Fine Grained Approach for Domain Specific Seed URL Extraction

Lalit Mohan S
IIIT Hyderabad
lalit.mohan@research.iiit.ac.in

Sourav Sarangi
IIIT Hyderabad
sourav.sarangi@students.iiit.ac.in

Y R Reddy
IIIT Hyderabad
raghu.reddy@iiit.ac.in

Vasudeva Varma
IIIT Hyderabad
vv@iiit.ac.in

Abstract

Domain Specific Search Engines are expected to provide relevant search results. Availability of enormous number of URLs across subdomains improves relevance of domain specific search engines. The current methods for seed URLs can be systematic ensuring representation of subdomains. We propose a fine grained approach for automatic extraction of seed URLs at subdomain level using Wikipedia and Twitter as repositories. A ***SeedRel*** *metric and a Diversity Index for seed URL relevance are proposed to measure subdomain coverage. We implemented our approach for 'Security - Information and Cyber' domain and identified 34,007 Seed URLs and 400,726 URLs across subdomains. The measured Diversity index value of 2.10 conforms that all subdomains are represented, hence, a relevant 'Security Search Engine' can be built. Our approach also extracted more URLs (seed and child) as compared to existing approaches for URL extraction.*

1. Introduction

There are Billion websites with more than 3+ Billion internet users accessing the content directly or through search engines. Google, Bing, Baidu, etc. continue to be popular generic search engines with some estimates stating that they indexed few exabytes of data and consumed million+ computing hours for crawling and indexing. Despite powerful ranking algorithms and word disambiguation techniques, the relevance of search results continues to be an area of concern for generic search engines. For example, a leading generic search engine provided search result containing '*real estate properties*' for a query on '*HTTP properties*'. To reduce ambiguity and the amount of content to process for relevant search, Domain Specific Search Engine (DSSE) also known as Vertical Search Engines are gaining importance [1]. It is well established fact that DSSEs have better Precision due to limited scope and focused corpus resulting in less load on network, storage and processor. The word 'domain' for a search engine can be an area of interest or territory such as 'Medicine', 'Finance', 'Canada', etc. or an internet domains (gTLD) such as '.com', '.org', etc. Yahoo Finance, Google Scholar, etc. are some of the popular DSSEs. Some of the challenges impeding extensive adoption of DSSEs are (i) Selection of seed URL is manual and requires thorough investigation for determining seed URL usefulness (ii) Additional processing is required for removal of unrelated content that is not specific to the domain (iii) All subdomains of a domain may not be represented in the extracted content. To elaborate on subdomain importance, consider that there are 'n' subdomains (*SD*s) in a domain (*D*) such that

$$D = \{SD_1, SD_2, \ldots..SD_k, \ldots SD_n\}$$

if, only part of subdomains content is extracted, i.e.

$$\delta D' = \{SD'_1, SD'_2, \ldots..SD'_k, \ldots SD'_n\}$$

or, not all subdomains of the domain are represented

$$\nabla D' = \{SD_1, SD_2, \ldots..SD_k\}$$

This under representation ($\delta D' and\ \nabla D'$) of subdomains reduces the importance of DSSE. To further elaborate on importance of subdomain, consider a need to build a search engine for '*Education*' domain. If only part of curriculum of a course content is extracted or only few courses content is extracted, the DSSE relevance would be reduced. Hence, for subdomains content representation, availability of related seed URLs and the subdomain content containing child URLs is required apart from efficient indexing and ranking approaches. The research of Priyatam [2], Aggarwal [3], Mukherjea [4] and McCallum [5] also suggested the need for topic (subdomain) identification and seed URL representation at subdomain level for increased relevance of domain specific search engine.

Our proposed fine grained approach for URL extraction is based on subdomains (sub-topics/sub-groups) of a domain. In our approach, subdomains are identified in a systematic way with seed URLs and the underlying child URLs. For measuring seed URL

URI: http://hdl.handle.net/10125/50111
ISBN: 978-0-9981331-1-9

relevance of a subdomain, a *SeedRel* metric is calculated based on semantic similarity of child URLs metadata with the subdomain. To validate extracted seed URLs represent subdomains, we extended Shannon Diversity Index [6] for measuring URL diversity in the domain. The seed URL extraction approach and the measurement of *SeedRel* and *Diversity Index* are performed on 'Security - Information and Cyber' domain. With growing adoption of information technology across the world, the associated security concerns have also increased. Availability of prevalent domain specific - 'Security Search Engine' could improve the awareness and ease the keyword search in security domain, a research motivation for extracting URLs in 'Security' domain. Our approach contains an automated and continuous way for extracting seed URLs using Artificial Bee Colony (ABC) algorithm. ABC algorithm [7] [11] is known for its exploration and exploitation capability. The agents of ABC algorithm are used for extraction of seed URL, content/metadata from URLs and classification of URL content into different subdomains.

The focus of our research work is: (i) an automated seed URL extraction at subdomain level; (ii) extend a social sciences Diversity Index for measuring subdomain seed URL representation; and (iii) a new *SeedRel* metric to measure relevance of each seed URL for a subdomain. In the following sections of the paper, 2 - existing literature on seed URL extraction, 3 - approach for subdomain seed URL extraction, 4 - results and analysis to validate our approach on importance of subdomain seed URLs, comparison of our results with other approaches and finally conclusion is discussed.

2. Literature Survey

In this section, we reviewed existing work on identification of subdomain, approaches for seed URLs extraction, and scoring mechanisms on relevance of seed URLs. To identify subdomains in a systematic way, we studied research and industrial contributions on domain and subdomain classification. Some of the approaches for identification of subdomains are

- Standard Industrial Classification (SIC [8]) is widely used for classifying industries into industry sectors, major groups in a sector and divisions in a group. This is used primarily for business analysis and sharing of data across agencies. SIC has 9 divisions that are categorized into 83 major groups and each of the major group is further divided into an Industry group. For example, 'Computer and Office Equipment' industry group are classified into 80 groups. Though this is widely used and exhaustive, some of the emerging topics may not be represented.
- Formal Concept Analysis - FCA [9] is a principled/mathematical way of identifying concept hierarchy from a collection of objects based on their attributes and relations. FCA is used for text and data mining, knowledge management, etc. FCA method is rigorous but requires substantial manual effort for identifying concept hierarchy.
- Domain Ontologies are prevalent for most of the domains. Ontologies developed by domain experts are not hierarchical, contain subdomains of a domain [10] for classification. Ontologies are used primarily for sharing information, analyzing and reasoning on domain knowledge. As Ontologies are not hierarchical, identification of concepts and relationships of a subdomain would be a manual effort. Du et al. [11] used ontology created by domain experts for obtaining seed URLs. However, the work does not establish coverage across subdomains. Similar to ontology graphs, Zheng et al. [12] proposed a graph based framework for seed selection and compares with algorithms such as *MaxWeight, Maxout, MaxSCC* (strongly connected components), etc. However, the work does not discuss on subdomain coverage.
- Clustering is an exploratory data analysis technique for grouping data. Clustering techniques are used for identification of subdomains, groups, clusters based on distance/similarity measures with no prior training data, this technique can also be used for identifying newer trends or themes. Bergmark [13] and Weiss et al. [14] used distance/similarity measures for identifying clusters/subdomains/ regions. While Bergmark's approach is based on available webpage content, it does not validate representation of all subdomains.
- Topic modeling [15] discovers semantic structures in a textual body. The method automatically organizes and summarizes large web archives into sub-topics. Blei et al. [16] researched on using Latent Dirichlet Allocation with three-level hierarchical Bayesian model for identification of topics. However, like clustering techniques this require substantial relevant corpus for identification of topics. A recent work on URL extraction by Cheng et al. [17] for identification of topics from short text (twitter postings or anchor text) is encouraging but requires longer duration to crawl and get corpus representing all topics/ subdomains.
- Standards like ISO 27001: 2013 [18], contain Information Security groups and can be used for representing subdomains of security. These groups can be easily identified from the manuals by basic text parsing techniques. However, there will not be ISO standard for all domains.

Some of the other work on seed URL extraction and topic mapping are (i) Pappas et al. [19] identified topics using dynamic seed URLs and evaluated topic relevance. In this work, the identification of seed URLs is manual and does not confirm representation of all sub-topics of a topic. Prasath et al. [20] suggested the use of generic search engines for getting seed URLs. Prasath's approach would be difficult to reproduce as search results in most of the generic search engines are contextual - person, location, history, etc. The work of Aggarwal et al. [21] proposed an approach for topical discovery based on webpage content, URL tokens, linkage structure, etc. but did not focus on coverage of subdomains. In a more recent work, Sey et al. [22] performed an empirical evaluation to identify unvisited links of a page and classify them to more specific topics. Though this is a practical view to the world, the implementation is manual and requires site administrator's intervention for changing the links in a page. Almpanidis et al. [23] uses latent semantic analysis on URL and webpage content for identification of topics, this work also does not provide an insight on subdomains coverage. As available literature on seed URL extraction is sparse, we looked at the literature on focused crawlers for domains. Focused crawlers use Anchor text [24], webpage text [5], Page properties, Ontology [25], Graph [26], Key word base[27], Clickstream [28], Reinforcement Learning, etc. for extracting domain corpus. However, these approaches consider that seed URLs are already extracted either manually or automatically. The extracted seed URLs are used for crawling the internet for domain content.

To measure relevance of extracted seed URL, we extended the work of Chakrabarti et al. [24] on harvest rate, Garcia-molina et al. [29] for scoring seed URLs based on semantic similarity and Pant et al. [30] research on links/URLs attribution in our *SeedRel* metric for scoring seed URLs. The harvest rate of a seed URL is an extension of information retrieval metric - 'Precision' to measure relevant URLs among extracted URLs. While Precision and Recall are standard Information Retrieval metrics on relevance along with other Information Gain measures [31], they do not provide insight on URL/Content availability across subdomains. There is a possibility that some subdomains have more representation as compared to other subdomains. We studied Shannon Diversity Index [9] (similar to well published Shannon Entropy measure on randomness) and other diversity measures to measure subdomain representation as the index value quantifies the uncertainty (entropy or degree of surprise) associated with a prediction in a community.

3. Subdomain URL Extraction

Our fine grained approach of a domain identifies subdomain keyword phrases, extracts seed URLs from URL repositories, classifies seed URLs and its related child URLs into subdomains and measures seed URLs relevance. To identify all subdomains of a domain, automated approaches such as Topic Modeling and Clustering need large corpus and the intent of this research is to identify the URLs that contain the corpus. While Ontologies are domain related and created by domain experts, the identification of subdomains is not obvious in an OWL/RDF file [32]. We adopt ISO 27001:2013 [18] classification for 'Security' subdomain URL extraction, the standard is widely accepted by industry, government and research community. SIC or any other classifications may also be used for subdomain representation.

In Security, there are 14 distinct subdomains/groups (Asset Management, Access Management, Business Continuity, Communications, Compliance, Cryptography, Human Resources, Incident Management, Operations, Organization, Physical and Environmental, Policies, Supplier Relationship, System Acquisition, Development and Maintenance). These 14 subdomains have 35 control objectives and are further sub-divided into 114 controls. As there is an overlap between Information Security and Cyber Security [33], we included 'Cyber Security' to the list of subdomains for seed URL extraction of security domain. These new subdomains are based on NIST Cyber Security Framework [34] as there is no ISO standard for Cyber Security. We also included recent updates to ISO 27000 series - ISO 27017 and 27018 for Cloud Computing. With these additions, there are 17 subdomains for obtaining security related seed URLs.

3.1. URL Extraction Process

To obtain subdomain seed URLs, we extracted keyword phrases of subdomains from ISO and NIST documents after POS tagging and removal of stop words. Though Wikipedia is a rich resource of information with adequate moderation and freely available, we did not extract ‘Security’ keyword phrases to avoid bias. Also, Wikipedia is used to identify seed URLs from it. The Figure 1 shows our fine grained approach for extraction of seed URLs and its related child URLs of subdomains.

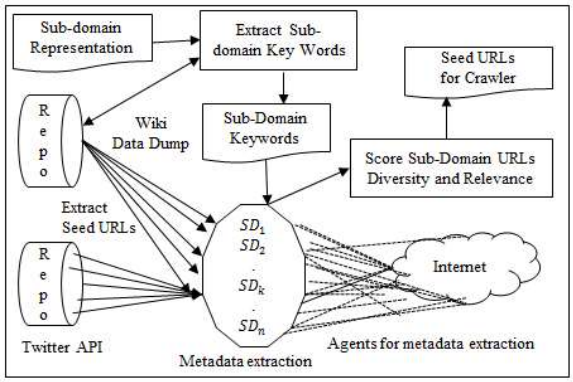

Figure 1. Fine Grained Approach for Extraction

The modified ABC algorithm for URL extraction is shown in Algorithm 1 [7] and its implementation is available at Github[1]. With the implementation of the ABC algorithm, seed URLs extraction is continuous and automated. This makes our approach scalable as other seed URL extraction approaches have static list of seed URLs [2], [20]. The algorithm has 3 types of agents/threads - Scouts, Onlookers and Employee. The scouts in the algorithm perform the initial exploration process of extracting the source of seed URL. Onlookers and employed agents/threads extract the webpage content of the URLs for relevance, thus performing an exploitation action. This approach also provides separation of duties and provides combinatorial and functional optimization [11].

```
Data: Domain Name, Subdomains[ ]
Result: Seed and Child URLs of a Domain
//Procedure Begins
Initialize All Variables
Initialize subdomain.maxCycle;
Initialize subdomain.runtime;
Set ChkSim = 0.75;
// Loop for identifying subdomain keyphrases
Loop Subdomains of Domain
{
  subdomain.KeyPhrases[] = Extract
            from WikiDump;
  subdomain.runtime =
  size of subdomain.KeyPhrases[ ];
  Loop KeyPhrases of subdomain.run ;
  {
    subdomain.initialize();
  }
  subdomain.maxCycle
            = subdomain.maxCycle +
            subdomain.runtime;
}
// Extraction of URLs with dynamic thread count
Loop All URLs of a Subdomain
 {
   subdomain.SendURLThreads();
   subdomain.CalculateSimilarity(ChkSim);
   subdomain.SendOnlookerThreads();
   subdomain.MemBestSeed();
  subdomain.SendScoutTheads();
}
```

Algorithm 1. URL Extraction

A brief description of ABC algorithm methods is as follows -

- **subdomain.KeyPhrases** : Contains the Keyword phrases of 'Security' subdomains.
- **subdomain.initialize** : Number of subdomains initialized. This task performed by the initial scout thread of a subdomain, supports exploration.
- **subdomain.maxCycle** : Total number of threads based on the Keyword phrases of all subdomains.
- **subdomain.MemBestSeed** : The best seed URL is memorized by the thread.
- **subdomain.SendURLThreads** : Thread generates a random score that is a mutant of the original solution. If the new URL contains higher similarity value to the subdomain, the earlier URL is replaced with the new URL.
- **subdomain.CalculateSimilarity** : Onlooker thread chooses a URL based on the similarity value associated with the seed URL. Based on accepted level of similarity, Onlooker threads are initiated. We used a random value >= 0.75 (scale 0 – 1, 1 – exact match) for validating the similarity of text.
- **subdomain.SendOnlookerThreads** : IP addresses of URLs are used for providing the direction for URL content extraction, clusters are formed based on sorted IP Address. Number of clusters determines the numbers of threads that should be running on ongoing basis.
- **subdomain.SendScoutTheads** : Trial parameter is defined for those solutions that are exhausted and not changing. This function determines those exhausted seed URLs and abandons them.

Wiki Python API[2] and Twitter Streaming API[3] are used for seed URL extraction. Two (Wikipedia and Twitter) repositories are used to avoid bias on subdomain representation. Wikipedia with

[1] https://github.com/souravsarangi/SeedURLWork

[2] https://pypi.python.org/pypi/wikipedia/

[3] https://dev.twitter.com/streaming/overview/

collaboration from communities is getting enriched and has become a rich source of information [35] . Twitter is another URL repository [2] with short text in the message as anchor text or meta data. Some of the other repositories such as DMOZ and Freebase are not considered as their support stopped and may not have current URLs. Metadata (AnchorText and meta description) of URL is extracted using Wget method instead of entire page/site content to reduce processing time for obtaining similarity value. We use Phrase2Vec[4] trained on large corpus – Google News Archives to provide semantic similarity of metadata phrases with Keyword phrases of subdomain. Based on Phrase2Vec similarity value (ranging from 0 - 1), URLs are assigned to subdomains.

3.2. Scoring

The extracted seed URLs and child URLs are scored for subdomain representation based on semantic similarity. After the classifications of URLs to subdomain, seed URL relevance (*SeedRel)* is measured. The metric for a subdomain extends the Harvest Rate metric suggested by Chakrabarti et al. [24]. Our metric contains a loss value (weeds/noise) to reduce seed URL relevance if it contains URLs that do not represent subdomain.

$$SeedRel = (1 - \gamma - \epsilon) * \frac{\sum_{i=0}^{k} \alpha_i}{\sum_{i=0}^{n} \alpha_i}$$

γ - Loss value, i.e., the ratio of URLs in a seed URL that have less than required similarity. If the number of URLs in a seed URL is 10 and URLs with less than 0.75 similarity is 3, the value would be 0.3. URLs that are not accessible because of HTTP error accumulate to the loss value.

ϵ – Is another loss value for duplicate domains in a Seed URL. This loss value is included as crawler that uses seed URL would crawl all pages in a domain and having URLs of a single domain does not enhance the relevance of seed URL.

α - Similarity value obtained using Phrase2Vec for each unique URL, the value ranges from 0 to 1.

i - URL that has similarity value greater or equal to a desired value.

k - For all URLs present in a seed URL of a subdomain.

n - For all URLs present in all seed URLs of a subdomain.

If there are URLs with same similarity value for multiple subdomains, URL is classified as belonging to all matching subdomains. If a URL in a seed URL is also another seed URL (identified by ABC algorithm function), the *SeedRel* score of initial seed URL is the summation of child seed URLs *SeedRel* score along with its own score. Hence, a seed URL containing another seed URLs would be scored higher.

The established Shannon Diversity index and the related values have experimental data to evaluate diversity in a community. When all species in a sample/population of interest are equally common, the Shannon index takes the value $ln(p_i)$. The more unequal the abundances of the species, the higher the weighted geometric mean of the p_i values, and the corresponding Shannon Index would be lesser. If all abundance is concentrated to one type, and the other types are very rare (even if there are many of them), the index value approaches zero. When there is only one species, Shannon index equals zero (as there is no uncertainty in predicting the type of the next randomly chosen entity).

$$SDDiversity = -\sum[p_i * \ln(p_i)]$$

p_i - Number of extracted URLs of a subdomain divided by total number of extracted URLs. The *SDDiversity* value is generally between 1 - 3.5. If the value is closer to 3.5, it suggests that subdomains are well represented.

4. Results and Analysis

Our experiment contains 11 subdomains of 'Information and Cyber' security domains identified from ISO 27001 and NIST Cyber Security Framework, these subdomains also represent the enterprise security architecture. The subdomain names are Access (includes Cryptography and Access), Management (includes Business Continuity, Communications, Compliance, Human Resources, Organization, Policies, Supplier Relationship), Operations Control (includes Incident Management, Operations), Network, Application, Endpoint, Hardware, Cloud Computing, Cyber and Attacks. The security architectural view for asset protection in an organization is shown in Figure 2, this view maps to the subdomain list.

With eleven subdomains of 'Security' and two (Wikipedia and Twitter) seed URL repositories, our algorithm extracted[5] 45,319 seed URLs and 1,029,466 child URLs using 8 GB Quadcore machine. Seed URLs from Twitter are extracted for a duration of 120 hours spread over 30 days to avoid topic domination. As part of the approach, URL duplicates at IP level are removed as crawlers would typically crawl through all accessible

[4] https://radimrehurek.com/gensim/models/phrases.html

[5] http://tinyurl.com/SubdomainSeedURL

links for an IP address. The representation of seed URLs and URLs across subdomains is represented in Table 1.

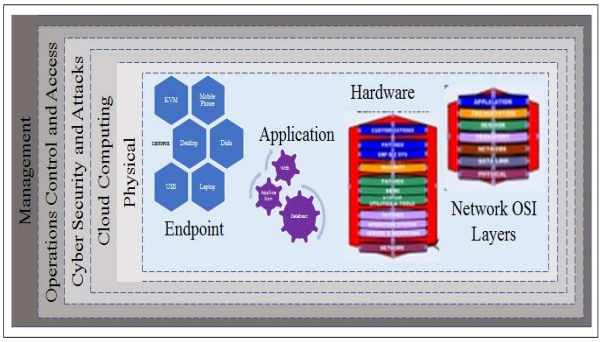

Figure 2. Security Architecture

Private Network'. We prepared an acronym mapper [36] based on existing security dictionaries (NIST, ENASE, etc.) for obtaining metadata similarity. After we implemented the mapper, the similarity values increased by 5% for 'Network' subdomain, this subdomain has maximum number of acronyms.

To validate that our subdomain extraction process is fine grained and has more URLs extracted, we performed Cluster Analysis using Lingo3G [14]. We extracted content of 400 URLs on 'Information and Cyber Security' keyword phrase using Bing search engine and performed Cluster Analysis. Results in Figure 3 shows that some of the key areas such as 'Cryptography', 'Cloud Computing' and 'Access Controls' are not represented and most of the clusters represent 'Security Management'. This states that URL extraction using cluster analysis would not represent all the subdomains of a domain.

Table 1. Results of URL Extraction at Subdomain level using ABC Algorithm

Subdomain	Seed URLs		Child URLs		URLs/Seed		Unique URLs with Similarity		
	All	Unique	All	Unique	All	Unique	< 0.5	0.5 - 0.75	> 0.75
Access	646	638	13,411	8,240	21	13	1,116	163	6,961
Application	2,622	2,417	52,211	27,706	20	11	9,990	346	17,370
Attacks	13,235	9,381	2,48,432	78,719	19	8	26,554	5,856	46,309
Cloud Computing	1,820	1,526	51,087	18,519	28	12	13,693	2,472	2,354
Cyber	2,468	1,884	46,988	14,253	19	8	9,644	1,478	3,131
Endpoint	4,366	3,825	1,25,955	63,101	29	16	44,354	1,179	17,568
Hardware	417	409	8,978	5,389	22	13	1,631	300	3,458
Management	3,979	2,327	85,605	30,453	22	13	15,550	1,108	13,795
Network	8,140	6,159	2,19,156	88,256	27	14	41,070	22,630	24,556
Operations Control	1,327	907	27,840	11,716	21	13	3,659	312	7,745
Physical	6,299	4,534	1,49,803	54,374	24	12	19,359	20,650	14,365
Total	**45,319**	**34,007**	**10,29,466**	**4,00,726**	**23**	**12**	**1,86,620**	**56,494**	**1,57,612**

Following are some of the observations from results -

- 'Attacks' and 'Network' subdomain have more seed URLs whereas 'Hardware' has lower number of seed URLs and URLs. Wiki and Twitter have more URLs for 'Attacks' and 'Network' subdomains. This provides an insight that most of the attacks are related to Network - DOS/DDOS, IP Spoofing, etc. 'Network' and 'Cloud Computing' subdomains have more child URLs per seed URL.
- In the URL extraction process, we could not crawl 77,238 URLs due to 'HTTP' errors and obstruction due to robots.txt. Implementing a sleep mode and other crawler functions such as politeness could have improved the count of extracted URLs.
- We realized some of the phrases and metadata had acronyms related to security like VPN for 'Virtual

Figure 3. Cluster Analysis for Subdomains

We validated results of our approach for subdomain representation vis-a-vis LDA Topic Modeling. Topic Modeling was done on the responses extracted from 'Security StackExchange'. 50,001 Security related responses with more than 200 characters were extracted from StackExchange. LDA technique[6] with collapsed Gibbs sampling is used for identifying topics from StackExchange responses. The results are shown in LDA[7] and it can be observed that only 4 subdomains are represented through this process. Hence, it can be stated that extraction of seed URLs and related child URLs based on these topics is under-represented. We also compared our results with a recent work [2] on seed URLs extraction at domain level and checked representation of subdomains. We executed Twitter Streaming API for a period of 30 days to extract URLs and reduce topic domination. Majority of the subdomains are represented in the Twitter extract but with sparse subdomain representation as shown in Table 2. The count of seed URLs extracted is very less in number as compared to our work though source repositories are same.

Table 2. Subdomain representation

#	Subdomain	Wikipedia	Twitter
1	Access	0	191
2	Application	3	53
3	Attacks	6	191
4	Cloud Computing	6	372
5	Cyber	0	76
6	Endpoint	1	21
7	Hardware	4	59
8	Management	3	389
9	Network	3	202
10	Operations Control	20	3046
11	Physical	29	2339

Like Shannon index, Simpson index is another widely used measure for diversity measurement and has good discriminant ability [6]. Hence, we also measured Simpson Index as it is relatively less sensitive to sample size variation. For Seed URLs and child URL with an arbitrary > 0.75 (the value ranges from 0 - no similarity to 1 - highest similarity or complete match). The Diversity Index and Evenness are high for URLs extracted at subdomain level as against URLs that are extracted at domain level as shown in Table 4. This indicates that count of extracted URLs is high by fine grained (subdomain) approach.

Table 4. Diversity Score

URL Category	Diversity and Evenness Measures			
	Shannon	Evenness	Simpson	Dominance
Seed URL	2.07	0.86	0.16	0.85
URL	2.25	0.94	0.16	0.85
< 0.5	2.05	0.85	0.15	0.85
0.5 -7.5	1.46	0.61	0.30	0.69
> 0.75	2.08	0.87	0.16	0.84
Domain Level	1.50	0.62	0.31	0.68

5. Conclusions and Future Work

Our approach for extraction of seed URLs provides coverage across subdomains irrespective of a domain. Except for identification of repositories (Wiki and Twitter), the entire process for extracting seed URLs and Child URLs, scoring for assessment of seed URL relevance (*SeedRel)* is automated. The given similarity value of 0.75 can be system generated using optimization techniques or empirical studies. The approach demonstrated extraction of 34,007 unique seed URLs representing security subdomains. These extracted URLs could be used for building 'Security' domain specific search engine that could be used by industry and academic security interest groups. The crawler for building search engine could use seed URLs with higher *SeedRel* for content enrichment. We plan to increase the duration of Twitter search and validation of content similarity based on in-domain webpages to increase the extracted URL count. The quantum of seed and child URLs can be enhanced by incorporating search on non-English, deep and dark web. We also plan to enhance our modified ABC algorithm for other features of a crawler such as politeness, re-visit, duplicate removal, etc. and make it more scalable. Our approach could be extended to other domains such as Agriculture, Banking, etc.

6. References

[1] K. Wöber, "Domain specific search engines", *Travel Destination Recommendation Systems : Behavioral Foundations and Applications*, CAB International Cambridge, MA, pp. 205–226, 2006.

[2] P. N. Priyatam, A. Dubey, K. Perumal, S. Praneeth, D. Kakadia and V. Varma, "Seed Selection for Domain-Specific Search", International Conference on *World Wide Web*, ACM, 2014, pp. 923–928.

[3] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu, "On the design of a learning crawler for topical resource discovery", *ACM Transactions on Information Systems*, vol. 19, no. 3, 2001, pp. 286–309.

[4] S. Mukherjea and S. Jose, "Organizing Topic-Specific Web Information", Proceedings of the Eleventh ACM on Hypertext and Hypermedia, 2000, pp. 133-141.

[6] https://pypi.python.org/pypi/lda

[7] http://tinyurl.com/FineGrainedLDA

[5] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, “A Machine Learning Approach to Building Domain-Specific Search Engines”, *International Joint Conference Artificial Intelligence*, vol. 16, 1999, pp. 662–667.

[6] I. F. Spellerberg and P. J. Fedor, “A Tribute to Claude Shannon and a plea for more rigorous use of species richness and diversity and the ‘Shannon--Wiener’ Index”, *Global Ecology Biogeography*, vol. 12, no. 3, Wiley Online Library, 2003, pp. 177–179.

[7] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, “A Comprehensive Survey: Artificial Bee Colony (ABC) Algorithm and Applications”, *Artificial Intelligence Review*, vol. 42, no. 1, Springer, 2014, pp. 21–57.

[8] "SIC Division Structure", URL - " http://tinyurl.com/SICSD", Sep 22nd 2017 .

[9] R. Wille, “Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies”, *Formal Concept Analysis*, Springer, 2005, pp. 1–33.

[10] N. F. Noy, D. L. McGuinness, and others, “Ontology development 101: A Guide to Creating your First Ontology”, Stanford knowledge systems Laboratory Technical report KSL-01-05 and Stanford medical informatics Technical report SMI-2001-0880, Stanford, CA, 2001.

[11] Y. Du, Y. Hai, C. Xie, and X. Wang, “An approach for selecting seed URLs of focused crawler based on user-interest ontology”, *Applied Soft Computing Journal*, vol. 14, no. 7224, 2014, pp. 663–676.

[12] S. Zheng, P. Dmitriev, C. L. Giles, H. I. Storage, and R. Information, “Graph-based Seed Selection for Web-scale Crawlers”, 2009, pp. 1967–1970.

[13] D. Bergmark, “Collection Synthesis”, 2nd ACM/IEEE-CS joint Conference on Digital Libraries, ACM, 2002, pp. 253–262.

[14] S. Osinski, J. Stefanowski, and D. Weiss, “Lingo : Search Results Clustering Algorithm Based on Singular Value Decomposition”, Intelligent Information Processing and Web mining, Springer, 2004, pp. 359–368.

[15] T. Hofmann, “Probabilistic Latent Semantic Indexing”, *Proceedings of 22nd Annual International ACM SIGIR Conference*, 1999,pp. 50–57.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *CrossRef List. Deleted DOIs*, vol. 1, no. 1, 2000, pp. 993–1022.

[17] X. Cheng, X. Yan, Y. Lan, and J. Guo, “BTM: Topic modeling over short texts,” *IEEE Transactions on Knowledge Data Engineering*, vol. 26, no. 12, 2014, pp. 2928–2941.

[18] "ISO 27001 Series Security Standards", URL - "https://www.iso.org/isoiec-27001-information-security.html", Sep 3rd 2017.

[19] N. Pappas, G. Katsimpras, and E. Stamatatos, “An Agent-based Focused Crawling Framework for Topic and Genre-related Web Document Discovery”, *International Conference on Tools with Artificial Intelligence*, vol. 1, 2012, pp. 508–515.

[20] R. Prasath and P. Öztürk, “Finding Potential Seeds through Rank Aggregation of Web Searches”, International Conference on Pattern Recognition and Machine Intelligence, Springer, 2011, pp. 227–234.

[21] I. B. M. T. J. Watson, “On the Design of a Learning Crawler for Topical Resource Discovery”, vol. 19, no. 3, 2001, pp. 286–309.

[22] A. Sey, A. Patel, and J. Celestino, “Empirical Evaluation of the Link and Content-based Focused Treasure-Crawler”, vol. 44, 2016, pp. 54–62.

[23] G. Almpanidis, C. Ã. Kotropoulos, and I. Pitas, “Combining Text and Link Analysis for Focused Crawling — An application for Vertical Search Engines”, vol. 32, 2007, pp. 886–908.

[24] S. Chakrabarti, K. Punera and M. Subramanyam, “Accelerated Focused Crawling through Online Relevance Feedback”, International Conference on World Wide Web, ACM, 2002, pp. 148–159.

[25] P. Bedi, A. Thukral, and H. Banati, “Focused Crawling of Tagged Web Resources using Ontology” , *Computers and Electrical Engineering*, vol. 39, no. 2, Elsevier, 2013, pp. 613–628.

[26] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, “Focused Crawling Using Context Graphs”, 2000, pp. 527–534.

[27] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-manning, “Domain-Specific Keyphrase Extraction”, *International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc., pp. 668--673,1999.

[28] F. Ahmadi-abkenari and A. Selamat, “An Architecture for a Focused trend Parallel Web Crawler with the Application of Clickstream Analysis”, *Information Sciences*, vol. 184, no. 1,Elsevier, 2012, pp. 266–281.

[29] H. Garcia-molina and L. Page, “Efficient crawling through URL ordering,” vol. 30, no. 98, 1998.

[30] G. Pant, “Exploration versus Exploitation in Topic Driven Crawlers” , pp. 1–20, 2002.

[31] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, “Novelty and diversity in information retrieval evaluation” , *Proceedings of 31st Annual International ACM SIGIR Conference*, 2008, p. 659.

[32] A. Ekelhart, S. Fenz, M. D. Klemen, and E. R. Weippl, “Security Ontology : Simulating Threats to Corporate Assets” , *Proceedings of 2nd International Conference on Information Systems Security*, 2006, pp. 249–259.

[33] R. Reid and J. Van Niekerk, “From information security to cyber security cultures”, *Information Security Proceedings ISSA Conference*, vol. 38, *South Africa,* 2014, pp. 97–102.

[34] “NIST Cyber Security Framework”, URL - https://www.nist.gov/cyberframework, Sep 3rd 2017.

[35] A. Barbaresi, “Finding viable seed URLs for web corpora : a scouting approach and comparative study of available sources”, 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 1–8.

[36] “Security Acronyms”, URL - http://tinyurl.com/SecArconym, Sep 22nd 2017 .