INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2021
Technology, Social Impact

A Fine Grained Approach to Develop Domain Specific Search Engine

Demonstration – Security Information and Retrieval Extraction eNgine (SIREN)

https://serc.iiit.ac.in/Bhompoo/infosec.html

Motivation

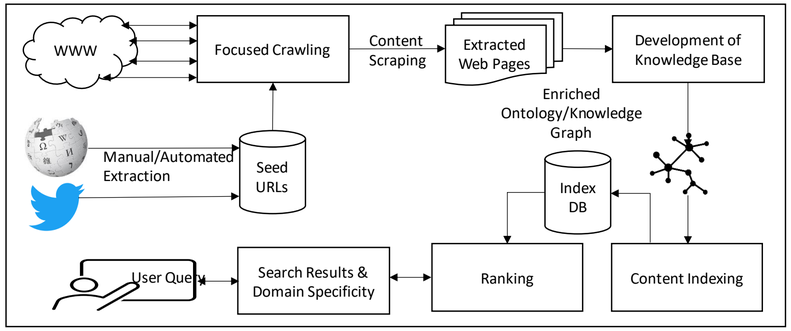
Relevance, Bias, Content Filtering, Click Spam are some of the issues in generic search engines. Specifically, for search results in knowledge intensive domains such as Health, Information Security and others.

A need for domain specific search engine. With following as research questions

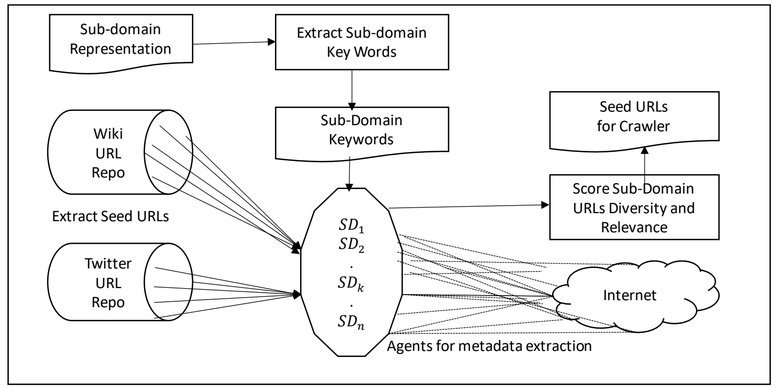
RQ1) What are sub-domains and how to systematically identify sub-domains in a domain? Are there enough URLs that represent sub-domains of a domain? What approach provides efficient (compute and network resources) crawling while ensuring search quality?

RQ2) What are current methods in ontology enrichment? How to enrich a seed ontology with multi-words and instances from domain specific web content while maintaining context?

RQ3) What are existing approaches and their gaps to assess search results quality? What factors influence web page credibility and how can it be used for ranking with flexibility for user intervention?



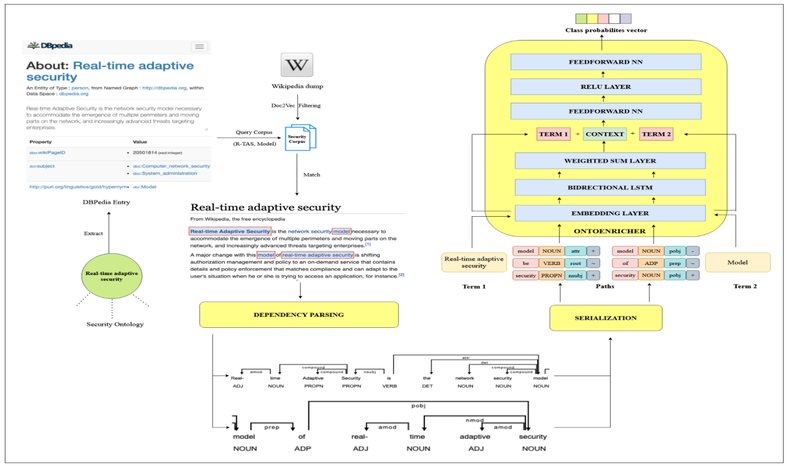
Seed URL Extraction



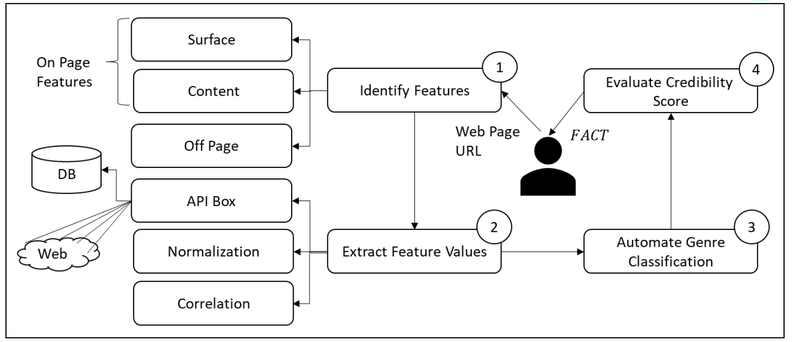
Extended a Meteheuristic algorithm (Artificial Bee Colony) for crawl efficiency. Proposed a metric on Seed URL and Domain diversity.
Approach better than industry scale open source crawlers, extracted more URLs representing more sub-domains

	Seed URLs		Child URLs		URLs/Seed		Unique URLs with Similarity		
Subdomain	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	248,432	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	125,955	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	219,156	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	149,803	54,374	24	12	19,359	20,650	14,365
Total	**45,319**	**34,007**	**1,029,466**	**400,726**	**23**	**12**	**186,620**	**56,494**	**157,612**

Ontology Enrichment



Credibility Assessment



Lalit Mohan Sanagavarapu, Prof Y Raghu Reddy

Software Engineering Research Centre