Received: August 2, 2023. Revised: September 23, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Text Matching Technique-based Intelligent Web Crawler in Hybrid Mode

Amol Subhash Dange[1]* Manjunath Swamy Byranahalli Eraiah[1]
Asha Kethaganahalli Hanumanthaiah[2] Manju More Eshwar Rao[3]
Sunil Kumar Ganganayaka[4]

[1]Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India, and Visvesvaraya Technological University, Belagavi.
[2] Department of Information Science and Engineering, Global Academy of Technology, Bengaluru, India, and Visvesvaraya Technological University, Belagavi.
[3]Department of Computer Science and Engineering, PES University, Bengaluru, India
[4]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bengaluru, India
* Corresponding author's Email: amoldange_cse@adcet.in

Abstract: Web crawlers gather and analyze a large amount of data available online to obtain specific forms of objective data, such as news. Web crawlers are becoming more important since big data is used in numerous different sectors and web data is rising dramatically each year. However, when analyzing large volumes of information and making rapid decisions, the organization frequently uses minimal data, which leads to inefficient choices. In this paper, the minibatch stochastic gradient descent (SGD) optimization and radial basis function SVM are proposed to assist organizations in the targeted crawling of relevant online artifacts and semantically matching them against internal big data for better strategy decisions. The proposed method has been used and extensively evaluated in the e-procurement field. The minibatch SGD optimization and radial basis function SVM has gradually been expanded to include more fields such as robot programming and cloud hosting. The existing methods of web crawler for pharmacokinetics (WCPK), automatic extraction method, malicious website detection techniques, and BERT with softmax layer method are used to justify the effectiveness of the proposed minibatch SGD optimization and the radial basis function SVM method. The proposed method achieves better precision, recall, and f1-measure of 99.25%, 98.91%, and 99.57% on DMOZ dataset and 96.23%, 94.71%, and 97.53% on synthesized dataset when compared to the existing methods.

Keywords: Web crawler, Intelligent crawler, Text matching, Score of relevance, Semantic similarity.

1. Introduction

The web crawler, search engine bot, or spider plays a crucial role in downloading and indexing vast amounts of information from the world wide web. Its main objective is to comprehend the content of web pages and gather relevant information whenever necessary. In recent years, websites have witnessed substantial growth in size due to the incorporation of diverse data formats, particularly multimedia components. According to recent web structure and content research, the indexed web has at least 4.49 billion pages, while the deep web is thought to be 500 times larger than the surface web [1]. Research and development efforts utilize Artificial intelligence, which includes the integration of natural language processing and web crawling, to create a recommender system [2]. Web applications have been the primary force behind global networking, collaboration, and social engagement [3]. Document summary or compressing the contents of a text to a brief form that captures the significant concepts in the document is an important task in the natural language domain [4]. The vertical search engine and domain-specific information-sharing platform can both be created using the focused crawler [5]. Additionally, one of the most significant attack surfaces is web

pages where hackers pretend to be a well-known company's website to collect data from users, typically through a login or sign-up form [6]. The attacker creates a fake website and distributes links to social media sites like Twitter, Facebook, as well as emails, instructing the receiver to act immediately and creating a sense of urgency or panic [7].

With the growth and popularity of the Internet, a growing number of social media sites like GoWhere, Douban, Weibo, and Douban enable users to express their thoughts and send text comments web [8]. Currently, the majority of web-based information services, including news portals, web-based information services, and repository management of public scientific data are built on database (DB) applications [9]. The use of internet of things (IoT) devices for a variety of purposes, from home automation to the monitoring of vital infrastructures, has complicated the cyber-defense environment [10]. Online investigation and digital forensics, which make use of open-source intelligence (OSINT) are some of the most popular technologies and techniques for identifying cybercriminals [11]. The majority of text data hiding techniques are classified into three categories: format-based text information hiding, image-based text information hiding, and natural information hiding [12]. Web pages are made up of HTML elements and the data that is passed between them. Web scraping is the process of obtaining specific data from these elements to generate data for other applications such as product review analysis, online price change monitoring, weather data monitoring, article gathering, and tracking online presence [13]. However, when analyzing large volumes of information and making rapid decisions, the organization frequently uses minimal data, which leads to inefficient choices. To overcome these issues, the primary contributions of the paper are summarized below:

- The main goal of the minibatch SGD optimization and radial basis function SVM is to assist organizations in the targeted crawling of relevant online artifacts and semantically matching them against internal big data for better strategy decisions.
- This work combines a semantic equality-depend method with a probabilistic similarity mechanism, utilizing title text, bold terms, anchor text, and title as feature documents.
- The minibatch SGD optimization and radial basis function SVM are evaluated based on precision, recall, and f1-measure.

The rest of the paper is organized as follows: The literature survey is explained in section 2. Section 3 explains the methodology. Section 4 describes the results & discussion. And section 5 discusses the conclusion.

2. Literature survey

Kaur [14] implemented an intelligent hidden web crawler (IHWC) method to solve the relevant issues such as domain classification, prioritizing URLs, and avoiding exhaustive searching. Using rejection rules, the crawler selects appropriate web pages and ignores the insignificant ones. The IHWC method accurately and effectively harvests hidden web interfaces from large-scale websites and achieves higher rates. However, the implemented method requires a limited amount of data coverage to obtain all websites in urban domains.

Bifulco [15] presented crawling artefacts of interest and matching them against eNterprise sources (CAIMANS) to support businesses in choosing web artifacts based on their backgrounds and experience, while also providing alignment with their data and information sources. The presented method uses a K-means algorithm extension along with a semantic module to identify significant information inside crawled artifacts. CAIMANS increases manual search in terms of search time and extracted result quality across a variety of application domains. However, sharing sensitive data between several systems during the process of matching artifacts with enterprise source increase the risk of data breaches and safety risk.

Hosseinkhani [16] introduced an anti-based crawler method (ANTON) to create an optimal ontology-based method for web crime mining. The ANTON was built on an improved crime ontology by employing an ant-miner-focused crawler, which was formed by research on ant foraging behavior. The ANTON method was optimized using an ant colony optimization. The use of crime ontologies and an improved ant-based crawler achieves better accuracy. However, creating ant foraging behavior and updating the ontology needs a large amount of computing time and resources.

Remya Ampadi Ramachandran [17] implemented a web crawler for pharmacokinetics (WCPK) method for PK analysis provided a new to web crawling. The implemented WCPK method was established to fill the gap of web crawler capable of handled full-text downloads. The implemented WCPK method was automatically handled the full-text retrieval module from results of metadata search. However, when it comes to the analysis of large data

the implemented WCPK method was not be feasible.

Zhinian Shu & Xiaorong Li [18] implemented an automatic extraction method of web text information based on network topology coincidence degree. Web crawler, hypertext tag, and search engine were utilized for web text information classification, and the reduction of dimensionality was carried out. The implemented automatic extraction method had promoted the information of web text had high authority and quality. However, the implemented method's web page structure was changeable and complex, so it was difficult to extracted the text of surrounding web pages.

Liu and Lee [19] presented a malicious website detection technique that uses a convolutional neural network (CNN) to learn and identify screenshot images of webpages. As a classification procedure, the presented method employs a CNN, which is a type of deep neural network. The experimental outcomes demonstrate that the malicious website detection technique performs better and is suitable for a real-world web context. However, the distribution of malicious websites in the dataset was significantly imbalanced with a large number of legitimate websites.

Amit Kumar Nandanwar & Jaytrilok Choudhary [20] implemented a BERT with Softmax layer, which was a fine-tuned model employed to classify the web pages as the model of categorization. The implemented model utilized contextual embeddings developed by symmetry multi-head encoder layer of the bidirectional encoder representations from transformers (BERT). The implemented model effectively increases performance by memorizing critical information and finding patters from unlabeled text data. However, the implemented model required to increase the classification performance in various feature-combination method including semantic and contextual features.

3. Methodology

The proposed minibatch SGD optimization and the radial basis function SVM method are proposed to assist organizations in the targeted crawling of relevant online artifacts and semantically matching them against internal big data for better strategy decisions. The proposed method has significant parts such as a fetcher of webpages for fetching and downloading the relevant pages, and a repository for crawlers through which webpages downloaded are

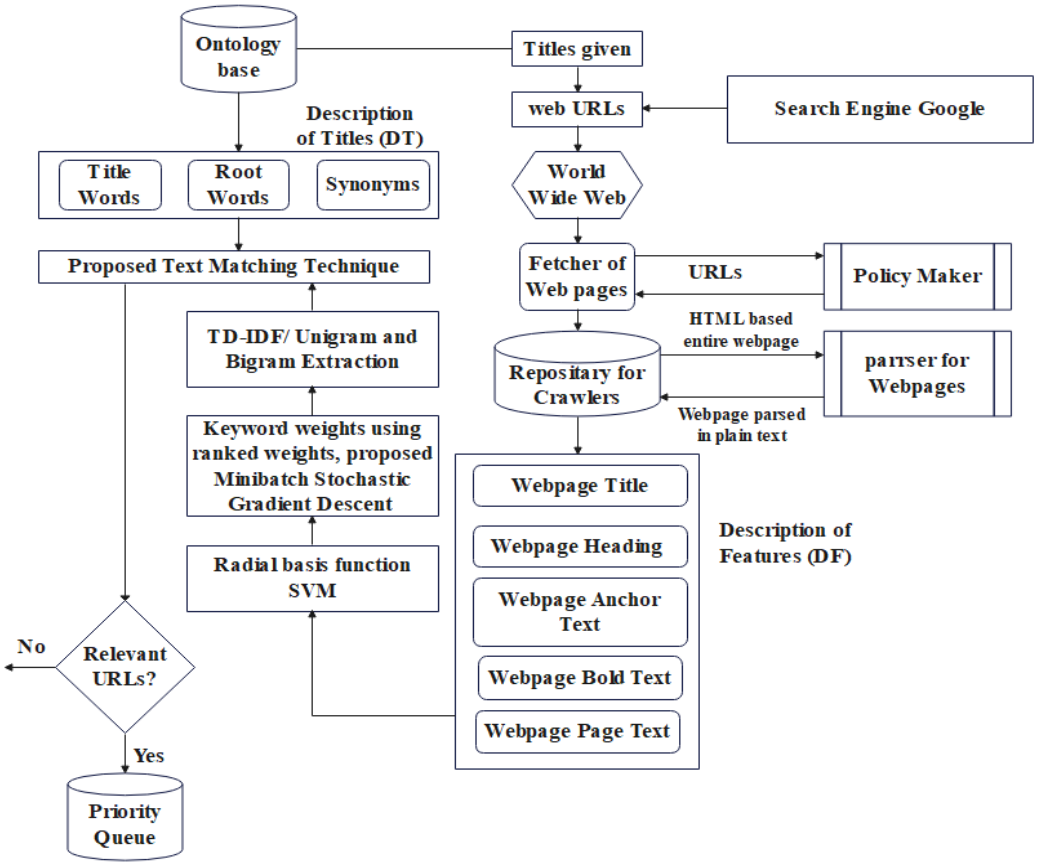

Figure. 1 Block diagram of the proposed method

stored. Then, policy maker inspects the strategies for web pages downloadable, parser for webpages parses the pages downloaded, and computations for calculation of scores among description of titles (DT) and description of features (DF) by making use of the proposed algorithm.

The user initializes the titles and the URLs of these titles will be fetched from the results of the search engine. The DT values are generated by obtaining the words and synonyms. To download the webpages, the twelve URLs are sent to the fetcher of webpages one after the other and then sent to the policy maker, where verifies the predefined policies and returns the information to the fetcher and the downloading of webpages will begin after the acceptance of the webpages. The downloaded pages will be stored by the repository and parsed by the parser of webpages by returning webpages in the form of plain text by eliminating the HTML tags. Then the extraction of features including titles, headings, texts that are anchored and bold, and so on are done from the pages parsed. The DF values are generated by making use of mechanisms for processing features such as stemming, tokenization of the words, and speech tagging. The DT and DF term similarities are computed by making use of the text matching technique in hybrid mode. If this similarity is below a threshold value, then the web pages are treated as irrelevant. Based on the web page's score of relevance, every URL is assigned priority before the priority queue stores it.

To find the similarities between DT and DF, the above-mentioned five parameters are used. Based on the pair of words used, the similarity value, and the score of relevance will be varied. Some of the word pairs will have very low scores of relevance. The proposed text-matching technique in hybrid mode will help to find the solution for these challenges. The overview of the proposed methodology is represented in Fig. 1.

The mechanism implemented will have notable parts such as a fetcher of webpages for fetching and installing similar pages, a repository for crawlers by which webpages installed are saved, a policy maker that check the strategies for web pages installed, and measurements for calculations of the vector between the description of titles (DT) and description of features (DF) through making utilize of the implemented algorithm. Fig. 2 illustrates these steps utilized in the introduced crawler methodology.

The user starts the titles and the URLs of these titles will be fetched from a solution of google. The DT values are created by getting the terms and synonyms. To install the webpages, the twelve URLs are sent to the fetcher of webpages one after the other and then sent to the policy maker, who, checks the predetermined policies and returns the data to the fetcher and the installing of webpages will begin after the acceptance of the webpages. The installed pages will be saved through the repository and parsed through the parser of webpages by returning webpages in the form of plain texts by removing the HTML tags. Then the extraction of features containing titles, headings, texts that are anchored and bold, and so on are done from the webpages parsed. The DF values are created through making utilization of methods for processing of features such as stemming, tokenization of the words, and speech tagging. The DT and DF term equalities are calculated by making utilization of a text-matching mechanism in hybrid mode. If this similarity is less than the threshold value, then the web pages are treated as irrelevant. Depending on webpage's vector of relevance, all URL is allocated with the priority before the priority queue saves it.

To get the similarities between DT and DF, the above derived 5 parameters are utilized. Depending on the pair of words utilized, the similarities value, and the vector of relevance will be differed. Some of the term pairs will have less score of relevance. The implemented text matching technique in hybrid mode will be used to find the result for these difficulties.

3.1 Term frequency–inverse document frequency (TF-IDF)

The current relevance of net records associated with their collection identification will be accessed by comparing the word content of records to their most recent extent of collection explanation. The recent scope is primarily explained using some reference records explaining an event. Its scope has to be narrowed down when such records have an uncertain topic even more and key terms are distributed to simplify their topical purpose. This allows a potential also an instinctive topical description. The TD-IDF equations are expressed in the following Eqs. (1), (2), and (3).

$$tf(t,d) = \sum_{c \in d} f(c,t) \tag{1}$$

$$idf\ (t,d) = log\ \frac{|D|}{f(t,d)} \tag{2}$$

$$t\ f\ idf(t,f,d) = tf(t,d) \times idf\ (t,d) \tag{3}$$

Were,
d - whole record
f - frequency
c - count in each data

D - whole document

$|D|$ - corpus size

$f(t,d)$ - represent the number of times a term 't' appears in the document 'd'

3.2 Proposed minibatch stochastic gradient descent (SGD) optimization

The process begins by extracting the stop words and analyzing web pages. The remaining terms are then stemmed, and the TF-IDF of every term is measured. Both bigrams, and unigrams are efficient for crawling. Each term is assigned a weight based on its TF-IDF, and this weight is multiplied by the term's TF-IDF value. The terms with higher weights are topic. During the crawling process, the crawler enlarges the group of title key terms by containing relevant terms that are discovered. Pages that contain similar vectors to 0.9 are considered to have similar content, and relevant terms are removed from these pages. The empirical risk gradient is calculated through SGD, and each iteration predicts the gradient using a randomly chosen instance. It modifies the training components $m^{(i)}$ and $n^{(i)}$ as in Eq. (4),

$$\theta_{t+1} = \theta_t - \eta.\nabla_\theta J(\theta_t; m^{(i)}; n^{(i)}) \quad (4)$$

Where, $\nabla_\theta J(\theta_t)$ is the gradient of the loss function $J(\theta_t)$ with respect to θ_t, η is the learning rate, and $m^{(i)}$ and $n^{(i)}$ shows the training data in the form of inputs-outputs pairs.

Batch gradient descent involves unnecessary computations for large datasets. Stochastic gradient descent eliminates this overhead by updating parameters one at a time. Due to its frequent updates with substantial differences, the objective function changes gradually. During bat descent coverages to lesser, SGD's variation allows elements to potentially omit new and possibly better local minima. However, it has been shown that SGD displays the same convergence behavior as batch gradient descent when the learning rate slows down gradually.

3.3 Radial bias function SVM

SVM design identification relies on feature studying thesis mechanisms and finds important applications in computer pattern identification. However, researchers have explored the application of neural networks, leading to current ML mechanisms facing significant difficulties. As a result, efforts have been made to enhance SVM classification algorithms by incorporating web structure analysis. One notable development is the implementation of an SVM classifier in topic crawling, attracting considerable attention from scholars. The cause behind learning is to determine $\emptyset_\alpha(x)$ from the enable characteristics $\{\emptyset_\alpha : \alpha \in \Omega\}$

so that $\emptyset_\alpha(x_i) \to y_i$ for $(i = 1, 2, \dots, n)$ and the expected risk $E(\alpha)$ is minimum. Minimization risk has been expressed in Eq. (5).

$$E_p(\alpha) \approx \frac{1}{2}\int |\emptyset_\alpha(x) - y|\, dQ(x,y) \quad (5)$$

Where, $Q(x,y)$ describes an irrelational probability distribution that is difficult to calculate $E(\alpha)$ straightly. One more simple approach is an actual risk is expressed in Eq. (6):

$$E_p(\alpha) = \frac{1}{2n}\sum_{i=1}^{n} |\emptyset_\alpha(x_i) - y_i| \quad (6)$$

However, a major concern with this method is that even a slight error in the training set cannot guarantee error-free predictions, especially when the number of training instances (n) is limited. In such cases, Vapnik bound for can be used to represent the minimum probability (1-p) of any errors occurring as derived in Eqs. (7) and (8).

$$E(\alpha) \le R_p(\alpha) + \Psi\left(\frac{h}{n}, \frac{log(p)}{n}\right) \quad (7)$$

Where,

$$\Psi\left(\frac{h}{n}, \frac{log(p)}{n}\right) = \sqrt{\frac{1}{n}\left[h\left(log\frac{2n}{h} + 1\right) - log\left(\frac{p}{4}\right)\right]} \quad (8)$$

In this context, the element 'h' represents the Vapnik-Chervonenkis dimension, commonly referred to as VC-dimension. It characterizes the predictive function set $\emptyset_\alpha$. Specifically, in case of linear SVM, it aims for construct two separating hyperplanes that are maximally distant from each other and contain no samples between them. This can be expressed in Eq. (9).

$$w.x + b = \pm 1 \quad (9)$$

For major problems in the non-linear SVMs, $k(x, x_i) = (x, x_i)d$ can be utilized for the polynomial classifiers, $k(x, x_i) = \tanh[k(x, x_i) + \Theta]$ and for neural networks, immensely employed kernel was the Gaussian RBF in Eq. (10).

$$k(x, x_i) = exp\left[\frac{||x - x_i||^2}{((2\sigma)^2)}\right] = exp[-\gamma||x - x_i||^2] \quad (10)$$

For every non-linear classifier, this kernel may be expanded simply to any kind of huge direction. In this σ^2 indicates difference, and $\gamma = 1/2\,\sigma^2$ would be unchangeable. Commonly, a simple bound of $0 < \gamma \leq C$ is used; C is persistent.

The content of information (CI) of DT and DF are separated by the maximum of differences and common terms between them. The score of relevance between these two is described in Eqs. (11), (12), (13) and (14).

$$S_{ci}(DT, DF) = \frac{2*CI(CLS(DT,DF))}{CI(DT)+CI(DF)} \quad (11)$$

where content of information of DT is computed by:

$$CI(DT) = -\log(P(DT)) \quad (12)$$

Content of information of DF is computed by:

$$CI(DF) = -\log(P(DF)) \quad (13)$$

and common lowest subsumer (CLS)among DF and DT is calculated as:

$$CLS(DF, DT) = -\log(P(DF) \vee P(DT)) \quad (14)$$

The value of $P(DF)$ and $P(DT)$ are computed in the Eqs. (15) and (16).

$$P(DF) = \frac{Count\ of\ Wordnet\ concepts\ subsumed\ by\ DF}{Count\ of\ Wordnet\ concepts} \quad (15)$$

$$P(DT) = \frac{Count\ of\ Wordnet\ concepts\ subsumed\ by\ DT}{Count\ of\ Wordnet\ concepts} \quad (16)$$

To increase the similarities among DT and DF, the model proposes a novel text-matching technique in hybrid mode, by making the combination of semantic text matching and probability-based text matching, to satisfy our considerations in the work. The mathematical modeling of text matching technique is given below in Eq. (17).

$$S_{tm}(DT, DF) = Maximum(S_{ci}(DT, DF), S_{bm}(DT, DF)) \quad (17)$$

The probability-based text matching $S_{bm}(DT, DF)$ is computed by making use of the BestMatch25 (BM25) approach, which is a probabilistic-based model that works according to inverse document frequency, doc length normalization and terms frequency. The score of relevance of texts of the entire page, title, font, and headings are utilized to compute the priority of webpage which are unvisited. The mathematical model $S_{pr}(webURL)$ for computing priority of webpages is expressed in Eq. (18).

$$S_{pr}(webURL) = Average(S_{Fpr}(DT, DF), S_{Tpr}(DT, DF), S_{TFpr}(DT, DF), S_{Hpr}(DT, DF)) \quad (18)$$

Where $S_{Fpr}(DT, DF)$ score of entire webpage text, $S_{Tpr}(DT, DF)$ is the score of title texts, $S_{TFpr}(DT, DF)$ is the score of text font and $S_{Hpr}(DT, DF)$ is the score of the headings text. The algorithmic representation of a proposed technique is as follows:

Algorithm: Proposed methodology

Input: Title
Output: $S_{tm}(DT, DF)$
p=1;
while (p<n) do
Fetch the terms of Title and store in DT
q=1
while (q<m) do
Fetch values of Feature sets and store in DF
Compute $S_{tm}(DT, DF) = Maximum(S_{ci}(DT, DF), S_{bm}(DT, DF))$
If $S_{tm}(DT, DF)$> threshold_value
Download webpages and store URLs available in priority queue
End If
End While_m
End While_n

4. Results and discussions

4.1 Experimental setup

In this experiment, the PC used for all trials has a Core i5 processor, 16 GB of RAM, and an NIVIDIA GeFoce TX 2080 Ti graphics card. The model training time is shortened by using GPU computing Tensorflow programming. The implemented method uses DMOZ data and synthesized dataset, search engine, and data manually labelled through access to random hosts are legitimate website data. Its complexity lies in the face that many positive samples employ hyperlink/text hiding strategies and sometimes include sensitive terms as well. As a result, detection is frequently misleading and a high false positive rete results. In this experiment, the dataset has analysed for one month with total of 2500 data.

In that 1100 were simple extraction, 1000 were complex extraction, and 400 were contained new words.

4.2 Performance metrics

- Harvest rate: The harvest rate HR is ratio of relevant web pages recovered (R_{wp}) out of whole pages retrieved (N_{wp}). It is expressed in Eq. (19)

$$HR = \frac{R_{wp}}{N_{wp}} \quad (19)$$

- Precision: Precision (P) is the ratio of properly installed web pages for the subject (r_i) and total installed web pages for the topic (n_i) is expressed in Eq. (20)

$$P = \frac{r_i \cap n_i}{n_i} \quad (20)$$

- Irrelevance ratio (IR): It is measured by the quantity of irrelevant web pages recovered $N_{wp} - R_{wp}$ through a total number of web pages retrieved N_{wp}. It is expressed in Eq. (21).

$$IR = \frac{N_{wp} - R_{wp}}{N_{wp}} \quad (21)$$

- Recall – recall is calculated as the sum of the true positives and the positive class images in Eq. (22)

$$Recall = \frac{TP}{TP+FN} \quad (22)$$

- F1-Score – It is also known as the harmonic mean, which seeks a balance between recall and precision. It is expressed in Eq. (23).

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \quad (23)$$

4.3 Analysis

4.3.1. Performance evaluation using harvest rate

Priority crawler focuses solely on calculating lexical equality among title and web pages disregarding semantic similarity. Consequently, this approach yields a low rate, with an average HR of 0.21 after crawling 5000 pages. On the other hand, the SGD crawler enhances relevance scoring by multiplying the semantic equality vector with TF-IDF score. Poor results are indicated when the crawler achieves an average HR of 0.27. In contrast, the SVM crawler merges the advantages of cosine and semantic similarity. The integration leads to superior results, as demonstrated by the crawler achieving the highest harvest rate and installing the maximum number of web pages compared to more crawlers. In conclusion, the implemented hybrid crawler is the most effective solution. Table 1 shows the average for all topics average HR and Fig. 2 shows a comparison of harvest results for the crawlers.

4.3.2. Performance evaluation through precision

Fig. 3 Presents the comparison of precision rates of the SGD crawler, SVM crawler, priority ranking, and the proposed hybrid crawler. After crawling 5000 web pages, the priority ranking achieved an average precision of 17%, the SGD crawler achieved 23%, the SVM crawler achieved 36%, and the proposed hybrid crawler achieved the highest average precision rate of 36. Its average precision rate of 36% showcases its effectiveness, indicating that the proposed crawler is well-suited for dynamic crawling environments. Table 2. shows the average precision of crawlers.

Table 1. For all Topics average HR

Number of web pages	Priority ranking	SGD crawler	SVM crawler	Hybrid crawler (SGD+SVM)
100	0.36	0.39	0.431	0.514
200	0.34	0.37	0.417	0.526
300	0.34	0.37	0,397	0.567
400	0.34	0.37	0.426	0.542
500	0.33	0.37	0.436	0.497
600	0.33	0.36	0.391	0.483
700	0.33	0.34	0.372	0.471
800	0.32	0.34	0.381	0.462
900	0.32	0.34	0.363	0.431
1000	0.31	0.33	0.347	0.447
2000	0.29	0.32	0.337	0.431
3000	0.27	0.31	0.321	0.414
4000	0.24	0.29	0.317	0.407
5000	0.21	0.27	0.312	0.389

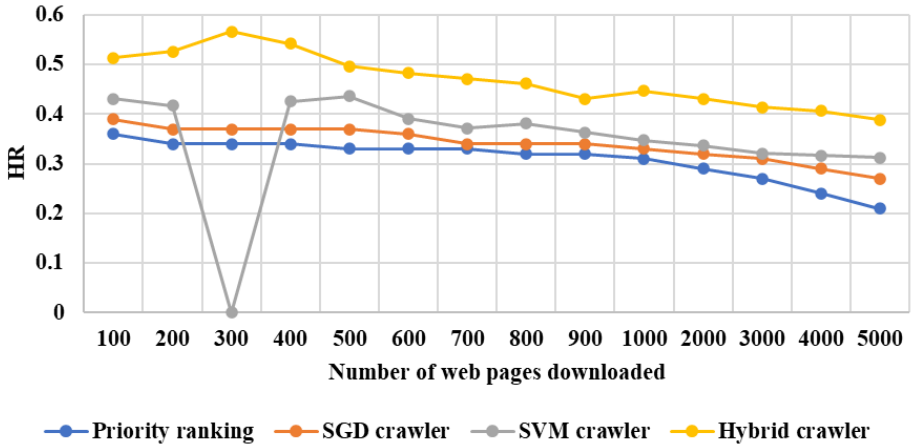

Figure. 2 Comparison of Harvest result for the crawlers

Table 2. Average precision of crawlers

Number of web pages	Priority ranking	SGD crawler	SVM crawler	Hybrid crawler (SGD+SVM)
100	37	32	43	46
200	34	34	44	41
300	34	26	53	42
400	33	37	43	42
500	32	37	42	41
600	31	34	41	41
700	31	32	42	40
800	27	31	39	41
900	24	31	37	40
1000	23	30	36	39
2000	24	31	37	40
3000	21	33	39	41
4000	19	27	33	38
5000	17	23	29	36

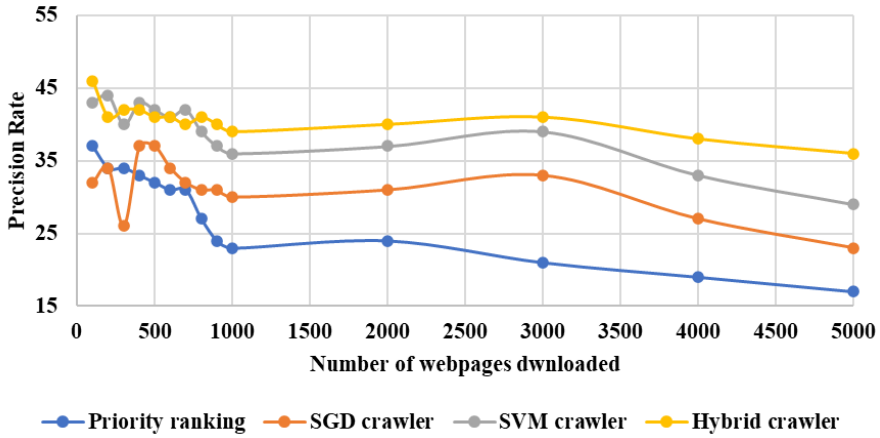

Figure. 3 Comparison of precision rate results for focused crawlers

4.3.3. Performance evaluation by irrelevance ratio

Fig. 4 Illustrate comparison of irrelevance ratio result for SVM, SGD crawler, and proposed hybrid crawler of the average irrelevance ratio. After crawling 5000 web pages, the priority achieved an average IR of 0.79, the SVM crawler achieved 0.688, the SGD crawler achieved 0.73, and the implemented hybrid crawler achieved the lowest average irrelevance ratio of 0.611. The results indicate that the implemented hybrid crawler excels in filtering out highly irrelevant web pages, with an average IR of 0.611. It outperforms the other crawlers in terms of irrelevance ratio. This further emphasizes the capability of the implemented crawler to effectively

Table 3. Average irrelevance ratio of crawlers

Number of web pages	Priority ranking	SGD crawler	SVM crawler	Hybrid crawler (SGD+SVM)
100	0.64	0.61	0.569	0.486
200	0.66	0.63	0.583	0.474
300	0.66	0.63	0.603	0.433
400	0.66	0.63	0.574	0.458
500	0.67	0.63	0.564	0.503
600	0.67	0.64	0.609	0.517
700	0.67	0.66	0.628	0.529
800	0.68	0.66	0.619	0.538
900	0.68	0.66	0.637	0.569
1000	0.69	0.67	0.653	0.553
2000	0.71	0.68	0.663	0.569
3000	0.73	0.69	0.679	0.586
4000	0.76	0.71	0.683	0.593
5000	0.79	0.73	0.688	0.611

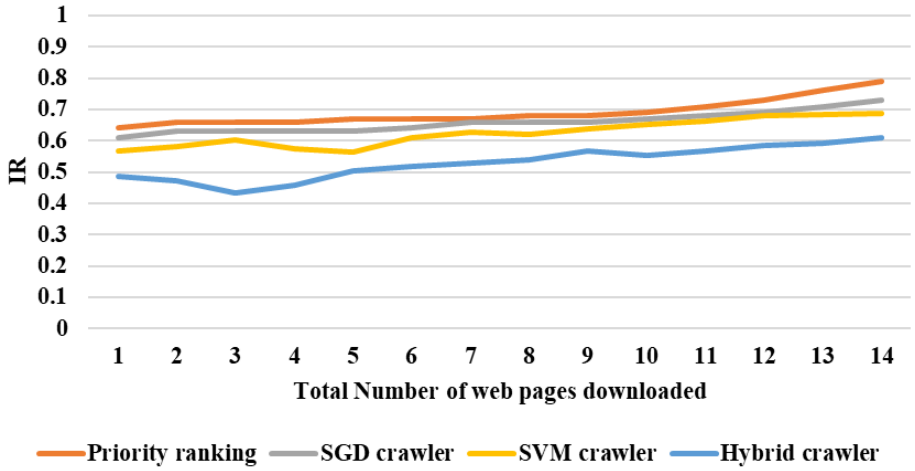

Figure. 4 Comparison of irrelevance ratio result for the crawlers

Table 4. Comparative analysis with existing methods

Author	Methods	Dataset	Precision (%)	Recall (%)	F1-measure (%)
Remya Ampadi Ramachandran [17]	WCPK	Synthesized dataset	92.03	94.42	95.17
Zhinian Shu & Xiaorong Li [18]	Automatic extraction method	Synthesized dataset	94.53	91.20	93.64
Liu and Lee [19]	Malicious website detection technique	DMOZ dataset	91.84	93.61	92.71
Amit Kumar Nandanwar & Jaytrilok Choudhary [20]	BERT, SoftMax	DMOZ dataset	84	83	84
Proposed Method	Minibatch SGD optimization and Radial basis function SVM	Synthesized dataset	96.23	94.71	97.53
		DMOZ dataset	99.25	98.91	99.57

eliminate irrelevant content. Table 3. shows the average irrelevance ratio of crawlers.

4.4 Comparative analysis

The comparative analysis includes methods, precision, recall, and f1-measure. In this scenario, [17

and 18] were used synthesized dataset. And in the similar way, proposed method also synthesized dataset for the comparison. Table 4 shows that the comparative analysis with the existing methods of WCPK [17], Automatic extraction method [18], Malicious website detection technique [19], and BERT, SoftMax [20].

The existing methods WCPK [17] has a 92.03% precision, 94.42% recall, and 95.17% f1-measure, Automatic extraction method [18] has a 94.53% precision, 91.20% recall, and 93.64% f1-measure, Malicious website detection technique [19] has a 91.84% precision, 93.61 % recall, and 92.71% f1-measure. And BERT, SoftMax [20] has an 84% precision, 83% recall, and 84% f1-measure. When compared with the existing methods, the proposed method Minibatch SGD optimization and Radial basis function SVM achieves better precision, recall, and f1-measure of 99.25%, 98.91%, and 99.57% on DMOZ dataset and 96.23%, 94.71%, and 97.53% on synthesized dataset.

4.5 Discussion

This section provides a discussion about the minibatch SGD optimization and radial basis function SVM and compares those results with existing methods in comparative analysis section 4.3. The major goal of this study is to assist organizations in the targeted crawling of relevant online artifacts and semantically matching them against internal big data for better strategy decisions. The proposed method has gradually been expanded to include more fields such as robot programming and cloud hosting. The inclusion of a stop word recognition, stemming module ensures exact keyword identification, underscoring the significance of stop word stemming in an algorithm of a semantic web crawler. This work combines a semantic equality-depend method with a probabilistic similarity mechanism, utilizing title bold terms, text, anchor text, and feature document titles. The DT and DF term similarities are computed by making use of the text-matching technique in hybrid mode. If this similarity is below threshold value, then the web pages are treated as irrelevant. Based on the webpage's score of relevance, every URL is assigned with the priority before the priority queue stores it. The performance evaluation through harvest rate, precision, and irrelevance ratio are determined in the result section. When compared with the existing methods such as the WCPK [17], automatic extraction method [18], malicious website detection technique [19], and BERT, SoftMax [20], the proposed method SGD optimization and radial basis function SVM achieves better precision, recall, and f1-measure of 99.25%, 98.91%, and 99.57% on DMOZ dataset and 96.23%, 94.71%, and 97.53% on synthesized dataset.

4. Conclusion

In this paper, the minibatch SGD optimization and radial basis function SVM is proposed to assist organizations in the targeted crawling of relevant online artifacts and semantically matching them against internal big data for better strategy decisions. This work combines a semantic equality-depend method with a probabilistic similarity mechanism, utilizing title bold terms, text, anchor text, and title as feature documents. The proposed method has been used and extensively evaluated in the e-procurement field. The minibatch SGD optimization and radial basis function SVM has gradually been expanded to include more fields such as robot programming and cloud hosting. The inclusion of a stop word recognition, and stemming module ensures exact keyword identification, underscoring the significance of stop word stemming in a semantic web crawler algorithm. When compared with the existing methods such as the Malicious website detection technique, and PG-VTDM, the proposed method Minibatch SGD optimization and Radial basis function SVM achieves better precision, recall, and f1-measure of 99.25%, 98.91%, and 99.57% on DMOZ dataset and 96.23%, 94.71%, and 97.53% on synthesized dataset.

Notation

symbol	Description
d	whole record
f	Frequency
c	count in each data
D	whole document
$\|D\|$	corpus size
$f(t,d)$	the number of times a term 't' appears in the document 'd'
$m^{(i)}$ and $n^{(i)}$	Training data form of inputs-outputs pairs
$E\ (\alpha)$	expected risk
$Q(x,y)$	irrelational probability distribution
n	number of training instances
h	Vapnik-Chervonenkis dimension
$\emptyset_\alpha$	predictive function set
σ^2	difference
C	persistent
CI	content of Information
CLS	Common Lowest Subsumer
DT	Description of Titles
DF	Description of Features
$CI(DT)$	Content of Information of DT

$CI(DF)$	Content of information of DF
$P(DF)$	Probability of Description of Features
$P(DT)$	Probability of Description of Titles
$S_{Fpr}(DT, DF)$	score of entire webpage text
$S_{Tpr}(DT, DF)$	score of Title texts
$S_{TFpr}(DT, DF)$	score of text font
$S_{Hpr}(DT, DF)$	score of the headings text
HR	Harvest rate
R_{wp}	ratio of relevant web pages recovered
N_{wp}	whole pages retrieved
P	Precision
r_i	ratio of properly installed web pages for the subject
n_i	total installed web pages for the topic
IR	Irrelevance Ratio
$\nabla_\theta J(\theta_t)$	the gradient of the loss function $J(\theta_t)$ with respect to θ_t
η	learning rate

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

For this research work all authors' have equally contributed in conceptualization, methodology, validation, resources, writing—original draft preparation, writing—review and editing.

References

[1] A. Capuano, A. M. Rinaldi, and C. Russo, "An ontology-driven multimedia focused crawler based on linked open data and deep learning techniques", *Multimedia Tools and Applications*, Vol. 79, No. 11-12, pp. 7577-7598, 2020.

[2] N. C. Macias, W. Düggelin, Y. Ruf, and T. Hanne, "Building a Technology Recommender System Using Web Crawling and Natural Language Processing Technology", *Algorithms*, Vol. 15, No. 8, p. 272, 2022.

[3] H. Khatter and A. K. Ahlawat, "An intelligent personalized web blog searching technique using fuzzy-based feedback recurrent neural network", *Soft Computing*, Vol. 24, No. 12, pp. 9321-9333, 2020.

[4] D. Anand and R. Wagh, "Effective deep learning approaches for summarization of legal texts", *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 5, pp. 2141-2150, 2022.

[5] J. Liu, Z. Yang, X. Yan, and D. Chen, "Applying particle swarm optimization-based dynamic adaptive hyperlink evaluation to focused crawler for meteorological disasters", *Complex & Intelligent Systems*, 2023.

[6] M. S. Paniagua, E. Fidalgo, E. Alegre, and R. A. Rodríguez, "Phishing websites detection using a novel multipurpose dataset and web technologies features", *Expert Systems with Applications*, Vol. 207, p. 118010, 2022.

[7] S. D. Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques", *Annals of Data Science*, 2022.

[8] J. Wang, S. Li, and X. Zhou, "A Novel GDMD-PROMETHEE Algorithm Based on the Maximizing Deviation Method and Social Media Data Mining for Large Group Decision Making", *Symmetry*, Vol. 15, No. 2, p. 387, 2023.

[9] T. Bai, Y. Ge, S. Guo, Z. Zhang, and L. Gong, "Enhanced Natural Language Interface for Web-Based Information Retrieval", *IEEE Access*, Vol. 9, pp. 4233-4241, 2021.

[10] P. Koloveas, T. Chantzios, S. Alevizopoulou, S. Skiadopoulos, and C. Tryfonopoulos, "inTIME: A Machine Learning-Based Framework for Gathering and Leveraging Web Data to Cyber-Threat Intelligence", *Electronics*, Vol. 10, No. 7, p. 818, 2021.

[11] C. Horan and H. Saiedian, "Cyber Crime Investigation: Landscape, Challenges, and Future Research Directions", *Journal of Cybersecurity and Privacy*, Vol. 1, No. 4, pp. 580-596, 2021.

[12] Y. Long, Y. Liu, Y. Zhang, X. Ba, and J. Qin, "Coverless Information Hiding Method Based on Web Text", *IEEE Access*, Vol. 7, pp. 31926-31933, 2019.

[13] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages", *IEEE Access*, Vol. 8, pp. 61726-61740, 2020.

[14] S. Kaur, A. Singh, G. Geetha, and X. Cheng, "IHWC: intelligent hidden web crawler for harvesting data in urban domains", *Complex & Intelligent Systems*, Vol. 9, No. 4, pp. 3635-3653, 2023.

[15] I. Bifulco, S. Cirillo, C. Esposito, R. Guadagni, and G. Polese, "An intelligent system for focused crawling from Big Data sources", *Expert Systems with Applications*, Vol. 184, p. 115560, 2021.

[16] J. Hosseinkhani, H. Taherdoost, and S. Keikhaee, "ANTON framework based on semantic focused crawler to support web crime mining using

SVM", *Annals of Data Science*, Vol. 8, No. 2, pp. 227-240, 2021.

[17] R. A. Ramachandran, L. A. Tell, S. Rai, N. I. M. Gedara, X. Xu, J. E. Riviere, and M. J. Douraki, "An Automated Customizable Live Web Crawler for Curation of Comparative Pharmacokinetic Data: An Intelligent Compilation of Research-Based Comprehensive Article Repository", *Pharmaceutics*, Vol. 15, No. 5, p. 1384, 2023.

[18] Z. Shu and X. Li, "Automatic extraction of web page text information based on network topology coincidence degree", *Wireless Communications and Mobile Computing*, Vol. 2022, 2022.

[19] D. Liu and J. H. Lee, "CNN Based Malicious Website Detection by Invalidating Multiple Web Spams", *IEEE Access*, Vol. 8, pp. 97258-97266, 2020.

[20] A. K. Nandanwar and J. Choudhary, "Contextual Embeddings-Based Web Page Categorization Using the Fine-Tune BERT Model", *Symmetry*, Vol. 15, No. 2, p. 395, 2023.