Comprehensive Analysis of Web Page Classifier for Fsocused Crawler

Gourav Kumar Shrivastava, Praveen Kaushik, Rajesh Kumar Pateriya

Abstract: ***Focused Crawler collects domain specific web page from the internet. However, the performance of focused web crawler depends upon the multidimensional nature of the web page. This paper presents a comprehensive analysis of recent web page classifiers for focused crawlers and also explores the impact of web-based feature in collaboration with web classifier. It also evaluates the performance of classification technique such as Support vector machine, Naive Bayes, Linear Regression and Random Forest over web page classification. Along with that it examines the impact of web feature i.e. anchor text, Page content and link over web page classification. Finally the paper yield interesting result about the collective response of web feature and classification technique for web page classification as a relevant class and irrelevant class.***

Keywords: Focused Crawler, Feature Extraction Technique, Anchor text, Page Content, Link Priority, Naive Bayes, Linear Regression, Random Forest, SVM

I. INTRODUCTION

With the rapid development of the World Wide Web, information over the internet has been exponentially increased. Recently researchers focused to organize this massive information in such a way that helps the end user to extract the information efficiently and accurately. Search engines such as Google, Bing and Yahoo etc. have been working to overcome this problem. Search engines have employed web crawler to download web page and make data repository system over the local server.

A web crawler is an automated system run by the search engine to collect the Metadata about the web page and assemble in a corpus of the web after indexing, by traversing and downloading the web page. The main objective of the web crawler is to gather the web page and establish a link structure among them to provide rapid and efficient search results to the user's request.

Crawling behavior of Web crawler (WC) is either restricted or unrestricted to search domain. Unrestricted web crawler retrieves an enormous amount of domain independent web page for normal search engine (NSE) with high time and space complexity. Whereas restricted web crawler also known as focused crawler (FC), captures the finite source of the web page within the specific domain for vertical search engines (VSE) with obvious lower time and space complexity (Lu et al., 2016; Du et al., 2015). Over unrestricted web crawler, restricted focused crawlers have diversity of applications such as search engines, information extraction, digital libraries, and text classification.

Focused crawler was introduced by Chakrabarti et al. (2000) and it is a domain specific crawler that is used to retrieve domain specific web page based on content and link structure. The basic theme behind the focused crawler is to classify the crawler page with respective topic taxonomy. Selecting relevant web page or classifying web page over the relevancy of URL is the most important task of focused crawler.

This paper presents a comprehensive study of focused crawler for web page classification techniques and presents a framework to evaluate the performance of web classifier with focused crawler for web page classification. It also summarizes the effect of web feature over Web page Classification through Focused crawler. This paper presents a framework to scrutinize and pre-process Web URL and then formulates the supervised classification technique for Focused Crawler. It also provide a platform for comparative analysis of web feature extraction technique with supervised classification approach and yields interesting facts about the capabilities and deficiency of Focused Crawler to categorize the web page.

The rest of the paper is organized as follows: Sect. 2 presents an overview of web crawler; Sect. 3 covers description about focused crawler, Sect. 4 covers related work on web page classification and Focused crawler for the different task. Sect. 5 present a framework for scrutinizing and pre-processing web page data set and discusses classification procedure of web page via SVM, NB, LR, RF. Sect. 6 describes the experimental setup for comparative evaluation of different web feature extraction technique with classification approach for web page classification and finally, Sect. 7 concludes the paper and outlines the founding and future work.

II. WEB CRAWLER

A web crawler is an automated process to exploit the graphical structure of Web, in order to link and index them for browsing in a systematic and efficient manner over the search engine. In their initial stage such programs were also known as wanderers, robots, spiders, fish, and worms, words for clear understanding the Web imagery (Kim and Pant, 2018; Wang et al., 2017).

In graphical View, WWW represents a directed graph where the webpage represents a node and hyperlink represents an edge and search operation is summarized as traversing of the directed web graph. Web crawler uses the graphical structure of web for gathering the Metadata of all visited page after traversing over them via one

Revised Manuscript Received on July 05, 2019

Gourav Kumar Shrivastava, Department of CSE, Maulana Azad National Institute of technology, Bhopal, India.

Praveen Kaushik, Department of CSE, Maulana Azad National Institute of technology, Bhopal, India.

Rajesh Kumar Pateriya, Department of CSE, Maulana Azad National Institute of technology, Bhopal, India.

page to another page. A web crawler is used by the search engine to retrieve the web pages and insert the replica of them to local server repository to enhance the experience of the web search engine.

The simplest form of normal web crawler begins with a set of one or more URLs as a seed that is maintained in the unvisited list of URLs known as a frontier. Each crawling loop involves picking up next URL from seed set and fetch the HTTP information and parse them as shown in figure 1. Parsing extracts both text and link information from the HTTP page. The extracted text is fed into text indexer and extracted URLs are added as unvisited URLs in the frontier. Before adding to the frontier, relative benefit score of the extracted URL corresponding to seed URL are evaluated. The crawling loop is terminated after crawling a sufficient number of URLs as per software and hardware specification.

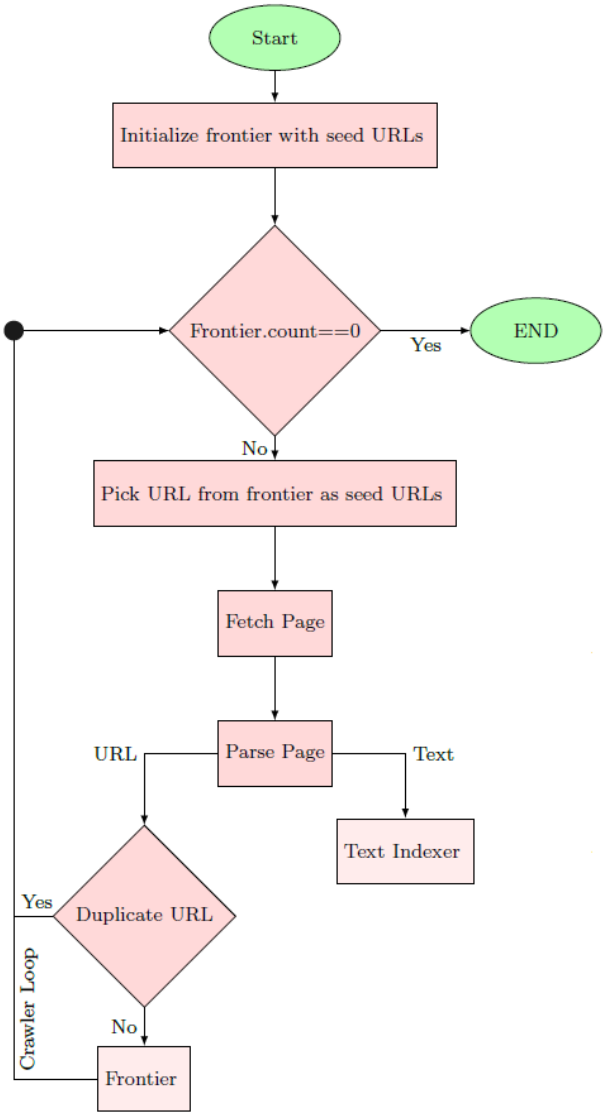

Fig. 1 Basic Web Crawling Procedure

Besides URL, Link and Content, web crawler having following feature standard that must be incorporated to enrich web data ethics (Pant et al., 2004), as shown in figure 2.

(A) **Robustness:-** Certain Web server launch one spider trap to create an illusion to mislead the crawler for getting finite number of web page in any specific domain and leads to stuck it down. It's a mandatory feature of every web crawler to be designed in such a manner that resists such a trap. However not every trap are malicious but some are inadvertent side-effect to fascinate website development.

(B) **Politeness:-** Politeness policy must state the rate of web page download by web crawler to control portion of the bandwidth of a website server to be used in crawling.

(C) **Distributed:-** Crawling should be executed globally and distributed over different machine simultaneously to enrich the speed of crawling.

(D) **Scalable:-** For managing the periodic speed and load over web, crawler architecture should be flexible to scale up the crawling rate by adding extra bandwidth and machine.

(E) **Performance and Efficiency:-** Design of Crawler should be system efficient that increase the throughput of processor, network bandwidth and storage capacity.

(F) **Quality:-** Crawler should have higher harvesting ratio that indicates and fetch useful page first.

(G) **Freshness:-** Crawler should be in contiguous updating mode to fetch fresh copy of previous stored page.

(H) **Extensible:-** Crawler should be designed in such a manner such that it is extensible for the new data format and protocol.

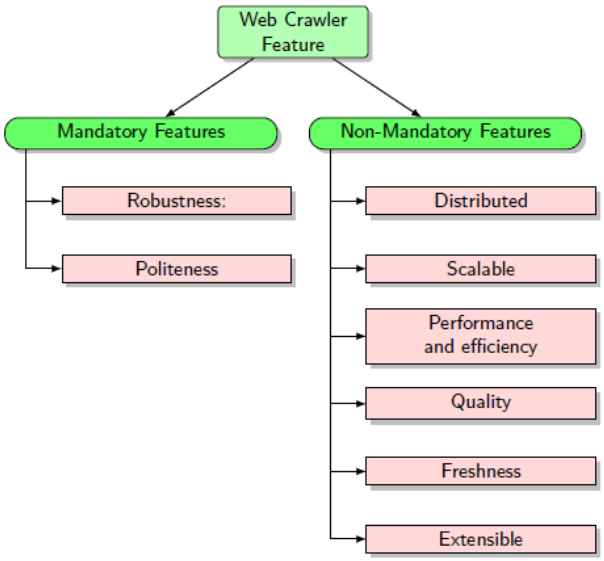

Fig. 2 Web Crawler Features

III. FOCUSED CRAWLER

Chakrabarti et al. (2000) introduced Focused Crawler (Topical Crawler) that seeks, acquires, indexes, maintains pages and respond to the specific requirements expressed by the topical queries or interested profiles to narrow the segment of web, hardware, and network resources. Focused crawler driven by a rich context such as a web page's content, URL extensions and the hyperlink structure in a decentralized manner. For instance, if a web crawler needs to extract web page for a use specific topic like “Higher education in India" from a specific domain (.in,.edu,.com) and in a particular language (Hindi, English), then it needs to employ topic oriented crawler or focused crawler. Focused Crawler, try to bias crawled pages in specific categories as per end user interest. Apart from normal Web crawler, Focused Crawler builds a text classifier using labeled example pages and supervises the crawler by preferentially selecting from

frontier pages that appear most likely to belong to categories of interest, according to classifiers prediction. Focused crawler is used to check the relevance score of the crawled page to score the unvisited URLs extracted from it. The scored URLs are then added to the frontier.

IV. RELATED WORK

Web crawlers-also known as robots, spiders, worms, walkers, and wanderers- are almost as old as the web itself. The first crawler, Matthew Gray's Wandered, was written in the spring of 1993, roughly coinciding with the first release of NCSA mosaic. Web crawler has become an increasingly important application in recent years, because of its unique ability to search and store web page over the internet. An increased level of interest is seen in the field of search engine perhaps domain specific search engine for information retrieval (Malhotra and Sharma, 2017), web page classification (Lu et al., 2016; Wang et al., 2010; Zhao et al., 2016; Saleh et al., 2017a), marketing (Zhou et al., 2018; Garcia-Nunes and da Silva, 2019), extracting public sentiment over any global issue like food quality (Geng et al., 2017). Lu et al. (2016) present Link priority evaluation model based focused crawler for web page classification. Whereas Saleh et al. (2017a) present disambiguation model, Ahmadi-Abkenari and Selamat (2012) present treasure graph, Seyfi et al. (2016) treasure crawler for web page classification that based on link and page content feature.

Along with that recently researchers use to design specific task oriented focused crawler for Relevant Web Page Selection (Wang et al., 2010), A fast distributed focused crawler (Achsan et al., 2013), Harvesting Deep Web Interface (Zhao et al., 2016), Duplicate Content Detection (Khalil and Fakir, 2017), Web matrix (Ahmadi- Abkenari and Selamat, 2012), Analysis Dark Web for Child Pornography (Dalins et al., 2018) and Business threats and opportunities analysis (Garcia-Nunes and da Silva, 2019).Apart from search domain focused crawler also being employed for vulnerability and security analysis, Kim and Pant (2018) design focused crawler for Detection of Malicious Web Page with the help of machine learning technique. Along with that total eighteen articles (published in 2010 to 2018) presented in this survey are summarized in Table 1 that contains six columns. The main task of the articles is illustrated in the second column. Column third illustrates crawler type i.e. either web crawler or focused crawler. Where WC and FC is used to represent web crawler and focused crawler respectively. Column fifth and sixth illustrate method and algorithm used for crawling web in different application. Whereas fourth column describes the name of web feature used for analysis web page that has been used for evaluating different methodology.

Reference	Task	CT	Feature	Methods	Algorithm
Lu et al. (2016)	Web Page Classification	FC	Anchor text, Link Content	Improved Term Weight	Link Priority Evaluation, Page Content Block Partition , Joint Feature Evaluation
Wang et al. (2010)	Relevant Web Page Selection	FC	Kernal Density, Normal density ,Page Rank , BFS	TF-IDF	Naive Bayes
Zhao et al. (2016)	Harvesting Deep Web Interface	FC	Path, Link, Page Content, Anchor	Reverse and Incremental Searching	Form Classifier
Du et al. (2015)	Semantic Focused Crawler	FC	Anchor text	Semantic Similarity Retrieval Model	Cosine Similarity
Seyfi et al. (2016)	Treasure Crawler	FC	Link, Content, Anchor, Heading, URL	Topic Boundary	T-graph, D-tree
Geng et al. (2017)	Public opinion over Food Safety	FC	Page Content	Similarity computation based Multiple Reference Factor	Term Frequency , Inverse Document Frequency , Best First , Breadth-First
Kumar et al. (2018)	Keyword Query Based Focused Crawler	FC	URL, Link	Query-based Crawler	Dom Structure, K level, Max Ancestor
Yan and Pan (2018)	Evolutionary Focused Crawler	FC	Page Rank	Vector Space Model	Genetic Algorithm, Similarity Function
Zhou et al. (2018)	Agricultural Market Analysis	FC	Theme Relevancy crawling	Key Multi-Mode Matching	Aho Corasick Algorithm, Term frequency and Inverse Document Frequency
Khalil and Fakir (2017)	Duplicate Content Detection	WC	URL, Page Content and Depth level	Parallel Web crawling and Scraping	Similarity hash Function
Liu and Hu (2019)	Public Sentiment Analysis towards Green Building	FC	Keywords Search and Depth level	Ontology and text mining	Part of Speech, dictionary-based approach
Saleh et al. (2017a)	Web page classification	FC	Keywords Search and Wordnet	Disambiguation Domain ontology	Naive Bayes, Support Vector Machine, Genetic Algorithm

Dalins et al. (2018)	Analysis Dark Web for Child Pornography	FC	Labeling and Page Content	TOR used Motivation Model	SHA-1
Garcia-Nunes and da Silva (2019)	Business threats and opportunities Analysis	FC	Pattern recognition	Ontology and Weak Signal Monitoring	Part of Speech tagging , SVO Typology
Wang et al. (2017)	Detection of Malicious Web Page	WC	Correlation Based Feature	Machine Learning	Decision Tree, DOM
Kim and Pant (2018)	Website Audience Demography	WC	Design And Content	Machine Learning and Statistical Analysis	Naive bayes , SVM, Logistic Regression
Seyfi and Patel (2016)	Topically Relevant Harvesting	FC	Link and Content	Treasure Graph	Best First and Similarity Function
Ahmadi-Abkenari et al. (2012)	Web importance metrics Analysis	FC	Link and Content	Treasure Graph	Best First and Similarity Function

Table 1. Article Summary

V. COMPARATIVE ANALYSIS

Comparative analysis of recent research for the design of focused crawler for web page classification present interesting and useful facts regarding the state-of-the art of web-based feature extraction technique. This paper presents a three-tier framework for comparing the performance of classification techniques over focused crawler for web page classification as shown in figure 3.

A. Parsing

Once the web page is extracted, parsing is used to extract link, content and URL information in order to build HTML Tag Tree. Parsing also involves prepossessing step like extraction of canonical form, removal of stop words as stop listing and stemming. Proposed Parsing of URL include following canonicalization procedures:

(i) Case Conversion:- Presented Framework convert the protocol and host name to lowercase in order to eliminate duplicacy. For instance https://www.MAKEMYTRIP.com are prepossessed as https://www. makemytrip.com/

(ii) Size Reduction:- For reducing the size of url presented framework remove the `anchor' or `reference' part of the URL. for instance https://www.makemytrip.com/cabs/#what are reduce to https://www.makemytrip.com/cabs/

(iii) URL encoding:- Parser encode the commonly used character as ` ' in similarity pattern to control duplicacy.

(iv) Pattern Recognition:- Convert the URL in similar canonical form for pattern recognition.

(v) Home Page Recognition:- Presented Framework used heuristics approach to recognize default Web pages.

(vi) Remove double dots:- To reduce the size of URL, remove `..' from parent directory of URL path.

(vii) Port Number Treatment:- if port number is not mentioned in URL include the default port number as 80.

B. Feature Extraction

Presented Framework work over three web feature i.e. Anchor text, Page content, and link. For efficient classification, web feature of every web URL needs to be extracted and evaluate their relevancy score before classification.

C. Classification of Web Page

After examining the Hyperlink, Page content and Anchor text of web page focused crawler framework classifies relevant data set into two different classes as relevance, irrelevance web page. This paper evaluates the performance of Classifiers SVM, Naive Bayes, Linear regression and random Forest with different web feature extraction technique such as breadth-first, best first, Anchor only, Link Context, Page Content block partition algorithm (CBP) and link priority evaluation (LPE).

(i) Support Vector Machine: Support vector machine maximizes the margin of separator hyperplane to classify the web page with class label relevance and irrelevance web page. Whereas after incorporating web-based feature such as anchor text, hyperlink, and page content the relevancy of relevant web page is increased. SVM treat all the feature as token of web vector space as shown in equation (1).

$$W_{vs}^{f} = \{(W_{Anchor,}W_{hlink,}W_{Keywords})w_t \epsilon\, f_{url}\} \qquad (1)$$

Where,

- f_{url} is seed URL
- W_{vs}^{f} is web vector space
- $W_{Anchor,}$ is the set of anchor text
- $W_{hlink,}$ is the set of hyperlinks
- $W_{keywords}$ is the set of keywords
- w_t is web feature based token and finally creates an optimal hyper plane as shown in equation 2, 3.

$(W_{vs}^{f} + High) \geq$

$$R_{webpage} \forall Web\ url\ over\ positive\ hyperplane \quad (2)$$

$$(W_{vs}^{f} + Low) \geq NR_{webpage} \forall Web\ url\ over\ negative\ hyperplane \quad (3)$$

(ii) Naive Bayes: Naive bayes return class label to web page (relevant, Irrelevant) on the basis of maximum posterior probability as shown in equation 4 and 5.

$$C_{wp} = argmax_{p\in(Relevant,Irrelevant)} P(r|wp) \quad (4)$$

$$P(r|wp) = \frac{P(wp|r)P(r)}{P(wp)} \quad (5)$$

Where, $P(r|wp)$ is final posterior probability and $P(wp|r)$ is the probability of web page 'wp' belong to relevant class 'r'. Whereas $P(r)$ and $P(wp)$ is the independent probability for relevant class 'r' and web page 'wp'.

$$P(wf|score) = P(wf|anchor) * P(wf|link) * P(wf|content) \quad (6)$$

Where $P(wf|score)$ are independent given the relevance class (R) and each web feature substitute their individual probability for exploring relevance class.

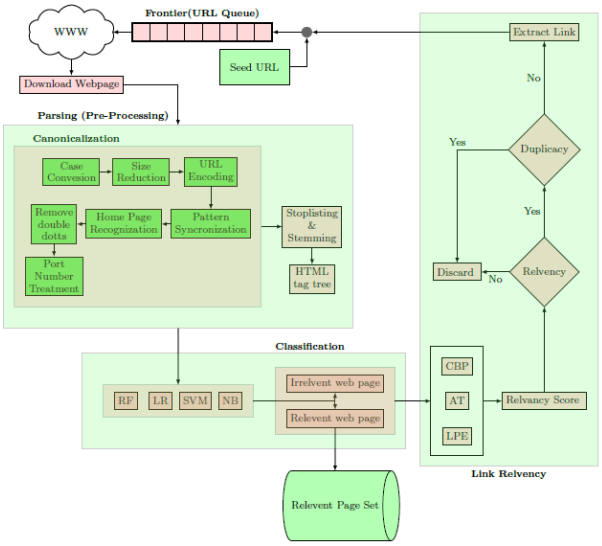

Fig.3 Proposed Framework of Focused Crawler for web page classification.

(iii) Random forest: Random forest predicts the class level for web page by building randomized regression trees $\{r_n(x, pc, ds)m \geq 1\}$ based relationship between web class level and web page as shown in equation (7). Where E_{wc} is exception on web class (ws) classification with random parameter (r) on condition x and frontier seed set (F_{ss}).

$$\overline{r_n}(x, f_{ss}) = E_{wc}\ [r_n\ (x, wc, f_{ss})] \quad (7)$$

Whereas after incorporating web feature as conditional parameter `x' lead to a minimized exception (E_{wc}) on web class and increase classification rate.

(iv) Linear regression: Linear function separate relevant and irrelevant web page into two different classes by finding a decision boundary that linearly separates the frontier URL seed as shown in equation 8. Where, W passing the web feature function w*x through the threshold function as shown in equation 9.

$$W_c(x) = \begin{cases} relevant & if\ C * x \geq 0 (relevancy\ score) \\ irrelevant & if\ C * x < 0 (relevancy\ score) \end{cases} \quad (8)$$

$$W_c\ (x) = threshhold\ W * x \quad (9)$$

VI. EXPERIMENTAL SETUP AND RESULT ANALYSIS

For Comparative analysis of classification technique used in focused crawler for web page classification, two different experiments has been carried out over two different Data Source. First evaluation has been carried out over the Reuters-21578 Corpus data set and other evaluation has been carried out over 20 Newsgroups data set.

Performance of web page classification technique is evaluated using Precision, Recall, and F-Measure. Precision for web page classification is the fraction of web page assigned that are relevant for the focused crawler, which evaluate rejection accuracy for irrelevant web pages as shown in equation 10. Recall is the proportion of relevant web pages assigned by classifier, which evaluate selection accuracy of relevant web pages as shown in equation. Consider (X) is the set of relevant web pages in test dataset and (Y) is the set of relevant web pages suggested by classifier.

$$Precision = \frac{|X \cap Y|}{|Y|} * 100 \quad (10)$$

$$Recall = \frac{|X \cap Y|}{|X|} * 100 \quad (11)$$

F-Measure evaluate the harmonic mean of Precision and Recall that used to measure the performance of proposed framework for web page classification according to their relative importance parameter β as shown in equation 12 & 13.

$$F - Measure = \frac{(\beta^2 + 1)\ Precision * Recall}{\beta^2 * Precision + Recall} * 100 \quad (12)$$

$$\begin{cases} if\ \beta > 1\ (\text{Recall is more important}) \\ if\ 0 < \beta < 1\ (\text{Precision is more important}) \\ if\ \beta = 1\ (Both\ are\ equally\ important) \end{cases} \quad (13)$$

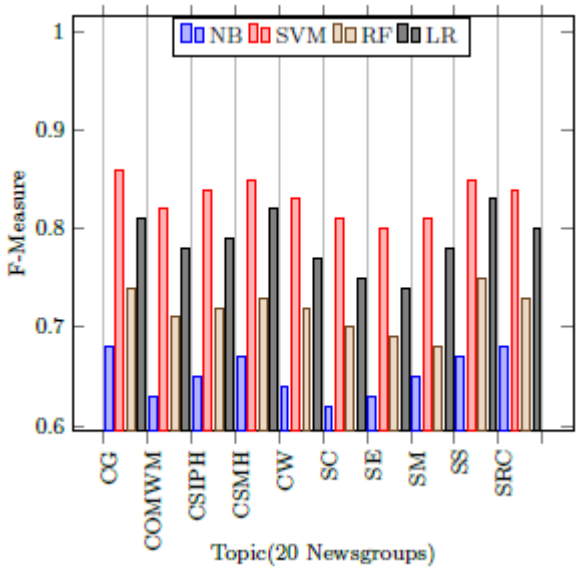

Fig. 4 Performance of web classifier for 20 Newsgroups dataset

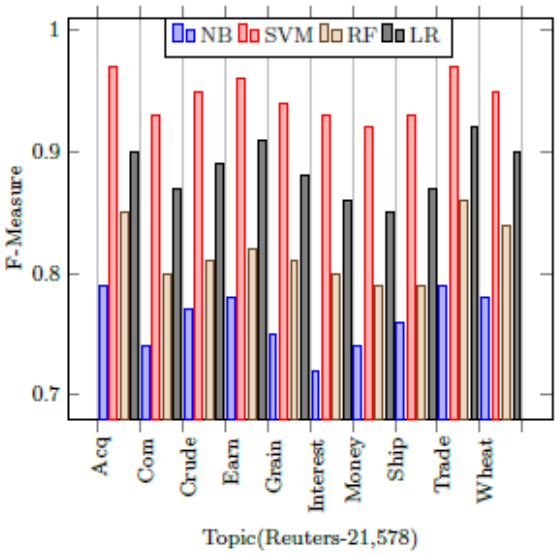

Fig. 5 Performance of web classifier for Reuters-21578 dataset

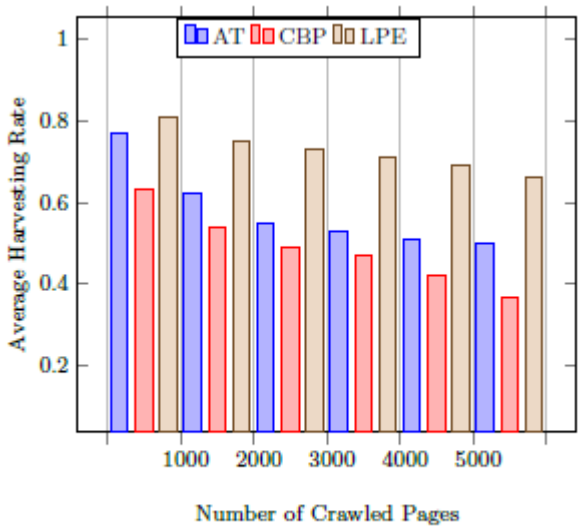

Fig. 6 Harvesting rate of proposed framework with NB

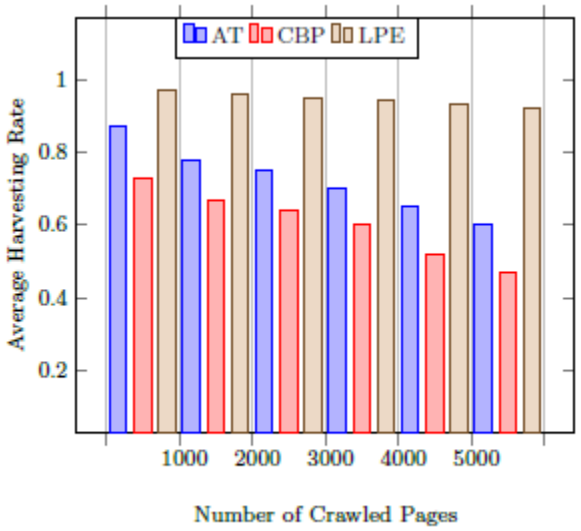

Fig. 7 Harvesting rate of proposed framework with SVM

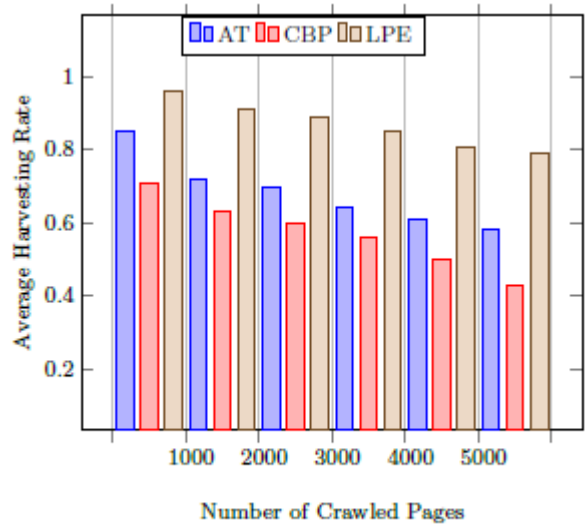

Fig. 8 Harvesting rate of proposed framework with LR

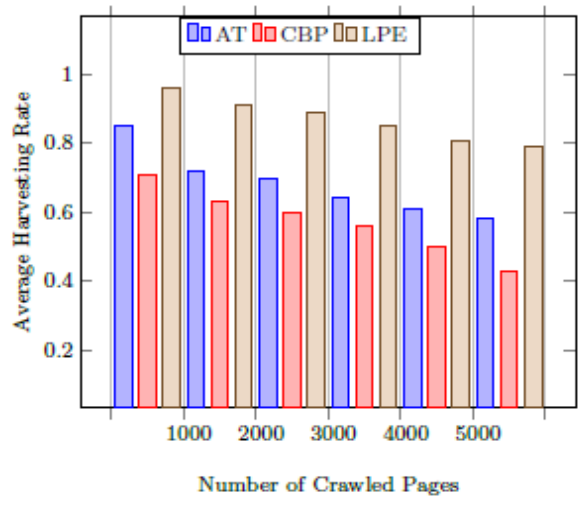

Fig. 9 Harvesting rate of proposed framework with RF

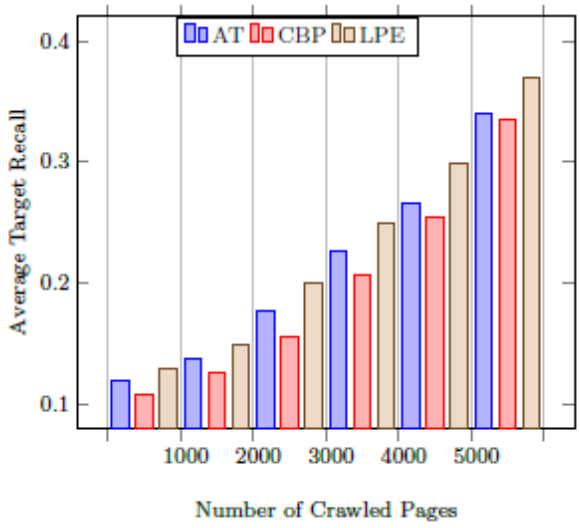

Fig. 10 Target recall of proposed framework with SVM

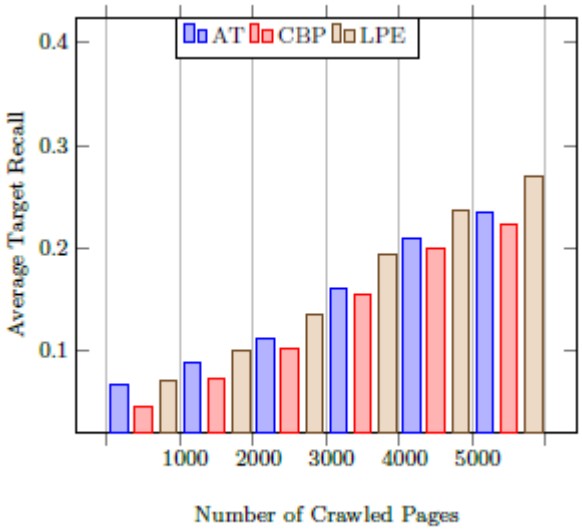

Fig. 11 Target recall of proposed framework with NB

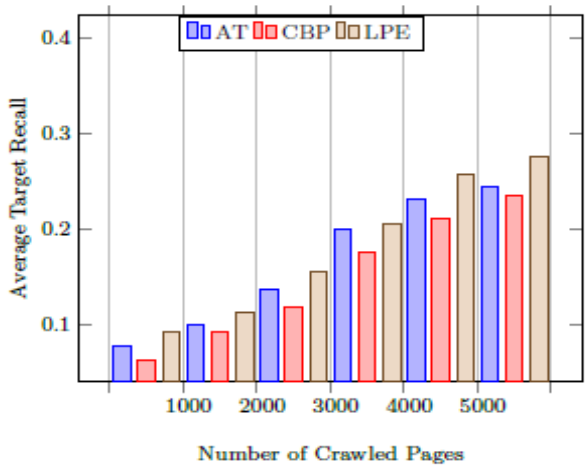

Fig. 11 Target recall of proposed framework with LR

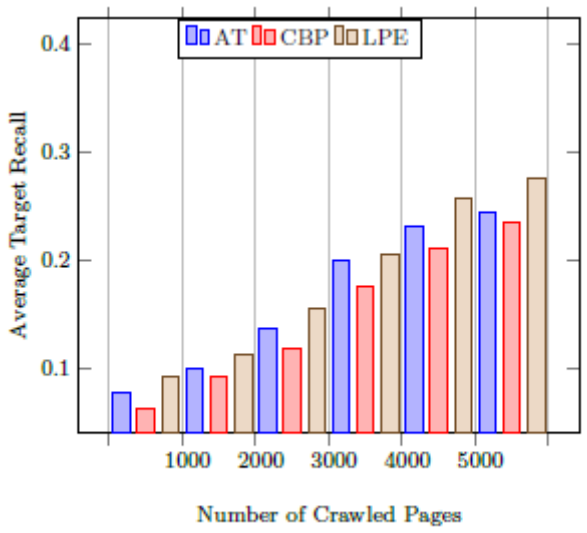

Fig. 11 Target recall of proposed framework with RF

The baseline web page classifier Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF) and Linear Regression (LR)can yield approximate 0.76- 0.68, 0.97 - 0.88 ,0.85-0.74 and 0.90-0.82 F-measure respectively over Routers-21578 and 20 Newsgroup data set as shown in figure 4 & 5. Where SVM lead by approximate over both the data set.

Performance of proposed framework for crawling relevant web page as focused crawler is evaluated by using harvest rate and target recall. Harvesting rate is the fraction of web page crawled that are relevant to crawling topic i.e. use to measure rejection accuracy of irrelevant web page and target recall is the fraction of relevant web page crawled i.e. use to measure selection accuracy of relevant web page as shown in equation 14 & 15.

$$Harvesting\ Rate = \frac{|S_{(x)} \cap R|}{S_{(x)}} * 100 \quad (14)$$

$$Target\ Recall = \frac{|S_{(x)} \cap R|}{R} * 100 \quad (15)$$

Consider the target set R is the relevant set in the virtual Web, S(x) is the set of first x pages crawled. Proposed framework with SVM and web content extraction technique (Anchor text (AT), Content Block Partition (CBP) and Link Priority evaluation (LPE)) significantly lead the crawling performance over Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR) as shown in figure 6, 7, 8 & 9. Harvesting rate of proposed framework with SVM classifier achieve approximate 97% harvesting rate and 13% target recall over LPE as shown in figure 7 & 10. Whereas with LR , proposed framework achieve approximate 96% harvesting rate and 11% target recall over LPE as shown in figure 8 & 11. With NB, proposed framework achieve approximate 81% harvesting rate and 7% target recall over LPE as shown in figure 6 & 13. However RF gain approximate 91% harvesting rate and 9% target recall for focused crawler over LPE as shown in figure 9 & 12. After evaluating the performance web classifier with web content extraction technique following outcome has been acquired. LPE is best-suited

web content extraction technique for the focused crawler to classify web page. Whereas Anchor text and CBP gives the biased result. SVM is best-suited classifier for web page classification through focused crawler Whereas LR and RF gives the biased result.

VII. CONCLUSION

This paper incorporates a comparative analysis to evaluate the performance of classifier used in focused crawler for web page classification after examine the web-based feature. Web-based feature such as anchor text, page content, and link generate web feature vector that significantly improves the evaluation of relevancy score for selecting the most relevant web page for training the classifier. This paper also presents a three-tier framework to classify web page. This framework initially pre-processed web URLs to examine their feature for training purpose and yield interesting result about web classifier. For the feature extraction, LPE is best suited as a web content extraction technique over the web page. However, this paper also evaluates the performance of supervised classification technique to classify web page and it is observed that SVM is best-suited web page classifier.

REFERENCES

1. Achsan, Harry T. Yani, and Wahyu Catur Wibowo. A fast distributed focused-web crawling. Procedia Engineering, 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.
2. Ahmadi-Abkenari, Fatemeh, and Ali Selamat. An architecture for a focused trend parallel web crawler with the application of clickstream analysis. Information Sciences, 2012, 184(1).pp. 266-281
3. Barros, R., J. A. Rodrigues Nt. , H. J. A. Carneiro Filho, F. R. S. Ferreira, O. C. Fernandes, C. E. P. Silva, A. L. G. Ribeiro, G. B. Xexeo, and J. M. de Souza. 2009. A collaborative approach to building evaluated web pages datasets. In 2009 13th international conference on computer supported cooperative work in design,2009, pp.668-673
4. Batsakis, Sotiris, Euripides G. M. Petrakis, and Evangelos Milios. Improving the performance of focused web crawlers. Data and Knowledge Engineering, 2009, pp. 1001-1013.
5. Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. Focused crawling: A new approach to topic-specific web resource discovery. Computer Networks, 2000, pp.1623-1640
6. Chen, X., and X. Zhang. Hawk: A focused crawler with IEEE content and link analysis. In 2008 international conference on e-business engineering, 2008, pp. 677-680.
7. Dalins, Janis, Campbell Wilson, and Mark Carman. Criminal motivation on the dark web: A categorization model for law enforcement. Digital Investigation, 2018, 24, pp.62-71.
8. Deka, Ganesh Chandra. Chapter three – nosql web crawler application. In A deep dive into nosql databases: The use cases and applications, eds. Pethuru Raj and Ganesh Chandra Deka. Vol. 109 of Advances in computers, 2018,pp.77-100.
9. Dimri, Neha, Himanshu Kaul, and Daya Gupta. Metaxplorer: an intelligent and adaptable metasearch engine using a novel ordered weighted averaging operator. International Journal of System Assurance Engineering and Management, 2018, vol.9, pp.1-11.
10. Dong, H., and F. K. Hussain. Self-adaptive semantic focused crawler for mining services information discovery. IEEE Transactions on Industrial Informatics. 2014, vol. 10(2), pp.1616-1626.
11. Du, Yajun, Wenjun Liu, Xianjing Lv, and Guoli Peng. An improved focused crawler based on semantic similarity vector space model. Applied Soft Computing. 2018, vol.36, pp. 392-407.
12. Garcia-Nunes, Pedro Ivo, and Ana Estela An tunes da Silva. Using a conceptual system for weak signals classification to detect threats and opportunities from web. Futures. 2019, vol. 107, pp.1-16.
13. Geng, Z., D. Shang, Q. Zhu, Q. Wu, and Y. Han. Research on improved focused crawler and its application in food safety public opinion analysis. In 2017 chinese automation congress (cac).2017, pp.2847-2852
14. Khalil, Salim, and Mohamed Fakir. Rcrawler: An r package for parallel web crawling and scraping. SoftwareX, 2017, vol.6, pp. 98-106.
15. Kim, Iljoo, and Gautam Pant. Predicting web site audience demographics using content and design cues. Information and Management. 2018.
16. Kumar, Manish, Ankit Bindal, Robin Gautam, and Rajesh Bhatia. Keyword query based focused web crawler. Procedia Computer Science. 2018, pp. 584-590.
17. Kumar, Mukesh, and Vig Renu. Learnable focused meta crawling through web. Procedia Technology. 2012, pp. 606-611.
18. Lee, Ji-Hyun, Wei-Chang Yeh, and Mei-Chi Chuang. Web page classification based on a simplified swarm optimization. Applied Mathematics and Computation. 2015, vol.270, pp. 13-24
19. Liu, Xiaojun, and Wei Hu. Attention and sentiment of chinese public toward green buildings based on sina weibo. Sustainable Cities and Society. 2019, vol. 44, pp. 550-558
20. Lu, Houqing, Donghui Zhan, Lei Zhou, and Dengchao He. An improved focused crawler: Using web page classification and link priority evaluation. Mathematical Problems in Engineering, 2016, pp-1-10
21. Malhotra, Ruchika, and Anjali Sharma. Quantitative evaluation of web metrics for automatic genre classification of web pages. International Journal of System Assurance Engineering and Management.2017
22. Olston, Christopher, and Marc Najork. Web crawling. Foundations and Trends R in Information Retrieval. 2010, Vol. 4(3), pp. 175-246
23. Pant, Gautam, Padmini Srinivasan, and Filippo Menczer. Crawling the web. In In web dynamics: Adapting to change in content, size, topology anduse. edited by m. levene and a. poulovassilis, 153-178.Springer.2004, pp. 153-178.
24. Patel, Ahmed, and Nikita Schmidt. Application of structured document parsing to focused web crawling. Computer Standards and Interfaces. 2011, vol. 33(3), pp.325-331.
25. Qi, Xiaoguang, and Brian Davison. Web page classification: Features and algorithms. ACM Comput. Surv. 2009. 41.
26. Ricca, Filippo, Maurizio Leotta, and Andrea Stocco. Three open problems in the context of e2e web testing and a vision. 2018,
27. Saleh, Ahmed I., Arwa E. Abulwafa, and Mohammed F. Al Rahmawy. 2017a. A web page distillation strategy for efficient focused crawling based on optimized naive bayes (onb) classifier. Applied Soft Computing. 2017, vol. 53,pp. 181-204.
28. Saleh, Ahmed I., Arwa E. Abulwafa, and Mohammed F. Al Rahmawy. 2017b. A web page distillation strategy for efficient focused crawling based on optimized naive bayes (onb) classifier. Applied Soft Computing.2017, vol. 53, pp.181-204.
29. Seyfi, Ali, and Ahmed Patel. A focused crawler combinatory link and content model based on t-graph principles. Computer Standards and Interfaces .2016, vol. 43, pp. 1-11.
30. Seyfi, Ali, Ahmed Patel, and Joaquim Celestino J_unior. Empirical evaluation of the link and content-based focused treasure-crawler. Computer Standards and Interfaces.2016, vol. 44, pp. 54-62.
31. Shemshadi, Ali, Quan Z. Sheng, and Yongrui Qin. Chapter 2 - the anatomy of an intent based search and crawler engine for the web of things. In Managing the web of things, eds. Quan Z. Sheng, Yongrui Qin, Lina Yao, and Boualem Benatallah.2017, pp. 37-72.
32. Sirisha Gadiraju, N. V. G., R. Krishna Chaitanya, and G. V. Padma Raju. Effect of feature selection method on the performance of focused crawlers|a case study on traditional and accelerated focused crawlers. In 2010 international conference on networking and information technology,2010, pp. 482-487.
33. Stevanovic, Dusan, Aijun An, and Natalija Vlajic. Feature evaluation for web crawler detection with data mining techniques. Expert Systems with Applications.2012. vol. 39 (10), pp. 8707-8717.
34. V. Udapure, Trupti, Ravindra D. Kale, and Rajesh C. Dharmik. Study of web crawler and its different types. IOSR Journal of Computer Engineering. 2014. vol. 16, pp. 01-05.
35. Wang, Rong, Yan Zhu, Jiefan Tan, and Binbin Zhou. Detection of malicious web pages based on hybrid analysis. Journal of Information Security and Applications. 2017, vol. 35, pp. 68-74.
36. Wang, W., X. Chen, Y. Zou, H.Wang, and Z. Dai. A focused crawler based on naive bayes classifier. In 2010 third international symposium on intelligent information technology and security informatics. 2010. pp. 517-521.
37. Yan, W., and L. Pan. Designing focused crawler based on improved genetic algorithm.In 2018 tenth international conference on advanced computational intelligence (icaci). 2018, pp. 319-323.
38. Zhang, Huaxiang, and Jing Lu. Sctwc: An online semi-supervised clustering approach to topical web crawlers. Applied Soft Computing, 2010, vol. 10 (2), pp. 490-495.
39. Zhao, F., J. Zhou, C. Nie, H. Huang, and H. Jin. Smartcrawler: A two-stage crawler for efficiently

harvesting deep-web interfaces. IEEE Transactions on Services Computing, 2016, vol. 9 (4), pp. 608-620.
40. Zheng, Hai-Tao, Bo-Yeong Kang, and Hong-Gee Kim. An ontology-based approach to learnable focused crawling. Information Sciences. 2008, vol. 178, pp. 4512-4522.
41. Zhou, Jianghui, Chunming Cheng, Li Kang, and Ruizhi Sun. Integration and analysis of agricultural market information based on web mining. IFAC-PapersOnLine. 2018, vol. 51 (17), pp.778-783.
42. Zhu, Q. 2007. An algorithm ofc for the focused web crawler. In 2007 international conference on machine learning and cybernetics.2007,pp. 4059-4063.