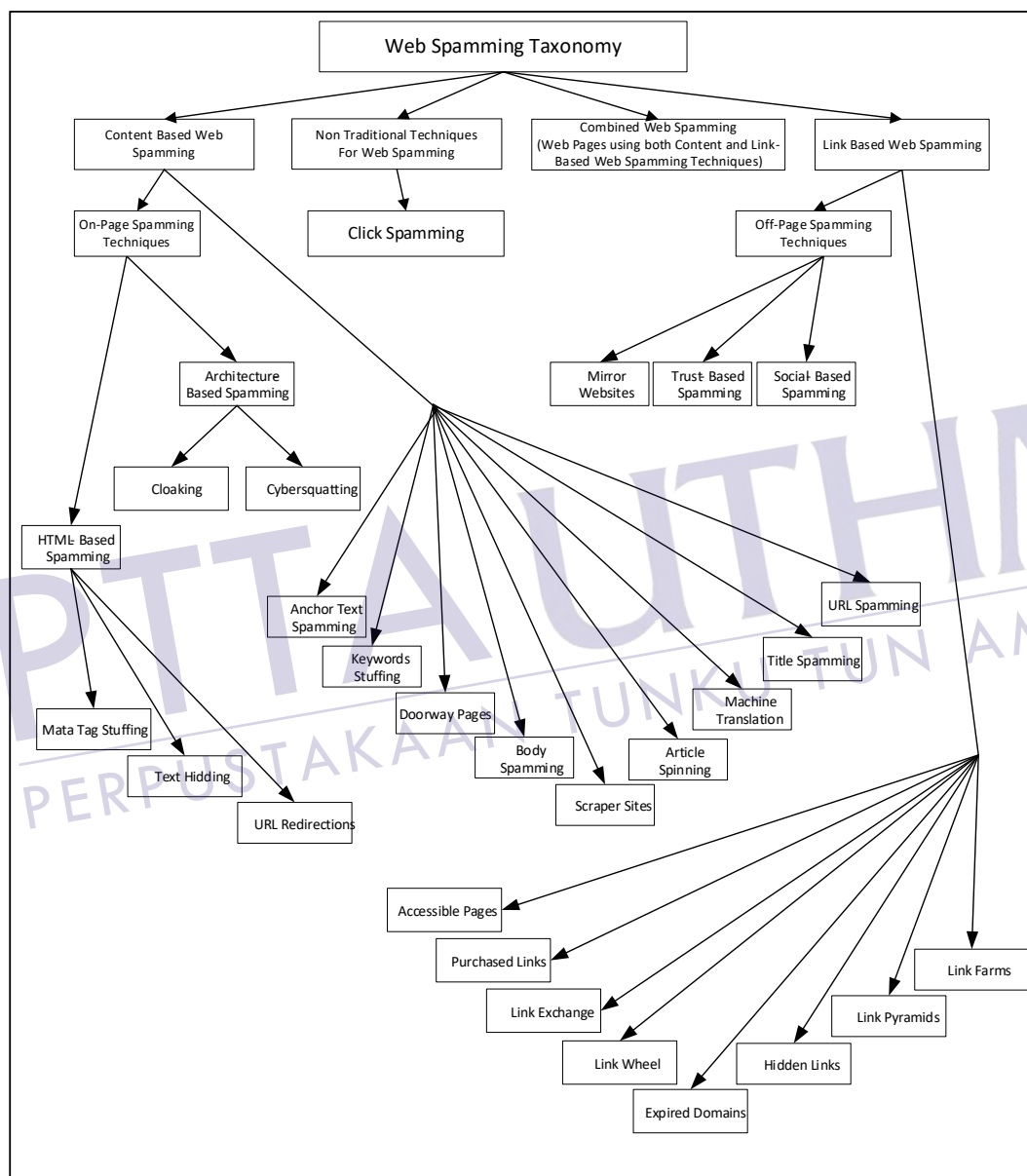


Figure 2.1: Categorization of web spamming techniques

In order to identify and resist the web spamming techniques as depicted in Figure 2.1, researchers proposed several algorithms and methods (Chandra, A; Suaib, 2014). All

proposed techniques can be categorized into four different groups. Figure 2.2 depicts the categorization of techniques combating web spamming.

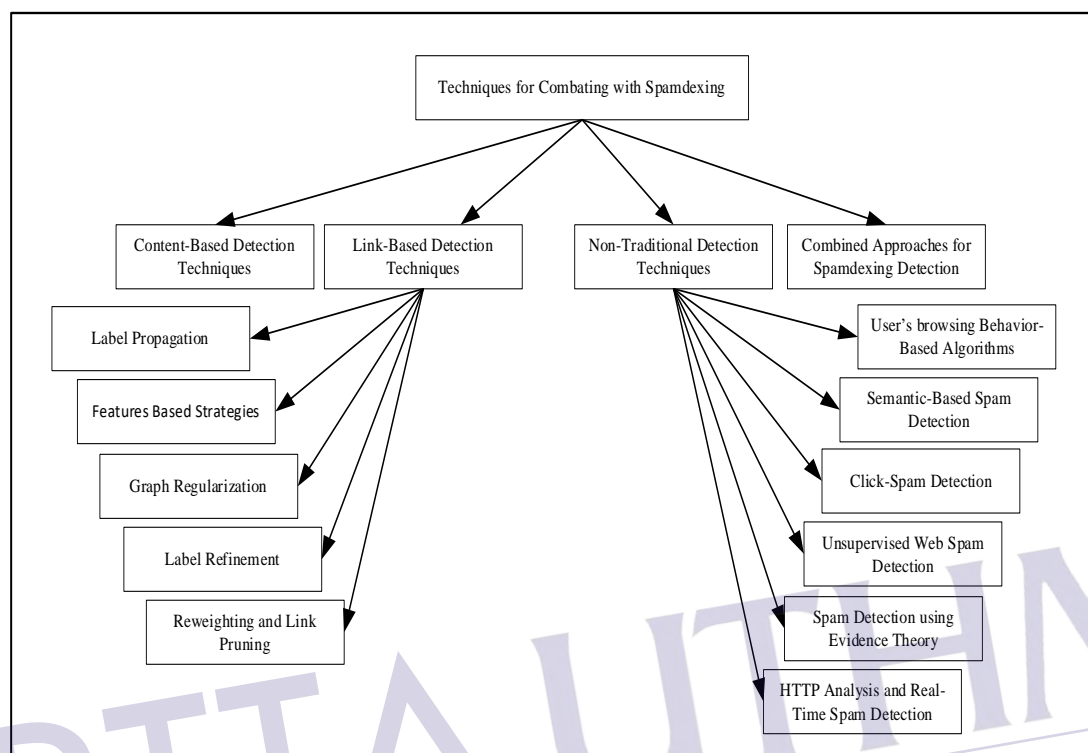


Figure 2.2: Categorization of anti-web spamming techniques

The first group consists of techniques that analyze content features, such as content duplication, word counts, or language models. The second group of techniques includes link-based information such as performs link-based trust and distrust propagation, link pruning, neighbor graph connectivity, graph-based label smoothing, and study statistical anomalies. The third group consists of non-traditional techniques for web spam detection, such as exploiting clickstream data, HTTP sessions information, and user behavior data. Finally, the last group practicing the combined approach for web spam detection. The web spam combating algorithms in each group are discussed in detail in the next sections.

2.3 Algorithms for Content-Based Web Page Spam Detection

Fetterly *et al.* (2004) did the fundamental research on content-based spam detection, and they provided the base for content-based spam detection and prevention

algorithms (Fetterly et al., 2003, 2004, 2005; Jindal & Liu, 2008). Usually, in content-based techniques, statistical analysis is used for spam detection (Fetterly *et al.*, 2004). Spammers are mostly using the software to generate spam web pages (Gan & Suel, 2007). They apply software in the weaving and phrase stitching techniques and are not meant for real users. These pages present the unusual properties (Gyongyi & Garcia-Molina, 2005). In different studies conducted by researchers in the area, they have identified that spam page URLs have an unprecedented number of dashes, dots, length, and digits (Spirin & Han, 2012). The authors reported that 80 percent of the longest identified hostnames point towards the adult web pages, while 11 percent refer to financial credit web pages. The researchers also determined that these web pages themselves have a replicated nature. Most of the spam web pages on the same host will have a low word count variation.

Another very fascinating observation by researchers is that spam web pages' change quickly. They mainly observed the average amount of weekly changes on a given host for all the pages and identified that most of the very active spam hosts could be detected using this feature. All recommended features can be seen in the article (Fetterly *et al.*, 2004). In other work done by them (Fetterly et al., 2003, 2005), they worked on content duplication and identified that massive clusters with identical content are spam. (Fetterly *et al.*, 2004) applied the shingling technique (Broder *et al.*, 1997) for finding the duplicate content and such clusters, which is the Rabin fingerprint-based method

In the research article, Jindal & Liu, (2008) performed a more in-depth analysis and came up with few other content-based features. Finally, they combined all these features in a classification model within boosting, bagging frameworks, and C4.5. In another work, Gyöngyi *et al.*, (2004) presented a detailed report on how machine learning models and features can improve web spam detection algorithms' quality. The researchers obtained excellent classification outcomes using state of the art learning models, RandomForest and LogitBoost, and low computational content features. Gyöngyi *et al.*, also identified the global and computationally demanded features such as, PageRank (PR) yield only little additional improvement in the quality. Therefore, the proper and careful choice of a machine learning model is critical.

Piskorski *et al.*, (2008) and Sydow *et al.*, (2007) analyzed the linguistic features for identifying web spam. They considered various Natural Language Processing



(NLP) features. Such as lexical and content diversity, emotiveness, lexical validity, usage of active and passive voices, entropy and syntactical diversity, and several other features. Finally, several features are proposed by Benczúr *et al.*, (2007) based on the appearance of keywords on a webpage that is either hugely spammed or excellent advertising value. Authors also investigate the discriminative influence of the following features: Yahoo Mindset Classification of a web page as either non-commercial or commercial, Online Commercial Intention value allocated to an URL in a Microsoft adCenter, number of Google AdSense ads on a web page, and popular keywords by Google AdWords (Spirin & Han, 2012). The accuracy of spam detection is 3% more than the work done by Castillo *et al.*, (2006), and they did not consider these features.

Chellapilla and Chickering, (2006) analyzed the advertising click-through logs and Search Engine (SE) query logs for monetizability and query popularity. Monetizability is that income generated by all advertisements presented to the users in response to their search keywords, while query popularity is popular keywords used by the users for searching relevant information on the internet. Out of all keyword categories, authors used the top five thousand keywords, and then, they requested the top two hundred links for every keyword four times by using the several agent-fields to mimic requests by a Crawler (c) and a User (u). Moreover, they applied the cloaking test using Equation 2.1, the revised version of the cloaking detection test proposed in the initial work (Wu & Davison, 2005, 2006).

$$CloakingScore(p) = \frac{\min[D(c_1, u_1), D(c_2, u_2)]}{\max[D(c_1, c_2), D(u_1, u_2)]} \quad (2.1)$$

Where

$$D(a_1, a_2) = 1 - 2 \frac{a_1 \cap a_2}{a_1 \cup a_2} \quad (2.2)$$

Equation 2.2 represents the normalized-term-frequency difference for two copies of a web page described as a set of terms. Chellapilla and Chickering, (2006) reported 0.75 and 0.985 accuracies at high recall values for monetizable and popular queries, which apparently suggests that their proposed method is beneficial for cloaking detection. However, there are some weaknesses in their work. The proposed technique can have a slightly high false-positive rate. For instance, legitimately generated dynamic web pages also include several links and terms on each access. To overcome this weakness,

they enhanced their previously proposed method by using the structural properties of a web page (Lin, 2009). The enhanced method used tags instead of links and words on web pages to calculate the cloaking score.

Ji and Zhang, (2015) analyzed the content features of spam and non-spam web pages. The content features of non-spam web pages contain several statistical regularities. While spam web pages contain only a few statistical regularities due to the reason that spam web pages are created randomly with a lot of duplicate content to boost the PR in SERPs. Ji and Zhang explored the content features of pages and identified huge differences between spam and no-spam web pages of content features. Authors examined the content features includes the number of words in the title, number of words on the page, the average length of the word, the anchor's text fraction, the fraction of the visible text, corpus precision, compression rate, corpus recall, query recall, query precision, entropy, and independent LH. After examining the features above, it was identified that many regularities exist in the content features for non-spam web pages, while a few regularities exist in spam web pages (Ji & Zhang, 2015).

Rubinstein *et al.*, (2017) introduced spam detection and prevention strategy in social media networks. This social networking system is consisting of two modules, detection, and a prevention module. The detection module detects the spam comments posted by the users. At the same time, the prevention module extracts the content signals to analyze and determine whether the comment includes spam content or not. After identifying the spam content based on social signal analysis and content signal analysis, the prevention module acts by blocking the spam comment, and it also educates the user who posted the spam comment. In (Jain et al., 2018), the authors discussed the importance of spam detection strategies in a social media text. The authors proposed using Convolutional Neural Network (CNN), a deep learning technology for spam detection and suggested adding an extra semantic layer on the top of it. The proposed model is known as the Semantic Convolutional Neural Network (SCNN). After testing the model using the Twitter dataset and SMS Spam dataset, they reported very high spam detection accuracy. Automated article spinning tools are more favorite among the web spammers and they are using spinning tools to avoid the detection of duplicate contents.

In any input article, the spinning tool can replace the phrases or words with a synonym to cheat the plagiarism detectors. Spinning tools are easy to use, and with

very few clicks, spammers can spin the given article hundreds of times, and then using web proxies, they are posting their spam articles on target websites. Zhang *et al.*, (2014) introduced a method for the identification of spun content. This technique is directly attached to the underlying mechanism used by article spinning tools. The author's method was based on the phrases or words which do not change when the spinning tools generate spun content. Based on the idea above, they developed a tool Dspin for the identification of spun articles. For experiments, they used two datasets of crawled articles to identify automatically spun articles and spamming behavior of the spammer. The researcher also identified link-based web spam, and they proposed several algorithms and techniques for link-based web spam detection. All web spam combating algorithms which are using link-based features are categorized and discussed in the next sections.

2.4 Algorithms for Link-Based Web Page Spam Detection

Based on the working mechanism, all web page spam detection link-based algorithms can be divided into five categories. The first category focuses on the recognition of suspicious nodes, links, and their succeeding down-weighting. Web spam detection algorithms in the second category deal with the topological relationship between a set of web pages with known labels and web pages with unknown labels. Graph regularization methods are used by the third category of link-based web spam detection algorithms. The concept of label refinement based on web graph topology is used by the fourth category of link-based spam detection algorithms. The algorithms in the final category extract the link-based features for every node and apply several machine learning techniques to detect web spam.

Table 2.1 shows the category of link-based techniques. The comparison criteria are based on the working mechanism, algorithms used, complexity, and information type used. As most of the link-based techniques support the Hyperlink-Induced Topic Search (HITS) and PageRank (PR) algorithms, among all these algorithms, PR and HITS are the most critical ones. Each category of link-based web spam detection algorithms is discussed in detail in the next subsections.

Table 2. 1: The comparison of link-based detection algorithms

	<i>Techniques</i>				
	Label Propagation Strategies	Feature-Based Strategies	Graph Regularization Strategies	Label Refinement Strategies	Reweighting and Link Pruning
<i>Algorithms used by spam detection Strategies</i>	PageRank TrustPage	Truncated PageRank	PageRank	Clustering Algorithms	HITS PageRank
<i>The working mechanism of Spam detection Strategies</i>	Exploits the topological relationship between the web pages	Work-based on graph regularization method	Uses the idea of label refinement based on the web graph topology	Extracting link-based features for each node and use various machine learning algorithms	Identify the suspicious nodes and links and their subsequent down weighting
<i>Techniques used for Mining</i>	Web Structure Mining	Web structure Mining	Web Structure Mining	Web Content Mining	Web Structure Mining and Web Content Mining
<i>Information Type used by the web spam detection Strategy</i>	Topological Relationship	Structural Patterns	URL	Link base features of each node	Down weighting of links and nodes
<i>Complexity</i>	Internal Structure	-	URL Classification	Limited data set are allowed	Relationship between the nodes

2.4.1 Algorithms Based on Label Propagation

The first category of Link-based web spam detection is Label propagation. The fundamental concept behind the label propagation-based algorithms is to analyze the group of web pages on the internet for which labels are already known and then by applying the several propagation rules for computing the labels of other nodes. TrustRank (TR) is one of the earliest algorithms from this group. It uses the personalized PageRank for propagating the trust from a tiny seed set of excellent web pages (Gyöngyi *et al.*, 2004). The general perception is that excellent web pages point mostly to the best quality websites. TrustRank relies on the principle of the relative isolation of a perfect set of pages.

The authors suggested the inverse PageRank for selecting the seed set of legitimate famous pages. It works on the graph with all edges reversed. After

calculating the inverse PageRank values for all the web pages on the World Wide Web, they took Top-P web pages and asked the human experts to judge these web pages' status. The personalization vector is constructed by them where components corresponding to creditable judged web pages are non-zero. Ultimately, Personalized Page Rank is computed. As compare to PageRank, TrustRank shows good properties for web spam status. Anti-TrustRank (ATR) is the follow-up work on trust propagation (Krishnan & Raj, 2006). Manaskasemsak & Rungsawang, (2015) Worked on a web spam detection problem and proposed a novel technique that adopts the ACO learning to construct a rule-based classifier. The authors proposed three different strategies, Trust-ACO, Distrust ACO, and Combine ACO, which depend on distrust and trust hypotheses. The first strategy is designed for constructing the non-spam classifier, which detects the non-spam from spam pages.

Besides the TrustRank, there is another technique known as distrust propagation, and researchers used this technique on a group of spam web pages on an inverted graph. They selected the seed set of web pages having high PR values. This approach to finding the spam web pages outperformed the TrustRank with high accuracy. Some more researchers worked on Trust and Anti-TrustRank (Leng *et al.*, 2014; Smitha, 2017; Whang *et al.*, 2018). In Zhang *et al.*, (2014), another group of researchers proposed a new semi-automatic anti-spam algorithm, TDR. This method is taking advantage of trust and distrust propagation and also implements differential trust and distrust propagation. To each page, TDR assigns a D-Rank and T-Rank Score for simultaneous propagation from seed sets using bidirectional links to the whole web. D-Rank/T-Rank propagation is penalized by the target's current T-Rank/D-Rank, i.e., trustworthy/untrustworthy web pages received more trust/distrust propagation than an untrustworthy/trustworthy page from a similar source page. TRD overcomes TrustRank and Anti-TrustRank limitations because differential trust/distrust propagation decreases much bad-to-good distrust propagation and good-to-bad trust propagation, which cause adverse impacts to Anti-TrustRank and TrustRank. Authors claimed that their experimental results outperform the existing anti-spam techniques for spam detection and spam demotion tasks (Zhang *et al.*, 2014).

Another algorithm is used for computing the badness of a web page using the inverse page rank calculation which is known as BadRank (BR) (Sobek, 2002). The correlation between page rank and trust rank is the same as the relationship between

bad rank and antitrust rank. Wu et al. worked more on the concept of propagation by investigating how distrust and trust propagation approaches can work together (Davison, 2006). First, the TR algorithm's trust propagation method is challenged; each child receives the equivalent share of trust from a parent $\frac{TR(p)}{|Out(p)|}$. Moreover, they suggested two more procedures:

- i. Logarithmic splitting: In logarithmic splitting, every child receives an equivalent share of the score from a parent normalized by the log of the number of children.

$$c \cdot \frac{TR(p)}{\log(1+|Out(p)|)} \quad (2.3)$$

- ii. Constant Splitting: While in constant splitting, every child receives a similar discounted share of trust from parent $c \cdot TR(p)$ regardless of the number of children.

Several partial trust aggregation strategies are also analyzed by them, while TR only recapitulates every parent's trust score. Finally, they came up with a linear combination of distrust and trust values:

$$TotalScore(p) = \eta \cdot TR(p) - \beta \cdot AntiTR(p) \quad (2.4)$$

Where $\eta, \beta \in (0, 1)$. The investigation shows that the combination of both propagation techniques performs better in identifying web spam detection.

Guha et al., (2004) worked on the concept of trust and distrust propagation, keeping reputation systems in mind. The disintegration property of page rank is utilized by the two algorithms (Benczur *et al.*, 2005; Gyongyi *et al.*, 2006) to determine the underserved number of page ranks coming from dubious pages.

Benczur *et al.*, (2005) proposed the SpamRank (SR) algorithm, to find the sponsors of a webpage, they used Monte Carlo Simulations (Berman & Plemmons, 1994). Benczur *et al.*, introduced the concept of penalty score. A penalty score will be assigned to every web page after analyzing whether PPR score $PPR(\chi_j)I$ $PPR(\rightarrow)_i$ χ_j is shared with dubious web pages. And finally calculates the spam rank for every web page. The fundamental task of this algorithm is assigning the penalty scores. This

proposed technique's primary perception is that page rank obeys the power-law distribution (Pandurangan *et al.*, 2006).

Gyongyi *et al.*, (2006) also introduced the concept of Spam Mass (SM). It calculates the value of page rank coming from spam web pages. Like trust rank, it also requires the core of useful known web pages for calculating the score of page rank coming from the excellent web pages. The algorithm operates in two steps. In the first step, it calculates the PR $\vec{\pi}$ and TR $\vec{\pi}'$ vectors and determines the score of spam mass for every web page by applying the formula $\vec{m} = \frac{\vec{\pi} - \vec{\pi}'}{\vec{\pi}}$. The second step is threshold decision. It depends on the score of spam mass, is made. It is worth mentioning that the algorithm can efficiently use the information about spam web pages.

Page *et al.*, (1999) described the credibility-based link analysis and discussed the idea of each page's k-Scope credibility for its estimation. Pages *et al.*, proposed various techniques and showed how to use them for web spam detection. Explicitly the authors described the idea of Badpath (BP). It is a k-hop random walk beginning from the current webpage and stopping at a spam webpage, and then calculate the tuned k-Scoped Credibility value as

$$C_k(p) = \{1 - \sum_{l=1}^k [\sum_{Path_l(p) \in BadPath_l(p)} P(path_l(p))]\} \quad (2.5)$$

Where k is a parameter which specifies the range of a random walk, $\gamma(p)$ represents the credibility penalty factor which is required to deal with only incomplete information of all spam web pages on World Wide Web and $P(path_l(p)) = \prod_{i=0}^{l-1} w_{ii+1}$. Before doing the link-based ranking or altering the personalization vector in trust rank, antitrust rank, or personalized page rank, a credibility score can be used to prune or down-weight the low trustworthy links.

The author described the anchor's idea as a subset of web pages with already recognized labels, and several anchor-based measures on graphs are investigated. They discussed Personalized Page Rank (PPR), Harmonic Rank (HR), and Nonconserving Rank (NR) in this work. Harmonic rank is defined through a random walk on a transformed graph with an added sink and source such that every anchor vertex is connected to a source, and the sink is connected to every vertex with probability c . Nonconserving rank is the generalization of PPR satisfying the equation

$$\vec{\pi} = (I - (1 - c).M^T)^{-1}\vec{r}. \quad (2.6)$$

The authors concluded that harmonic rank is best for distrust propagation, while non-conserving rank is best for trust propagation. Benczúr *et al.*, (2006) introduced the spam detection algorithm. The algorithm utilizes the web page's similarity. For computing a spam value for a new web page, they used similarity-based top-K lists. For computing the similarity between web pages, authors considered the different methods like CompanionRank (CR), co-citation, KNN-SVD projections, and SimRank (SR) (Jeh & Widom, 2002). Based on the experiments, It can be concluded that similarity-based spam detection produces excellent results at high recall levels, while at low levels of recall, combined trust-distrust (Davison, 2006) and ATR (Krishnan & Raj, 2006) show the higher precision.

Some other researchers did fundamental research in the area and provided the base for further research (Baeza-Yates *et al.*, 2006; Becchetti *et al.*, 2006a, 2008b). They introduced the concept of Functional Rank (FR), by using the several damping functions, they generalized the PageRank. They considered the general formula for ranking:

$$\vec{p} = \frac{1}{N} \vec{1} \sum_{j=0}^{\infty} \text{damping}(j)(M^T)^j \quad (2.7)$$

The authors proved the theorem that whatever is the damping function, for instance 1 is the sum of damping's yields a best defined normalized functional ranking. For proposing the practical techniques of rank calculation, they studied different damping functions like linear, exponential (PR), General Hyperbolic (HyperRank), and quadratic hyperbolic (TotalRank). In another research work done by them, they worked on the application of general damping functions for detecting web spam, and they proposed a truncated PR algorithm (Becchetti *et al.*, 2006). This algorithm is using the truncated exponential model. As spam web pages have a massive number of different supporters at the shorter distance, while at the longer distances, the number of supporters is less. Thus, they suggested the use of the damping function, which ignores the direct participation of in-links for the first several levels.

$$damping(j) = \begin{cases} 0 & \text{if } j \leq J, \\ D(1 - c)^j & \text{otherwise} \end{cases} \quad (2.8)$$

They also propose the probabilistic counting algorithm in the same work. The primary purpose of this algorithm is to determine the supporters for a web page efficiently.

2.4.2 Algorithms Based on Link-based Features

In this second category of link-based web spam detection algorithms, the web pages are represented as feature vectors to perform the clustering analysis or standard classification. Amitay *et al.*, (2003) worked on link-based features to accomplish website categorization based on their functionality. The authors assumed that websites that share that structural pattern, such as the number of outgoing links per web page or average page level, share similar roles on the web. For example, online web directories (Yelp, Yellow Pages, Google My Business, Bing Places) frequently consist of web pages with the high ratio of the outgoing links to inlinks, like a structure of a tree, with the depth of a web page the number of outgoing links increases. While spam websites consist of particular topologies with the primary aim of boosting PR using prohibited techniques. Overall, the researchers represented every website as a vector of sixteen connectivity features and then performed the clustering using cosine for similarity measure. In a dataset of eleven hundred websites, the researchers claimed that they managed to recognize a hundred and eighty-three web spam rings forming thirty-one clusters. Using TrustRank (TR), PageRank (PR), and Truncated PageRank (TPR) computations, several link-based features are derived by Becchetti *et al.*, (2006b).

Kumar *et al.*, (2016), proposed a fascinating strategy, Dual-Margin Multi-Class Hypersphere Support Vector Machine (DMMH- SVM), which can classify the web spam automatically based on web spam type. They also introduced cloaking-based spam features to support their classifier model and achieve high precision and recall rate. The proposed classifier DMMH-SVM classifies the web pages into four different categories, i.e., cloaking spam, link spam, content spam, and combined spam. The authors reported that their classifier model could effectively categorize web spam and achieve high precision, accuracy, and recall. In Iqbal *et al.*, (2017), the authors compared different machine learning classifiers currently used for web spam

classification. The authors evaluated the efficiency of current machine learning classifiers and discussed the ideas about new features, which can be helpful for web spam detection.

Patil and Bhadane, (2016) studied the different efficient spam identification methods based on a classifier that joins language models with new link-based features. Singh and Singh, (2018) proposed CFS-PSO, a Swarm-Based hybrid technique, and this hybrid technique consolidates the characteristics of the Particle Swarm Optimization (PSO) strategy and Correlation-Based Feature Selection (CFS). For Machine Learning and Data Mining, the crucial part is feature selection (pre-processing strategy). The main goal of feature selection is to build a simple and more logical model to improve the performance regarding increasing the accuracy and decreasing the time to develop the learning model. The authors evaluated the performance of their technique on WEB SPAM-UK2006 with five classifiers. Their empirical results showed a decrease in original features and an increase in F-measure up to 88% and 45.83% sequentially.

2.4.3 Algorithms Based on Graph Regularization

The third group of the Link-based web spam detection algorithm is more effective because it utilizes the web graph to smoothen the predicted labels. Some empirical analysis and studies proved that using graph regularization algorithms for web spam detection is more effective. Tikhonov *et al.*, (1977) and Wahba, (1990) used the regularization theory for spam detection as described in Abernethy *et al.*, (2010). The main reasons behind it are that it is addressing the fact that spammers are not placing the hyperlinks randomly, and to some extent, there is a similarity between linking web pages Chakrabarti, (2002) and Davison, (2000b). This motivated them to add the regularizer to the objective function for even predictions. In addition, the second reason is that it is using the technique of approximate isolation of genuine web pages that argues for asymmetric regularizer. They come up with the following final objective function:

$$\Omega(\vec{w}, \vec{z}) = \frac{1}{l} \sum_{i=1}^l L(\vec{w}^T \vec{x}_i + z_i, y_i) + \lambda_1 \|\vec{w}\|^2 + \lambda_1 \|\vec{z}\|^2 + \gamma \sum_{(i,j) \in \mathcal{E}} a_{ij} \Phi(\vec{w}^T \vec{x}_i + z_i, \vec{w}^T \vec{x}_j + z_j), \quad (2.9)$$

Where \vec{w} expresses the vector of coefficients, \vec{x}_i and y_i representing the features and a real label correspondingly, $L(a, b)$ denotes the loss function, bias term is z_i , the weight of the link $(i, j) \in \mathcal{E}$ is represented by a_{ij} , and the regularization function is $\Phi(a, b) = \max[0, b - a]^2$. To find the optimization problem solution, the authors provided two different methods, alternating optimization and the conjugate gradient. For the host graph, the problem of weights setting is also studied by these researchers and concludes that the best results can be achieved by the logarithm of the number of links. Finally, their experimental study proved that the algorithm got excellent scalability properties. A discrete analog of classification regularization theory (Tikhonov *et al.*, 1977; Wahba, 1990), is built by (Dengyong Zhou *et al.*, 2007) by determining discrete operators of Divergence, gradient, and Laplacian on the directed graphs, and they propose the following algorithm. Initially, the inverse weighted PR is computed with transition probabilities, which are defined as $a_{ij} = \frac{w_{ji}}{\ln(p_i)}$. In their second step, they created the graph Laplacian.

$$L = \Pi - \alpha \frac{\Pi A + A^T \Pi}{2} \quad (2.10)$$

Here α represents the user-specified parameter in $[0, 1]$, the transition matrix is represented with A , and Π represents the diagonal matrix with the PR score over diagonal. Then, they solved the matrix equation below.

$$L\vec{\varphi} = \Pi\vec{y} \quad (2.11)$$

Where vector \vec{y} is consisting of three values $\{-1, 0, 1\}$, if the page is normal then the value of $\vec{y}_i = 1$, if the page is spam the value of \vec{y}_i is 0 and if the page's label is unknown, then the value of the vector \vec{y}_i is 0. Finally, they used the sign of the corresponding component of the vector $\vec{\varphi}$ for the classification decision. The algorithm works excellent on graphs that are strongly connected.

Some more researchers conducted different studies on the algorithms based on graph regularization and come up with different interesting results. (Cheng *et al.*, 2011) gave the idea of extracting web spam URLs from search engine optimization

REFERENCES

- Abernethy, J., Chapelle, O., & Castillo, C. (2010). Graph regularization methods for Web spam detection. *Mach. Learn.*, 81(2), 207–225. <https://doi.org/10.1007/s10994-010-5171-1>
- Adams, C., LaRiviere, K., & Jones, J. (2017, April). *Method and system of optimizing a web page for search engines*. Google Patents.
- Admin, T. (2018). FBI 2017 Internet Crime Report.
- Agrawal, M., & Velusamy, R. L. (2016). Unsupervised spam detection in hyves using SALSA. *Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015*, 517–526. Springer.
- Amitay, E., Carmel, D., Darlow, A., Lempel, R., & Soffer, A. (2003). The connectivity sonar: detecting site functionality by structural patterns. *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, 38–47. ACM.
- Amrutkar, C., Kim, Y. S., & Traynor, P. (2017). Detecting mobile malicious webpages in real time. *IEEE Transactions on Mobile Computing*, (8), 2184–2197.
- Ardi, C., & Heidemann, J. (2019). Precise detection of content reuse in the web. *ACM SIGCOMM Computer Communication Review*, 49(2), 9–24.
- Asdaghi, F., & Soleimani, A. (2019). An effective feature selection method for web spam detection. *Knowledge-Based Systems*, 166, 198–206.
- Baeza-Yates, R. A., Castillo, C., López, V., & Telefónica, C. (2005). Pagerank Increase under Different Collusion Topologies. *AIRWeb*, 5, 25–32.
- Baeza-Yates, R., Boldi, P., & Castillo, C. (2006). Generalizing pagerank: Damping functions for link-based ranking algorithms. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in*



PTTA UTHM
PERPUSTAKAAN TUN AMINAH

Information Retrieval, 308–315. ACM.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2006a). Using rank propagation and probabilistic counting for link-based spam detection. *Proc. of WebKDD*, 6.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2008). Web Spam Detection: Link-based and Content-based Techniques. *The European Integrated Project Dynamically Evolving Large Scale Information Systems DELIS Proceedings of the Final Workshop*, 222, 99–113. Retrieved from http://www.chato.cl/papers/becchetti_2008_link_spam_techniques.pdf

Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. A. (2006b). Link-based characterization and detection of web spam. *AIRWeb*, 1–8.

Benczúr, A. A., Csalogány, K., & Sarlós, T. (2006). Link-based similarity search to fight web spam. *In AIRWEB*. Citeseer.

Benczur, A. A., Csalogany, K., Sarlos, T., & Uher, M. (2005). Spamrank—fully automatic link spam detection work in progress. *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 1–14.

Benczúr, A., Bíró, I., Csalogány, K., & Sarlós, T. (2007). Web spam detection via commercial intent analysis. *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, 89–92. ACM.

Benoit, Kenneth, David Muhr, and K. W. (2017). Stopwords: Multilingual Stopword Lists. Retrieved August 7, 2020, from <https://cran.r-project.org/web/packages/stopwords/index.html>

Berman, A., & Plemmons, R. J. (1994). *Nonnegative matrices in the mathematical sciences* (Vol. 9). Siam.

Bharat, K., & Henzinger, M. R. (2017). Improved Algorithms for Topic Distillation in a Hyperlinked Environment. *ACM SIGIR Forum*, 51(2), 194–201. ACM.

Bhattacharjee, R., & Goel, A. (2007). Algorithms and incentives for robust ranking. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 425–433. Society for Industrial and Applied Mathematics.

Bianchini, M., Gori, M., & Scarselli, F. (2005). Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1), 92–128.

- Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic clustering of the web. *Computer Networks and ISDN Systems*, 29(8), 1157–1166.
- Carraher, T. R., & Palmer, J. (2017, April). *Search engine optimization using page anchors*. Google Patents.
- Castillo, C. (2008). Datasets for Research on Web Spam Detection.
- Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., & Vigna, S. (2006). A reference collection for web spam. *ACM SIGIR Forum*, 40(2), 11–24. <https://doi.org/10.1145/1189702.1189703>
- Castillo, C., Donato, D., Gionis, A., Murdock, V., & Silvestri, F. (2007). Know your neighbors: Web spam detection using the web topology. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 423–430. ACM.
- Chakrabarti, S. (2002). *Mining the Web: Discovering knowledge from hypertext data*. Elsevier.
- Chandra, A; Suaib, M. (2014). A Survey on Web Spam and Spam 2.0 - ProQuest. *International Journal of Advanced Computer Research*, (2).
- Chatterjee, M., & Namin, A. S. (2018). Detecting Web Spams Using Evidence Theory. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 01, 695–700. <https://doi.org/10.1109/COMPSAC.2018.10321>
- Chellapilla, K., & Chickering, D. M. (2006). Improving Cloaking Detection using Search Query Popularity and Monetizability. *AIRWeb*, 17–23.
- Chen, C., Wang, Y., Zhang, J., Xiang, Y., Zhou, W., & Min, G. (2017). Statistical features-based real-time detection of drifted twitter spam. *IEEE Transactions on Information Forensics and Security*, 12(4), 914–925.
- Cheng, Z., Gao, B., Sun, C., Jiang, Y., & Liu, T.-Y. (2011). Let web spammers expose themselves. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 525–534. ACM.
- Cooley, S. (2015, January). *Systems and methods for associating website browsing behavior with a spam mailing list*. Google Patents.
- Dai, N., Davison, B. D., & Qi, X. (2009). Looking into the past to better classify web spam. *In Proceedings of the 5th International Workshop on Adversarial*



Information Retrieval on the Web, 1–8.

Daswani, N., & Stoppelman, M. (2007). The anatomy of Clickbot. A. *Proceedings of the First Conference on First Workshop on Hot Topics in Understanding Botnets*, 11. USENIX Association.

Davison, B. (2000a). Recognizing nepotistic links on the web. *Artificial Intelligence for Web Search*, 23–28.

Davison, B. (2000b). Topical locality in the web. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 272–279. ACM.

Davison, B. (2006). *Propagating trust and distrust to demote web spam*.

Dou, Z., Song, R., Yuan, X., & Wen, J.-R. (2008). Are click-through data adequate for learning web search rankings? *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 73–82. ACM.

Egele, M., Kolbitsch, C., & Platzer, C. (2011). Removing web spam links from search engine results. *Journal in Computer Virology*, 7(1), 51–62. <https://doi.org/10.1007/s11416-009-0132-6>

El-Mawass, N., & Alaboodi, S. (2017). Data Quality Challenges in Social Spam Research. *J. Data and Information Quality*, 9(1), 4:1--4:4. <https://doi.org/10.1145/3090057>

Fdez-Glez, J., Ruano-Ordás, D., Laza, R., Méndez, J. R., Pavón, R., & Fdez-Riverola, F. (2016). WSF2: A Novel Framework for Filtering Web Spam. *Scientific Programming*, 2016. <https://doi.org/10.1155/2016/6091385>

Fdez-Glez, Jorge, Ruano-Ordás, D., Laza, R., Méndez, J. R., Pavón, R., & Fdez-Riverola, F. (2016). WSF2: a novel framework for filtering web spam. *Scientific Programming*, 2016.

Fetterly, D., Manasse, M., & Najork, M. (2003). On the evolution of clusters of near-duplicate web pages. *Web Congress, 2003. Proceedings. First Latin American*, 37–45. IEEE.

Fetterly, D., Manasse, M., & Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. *Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM*



SIGMOD/PODS 2004, 1–6. ACM.

Fetterly, D., Manasse, M., & Najork, M. (2005). Detecting phrase-level duplication on the world wide web. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 170–177. ACM.

Gan, Q., & Suel, T. (2007). Improving web spam classifiers using link structure. *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, 17–20. ACM.

Geng, G.-G., Li, Q., & Zhang, X. (2009). Link based small sample learning for web spam detection. *Proceedings of the 18th International Conference on World Wide Web*, 1185–1186. <https://doi.org/10.1145/1526709.1526920>

Geng, G., Wang, C., & Li, Q. (2008). Improving web spam detection with re-extracted features. *Proceedings of the 17th International Conference on World Wide Web*, 1119–1120. ACM.

Giomelakis, D., & Veglis, A. (2016). Investigating search engine optimization factors in media websites: the case of Greece. *Digital Journalism*, 4(3), 379–400.

Google. (2017). Search Engine Optimization (SEO) Starter Guide.

Guha, R., Kumar, R., Raghavan, P., & Tomkins, A. (2004). Propagation of trust and distrust. *Proceedings of the 13th International Conference on World Wide Web*, 403–412. ACM.

Guo, Z., & Guan, Y. (2018). Active Probing-Based Schemes and Data Analytics for Investigating Malicious Fast-Flux Web-Cloaking Based Domains. *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, 1–9. IEEE.

Gyongyi, Z., Berkhin, P., Garcia-Molina, H., & Pedersen, J. (2006). Link spam detection based on mass estimation. *Proceedings of the 32nd International Conference on Very Large Data Bases*, 439–450. VLDB Endowment.

Gyongyi, Z., & Garcia-Molina, H. (2005). Web spam taxonomy. *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*.

Gyöngyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with trustrank. *Proceedings of the Thirtieth International Conference on Very Large*



Data Bases-Volume 30, 576–587. VLDB Endowment.

- Heymann, P., Koutrika, G., & Garcia-Molina, H. (2007). Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6).
- Hsu, C.-C., Lai, Y.-A., Chen, W.-H., Feng, M.-H., & Lin, S.-D. (2017). Unsupervised Ranking using Graph Structures and Node Attributes. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 771–779. ACM.
- Hu, W., Du, J., & Xing, Y. (2016). Spam filtering by semantics-based text classification. *Advanced Computational Intelligence (ICACI), 2016 Eighth International Conference On*, 89–94. IEEE.
- Hvitfeldt, E., & Silge, J. (2020). *Supervised Machine Learning for Text Analysis in R* (First Edit). Retrieved from <https://smltar.com/>
- Iqbal, M., Abid, M. M., Waheed, U., & Alam Kazmi, S. H. (2017). *Classification of Malicious Web Pages through a J48 Decision Tree, a Naïve Bayes, a RBF Network and a Random Forest Classifier for WebSpam Detection.*
- Jain, G., Sharma, M., & Agarwal, B. (2018). Spam detection on social media using semantic convolutional neural network. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, 8(1), 12–26.
- Jakubiček, M., Kovář, V., Rychlý, P., & Suchomel, V. (2020). Current challenges in Web corpus building. *Proceedings of the 12th Web as Corpus Workshop*, 1–4.
- Jeh, G., & Widom, J. (2002). SimRank: a measure of structural-context similarity. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 538–543. ACM.
- Ji, H., & Zhang, H. (2015). Analysis on the content features and their correlation of web pages for spam detection. *China Communications*, 12(3), 84–94. <https://doi.org/10.1109/CC.2015.7084367>
- Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 219–230. ACM.
- Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1), 359–



392.

- Keyaki, A., & Miyazaki, J. (2017). Part-of-speech tagging for web search queries using a large-scale web corpus. *Proceedings of the Symposium on Applied Computing - SAC '17*, 931–937. <https://doi.org/10.1145/3019612.3019694>
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), 604–632.
- Kou, Z., & Cohen, W. W. (2007). Stacked graphical models for efficient inference in markov random fields. *Proceedings of the 2007 SIAM International Conference on Data Mining*, 533–538. SIAM.
- Krishnan, V., & Raj, R. (2006). Web Spam Detection with Anti-Trust Rank. *AIRWeb*, 6, 37–40. Retrieved from <http://www.ra.ethz.ch/cdstore/www2008/airweb.cse.lehigh.edu/2006/proceedings.pdf#page=45>
- Kumar, S., Gao, X., Welch, I., & Mansoori, M. (2016). A machine learning based web spam filtering approach. *Proceedings - International Conference on Advanced Information Networking and Applications, AINA, 2016-May*, 973–980. <https://doi.org/10.1109/AINA.2016.177>
- Kunder, M. de. (2021). The size of the World Wide Web (The Internet). Retrieved from <https://www.worldwidewebsite.com/>
- Ledford, J. L. (2015). *Search engine optimization bible* (Vol. 584). John Wiley & Sons.
- Lempel, R., & Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2), 131–160.
- Leng, A. G. K., Singh, A. K., Kumar, P. R., & Mohan, A. (2014). TPRank: Contend with web spam using trust propagation. *Cybernetics and Systems*, 45(4), 307–323. <https://doi.org/10.1080/01969722.2014.887938>
- Li, L., Shang, Y., & Zhang, W. (2002). Improvement of HITS-based algorithms on web documents. *Proceedings of the 11th International Conference on World Wide Web*, 527–535. ACM.
- Li, X., Zhang, M., Liu, Y., Ma, S., Jin, Y., & Ru, L. (2014). Search engine click spam detection based on bipartite graph propagation. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining - WSDM '14*, 93–102.



PTTA UTHM
PERPUSTAKAAN TUN AMINAH

<https://doi.org/10.1145/2556195.2556214>

- Li, Y., Nie, X., & Huang, R. (2018). Web spam classification method based on deep belief networks. *Expert Systems with Applications*, 96, 261–270.
- Li, Z., Zhang, K., Xie, Y., Yu, F., & Wang, X. (2012). Knowing your enemy: understanding and detecting malicious web advertising. *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 674–686. ACM.
- Lin, J.-L. (2009). Detection of cloaked web spam by using tag-based methods. *Expert Systems with Applications*, 36(4), 7493–7499.
- Liu, J., Su, Y., Lv, S., & Huang, C. (2020). Detecting Web Spam Based on Novel Features from Web Page Source Code. *Security and Communication Networks*, 2020.
- Liu, Yiqun, Zhang, M., Ma, S., & Ru, L. (2008). *User Behavior Oriented Web Spam Detection*. 1039–1040.
- Liu, Yuting, Gao, B., Liu, T.-Y., Zhang, Y., Ma, Z., He, S., & Li, H. (2008). BrowseRank: letting web users vote for page importance. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 451–458. ACM.
- Makkar, A., & Kumar, N. (2020). An efficient deep learning-based scheme for web spam detection in IoT environment. *Future Generation Computer Systems*, 108, 467–487.
- Manaskasemsak, B., & Rungsawang, A. (2015). Web spam detection using trust and distrust-based ant colony optimization learning. *International Journal of Web Information Systems*, 11(2), 142–161. <https://doi.org/10.1108/IJWIS-12-2014-0047>
- MCA, N. Z. J., & Prakash, P. (2016). *Document content based web spam detection using cosine similarity measure*. 7(June).
- Metaxas, P. T., & Pruksachatkun, Y. (2017). *Manipulation of search engine results during the 2016 US congressional elections*.
- Morgan, E. L. (2012). Foray's into parts-of-speech.
- Moz. (2020). Google Algorithms Update History. Retrieved from <https://moz.com/google-algorithm-change>



- Neria, M. Ben, Yacovzada, N.-S., & Ben-Gal, I. (2017). A Risk-Scoring Feedback Model for Webpages and Web Users Based on Browsing Behavior. *ACM Transactions on Intelligent Systems and Technology*, 8(4), 1–21. <https://doi.org/10.1145/2928274>
- Nomura, S., Oyama, S., Hayamizu, T., & Ishida, T. (2004). Analysis and improvement of hits algorithm for detecting web communities. *Systems and Computers in Japan*, 35(13), 32–42.
- Nothman, J., Qin, H., & Yurchak, R. (2018). Stop word lists in free open-source software packages. *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 7–12.
- Oentaryo, R., Lim, E.-P., Finegold, M., Lo, D., Zhu, F., Phua, C., ... Nguyen, M. N. (2014). Detecting click fraud in online advertising: a data mining approach. *The Journal of Machine Learning Research*, 15(1), 99–140.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*.
- Pandurangan, G., Raghavan, P., & Upfal, E. (2006). Using pagerank to characterize web structure. *Internet Mathematics*, 3(1), 1–20.
- Parmar, K., Trivedi, A., Chauhan, P., & Suchak, K. (2020). Webspam Detection Using Classification Algorithms and Optimizing the Performance of Classifiers by Selecting the Features. *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 124–129. IEEE.
- Patil, M. R. C., & Bhadane, M. V. R. (2016). Survey on Web Spam Detection using Link and Content Based Features. *International Journal on Recent and Innovation Trends in Computing and Communication*, 4(6), 467–470.
- Petkova, L. (2019). SECURITY'S LEAKS IN SEO SPAMMING. *Knowledge International Journal*, 35(3), 987–991.
- Piskorski, J., Sydow, M., & Weiss, D. (2008). Exploring linguistic features for web spam detection: a preliminary study. *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, 25–28. ACM.
- Poblete, B., Castillo, C., & Gionis, A. (2008). Dr. searcher and mr. browser: a unified hyperlink-click graph. *Proceedings of the 17th ACM Conference on Information*



and Knowledge Management, 1123–1132. ACM.

Practices, C. (2015). *Understanding Search-Engine Optimization*. (October), 43–52. <https://doi.org/10.1109/MC.2015.297>

Prieto, V. M., Álvarez, M., & Casheda, F. (2013). SAAD, a content based Web Spam Analyzer and Detector. *Journal of Systems and Software*, 86(11), 2906–2918. <https://doi.org/10.1016/j.jss.2013.07.007>

Prieto, V. M., Álvarez, M., López-García, R., & Casheda, F. (2012). Analysis and detection of web spam by means of web content. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7356 LNCS, 43–57. https://doi.org/10.1007/978-3-642-31274-8_4

Radlinski, F. (2007). Addressing malicious noise in clickthrough data. *Learning to Rank for Information Retrieval Workshop at SIGIR, 2007*.

Rao, J. M., & Reiley, D. H. (2012). The economics of spam. *Journal of Economic Perspectives*, 26(3), 87–110.

Redmiles, E. M., Chachra, N., & Waismeyer, B. (2018). Examining the Demand for Spam: Who Clicks? *Proc. of CHI*, 1–10. <https://doi.org/10.1145/3173574.3173786>

Regalado, D., Harris, S., Harper, A., Eagle, C., Ness, J., Spasojevic, B., ... Sims, S. (2015). *Gray Hat Hacking The Ethical Hacker's Handbook*. McGraw-Hill Education Group.

Roberts, G. O., & Rosenthal, J. S. (2003). Downweighting tightly knit communities in world wide web rankings. *Advances and Applications in Statistics (ADAS)*, 3, 199–216.

Robertson, S., Zaragoza, H., & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, 42–49. ACM.

Roul, R. K., Asthana, S. R., & Shah, M. I. T. (2016). *Detection of spam web page using content and link-based techniques : A combined approach*. 41(2), 193–202.

Roy, D., Mitra, M., & Ganguly, D. (2018). To Clean or Not to Clean. *Journal of Data and Information Quality*, 10(4), 1–25. <https://doi.org/10.1145/3242180>

- Rubinstein, Y. D., Wiseman, J., & Choi, M. K.-S. (2017, January). *Spam detection and prevention in a social networking system*. Google Patents.
- Sadredini, E., Guo, D., Bo, C., Rahimi, R., Skadron, K., & Wang, H. (2018). A Scalable Solution for Rule-Based Part-of-Speech Tagging on Novel Hardware Accelerators. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, 665–674. <https://doi.org/10.1145/3219819.3219889>
- Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.
- Saraswathi, D., Kathiravan, A. V., & Kavitha, R. (2012). A new enhanced technique for link farm detection. *International Conference on Pattern Recognition, Informatics and Medical Engineering, PRIME 2012*, 74–81. <https://doi.org/10.1109/ICPRIME.2012.6208290>
- Sedhai, S., & Sun, A. (2018). Semi-supervised spam detection in Twitter stream. *IEEE Transactions on Computational Social Systems*, 5(1), 169–175.
- Shahzad, A., Mahdin, H., & Nawli, N. M. (n.d.). *An Improved Framework for Content-based Spamdexing Detection*.
- Sharma, D., Shukla, R., Giri, A. K., & Kumar, S. (2019). A Brief Review on Search Engine Optimization. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 687–692. IEEE.
- Singh, S., & Singh, A. K. (2018). Web-Spam Features Selection Using CFS-PSO. *Procedia Computer Science*, 125, 568–575.
- Smitha, L. (2017). *Topical and Trust Based Page Ranking Using Automatic seed selection*. (2006), 0–3. <https://doi.org/10.1109/IACC.2017.156>
- Sobek, M. (2002). *Pr0-Google's Pagerank 0 Penalty*. Accessed 25 February.
- Spirin, N., & Han, J. (2012). Survey on web spam detection: principles and algorithms. *Acm Sigkdd Explorations Newsletter*, 13(2), 50–64.
- Stanford. (2004). Stanford Log-linear Part-Of-Speech Tagger.
- Stone-Gross, B., Stevens, R., Zarras, A., Kemmerer, R., Kruegel, C., & Vigna, G. (2011). Understanding fraudulent activities in online ad exchanges. *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*,



279–294. ACM.

Summerville, A., Snodgrass, S., Guzdial, M., Holmgard, C., Hoover, A. K., Isaksen, A., ... Togelius, J. (2018). Procedural Content Generation via Machine Learning (PCGML). *IEEE Transactions on Games*, 10(3), 257–270. <https://doi.org/10.1109/TG.2018.2846639>

Sydow, M., Piskorski, J., Weiss, D., & Castillo, C. (2007). *Application of machine learning in combating web spam*. Submitted for publication in IOS Press.

Tikhonov, A. N., Arsenin, V. I., & John, F. (1977). Solutions of ill-posed problems (Vol. 14). *Washington, DC: Winston*.

Vasumati, D., Vani, M. S., Bhramaramba, R., & Babu, O. Y. (2015). Data Mining Approach to Filter Click-spam in Mobile Ad Networks. *Int'l Conference on Computer Science, Data Mining & Mechanical Engg.(ICCDMMME'2015) April*, 20–21.

Wahba, G. (1990). *Spline models for observational data*. *Society for Industrial and Applied Mathematics*.

Wan, J., Liu, M., Yi, J., & Zhang, X. (2015). Detecting spam webpages through topic and semantics analysis. *GSCIT 2015 - Global Summit on Computer and Information Technology - Proceedings*. <https://doi.org/10.1109/GSCIT.2015.7353328>

Webb, S., Caverlee, J., & Pu, C. (2007). Characterizing Web Spam Using Content and HTTP Session Analysis. *CEAS*. Citeseer.

Webb, S., Caverlee, J., & Pu, C. (2008). Predicting Web Spam with HTTP Session Information. *17th ACM Conference on Information and Knowledge Management*, 339–348. <https://doi.org/10.1145/1458082.1458129>

Wei, S., & Zhu, Y. (2017). Cleaning Out Web Spam by Entropy-Based Cascade Outlier Detection. *International Conference on Database and Expert Systems Applications*, 232–246. Springer.

Whang, J. J., Jeong, Y. S., Dhillon, I. S., Kang, S., & Lee, J. (2018). *Fast Asynchronous Anti-TrustRank for Web Spam Detection*.

Wu, B., & Davison, B. (2005). Cloaking and Redirection: A Preliminary Study. *AIRWeb*, 7–16.



- Wu, B., & Davison, B. D. (2006). Detecting semantic cloaking on the web. *Proceedings of the 15th International Conference on World Wide Web*, 819–828. ACM.
- Xu, H., Liu, D., Koehl, A., Wang, H., & Stavrou, A. (2014). Click fraud detection on the advertiser side. *European Symposium on Research in Computer Security*, 419–438. Springer.
- Zarras, A., Kapravelos, A., Stringhini, G., Holz, T., Kruegel, C., & Vigna, G. (2014). The dark alleys of madison avenue: Understanding malicious advertisements. *Proceedings of the 2014 Conference on Internet Measurement Conference*, 373–380. ACM.
- Zarras, A., Papadogiannakis, A., Ioannidis, S., & Holz, T. (2015). Revealing the relationship network behind link spam. *2015 13th Annual Conference on Privacy, Security and Trust, PST 2015*, 101–108. <https://doi.org/10.1109/PST.2015.7232960>
- Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1), 1–141.
- Zhang, H., Goel, A., Govindan, R., Mason, K., & Van Roy, B. (2004). Making eigenvector-based reputation systems robust to collusion. *International Workshop on Algorithms and Models for the Web-Graph*, 92–104. Springer.
- Zhang, Q., Wang, D. Y., & Voelker, G. M. (2014). DSpin: Detecting Automatically Spun Content on the Web. *Proceedings 2014 Network and Distributed System Security Symposium*, (February), 23–26. <https://doi.org/10.14722/ndss.2014.23004>
- Zhang, S., & Cabage, N. (2017). Search engine optimization: Comparison of link building and social sharing. *Journal of Computer Information Systems*, 57(2), 148–159. <https://doi.org/10.1080/08874417.2016.1183447>
- Zhang, X., Wang, Y., Mou, N., & Liang, W. (2014). Propagating Both Trust and Distrust with Target Differentiation for Combating Link-Based Web Spam. *ACM Transactions on the Web*, 8(3), 1–33. <https://doi.org/10.1145/2628440>
- Zhou, B., & Pei, J. (2007). Sketching landscapes of page farms. *Proceedings of the 2007 SIAM International Conference on Data Mining*, 593–598. SIAM.



- Zhou, B., & Pei, J. (2009a). Link spam target detection using page farms. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3), 13.
- Zhou, B., & Pei, J. (2009b). OSD: An online web spam detection system. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, 9.
- Zhou, B., Pei, J., & Tang, Z. (2008). A Spamicity Approach to Web Spam Detection. *Proceedings of the 2008 SIAM International Conference on Data Mining*, 277–288. <https://doi.org/10.1137/1.9781611972788.25>
- Zhou, Dengyong, Burges, C. J. C., & Tao, T. (2007). Transductive link spam detection. *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, 21–28. ACM.
- Zhou, Denny, Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 321–328.
- Zuze, H., & Weideman, M. (2013). Keyword stuffing and the big three search engines. *Online Information Review*.
- Zwicky, R. K. (2015, October). *Click fraud detection*. Google Patents.

