The Study of Multi-Search Services for Terrorist Network Mining

R. D. Gaharwar
Assistant. Professor
G. H. Patel Department of Computer Science and Technology,
Sardar Patel University,
Vallabh Vidyanagar, India

D. B. Shah
Professor
G. H. Patel Department of Computer Science and Technology,
Sardar Patel University,
Vallabh Vidyanagar, India

G. K. S. Gaharwar
Assistant. Professor
School of Business and Law,
Navrachana University,
Vadodara, India

ABSTRACT
Search engines are designed to make searching the enormous internet data easy but this digital data is growing exponentionally every year. Individual search engine can hardly cope up with this growth rate. There are number of standard search engines available today like Google, Bing, Ask, Dogpile etc but none of them is ideal. Every search services have different threshold ratios, different ranking algorithms leading to deviation in the output of each search result. Multi search services engaging different search engines can be solution to this problem. This paper presents Terrorist Meta Crawler, a web application that uses multiple search services to respond terrorist related user searches. Terrorist Meta Crawler sends web crawlers on different search engines to search terrorist information on web. Terrorist Network Mining can employ Terrorist Meta Crawler to mine vast ocean of data on web of the information of terrorist networks. This paper studies the Control Flow of Terrorist Meta Crawler for Terrorist Networking Mining application.

Keywords
Search Engine, Web Crawler, Search Services, Terrorist Network, Meta Crawler

1. INTRODUCTION

1.1 Search Engines
Search Engines are web services that make searching easy and efficient. End user can submit their query to search engine which uses its own indexes for each web page on internet. These indexes help search engine to respond to user query within no time but each search engine uses different proprietary algorithm for generating their indexes. Hence same user query may give different set of web pages as output for different web search engines [1].

1.2 Web Crawlers/ Web Spiders
In a manual search user types the search word on search engine user interface and as an output to it search engine shows the list for the references. The end user then clicks on any of the reference and it takes him/her to next related web document. This manual search process is automated by web crawler/ web spider programs. As in manual search user types the search word on search engine and the web pages for the links so gathered again parses for other related links and this process goes on. The popular search engines send out web crawler keeps on parsing through each links present on the web page and to collate all the relevant information. [2]

1.3 Terrorist Network and Terrorist Web Mining
Network is the set of nodes connected by some links which can be considered as edges. These nodes can be places, things, people etc. In Terrorist Networks nodes are the terrorist/terrorist organization and the relationship between them are considered as edges. [3] Network Analysis help to find out the patterns of behavioral interaction displayed by these nodes. The special network in which nodes are terrorists/ terrorist organizations and the edges are the relationship between terrorists/ terrorist organizations is called Terrorist Network. While interacting with each other, these terrorists/ terrorist organizations form a covert network which helps in their organized crime [4]. Hence secrecy of the terrorist network is the key for their effective working.

1.4 Previous Findings
Andrej Danko observed that there is huge gap between the user's view and the processor view about web search service. User views it as a simple search request and processor views it as a task that requires great computational speed and capability. The author classified the different search engines on the bases of the indexing methods and the classification is as follows:

1. Crawler-based search engines: These are the types of search engines that use robots for automatic searching and indexing.
2. Human-powered Directories: These search engines allow manually adding and indexing links for every web page. No robotic tools are engaged.
3. Hybrid search engines: These are the combination of above both.

Any crawler-based search engine may have 3 following elements:

Spider: Automatically parse web page, gets the link and downloads it.

Indexer: Automatically indexes the web page downloaded by the spider.

Search Engine: User Interface where user can type his/her query and search engine gives output in the form of relevant web pages on the bases of the end user query.

The author described in his article that the user entered search query is linked with a set of Boolean operators like OR, AND, NOT and NEAR. Boolean operators are useful but the

unstructured format of the search keywords is a big problem. When any user types 'India' on search engine user interface, search engine is unable to understand what exactly the user means; whether he/she means 'prime minister of india' or 'city' or 'state'. This problem arises due to the lack of semantic knowledge in a search engine. Finally user concluded that there is a lack of personalization in search engines. [5]

Rehman et al. studied the factors on which search engines like Google, Yahoo, Ask etc allocate the ranks to the web pages. The authors observed that coding and content of any web page predominately affects its rank. Hence by improving the quality of content and coding, the web page may upgrade its visibility during organic search. The authors studied that any search engine performs 4 important tasks:

1. Searching
2. Indexing
3. Hunting
4. User interfacing

While searching is being carried out, Search Engine Optimization (SEO) tries to optimize the rank for any web site or web page by upgrading its position in search result. The authors described that SEO can perform in 2 ways as follows [6]:

1. On-site Optimization
2. Off-site Optimization

Javad H. et al. states that the Information Retrieval Process is fairly difficult on internet due to the format of web pages. The web documents available on internet are heterogeneous collection of the different types of tags, links and actual content. [7] Different tags have different level of relevance in understanding of web documents for example title tag is very important as most of the search keywords appear in the title tag of the document. Meta tag contains the summary of the content of the web page. Anchor tag contains the textual link to the page to which that particular web page points.

2. TERRORIST META CRAWLER

A normal Meta Crawler tries to fire parallel search queries on multiple search services and then collate the results returned by them. In this paper presents a model of Terrorist Meta Crawler which fires the query of terrorist related information. A Terrorist Meta Crawler provides a user interface which accepts terrorist related query. In this model the emphasize is given on terrorist related queries only because when the relevant references are collected Terrorist Meta Crawler will make sure that only those references are considered as relevant which have any terrorist/ terrorist organization related data in it. Hence user is free to choose query about any terrorist/ terrorist organization to be typed on user interface. Terrorist Meta Crawler not only collects the relevant links but filters them too. It assigns ranks to each link that in result set. When any search engine is affected by spamming, it will give the output different from other search engines. Hence simple duplication removal technique for filtration mechanism will not be useful. Assigning Ranks to each will help in diluting the effects of spamming. [8]

The following figure presents a Control Flow Diagram of Terrorist Meta Crawler.

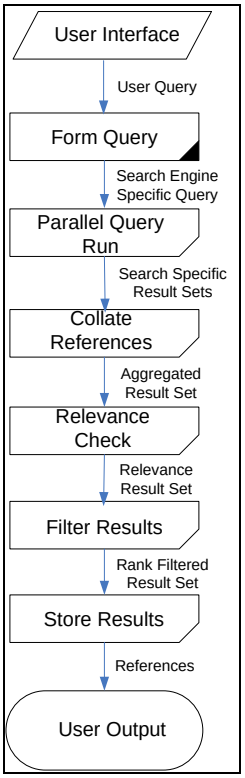

Fig 1: Terrorist Meta Crawler Control Flow Diagram

The Control Flow of Terrorist Meta Crawler is being broadly categorized in following 3 steps:

1. Form Query and Run Parallel
2. Check for Relevance of URL and Aggregation of Result
3. Filtration and Storage

Step 1: Form Query and Run Parallel

Terrorist Meta Crawler presents a user friendly interface where user enters the phrases to be searched. These query phrases are searched on popular search engines like Yahoo Search, Dogpile, Bing , Ask and MySearch. Hence Terrorist Meta Crawler generates appropriate query format for each search engine. For convenience Terrorist Meta Crawler generates query format for first 3 search web pages of each search engine. This is done by understanding the details of URL patterns of Yahoo Search, Dogpile, Bing , Ask and MySearch. The intensive study of these URL patterns showed that any generated URL by search engine can be divided into number of token strings. The search engine specific URL derived by

$$URL = prefix + keyword + index + \sum sufix$$

Where prefix = a fixed String

keyword = $\sum$query phrase

index = 1,2,…n

sufix = a fixed String

The URL generated from above formula can successfully give the web search page for any search engine. These URL are loaded on different threads and run parallel. For this purpose Terrorist Meta Crawler uses multithreaded architecture which improves the efficiency of performance.

Step2: Check for Relevance of URL and Aggregation of Result

Once the user query is successfully formatted in search engine specific URL they are fired for results. The output of these queries is not just collected and aggregated but Terrorist Meta Crawler will try to check for relevance of that particular URL. It parses the source code of the web page returned as result, for the links and contents. Contents will be thoroughly analyzed to find whether there is any terrorist / terrorist organization related information in it. When such information is found Terrorist Meta Crawler will try to find terrorist / terrorist organization related links in that web page and when such link is found it stores that page in database. For all the links found on web page Terrorist Meta Crawler will get the source code for that link and will again repeat the whole process of Relevance Check on it also. By repeating this whole process again and again Terrorist Meta Crawler will generate a set of web pages that do contain the linkage information between terrorist / terrorist organization only. This set of web pages which are collected from these 5 different search engines are then filtered into a singleton set of relevance web pages.

Step 3: Filtration and Storage

Many Meta Crawler applications use duplication removal method as a Filtration technique but this method has limitation of its own. Therefore Terrorist Meta Crawler uses Minimum Rank Removal method. Once a single set of web pages is created it is stored in database but along with the URL of relevant web pages their ranks are also stored in database. Allocation of rank to each web page is done by Terrorist Meta Crawler as follows:

Rank = number of times that particular web page appeared in the singleton set of relevant web pages.

Hence every URL stored in database will have a Rank associated with it. The web pages with minimum rank can be removed from the database. Minimum Rank Removal method works better than simple duplication removal technique because in duplication removal there is no way to indicate relevance of a reference. When any search engine is facing spamming issues it will give dishonest preference to some particular page and hence when its output is compared with the output of other search engine, both the outputs differ. The difference in the outputs will be because of all those web pages that are unethically given preference in output but these irrelevant web pages will be automatically removed by Minimum Rank Removal method. Hence the quality of the search service improves. Only the relevant and spamming free results are stored in database and will be given to end user as the output of their search query.

3. CONCLUSION

An individual search engine is optimized to respond to certain set of user queries only. Hence use of single search engine for all the user queries cannot be justified. This generates the need for a crawling application that may crawl on different search engines to give only an optimized set of relevant web pages. High end applications like Terrorist Network Mining requires such multi search service to respond to terrorist related user queries. Terrorist Network Mining requires terrorist related information to be collected from the open source. This paper presents the control flow for Terrorist Meta Crawler that gives a single interface to end user. Using this interface end user can fire their terrorist related query on multiple search engines at once. The end user will be provided with the list of relevant web pages that are fetched using different search services. Terrorist Meta Crawler can be used for the comparative study of search engines to check the variance in the output of multiple search engines.

4. REFERENCES

[1] Selberg, Erik, and Oren Etzioni. "The MetaCrawler architecture for resource aggregation on the Web." IEEE expert 12.1 (1997): 11-14.

[2] Heydon, Allan, and Marc Najork. "Mercator: A scalable, extensible web crawler." World Wide Web 2.4 (1999): 219-229.

[3] S. Azad and A. Gupta, "A Quantitative Assessment on 26/11 Mumbai Attack using Social NetworkAnalysis," Journal of Terrorism Research, vol. 2, no. 2, 2011.

[4] R. D. Gaharwar, D. B. Shah, and G.K.S. Gaharwar, "Terrorist Network Mining: Issues and Challenges," International Journal of Advance Research in Science and Engineering, vol. 4, no. 1, pp. 33-37, 2015.

[5] Danko, A. The concept of web search engines and metacrawlers within the Internet environment.

[6] ur Rehman, K., & Khan, M. N. A. (2013). The foremost guidelines for achieving higher ranking in search results through Search Engine Optimization. International Journal of Advanced Science and Technology,52, 101-110.

[7] Javad Hosseinkhani, Suriayati Chaprut, Hamed Taherdoost, and Amin Shahraki Moghaddam, "Propose a Framework for Criminal Mining by Web Structure and Content Mining," International Journal of Advanced Computer Science and Information Technology, vol. 1, no. 1, pp. 1-13, October 2012.

[8] R Gaharwar, D Shah, G Gaharwar, ”THE IMPLEMENTATION OF TERRORIST WEB MINER FOR THE TERRORIST NETWORKS”,IJRET 4 (11),pp 268-271