Comprehensive Analysis of Web Page Classifier for Focused 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
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.

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.

### 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. Ahmad Abkenari and Selamat (2012) present treasure graph. Seyfi et al. (2016) present 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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Task</th>
<th>CT</th>
<th>Feature</th>
<th>Methods</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lu et al. (2016)</td>
<td>Web Page Classification</td>
<td>FC</td>
<td>Anchor text, Link Content</td>
<td>Improved Term Weight</td>
<td>Link Priority Evaluation, Page Content Block Partition, Joint Feature Evaluation</td>
</tr>
<tr>
<td>Wang et al. (2010)</td>
<td>Relevant Web Page Selection</td>
<td>FC</td>
<td>Kernal Density, Normal density, Page Rank, BFS</td>
<td>TF-IDF</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>Zhao et al. (2016)</td>
<td>Harvesting Deep Web Interface</td>
<td>FC</td>
<td>Path, Link, Page Content, Anchor</td>
<td>Reverse and Incremental Searching</td>
<td>Form Classifier</td>
</tr>
<tr>
<td>Du et al. (2015)</td>
<td>Semantic Focused Crawler</td>
<td>FC</td>
<td>Anchor text</td>
<td>Semantic Similarity Retrieval Model</td>
<td>Cosine Similarity</td>
</tr>
<tr>
<td>Seyfi et al. (2016)</td>
<td>Treasure Crawler</td>
<td>FC</td>
<td>Link, Content, Anchor, Heading, URL</td>
<td>Topic Boundary</td>
<td>T-graph, D-tree</td>
</tr>
<tr>
<td>Geng et al. (2017)</td>
<td>Public opinion over Food Safety</td>
<td>FC</td>
<td>Page Content</td>
<td>Similarity computation based Multiple Reference Factor</td>
<td>Term Frequency, Inverse Document Frequency, Best First, Breadth-First</td>
</tr>
<tr>
<td>Kumar et al. (2018)</td>
<td>Keyword Query Based Focused Crawler</td>
<td>FC</td>
<td>URL, Link</td>
<td>Query-based Crawler</td>
<td>Dom Structure, K level, Max Ancestor</td>
</tr>
<tr>
<td>Yan and Pan (2018)</td>
<td>Evolutionary Focused Crawler</td>
<td>FC</td>
<td>Page Rank</td>
<td>Vector Space Model</td>
<td>Genetic Algorithm, Similarity Function</td>
</tr>
<tr>
<td>Zhou et al. (2018)</td>
<td>Agricultural Market Analysis</td>
<td>FC</td>
<td>Theme Relevancy crawling</td>
<td>Key Multi-Mode Matching</td>
<td>Aho Corasick Algorithm, Term frequency and Inverse Document Frequency</td>
</tr>
<tr>
<td>Khalil and Fakir (2017)</td>
<td>Duplicate Content Detection</td>
<td>WC</td>
<td>URL, Page Content and Depth level</td>
<td>Parallel Web crawling and Scraping</td>
<td>Similarity hash Function</td>
</tr>
<tr>
<td>Liu and Hu (2019)</td>
<td>Public Sentiment Analysis towards Green Building</td>
<td>FC</td>
<td>Keywords Search and Depth level</td>
<td>Ontology and text mining</td>
<td>Part of Speech, dictionary-based approach</td>
</tr>
<tr>
<td>Saleh et al. (2017a)</td>
<td>Web page classification</td>
<td>FC</td>
<td>Keywords Search and Wordnet</td>
<td>Disambiguation Domain ontology</td>
<td>Naive Bayes, Support Vector Machine, Genetic Algorithm</td>
</tr>
</tbody>
</table>
### Table 1. Article Summary

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>FC</th>
<th>Labeling and Page Content</th>
<th>TOR used Motivation Model</th>
<th>SHA-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dalins et al. (2018)</td>
<td>Analysis Dark Web for Child Pornography</td>
<td>FC</td>
<td>Labeling and Page Content</td>
<td>TOR used Motivation Model</td>
<td>SHA-1</td>
</tr>
<tr>
<td>Wang et al. (2017)</td>
<td>Detection of Malicious Web Page</td>
<td>WC</td>
<td>Correlation Based Feature</td>
<td>Machine Learning</td>
<td>Decision Tree, DOM</td>
</tr>
<tr>
<td>Seyfi and Patel (2016)</td>
<td>Topically Relevant Harvesting</td>
<td>FC</td>
<td>Link and Content</td>
<td>Treasure Graph</td>
<td>Best First and Similarity Function</td>
</tr>
<tr>
<td>Ahmadi-Abkenari et al. (2012)</td>
<td>Web importance metrics Analysis</td>
<td>FC</td>
<td>Link and Content</td>
<td>Treasure Graph</td>
<td>Best First and Similarity Function</td>
</tr>
</tbody>
</table>

**V. COMPARATIVE ANALYSIS**

Comparative analysis of recent research for the design of focused crawler for web page classification presents interesting and useful facts regarding the state-of-the-art of web-based feature extraction techniques. 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 reduced 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).

\[
W_{vs} = (W_{Anchor}, W_{Link}, W_{Keywords}) f_{url}
\]  

Where,

- \( f_{url} \) is seed URL
- \( W_{vs} \) is web vector space
- \( W_{Anchor} \) is the set of anchor text
- \( W_{Link} \) 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.

\[
(W_{vs} + High) \geq 0
\]
(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.

\[
C_{wp} = \arg \max_{p \in \{\text{Relevant, Irrelevant}\}} P(r|wp) \tag{4}
\]

\[
P(r|wp) = \frac{P(wp|r)P(r)}{P(wp)} \tag{5}
\]

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’.

\[
P(wf|\text{score}) = P(wf|\text{anchor}) \times P(wf|\text{link}) \times P(wf|\text{content}) \tag{6}
\]

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

(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 \(E_{wbc}\) is exception on web class (ws) classification with random parameter (r) on condition x and frontier seed set \(F_{ss}\).

\[
\tilde{r}_n(x, f_{ss}) = E_{wbc}[r_n(x, wc, f_{ss})] \tag{7}
\]

Whereas after incorporating web feature as conditional parameter ‘x’ lead to a minimized exception \(E_{wbc}\) 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.

\[
W_f(x) = \begin{cases} 
\text{relevant} & \text{if } C \ast x \geq 0 (\text{relevancy score}) \\
\text{irrelevant} & \text{if } C \ast x < 0 (\text{relevancy score})
\end{cases} \tag{8}
\]

\[
W_c(x) = \text{threshold } W \ast x \tag{9}
\]

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.

\[
\text{Precision} = \frac{|XY|}{|Y|} \ast 100 \tag{10}
\]

\[
\text{Recall} = \frac{|XY|}{|X|} \ast 100 \tag{11}
\]

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 \(\beta\) as shown in equation 12 & 13.

\[
F - \text{Measure} = \frac{(\beta^2 + 1) \ast \text{Precision} \ast \text{Recall}}{\beta^2 \ast \text{Precision} + \text{Recall}} \ast 100 \tag{12}
\]

\[
\begin{cases} 
\text{if } \beta > 1 & (\text{Recall is more important}) \\
\text{if } 0 < \beta < 1 & (\text{Precision is more important}) \\
\text{if } \beta = 1 & (\text{Both are equally important})
\end{cases} \tag{13}
\]
Comprehensive Analysis of Web Page Classifier for Focused Crawler

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

Retrieval Number: I7477078919/19©BEIESP
DOI:10.35940/ijitee.I7477.078919
Published By:
Blue Eyes Intelligence Engineering & Sciences Publication
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.

\[
Harvesting\ Rate = \frac{|S(x) \cap R|}{S(x)} \times 100
\]  

\[
Target\ Recall = \frac{|S(x) \cap R|}{R} \times 100
\]

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
P.2847

Comprehensive Analysis of Web Page Classifier for Focused Crawler

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
33. In 10th international conference on networking and information technology.2010, pp. 482-487.