Vol.64: e21210223, 2021
https://doi.org/10.1590/1678-4324-2021210223
ISSN 1678-4324 Online Edition

Review - Engineering, Technology and Techniques

A Critique Empirical Evaluation of Relevance Computation for Focused Web Crawlers

Joe Dhanith Pal Nesamony Rose Mary[1*]
https://orcid.org/0000-0002-9022-9145

Surendiran Balasubramanian[2]
http://orcid.org/0000-0001-5435-0880

Raja Soosaimarian Peter Raj[3]
http://orcid.org/0000-0002-7216-2207

[1]VelTech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Department of Information Technology, India;[2]National Institute of Technology Puducherry, Department of Computer Science and Engineering, Karaikal, India;[3]Vellore Institute of Technology, School of Computer Science and Engineering, Vellore, India.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Fabio Alessandro Guerra

Received: 2021.04.09; Accepted: 2021.05.25.

*Correspondence: joe.dhanith@gmail.com; Tel.: +91-8903939076 (J.D.P.N.R.M.).

HIGHLIGHTS

- This paper presents a survey on focused web crawlers.
- This paper presents the challenges in focused crawling research.
- This paper presents the highlights and hindrances of existing focused web crawlers.
- This paper also presents the future scope for research in focused web crawling.

Abstract: Analogous to the spectacular growth of information-superhighway, The Internet, demands for coherent and economical crawling methods are translucent to shoot up. Consequently, many innovative techniques have been put forth for efficient crawling. Among them the significant one is focused crawlers. The focused crawlers are capable in searching web pages that are suitable for the topics defined in advance. Focused crawlers attract several search engines on the grounds of efficient filtering, reduced memory and time consumption. This paper furnishes a relevance computation based survey on web crawling. A bunch of fifty two focused crawlers from the existing literature survey is categorized to four different classes - classic focused crawler, semantic focused crawler, learning focused crawler and ontology learning focused crawler. The prerequisite and the mastery of each metric with respect to harvest rate, target recall, precision and F1-score are discussed. Future outlooks, shortcomings and strategies are also suggested.

Keywords: Web Crawler; Focused Crawler; Semantic Crawler; Learning Crawler; Machine Learning; Ontology.

INTRODUCTION

The availability and usage of web pages in World-Wide-Web (WWW) has outrun 1.9 billion [1] gradually. Web content like statistics, multimedia and schedules also grows dynamically over this period. The gargantuan formation of data over the internet has become challenging to search the required information within a particular timestamp. Web crawlers alias internet robots, bots, or spiders, a system, which forms the prime part of a search engine, serves as the key parameter capable of facing these internet challenges [2]. The programmed bots or a script which are supposed to be an eminence grise in a search engine reacquires web sites repeatedly by accessing Uniform Resource Locator (URL).

The quantum leap in web contents brings about demanding stretch in maintaining the ongoing indices. Traditional crawlers actually gobble large amount of storage and bandwidth resources. The focused crawlers only download the most relevant web pages rather than downloading all the URLs randomly they visit. The focused crawlers work on the mutualism of the text content and the various URL links visited, to obtain the web pages of higher probability relevant to the topic [3]. This leads to the classification as classic focused crawler, semantic focused crawler, learning focused crawler and the ontology learning focused crawler.

Classic focused crawler searches, captures, indexes, and manages most relevant web pages on particular topic by using Vector Space Model (VSM) [4]. The Semantic focused crawlers are skilled software agents capable of traversing the Web and retrieving and downloading most relevant information from the web on particular topics with thesauri-based semantic similarity algorithms [5]. Learning focused crawler learns from the training set and predicts if the web pages are relevant to the topic. The Ontology learning focused crawler [6] integrates both the semantic technologies and the learning technologies needed to compute the relevance score of the web page.

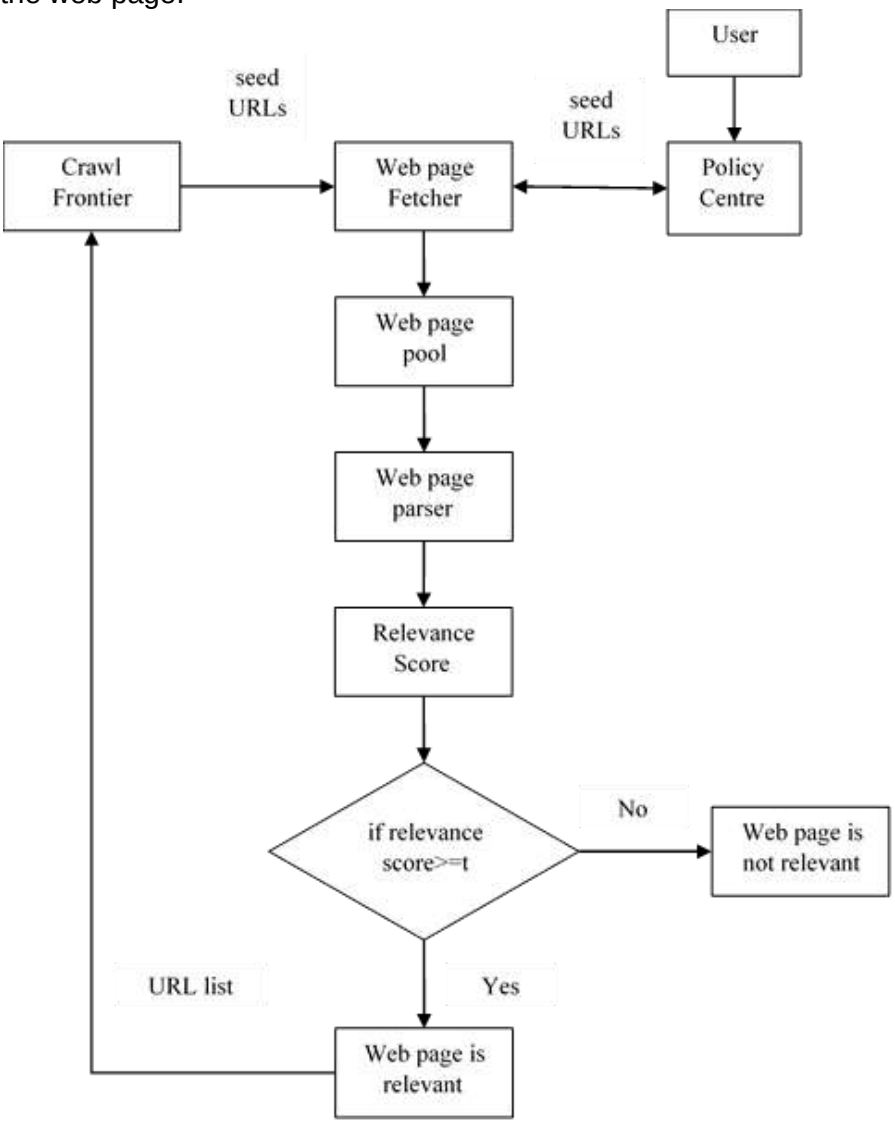

Figure 1. Architecture of Focused web crawler

The architecture of the focused web crawler is displayed in the Fig.1. The mode of action of focused web crawler is described beneath with the given flow diagram.

Step 1: The seed URLs and the depth of visiting the web pages are initialized in the policy centre by the user. After the initialization, the policy centre prepares itself to instruct the web page fetcher in order to download the web pages. The Seed URLs are the starting URLs which are relevant to the topic term. For Example, "https://www.apple.com/" is the Seed URL for the topic Apple.

Step 2: The web page fetcher downloads the web page and collects the URL list of the recently visited web pages and then sends successively to the policy centre. The policy centre checks correspondingly for all the downloadable URLs. Then those URLs are fed back to the web page fetcher and others to irrelevant list.

Step 3: The recently downloaded web pages by the web page fetcher are sent to the web page pool. All the HTML tags from the downloaded web pages that are stored in the web page pool are removed and stored as a plain text.

Step 4: The web page parser extracts only the meaningful information from this plain text.

Step 5: Subsequently this meaningful information extracted from the web page is sent to the relevance computation module to generate the relevance score of the web page. The relevance score above the threshold value is alone considered as relevant. In most of the existing works, threshold value was set as 0.7.

Step 6: The steps (2) to (5) iterates until the user defined depth is achieved.

The pinpointed challenges of the focused web crawling environment are:

(i) The dynamic nature of the information in the web pages, results in the inaccurate computation of the relevance score of the web page, (ii) The VSM based crawlers computes the relevance score exclusively for the web pages that have the topic term co-occurring in the target variables. The semantic similarity is obliterated by these crawlers, (iii) In this, manually predefined weights assigned to the target variables, used to compute the priority score of the web page, is insufficient to achieve a good harvest rate, (iv) The focused crawler also downloads irrelevant web pages because of the ambiguous words present in the web page which leads to the inefficient computation of the relevance score, (v) Priority assignment of the URL along the crawl path is a challenging task in the crawling environment, (vi) Full page text is alone not sufficient to efficiently retrieve the topical relevance of the web page, (vii) Due to the tremendous increase of the web pages in the internet, the number of irrelevant links dominate the relevant web pages. Only negligible links inside the webpage are considered as relevant (viii) Certain text information of the web pages are highly relevant to the topic while the major text are irrelevant. As a consequence, the overall relevance score of the web pages computed using full page text or anchor text or link context according to necessity is low. This may misguide the focused web crawler and produces inaccurate results, (ix) Diversity of services, globally distributed service registries, and the vast amount of information on the web steers to the poor indexing of web pages.

This survey crisps on the challenges and future enhancement of relevance search based crawlers. Fifty two focused crawlers have been explored and grouped into four different categories. A comprehensive assessment is thus done focusing on the four metrics (Harvest Rate, Target Recall, Precision and F1-Score). Harvest Rate is the ratio of count of relevant pages by the total pages downloaded, Target Recall is the ratio of relevant pages from a target set by the total pages downloaded, Precision is the ratio of count of relevant pages to the relevant pages from a target set downloaded and F1-Score is used to measure the aggregate performance of the crawler.

The remainder of this paper specifies that the section 2 projects the various accomplishments of the focused web crawlers, section 3 presents the highlights and hindrances for all the classes of focused web crawlers, section 4 speculates the future enhancements of this crawlers and section 5 presents the conclusion of our work.

Epitomes of focused web crawler

Focused web crawler is a relevance computation based crawler, which competes in downloading relevant web pages to a given topic.

VSM Crawler or Classic Focused Crawler

The VSM crawler is a type of focused crawler, which computes the relevance score of the web page applying the cosine similarity. In this crawler, various target variables discussed in section 2 are used to maneuver the relevance score. The cosine score is computed between the Term Frequency-Inverse Document Frequency (TF-IDF) vector of the target variables and the TF-IDF vector of the topic. The priority of each URL is assigned based on the average VSM score of different target variables.

The hyperlink based ranking considers only the hyperlink structure for the download of web pages, which leads to poor harvesting of web pages. The content based ranking considers only the text content to compute the similarity rank of the web pages, which leads to poor indexing. To solve these challenges [7] proposed a focused crawler by integrating cosine similarity, to compute the content based ranking, with the Bharat Hyperlink Induced Topic Search (BHITS) algorithm, to compute the hyperlink based ranking which evaluates the relevance of the web page. This work uses the knowledge base, incorporating a database of the crawling history, which supports to compute the web page to perform the next crawling.

Link analysis based crawling is inadequate to crawl the web pages accurately for the given topic. To handle these challenges [8] proposed a focused web crawler by combining both the content text and the link analysis. This work proposed a hyper text content link analysis (HCLA) algorithm to compute the relevance of the web page. The HCLA computes the Latent semantic indexing (LSI) weighted VSM for the full text context and the link analysis individually and combines it. The main aim of HCLA is to minimize the reconstruction cost of Singular Value Decomposition (SVD).

Only full page text and anchor text cannot capture the similarity of the web page accurately. To overcome this issue [9] proposed a focused crawler for cross language crawling, which adopts an algorithm, known as Focused crawling for Multiple Relevance Prediction Strategies (FCMRPS). The FCMRPS is an integration of the average similarity score of four target variables (full page text, anchor text, URL address and link structure) with the topic and shark search algorithm. This crawler implements cosine similarity algorithm to compute the similarity score of the target variables (full page text and anchor text) with the topic. The similarity score of the URL address is calculated appertained to the depth of the web page in the Open Directory Project (ODP) and the similarity score of the link structure is calculated dependant on the parent child relationship in the crawl path.

Manual assignment of weight values to the target variables during the computation of priority values of web pages spawning serious deviations in the results. To compute the optimal weight factors and to solve the deviation issue [10] proposed a focused crawler utilizing cell-like membrane-computing optimization algorithm (CMCOA). This work is amalgamation of both the optimal weight factor and the topical similarity. The CMCOA utilizes both the evolution and communication regulars to compute the optimal weight factors of full page text, anchor text, title text and surrounding text of paragraphs. The topical similarities of the full page text, anchor text, title text and surrounding text of paragraphs are computed by the VSM. They are then integrated with the optimal weight factors to compute the priority of the web page.

Seyfi et al. [11,12] proposed a focused crawler by using T-graph principles. This work gives solution to two problems in the focused crawler platform. One is identifying topical focus of the web page and the other is the priority of the web page. Dewey Decimal System (DDS) identifies the topical focus of the web page and T-graph computes the priority of the web page. T-graph is a tree structured graph where each node contains five important HTML attributes such as sub section heading (ISH), section heading which contains ISH, main heading, data around the link and target information. The average cosine score of the five attributes computes the cosine score of each node. If the average cosine measure is equal to 0.05, then the priority is calculated as the inverse of the minimum link distance in the T-graph. If the average cosine measure is exceeds the 0.05, the priority is calculated as the inverse of the graph levels in the T-graph.

The baseline VSM focused crawlers struggled to download the web pages related to recent events. The [13] proposed an intelligent focused crawler to effectively download and archive the web pages related to the recent events. This crawler utilizes three important target variables topic, date and location to effectively capture the recent information about the events. The date is extracted from the URL of the web page by regular expressions. The location vectors of the web pages are extracted by Named Entity Recognition (NER). The topic vector is generated using TF-IDF. These vectors are then used to compute the cosine

similarity. Then an average cosine similarity is computed for date, location and topic to compute the relevance of the web page.

The focused web crawlers encounters latency problem while crawling relevant web pages. The master-slave architecture of [14] helps to optimize the focused web crawler. The main objective is to ensure that, the relevance score of the web page is calculated only after the web page is downloaded. The TF-IDF based cosine similarity is used to compute the relevance score of the web page. The role of the master is to administer the crawl frontier and also the prioritization of the URLs in the crawl frontier. The role of the slave is to download the web page and computes the relevance score of the web page requested by the master. The slave module of the proposed work minimizes the latency of the crawler, by performing threading and parallelization. Table 1 depicts the comparison of VSM crawlers to various specifications proposed by different authors.

Table1. Comparison of VSM crawlers

Author	Algorithm	Target variables	Weighting Scheme	Seed URLs	Metrics	Harvest rate achieved after 5000 web page crawls
Seyfi et al., [11,12]	DDS, VSM	Full page text, anchor text, sub section heading (ISH), section heading which contains ISH, main heading, data around the link and target information	TF-IDF	14% generic seed URLs and 22% on-topic seed URLs	Recall, Harvest Rate	0.27
Yajun et al., [10]	CMCOA, VSM	full page text, anchor text, title text and surrounding text of paragraphs	Evolution regular, communication regular, TF-IDF	3 Seed URLs for each topic	Harvest Rate, Average Relevance and Average Errors	0.297
Almpanidis et al., [8]	VSM and HITS	Full page text and anchor text	LSI		Precision, Recall and Harvest rate	0.21
Chen et al., [9]	VSM	Full page text, anchor text, URL address and link structure	TF	3 Seed URLs for each topic	Harvest Rate, Sum of info, average running time	0.22
Mani Sekhar et al., [14]	VSM	Full page text and anchor text	TF-IDF		Operating time and harvest rate	0.21
Farag et al., [13]	VSM	Topic, Date and Location	TF-IDF, Regular Expression and NER	38 Seed URLs for each topic	Precision, Recall, F1-Measure and Harvest Rate	0.26
Singh et al., [15]	VSM	Full page text and link context	TF-IDF	10 Seed URLs for each topic	Harvest Rate	0.21
Geng et al., [16]	VSM and Multifactor correlation co-efficient	Full page text and crawler theme	TF-IDF		Harvest Rate, Precision and Recall	0.22
Xu et al., [17]	Particle Swarm Optimization	Full page text, anchor text, surrounding text and URL text	TF-IDF	100 Seed URLs for each topic	Harvest Rate	0.29
Rungsawang et al., [7]	VSM and BHITS	Title, full page text, anchor text and link context	TF-IDF	10 Seed URLs for each topic	Harvest Rate	0.27
Kumar et al., [18]	VSM, Hub score and Authority score	Full page text and anchor text	TF-IDF	Seed URLs are generated from ODP	Harvest Rate	0.26

Author	Algorithm	Target variables	Weighting Scheme	Seed URLs	Metrics	Harvest rate achieved after 5000 web page crawls
Goyal et al., [19]	Genetic algorithm	Title, full page text, anchor text, paragraph text, list text, bold text and heading text	Cosine similarity	http://www.stanford.edu/ is crawled up to depth 6.	Harvest Rate	0.26
Zhao et al. [20]	Cosine Similarity	Full page text, Context of URL, anchor text, and text around URL	TF-IDF	100 Seed URLs for each topic	Harvest Rate	0.29
Jung-ran Park et al. [21]	Cosine similarity and HITS	Full page text and anchor text	TF-IDF	15 Seed URLs for each topic	Harvest Rate	0.26
Chen et al. [22]	Cosine similarity and page rank	Full page text and anchor text	TF-IDF	Seed URLs are generated from ODP	Harvest Rate	0.29
Rawat et al., [23]	Cosine Similarity	Full page text and anchor text	TF-IDF		Harvest Rate	0.21
Hati et al. [24]	Cosine Similarity	Full page text, anchor text, cohesive text and also relevance score of parent pages	TF-IDF	1 Seed URL for each topic	Harvest Rate	0.27
Mangaravite et al. [25]	Cosine Similarity	Full page text, anchor text, title text and URL text	TF-IDF	3 Seed URL for each topic	Harvest Rate	0.27
Wei et al. [26]	VSM, cash gain and RVM	Full page text and link context	TF-IDF	10 Seed URL for each topic	Harvest rate, precision and recall	0.34
Gupta et al. [27]	Cosine Similarity	Full page text, keyword text and the title text	TF-IDF	15-20 Seed URL for each topic	Precision and Recall	

Semantic Focused Crawler

Semantic focused crawler, a category of focused crawler, computes the relevance score of the web page using thesauri-based semantic similarity algorithms. For computing the semantic similarity score of the web pages, these crawlers wield ontology. Ontology is domain specific and is designed by domain experts.

Diversity of services, globally distributed service registries, and the vast amount of information on the web are responsible for the poor indexing of web pages. [28] proposed a focused crawler by using a hybrid approach by combining ontology based crawlers and metadata based crawlers to improve poor indexing of web pages. The ontology based crawler captures the semantic meaning of the topic and the metadata based crawler fetches the descriptive text of the URLs, where Enhanced Case based Reasoning (ECBR) algorithm computes the relevance score of the web page. For further enhancement [5] proposed a self adaptive focused crawler based on semantic technologies. This work adopts a hybrid string matching algorithm which efficiently computes the relevance of the web page. The hybrid string matching algorithm is the integration of both the Resnik [29] semantic similarity algorithm and a statistics based similarity algorithm.

Bedi et al., [30] proposed a Social Semantic Focused crawler, to compute the relevance of the web page exercising concept ontology. This crawler scrutinizes only tagged web page for relevance computation. The topic semantic vector and the tagged web page semantic vector is computed by integrating TF-IDF and the semantic similarity score, which is a path length between two synsets (topic and the tagged web page) in concept ontology. Cosine similarity is computed by these two vectors. The web page is relevant if the cosine similarity score is greater than the threshold value or else is irrelevant.

The hyperlink based ranking considers exclusively the hyperlink structure to download the web pages, resulting in poor harvesting. The Content Based ranking considers only the text content to compute the similarity rank of the web pages, which produces poor indexing. To resolve these issues [31] proposed a focused crawler by integrating both the hyperlink ranking and content based ranking methodologies, as extension and intension similarity respectively. When user navigates a web page, certain hyperlinks clicked are carried to the appropriate pages which are considered to be semantically relevant. These semantically relevant web pages reflect in a web-log data and are referred as extension similarity. The intension similarity is referred as information content similarity (ICS) score between the web page and the topic.

Priority assignment for web pages at the crawl path is a challenging task in the crawling environment. [32] proposed a context graph algorithm to assign the download priority at the crawl path. Here, the web pages for the specific topic which the user intents is initially collected during the browsing session. After the user data is collected, a concept lattice is constructed by fast constructing lattice algorithm, henceforth arranging the web pages in descending order based on their TF-IDF weights. This concept lattice is a concept context-graph drawn by computing the semantic similarity between the core and non-core concepts. Based on the semantic similarity score the priority for the unvisited URL is assigned.

The VSM computes similarity score dependant on the co-occurrence of the topic term. Semantic similarity is ignored by VSM which worsen the harvest rate of crawlers. For further enhancement of this issue [33] introduced semantic similarity vector space model (SSVSM). Wu-palmer semantic similarity algorithm integrated over the TF-IDF for the topic term and the web page, to generate semantic vectors. These semantic vectors compute the cosine similarity. Higher the cosine similarity is, the more relevant the page is.

Table.2 depicts the comparison of semantic crawlers to various specifications proposed by different authors.

Table 2. Comparison of Semantic crawlers

Author	Algorithm	Target variables	Ontology	Seed URLs	Metrics	Harvest rate achieved after 5000 web page crawls
Bedi et al., [30]	Social semantic-VSM	Full page text	Concept ontology	10 Seed URLs for each topic	Harvest Rate	0.20
Yajun et al., [33]	SSVSM	Full page text and anchor text	WordNet	3 Seed URLs for each topic	Harvest rate, average similarity and average error	0.29
Dong et al., [5]	Hybrid string matching	Full page text, Meta data description, and link context	WordNet	Seed URLs are generated from ODP	Harvest Rate, Precision, Recall, Harmonic Mean	0.36
Dong et al., [28]	ECBR	Full page text, Meta data description, and link context	WordNet	Seed URLs are generated from ODP	Harvest Rate, Precision, Recall, Harmonic Mean	0.34
Yang et al., [34]	VSM	Full page text	WordNet	10 Seed URLs for each topic	Precision , Recall and Harvest Rate	0.34
Yajun et al., [31]	Information content similarity (ICS)	Full page text and anchor text	WordNet	3 Seed URLs for each topic	Precision , Recall and Harvest Rate	0.26
Dhanith et al., [35]	NPMI and Resnik semantic similarity	Full page text, anchor text, title text, bold text and heading text	WordNet	10 Seed URLs for each topic	Harvest Rate	0.23
Yajun et al., [32]	Formal concept Analysis	Full page text and anchor text	WordNet	3 Seed URLs for each topic	Precision , Recall, F1-Score and Harvest Rate	0.28
Yuvarani et al., [36]	distance based semantic similarity	Full page text, Link context and heading text	ARP Jena	First 10 Seed URLs generated from Google for each topic	Harvest Rate	0.26
Jalilian et al., [37]	distance based semantic similarity, term frequency and fuzzy inference system	Full page text	Ontology developed using protégé	2 Seed URLs for each topic	Harvest Rate	0.29
Hegade et al., [38]	Jaccard similarity, Lesk and Wu-Palmer semantic similarity	Full page text and anchor text	WordNet	https://soundcloud.com is the seed URL	Crawling Time	

Learning Focused Crawler

Learning focused crawler predicts the relevance of the web page on the topic by applying machine learning algorithms. These machine learning algorithms are trained by huge amount of training samples for learning. The trained algorithm is then utilized to predict the relevance of the web page. Most of the learning algorithms in the available literature use TF-IDF feature vectors for learning. The TF-IDF feature vectors are co-occurrence based and computes the similarity only if when the topic term co-occurs in the target variables of the web page.

Priority assignment at the crawl path is a challenging task in the crawling environment. [39] proposed a context graph based approach to assign priority score to the web pages at the crawl path. This work constructs the context graph for each seed document and finally merges the context graph of all the seed documents called merged context graph. The aim of the context graph is to capture the link hierarchies where web pages of relevant topic occur by availing the context information present in the web page. The TF-IDF vector representation of web pages present in this merged context graph, exclusively trains the Naive Bayes (NB) classifier. The NB classifier predicts the relevance of the web page.

Liu et al., [3,40] proposed a learning based focused crawler pertained to Hidden Markov Model (HMM). The user in the course of his browsing session collects useful web pages for a specific topic and a web graph is generated with these web pages. Latent Semantic Indexing (LSI) represents these web pages in a low-dimensional space and an X-means clustering algorithm is calculates the semantic relationship of the web pages, collected by the user. The cluster information and the web graph are incorporated to form a concept graph. From the concept graph, HMM predicts the relatedness of the current web page to the target page by calculating the distance between them.

Full page text is alone not sufficient to efficiently retrieve the topical relevance of the web page. Hence [41] proposed a focused crawler, by combining both the full page text and the link context to compute the topical relevance of the web page. This crawler adopts a four layer (networking, parsing and extraction, representation and intelligent) architecture. The Networking layer downloads the web page; Parsing and Extraction layer converts the html document into plain text and also extracts the full page text and the link context from the web page. The Representation layer converts the extracted documents into TF-IDF based features. These TF-IDF features are then used to train the Support Vector Machine (SVM) classifier. The trained SVM classifier predicts the relevance of the web page.

Only certain links inside the web page indicates relevant web pages while others do not. There is no efficient mechanism to categorize such links available in the web page. [42] proposed a learning based focused crawler using Maximum Entropy Markov Model (MEMM) and Conditional Random Fields (CRF). This work is a three layer architecture which includes data collection, pattern learning and focused crawling. The data collection phase is responsible for the collection of training samples. These training samples serve as input for pattern learning. The pattern learning phase then extracts useful features from the web page, to train the MEMM and CRF. The cosine similarity between the edge, full page text, Meta description, URL text, anchor text and the given topic are computed as a feature to train the MEMM and CRF. The MEMM and CRF then form an important component to predict the relevance of the web page.

Sentiment information grows rapidly day by day in the web. Modern focused crawlers cannot capture the sentiment information from the web. This is resolved by [43] and proposed a sentimental focused crawler to retrieve both the content based crawling and the sentiment based crawling. This work implements a new text classifier which combines both the topic and the sentiment classifiers. If both the classifiers predict the web page as relevant, then the web page is added into the repository. Or else, the web page is sent to the Graph based classifier to predict the relevance of the web page. The graph based classifier uses the Graph tunneling mechanism to predict the relevance of the web page. This is achieved by using Random Walk Path (RWP).

Identifying and separating the web pages with both the positive and negative sentiments is a challenging task. This disadvantage is reduced as [44] proposed a learning based sentimental focused crawler using Support Vector Regression (SVR). This work uses three main target variables Page URL, anchor text and the referring page. There are 21 features such as sentiment score of the URL, sentiment score of the host URL, Frequency of anchor text, sentiment score of the anchor text, average page size with and without HTML tags, DOM objects, number of images in the page, count of outbound links, frequency of sentences, frequency of words, count of unique words, length of sentence, count of self links, link and page size with and without HTML tags, sentiment score of sentences, words, meta data, and title, maximum sentiment score of sentence, and standard variation of sentiment score. These 21 features are extracted from the three target

variables Page URL, anchor text and the referring page to train the SVR. The trained SVR is then used to predict the relevance of the web page.

Certain parts of the web pages are highly relevant to the topic while others are not. Hence, the overall relevance score of the web pages computed using anchor text or link context is low. This may misguide the focused web crawler and produces inaccurate results. To improve the accuracy [45] computes the relevance score of the web page by partition the web pages into smaller parts. This work proposed a Content Block Partition-Selective Link Context (CBP-SLC) algorithm to compute the relevance of the web page. This algorithm utilizes four target variables full page text, anchor text, link context and content blocks (heading, paragraph, address, unordered list, table, table heading, table row, table values) to compute the relevance score of the web page. The sub-classifiers computes the relevance score of the web page by iteratively applying the SVM to construct a final classifier based on the voting method. The feature vector for the classifier was generated using Term Frequency Inverse Positive-Negative Document Frequency (TFIPNDF). The TFIPNDF computes the weight values for both the positive and negative examples. Another solution to improve accuracy proposed by [46] introduces improved Term Frequency Inverse Document Frequency (ITFIDF). This ITIDF uses Information gain metric to weight the terms for evaluating the proportion of feature distribution. Then the feature generated using ITFIDF trains the Naive Bayes classifier to predict the relevance of the web page.

Extraction of domain information for a specific topic is a challenging task. To handle this [47] proposed a semi-supervised learning based approach for focused web crawling. This crawler computes the cosine similarity of title text, full page text, URL text, anchor text and meta description text. These five cosine similarity values are then used to train the Naive Bayes classifier to predict the relevance of the web page.

The basic learning based crawlers repeatedly visit the web page that does not share any relevant website segments. This problem exhibits poor harvest rate. To encounter this challenge [48] proposed a focused crawler using history feature. The recent download-logs in this feature assigns high priority score to the web pages which download more relevant web pages. This work employs three classifiers, where one is trained by link context features, second is trained by linkage features and third trained by history features. These three classifiers adopt the Multinomial Naive Bayes classifier to predict the relevance of the web page. Finally an average combiner amalgamates the prediction results of three classifiers to produce the final prediction result. The connected irrelevant links to a particular web page is more than the relevant links as the internet era grows enormously.

Table.3 depicts the comparison of learning crawlers to various specifications proposed by different authors.

Table 3. Comparison of learning crawlers

Authors	Features	Algorithms	Target Variables	Training Data set	Metrics	Harvest Rate Achieved after 5000 web page crawls
Liu et al., [3,40]	Concept Graph generated using LSI and X-means	HMM	Full page text	ODP and Yahoo Directory	Harvest Rate, Precision and Recall	0.14
Peng et al., [45]	TFIPNDF	SVM	Full page text, anchor text, link context and content blocks	Reuters corpus, 20 Newsgroup corpus and ODP	Precision, Recall, F1-measure, Error rate, Harvest rate and target recall	0.36
Diligenti et al., [39]	TF-IDF	NB	Full page text	Manually collected dataset for 10 topics	Harvest Rate	0.29
Pant et al., [41]	TF-IDF	SVM	Full page text and link context	ODP	Harvest rate, Target recall	0.31
Houqing et al., [46]	ITFIDF	NB	Full page text and link context	Reuters corpus, 20 Newsgroup corpus and ODP	Harvest rate, Target recall	0.34
Pawar et al., [47]	Cosine similarity values	NB	Full page text, title text, anchor text, URL text and meta description text	Medicinal plant dataset	Precision, Recall, Accuracy and Harvest Rate	0.31
Amalia et al., [49]	TF-IDF	Multinomial NB	Full page text	Health and non-health dataset	Harvest Rate	0.29
Suebchua et al., [48]	TF and History feature	Multinomial NB	Full page text, link context, linkage features and recent crawl logs	ODP and Yahoo Japan directory	Harvest rate	0.34
Zheng et al., [50]	TF-IDF	NB	Full page text, anchor text and link context	50000 technical reports with the following fields ID, title, abstract, keywords	Harvest Rate	0.27
Liu et al., [42]	Cosine similarity	MEMM and CRF	edge, full page text, meta description, URL text, and anchor text	Yahoo Directory, ODP	Precision and Harvest rate	0.34
Illiou et al., [51]	TF-IDF	SVM with RBF kernel	Full page text, anchor text and link context	600 samples collected from Yahoo directory	Precision, Recall and Harvest Rate	0.31

Authors	Features	Algorithms	Target Variables	Training Data set	Metrics	Harvest Rate Achieved after 5000 web page crawls
Kaur et al. [52]	TF-IDF	Decision Tree	Full page text	Tel-8 and common crawl datasets	-	-
Fu et al., [43]	Term Frequency	Entropy based classifier and GBS classifier	Full page text and Sentiment data	corporate social responsibility (CSR) and post-marketing drug surveillance (PMDS)	Precision, Recall, F1-score and Harvest Rate	0.34
Vural et al., [44]	21 features	Support Vector regression	Page URL, Anchor text and referring page	ClueWeb09-B	Accumulated sentiment score, and average page rank	
Zowalla et al., [53]	TF-IDF	SVM	Full page text	87,562 web pages were collected from various medical sources	Harvest Rate	0.31
Dhanith et al., [54]	A-SGNS based cosine	RNN	Full page text and anchor text	Manually collected 360,000 topic and web page pairs	Harvest Rate and Irrelevance Ratio	0.42

Ontology Learning Based Crawler

Ontology learning focused crawler is a combination of both the semantic technologies and the learning technologies. The semantic technologies compute the relevant concepts of the given topic using the thesauri based ontology. Then the term frequencies of the relevant concepts are computed and given as an input to the machine learning algorithms for prediction.

Manual assignment of concept weights leads to poor harvest rate. To gain better harvest rate and also to obtain the optimal concept weights [6] proposed an ontology learning based focused crawler using Artificial Neural Network. The relevant concepts for the given topic are computed based on the distance between them in the domain specific UMLS ontology. Then the term frequency of the relevant concepts in the web page is calculated and given as input to the ANN for training it. The trained ANN predicts the relevance of the web page.

The focused crawler downloads irrelevant web pages because of the ambiguous words present in the web page. These ambiguous words steer to the inefficient computation of the relevance of the web page. To gain word ambiguity [55] proposed a semi supervised ontology learning based approach by implementing SVM. The Resnik semantic similarity [29] and the probability based similarity between the topic and the web page were calculated. These calculated similarity values are then used as features to train the SVM. The trained SVM predicts the relevance of the web page.

To gain more efficiency and to identify the unique sense of the words [56] proposed an ontology learning based crawler by using Word Sense Disambiguation (WSD). The WSD is implemented using Domain Disambiguation Ontology (D^2O). With the help of WSD, domain keywords are identified and its term frequencies are calculated. These term frequencies are then given as an input to the Optimized Naive Bayes (ONB) classifier to predict the relevance of the web page. The ONB is a combination of SVM, Genetic Algorithm and NB. The genetic algorithm optimized SVM removes the outliers from the positive and negative training samples. These samples are then used to train the NB classifier.

Only text content based similarity computation is not sufficient, to retrieve relevant web pages. To resolve this insufficiency [57] proposed an ontology learning focused crawler by integrating both the text and multimedia content, to compute the relevance score of the web page. Li semantic similarity algorithm [58] and the polysemy semantic similarity algorithm is applied in WordNet to compute the content based similarity score. The multimedia based similarity computation is performed using the Convolution Neural Network (CNN) algorithm. Then the text and image based similarity scores are integrated to compute the relevance of the web page.

Table 4 depicts the comparison of ontology learning crawlers to various specifications proposed by different authors.

Table 4. Comparison of various ontology learning crawlers

Author	Features	Algorithm	Target Variables	Ontology Used	Dataset Used	Metrics	Harvest Rate achieved after 5000 web page crawls
Zheng et al., [6]	Term Frequency of relevant concepts	ANN	Full page text	UMLS	Manually collected dataset	Harvest Rate, performanc e-cost ratio	0.2044
Saleh et al., [56]	Term frequencies of the domain keywords	ONB, Genetic Algorithm and SVM	Full page text	D^2O	Web Data Commons dataset	Precision, Accuracy, Error, and Harvest Rate	0.37
Dong et al., [55]	Resnik semantic similarity and the probability based similarity	SVM	Full page text	WordNet	Manually collected dataset	Precision, Recall, Harmonic Mean, Harvest Rate	0.36
Capuano et al. [57]	Features based on ImageNet	Hybrid semantic similarity algorithm and CNN	Full page text and images in the web page	WordNet and ImageNet	Manually collected dataset	Harvest Rate	0.31
Hassan et al. [59]	TF-IDF	hierarchical multi label classificatio n	Full page text	Ontology described knowledgeb ase designed using stardog	45k economic-related news	Harvest Rate and Average Similarity	

Highlights and hindrances

The review results reveal legitimately that the TF-IDF weighted cosine similarity score applied by VSM based crawlers computes the relevance of the web page. The rare words are assigned more weightage by the TF-IDF weighting scheme compared to frequent words. The TF-IDF computes the weights hinged on the co-occurrence of the topic word in the target variables, which forms the most significant factor to ascertain the relevance of the web page. Consequent to the computation of semantically relevant web pages as irrelevant, the VSM based crawlers evinces low harvest rate and high irrelevance ratio. The evolutionary optimization algorithms assigned optimal weights to various target variables to overcome the vast deviations and inaccurate results produced when, manually assigning weights to the target variables to calculate the priority value of the URL, and later is also considered to be a costlier process.

These stumbling blocks of VSM crawlers route to the Learning Based crawlers. As said earlier, the VSM crawlers that require a separate evolutionary optimization criteria to obtain optimal weights to various target variables, is overcome by the learning crawler which by itself automatically assigns required optimal weights to compute the priority value. Requirement of huge amount of data to train the machine learning algorithms, and collection of these data is a drastic process. Any untrained term occurs in the course of crawling, inaccurate results are yielded. Every classifier for training in learning based crawler utilizes TF-IDF feature vectors, whose dimension increases and decreases with the count of words present in the web page respectively. This variability is subject to the poor performance of this crawler, and hence incongruous for most of the studies related with dynamic crawling of web pages. Similar to VSM crawler this also avoids semantic similarity and hence the harvest rate is low.

These shortcomings directed to the invention of Semantic Focused Crawlers. The negligence of calculating semantic similarity caused low harvest rate by VSM and learning crawlers, is overcome in this crawler. To be specific, the semantic similarity score of the web page is computed using the domain specific ontology, even for the incident occurrence of the topic words in the target variables of the web page. This

subsequent potentiality of the semantic focused crawler produces high harvest rate. The major pitfalls of this type of crawlers are: (i) semantic focused crawlers require domain specific ontology specifically designed by domain experts to compute the relevance score of the web page. Any human error in the ontology leads to irrelevant results. (ii) Semantic similarity computation using ontology is a time consuming process in a dynamic web environment and (iii) These crawlers entails the manual assignment of weights to the target variables for priority computation , due to which vast deviations are highlighted that produces inaccurate results.

To resolve these issues, researchers journeyed their work with Ontology Learning Based Crawlers. This crawler fabricated high harvest rate and better crawling, as it is an integration of semantic and learning crawlers. The optimal assignment of weight values to each target variables in priority computation is the major advantage of this type of crawler. The only flaw is the usage of domain specific ontology in a dynamic internet, to compute the relevant concepts of the given topic, is most expensive.

Future work

At the outset, the literature survey and the performance assessment done for the various classes of crawlers gives an understanding that there are enormous areas to be improved and their disadvantages need to be resolved. The dimensionality problem caused by the TF-IDF vectors of learning focused crawler is yet to be sorted out. Variety of word embedding techniques [60–62] can decipher the complications in the computation of semantic similarity using ontology based approaches. Recent topics concerned with sentence embedding-based deep learning technology [63,64] may also resolve these issues. Diversity of services, globally distributed service registries, and the vast information categories on the web opens the door to the poor indexing of web pages. These issues caused the ambiguity, ubiquity and the heterogeneity problems during the dynamic crawling process. These problems are yet to be resolved.

CONCLUSION

This paper established a survey on the existing focused web crawlers. The available focused web crawlers are classified based on their working nature into four main classes namely Classic focused web crawler, Semantic focused web crawler, Learning focused web crawler and ontology learning focused crawler. Each class is scrutinized over their common crawling features based on the metrics such as harvest rate and irrelevance ratio. Every input and output is surveyed correspondingly enhancing possible future evaluations.

REFERENCES

1. Internet Live Stats [Internet]. 2020. Available from: https://www.internetlivestats.com/total-number-of-websites/
2. Badawi M, Mohamed A, Hussein A, Gheith M. Maintaining the search engine freshness using mobile agent. Egypt Informatics J [Internet]. 2013;14(1):27–36. Available from: http://dx.doi.org/10.1016/j.eij.2012.11.001
3. Batsakis S, Petrakis EGM, Milios E. Improving the performance of focused web crawlers. Data Knowl Eng [Internet]. 2009;68(10):1001–13. Available from: http://dx.doi.org/10.1016/j.datak.2009.04.002
4. Chakrabarti S, Van den Berg M, Dom B. Focused crawling: a new approach to top-specific Web source discovery. Comput Networks. 1999;31(11–16):1623–40.
5. Dong H, Hussain FK. Self-adaptive semantic focused crawler for mining services information discovery. IEEE Trans Ind Informatics. 2014;10(2):1616–26.
6. Zheng HT, Kang BY, Kim HG. An ontology-based approach to learnable focused crawling. Inf Sci (Ny) [Internet]. 2008;178(23):4512–22. Available from: http://dx.doi.org/10.1016/j.ins.2008.07.030
7. Rungsawang A, Angkawattanawit N. Learnable topic-specific web crawler. J Netw Comput Appl. 2005;28(2):97–114.
8. Almpanidis G, Kotropoulos C, Pitas I. Combining text and link analysis for focused crawling-An application for vertical search engines. Inf Syst. 2007;32(6):886–908.
9. Chen Z, Ma J, Lei J, Yuan B, Lian L, Song L. A cross-language focused crawling algorithm based on multiple relevance prediction strategies. Comput Math with Appl [Internet]. 2009;57(6):1057–72. Available from: http://dx.doi.org/10.1016/j.camwa.2008.09.021
10. Liu WJ, Du YJ. A novel focused crawler based on cell-like membrane computing optimization algorithm. Neurocomputing [Internet]. 2014;123:266–80. Available from: http://dx.doi.org/10.1016/j.neucom.2013.06.039
11. Seyfi A, Patel A. A focused crawler combinatory link and content model based on T-Graph principles. Comput Stand Interfaces. 2016;43:1–11.
12. Seyfi A, Patel A, Celestino Júnior J. Empirical evaluation of the link and content-based focused Treasure-Crawler. Comput Stand Interfaces. 2016;44:54–62.
13. Farag MMG, Lee S, Fox EA. Focused crawler for events. Int J Digit Libr. 2018;19(1):3–19.

14. Mani Sekhar SR, Siddesh GM, Manvi SS, Srinivasa KG. Optimized focused Web Crawler with Natural Language Processing based relevance measure in bioinformatics web sources. Cybern Inf Technol. 2019;19(2):146–58.
15. Singh B, Kumar Gupta D, Mohan Singh R. Improved Architecture of Focused Crawler on the basis of Content and Link Analysis. Int J Mod Educ Comput Sci. 2017;9(11):33–40.
16. Geng Z, Shang D, Zhu Q, Wu Q, Han Y. Research on improved focused crawler and its application in food safety public opinion analysis. 2017 Chinese Autom Congr [Internet]. 2017;2847–52. Available from: http://ieeexplore.ieee.org/document/8243261/
17. Xu G, Jiang P, Ma C, Daneshmand M. A Focused Crawler Model Based on Mutation Improving Particle Swarm Optimization Algorithm. Proc - 2018 IEEE Int Conf Ind Internet, ICII 2018. 2018;(Icii):173–4.
18. Kumar M, Vig R. Learnable Focused Meta Crawling Through Web. Procedia Technol [Internet]. 2012;6(1994):606–11. Available from: http://dx.doi.org/10.1016/j.protcy.2012.10.073
19. Goyal N, Bhatia R, Kumar M. A genetic algorithm based focused web crawler for automatic webpage classification. IET Conf Publ. 2016;2016(CP739).
20. Zhao F, Zhou J, Nie C, Huang H, Jin H. SmartCrawler: A two-stage crawler for efficiently harvesting deep-web interfaces. IEEE Trans Serv Comput. 2016;9(4):608–20.
21. Park JR, Yang C, Tosaka Y, Ping Q, Mimouni H El. Developing an automatic crawling system for populating a digital repository of professional development resources: A pilot study. J Electron Resour Librariansh. 2016;28(2):63–72.
22. Chen X, Zhang X. HAWK: A focused crawler with content and link analysis. IEEE Int Conf E-bus Eng ICEBE'08 - Work AiR'08, EM2I'08, SOAIC'08, SOKM'08, BIMA'08, DKEEE'08. 2008;677–80.
23. Rawat S, Patil DR. Efficient focused crawling based on best first search. Proc 2013 3rd IEEE Int Adv Comput Conf IACC 2013. 2013;908–11.
24. Hati D, Sahoo B, Kumar A. Adaptive focused crawling based on link analysis. ICETC 2010 - 2010 2nd Int Conf Educ Technol Comput. 2010;4:455–60.
25. Mangaravite V, Tavares De Assis G, Ferreira AA. Improving the efficiency of a genre-aware approach to focused crawling based on link context. Proc - 2012 8th Lat Am Web Congr LA-WEB 2012. 2012;17–23.
26. Zhao W, Guan Z, Cao Z, Liu Z. Mining and harvesting high quality topical resources from the web. Chinese J Electron. 2016;25(1):48–57.
27. Gupta S, Duhan N, Bansal P. An Approach for Focused Crawler to Harvest Digital Academic Documents in Online Digital Libraries. Int J Inf Retr Res. 2019;9(3):23–47.
28. Dong H, Hussain FK. Focused crawling for automatic service discovery, annotation, and classification in industrial Digital Ecosystems. IEEE Trans Ind Electron. 2011;58(6):2106–16.
29. Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. 1995;1. Available from: http://arxiv.org/abs/cmp-lg/9511007
30. Bedi P, Thukral A, Banati H. Focused crawling of tagged web resources using ontology. Comput Electr Eng [Internet]. 2013;39(2):613–28. Available from: http://dx.doi.org/10.1016/j.compeleceng.2012.09.009
31. Du Y, Hai Y. Semantic ranking of web pages based on formal concept analysis. J Syst Softw [Internet]. 2013;86(1):187–97. Available from: http://dx.doi.org/10.1016/j.jss.2012.07.040
32. Du Y, Pen Q, Gao Z. Data & Knowledge Engineering A topic-speci fi c crawling strategy based on semantics similarity. Datak [Internet]. 2013;88:75–93. Available from: http://dx.doi.org/10.1016/j.datak.2013.09.003
33. Du Y, Liu W, Lv X, Peng G. An improved focused crawler based on Semantic Similarity Vector Space Model. Appl Soft Comput J [Internet]. 2015;36:392–407. Available from: http://dx.doi.org/10.1016/j.asoc.2015.07.026
34. Yang SY. OntoPortal: An ontology-supported portal architecture with linguistically enhanced and focused crawler technologies. Expert Syst Appl [Internet]. 2009;36(6):10148–57. Available from: http://dx.doi.org/10.1016/j.eswa.2009.01.004
35. Joe Dhanith PR, Surendiran B. An ontology learning based approach for focused web crawling using combined normalized pointwise mutual information and Resnik algorithm. Int J Comput Appl [Internet]. 2019;0(0):1–7. Available from: https://doi.org/1206212X.2019.1684023
36. Yuvarani M, Iyengar NCSN, Kannan A. LSCrawler: A framework for an enhanced focused web crawler based on link semantics. Proc - 2006 IEEE/WIC/ACM Int Conf Web Intell (WI 2006 Main Conf Proceedings), WI'06. 2007;794–7.
37. Jalilian O, Khotanlou H. A new fuzzy-based method to weigh the related concepts in semantic focused web crawlers. ICCRD2011 - 2011 3rd Int Conf Comput Res Dev. 2011;3:23–7.
38. Hegade P, Shilpa R, Aigal P, Pai S, Shejekar P. Crawler by Inference. 2020;108–12.
39. Diligenti M, Coetzee F, Lawrence S, Giles CL, Gori M. Focused crawling using context graphs. Proc 26th … [Internet]. 2000;527–34. Available from: http://www.vldb.org/conf/2000/P527.pdf
40. Liu H, Janssen J, Milios E. Using HMM to learn user browsing patterns for focused Web crawling. Data Knowl Eng. 2006;59(2):270–91.
41. Pant G, Srinivasan P. Link contexts in classifier-guided topical crawlers. IEEE Trans Knowl Data Eng. 2006;18(1):107–22.
42. Liu H, Milios E. Probabilistic models for focused web crawling. Comput Intell [Internet]. 2012;28(3):289–328. Available from: http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=shwart&index=an&req=16352297&lang=en

43. Fu T, Abbasi A, Zeng D, Chen H. Sentimental Spidering. ACM Trans Inf Syst [Internet]. 2012;30(4):1–30. Available from: http://dl.acm.org/citation.cfm?doid=2382438.2382443
44. Vural AG, Cambazoglu BB, Senkul P. Sentiment-focused web crawling. ACM Trans web. 2014;8(4):22.1-22.21.
45. Peng T, Liu L. Focused crawling enhanced by CBP-SLC. Knowledge-Based Syst [Internet]. 2013;51:15–26. Available from: http://dx.doi.org/10.1016/j.knosys.2013.06.008
46. Lu H, Zhan D, Zhou L, He D. An Improved Focused Crawler: Using Web Page Classification and Link Priority Evaluation. Math Probl Eng. 2016;2016.
47. Pawar N, Rajeswari K, Joshi A. Implementation of an Efficient web crawler to search medicinal plants and relevant diseases. Proc - 2nd Int Conf Comput Commun Control Autom ICCUBEA 2016. 2017;48:87–92.
48. Suebchua T, Manaskasemsak B, Rungsawang A, Yamana H. History-enhanced focused website segment crawler. Int Conf Inf Netw. 2018;2018-Janua:80–5.
49. Amalia A, Gunawan D, Najwan A, Meirina F. Focused crawler for the acquisition of health articles. Proc 2016 Int Conf Data Softw Eng ICoDSE 2016. 2017;(October).
50. Zheng Z, Qian D. An improved focused crawler based on text keyword extraction. Proc 2016 5th Int Conf Comput Sci Netw Technol ICCSNT 2016. 2017;386–90.
51. Iliou C, Kalpakis G, Tsikrika T, Vrochidis S, Kompatsiaris I. Hybrid focused crawling for homemade explosives discovery on surface and dark Web. Proc - 2016 11th Int Conf Availability, Reliab Secur ARES 2016. 2016;229–34.
52. Geetha G, Kaur S. Smart Focused Web Crawler for Hidden Web [Internet]. Vol. 40, Information and Communication Technology for Competitive Strategies, Lecture Notes in Networks and Systems. Springer Singapore; 2019. 419–427 p. Available from: http://link.springer.com/10.1007/978-981-13-0586-3_42
53. Zowalla R, Wetter T, Math D, Pfeifer D. Crawling the German Health Web : Exploratory Study and Graph Analysis Corresponding Author : 2020;22:1–22.
54. Dhanith PRJ, Surendiran B, Raja SP. A Word Embedding Based Approach for Focused Web Crawling Using the Recurrent Neural Network. Int J Interact Multimed Artif Intell. 2020;In Press(In Press):1.
55. Hussain HD and FK. SOF: a semi-supervised ontology-learning-based focused crawler. Concurr Comput Pract Exp. 2013;25(6):1755–70.
56. Saleh AI, Abulwafa AE, Al Rahmawy MF. A web page distillation strategy for efficient focused crawling based on optimized Naïve bayes (ONB) classifier. Appl Soft Comput J [Internet]. 2017;53:181–204. Available from: http://dx.doi.org/10.1016/j.asoc.2016.12.028
57. Capuano A, Rinaldi AM, Russo C. An ontology-driven multimedia focused crawler based on linked open data and deep learning techniques. Multimed Tools Appl. 2019;
58. Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng. 2003;15(4):871–82.
59. Hassan T, Cruz C, Bertaux A. Predictive and evolutive cross-referencing for web textual sources. Proc Comput Conf 2017. 2018;2018-Janua(July):1114–22.
60. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado JD. Distributed Representations ofWords and Phrases and their Compositionality. EMNLP 2016 - Conf Empir Methods Nat Lang Process Proc. 2016;1389–99.
61. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 1st Int Conf Learn Represent ICLR 2013 - Work Track Proc. 2013;1–12.
62. Jeffrey Pennington, Richard Socher CDM. GloVe: Global Vectors for Word Representation Jeffrey. Proc 2014 Conf Empir Methods Nat Lang Process. 2017;1532–1543.
63. Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, et al. Deep Sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Trans Audio Speech Lang Process. 2016;24(4):694–707.
64. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst. 2014;4(January):3104–12.