SIREN - Security Information Retrieval and Extraction eNgine

by

LALIT Mohan Mohan, Neeraj Mathur, Shriyansh Agrawal, Y.Raghu Babu Reddy

in

The 40th European Conference on Information Retrieval

Report No: IIIT/TR/2018/-1

Centre for Software Engineering Research Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2018

SIREN - Security Information Retrieval and Extraction eNgine

Lalit Mohan S, Neeraj Mathur, Shriyansh Agrawal and Y R Reddy

International Institute of Information Technology,
Gachibowli,Hyderabad, India
lalit.mohan@research.iiit.ac.in,
neeraj.mathur@research.iiit.ac.in,
shriyansh.agrawal@research.iiit.ac.in,
raghu.reddy@iiit.ac.in

Abstract. Domain specific search engines (DSSE) are gaining popularity because of better search relevance and domain specificity. The growth of IT and internet led to increase of cyber attacks, however, lack of DSSE for Security is making users refer multiple sites for security information. We demonstrate SIREN, a search engine for 'Information and Cyber Security' with subdomain coverage, classification and site credibility for ranking search results. As part of our demonstration, we also automated identification of seed URLs (34,007) and related child URLs (400,726) of the security domain using Artificial Bee Colony algorithm. We also evaluated functional and non-functional parameters of available open source software stack that can be used for building other DSSEs.

Keywords: Security, Domain Specific Search Engine, Seed URL, Credibility

1 Introduction

There are 3.76+ Billion internet users accessing 1.27+ Billion websites and this is only expected to increase with technology innovation and affordable computing devices. This large number of websites are indexed and accessed using search engines. However, the relevance of search results is an area of concern due to ambiguity, bias, content filtering and other issues. For improved search relevance, Domain Specific Search Engines (DSSE) are gaining wider acceptance. 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. 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 relevance (ii) All subdomains of a domain may not be represented in the extracted content, (iii) Credibility of search results need to be domain specific (iv) Additional processing is required for removal of unrelated content that is not specific to the domain and (v)

Uniqueness/purpose of a domain needs to be well established.
With increased adoption of IT and internet, the concerns of security (confidentiality, integrity and availability) have also increased. Every day, thousands of websites are being compromised and millions of dollars are lost (hack, cyber extortion and other internet frauds) despite growing investment in security awareness and products. To address the concerns on security, there are increasing number of security related websites providing details on product comparisons, vulnerabilities, incidents, threats, controls, etc. However, with this growing internet information on security, relying on generic search engines that have relevance and ambiguity issues, bias, etc. can be detrimental to security knowledge. Thus, arises the need for a security search engine that could provide details on (i) vulnerabilities, threats, incidents, controls and advisories (ii) disambiguated and relevant search results ranked based on the credibility. Our interactions with Chief Information Security officers of Banking sector and other security experts reiterated the need for a security specific one stop site for information. We also draw the inspiration of building SIREN (**S**ecurity **I**nformation **R**etrieval and **E**xtraction e**N**gine) from the established PubMed[1] search engine for Health domain (critical for individual's bodily health like security is critical for non-bodily health of an organization/individual). Having a security search engine will benefit (i) Citizens : Security awareness; (ii) Government and related Bodies - To share advisories and controls leading to policy enablement and better collaboration with security agencies; (iii) Organizations using IT - reduces dependencies on paid security threat feeds; and (iv) Researchers - Access to content for furthering research in security related topics.

2 Demonstration

In the industry, existing work on security search has been in the form of threat feeds, public IP scan and not as a security webpage content. With the established need of building a SIREN, we reviewed existing literature and evaluated software components for building a search engine. Our study of existing literature on search engine components and the related gaps for building a domain specific search engine is available at [3]. The functional and non-functional parameters comparison of the opensource software components for building search engine is available at [4]. Our effort in building SIREN includes

- **Seed URL and Crawling** - In recent times, Twitter and Wikipedia are URL repositories with increased crowdsourcing and Social Media usage. However, the current methods for identification of seed URL has been manual. Inspired by Nature's Optimization algorithms, we developed an algorithm [8] based on Artificial Bee Colony (ABC) for exploration and exploitation of seed URLs. Our initial seed URLs count is 34,007 at an average of 12 child URLs per seed. The identified seed URLs are provided

[1] https://www.ncbi.nlm.nih.gov/pubmed/

to Storm-crawler for crawling the content with 400,726 URLs [7]. To obtain domain coverage, we used Groups and Controls of ISO 27001:2013 [1] standard, widely accepted in research and industry for identifying security gaps and implementing controls. We extended Claude Shannon Diversity index to ensure domain coverage, the Diversity Index measure of 2.1 suggests that we have coverage across all security groups/subdomains.

- **Noise Reduction** - Apache jsoup is used for removal of HTML tags of the crawled content. We used Python NLTK package for stopword removal and stemming to reduce the noise. To avoid loss of Security Acronyms while removing noise, the content is curated with a built in acronym list [5] developed using security focused regulators sites.
- **Classification** - For classifying security related content, we compared SVM, Naive Bayes, Phrase2Vec text processing algorithms. We agreed to use Phrase2Vec [2] considering the accuracy score (cut-off > 0.75 similarity score on a scale 0 - 1, with 1 being exact match of phrase) and the flexibility to parse ISO 27001:2013 manuals for relevant security related phrases.
- **Indexing and Ranking** - For indexing the content, we used Apache SOLR on content classified by Phrase2Vec algorithm. Though there are prevalent pageranking algorithms, we combined SOLR indexing with subdomain credibility of the page while displaying search results. Our credibility assessment to reduce search bias is based on Structural (Broken links, Page Loadtime, Spell errors and others) and Functional (Information source, Text cohesion, Content similarity, Webpage Genre and others) layers. A demo of credibility assessment is available at WebCred[2]. We normalized the values of credibility dimensions and assign a credibility score for each retrieved webpage.
- **Unique Features** - DSSEs are expected to have unique features beyond search results. SIREN uses the available web content to identify Vulnerabilities, Threats, Incidents and Controls. The categorization is based on security phrases from ISO 27001:2013 and related Security Ontology [6].

The demonstration of SIREN[3] to Chief Information Security Officers of Banking sector and the Department of Science and Technology, Government of India is well received, screenshot of SIREN is shown in Figure 1. The deployment is available on a cloud platform for scalability and ability to perform distributed crawling and query parse.

3 Conclusions

SIREN, a security search engine using available open source software has crawled 400,726 URLs and continues to crawl internet and can become one

[2] https://serc.iiit.ac.in/WEBCred/
[3] https://serc.iiit.ac.in/Bhompoo/infosec.html

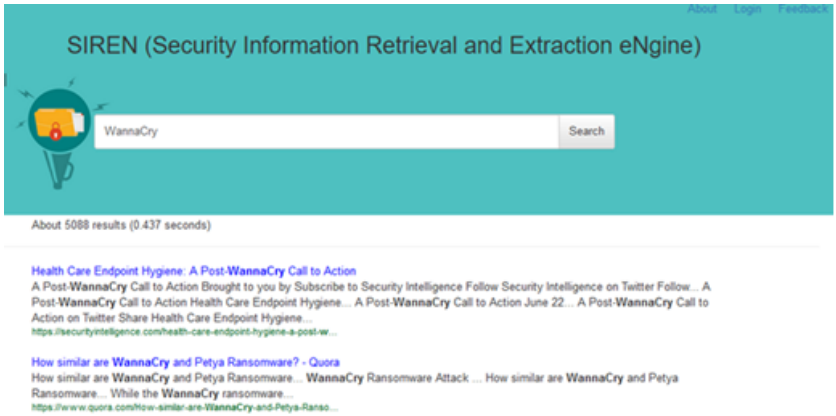

Fig. 1. SIREN - Security Search Engine

stop site for security related information. The usage of ABC algorithm, metric for seed URL relevance and the subdomain coverage improves the domain coverage of the search engine. The approach for credibility assessment in a domain reduces the search bias leading to increased usage of search engine. The evaluation of software stack for building search engine can be used as a package for building other domain specific search engines. We plan to extend SIREN to classify content based on evolving security ontology, explore Deep web, provide automated threat feeds and display visualization based on search criteria. The classified web content may be used for various security related analytics and recommendations.

References

1. ISO 27001 Series Security Standards, `https://www.iso.org/isoiec-27001-information-security.html`
2. Phrase2Vec Algorithm, `https://github.com/zseymour/phrase2vec`
3. Search Engine Components and related Gaps, `http://tinyurl.com/SearchComp`
4. Search Engine Software Comparison, `http://tinyurl.com/SearchToolComp`
5. Security Acronyms, `http://tinyurl.com/SecurityAcronym`
6. Ekelhart, A., Fenz, S., Klemen, M.D., Weippl, E.R.: Security ontology: Simulating threats to corporate assets. In: International Conference on Information Systems Security. pp. 249–259. Springer (2006)
7. Lalit Mohan S, Sourav Sarangi, Y.R.R., Varma, V.: Fine Grained Approach for Domain Specific Seed URL Extraction. In: HICSS. p. To be published. IEEE (2018)
8. Lalit Sanagavarapu, S.S., Reddy, Y.R.: ABC Algorithm for URL Extraction. In: ICWE, practi-O-web Workshop. p. To be published. Springer (2017)