Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 5 – No.8, August 2016 – www.caeaccess.org

Architectural Design of Meta Crawler 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

ABSTRACT

Nowadays world is flooded with the digital data as well as with different types of search services but none of the search service gives the complete view of internet data. Moreover there is scarcity of the Meta Crawlers that are application specific. This paper presents architectural design of Meta crawler designed especially for Terrorist Web Mining. This architectural design is use to create an effective Meta crawling service that span across search engines. Meta crawler collates the search results of multiple search engines to fetch the terrorist Networks related information. Search results from single search engine may prove insufficient and spamming affected hence there is need for Meta Crawler which filters the results aggregated. Meta Crawler proposed in this paper has a multithreaded architecture hence this Terrorist Meta Crawler design have advantages like performance efficiency and scalability. This architecture can be used to designed a terrorist network related dynamic Meta Crawler.

Keywords

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

1. INTRODUCTION

1.1 Search Engines

In modern times day to day information retrieval system is entirely depended on Search Engines. Search Engines uses enormous data available on web. Every search engine gives end user an interface to ask question and gain information through internet. Popular search engines likes Google, Ask, Wow, Bing, Dogpile, Yahoo Search, MyWebSearch etc uses different types of proprietary algorithms for indexing the web pages on internet with assumption that if web page is useful then the other links on same web page may also be related to the same topic. [1] This assumption helps to find the other web links in the related topic on the web.

1.2 Web Crawlers/ Web Spiders

Nowadays Search Engines employs crawlers/spiders to browse the World Wide Web. They automatically parse all the links on a web page and download the other related web pages which are known as crawling or spidering. These web crawlers separates the links on each web page and tries to find the other relevant web documents by following these links/URLs. Web crawlers or web spiders are just computer program or automatically running code scriptlet which indexes the web pages downloaded by search engines for fast search. The only issue here is the identifying the level of relevance of links. Simply following each web link may end the search to completely irrelevant reference hence timely checking for the relevance of the web page becomes crucial task of web crawler. [2]

1.3 Meta Crawlers

As the number of internet users are increasing so is the number and types of search engines. These Search engines differ in the way they search the World Wide Web for returning relevant results. Different types of search engines covers different areas on internet hence none can be considered ideal. Meta crawlers are the software which tries to employs diverse types of web crawlers to give user better and up-to-date search results. The search results from multiple search engines is collated by Meta crawler such that user gets the advantage of using multiple search services. Meta Crawler allows to search in expressive query language like English. Hence the user can search certain phrases in natural language. This query will run on multiple search engines in parallel to facilitate end user to query multiple search engines from single interface. Therefore Meta Crawler eliminates the end user need to search on different search engine and do not burden the end user to remember addresses and interfaces. [3]

1.4 Terrorist Web Mining

Terrorist Web Mining is a special kind of the Web Mining Technology which helps to detect the behavioral communication and structural patterns of these terrorists/ terrorist organizations in the terrorist network. The trivial task about Terrorist Web Mining is collecting information about terrorists/ terrorist organizations. Hence specialized data mining tools are needed to explore large number of heterogeneous web pages to detect the communication patterns of these terrorists/ terrorist organizations. [4] This mining technique studies terrorist networks which are formed by the terrorist organizations during the operations. The biggest challenge faced by Terrorist Web Mining technique is gathering data about these terrorists/ terrorist organizations.

1.5 SPAM Issues

Unethically and dishonestly increasing the rank of any web page so that it may appear in the result of a web search for any search engine is called spamming. Every search engine uses its own proprietary algorithm to generate ranks for the web pages. When any search is made search engine displays the output on the bases of these ranks. Web pages with high ranks appear at the top order in the search result. Hence if the rank of the web page is increased it will appear in the search result irrespective of whether it has relevant information or not. This leads to the inefficient search results. Hence spamming should be stopped or controlled.

Spamming cab be induced by search engine itself or by web page builders. When search engine unknowing generates the inappropriate search result, the reliability of search engine becomes questionable. Hence depending on a single search engine for searches may lead end user to wrong/ inappropriate outputs. Moreover web spammers undermine the search engines to deliberately and dishonestly increase the visibility of their web pages. [5]

2. PREVIOUS FINDINGS

Researchers all over the world are working in this area. Some of the previous researches are mentioned below:

Yang X. et al wrote that due to the availability of large data on internet information retrieval capability of any single search is insufficient. The authors described that working of any Meta search engine can be divided into following 6 steps:

1. Input a user query
2. Customized user query of particular search engine
3. Fire query parallel on different search engines
4. Collect output from all the search engines
5. Merge the output
6. Process the final output

On the bases of the above 6 steps working of any Meta search engine can be understood. [6]

Selbery E. et al. observe that although there are large numbers of search engines available, each one of them only covers small portion of the internet. Moreover due to the problem of web spamming many search engines give obsolete or irrelevant information for user query. Hence there is always a need of web crawler to overcome these drawbacks of individual search engine and helps to identify the diversification found among different search engines. Moreover Web crawler also provides the customization of user query, privacy and novel filtration of references. The authors describe that any web crawler should address following problems:

1. Any single search engine insufficient to address every user query
2. Search engines returns number of references among which many are outdated and irrelevant.
3. Obsolete and unrelated references could be removed easily
4. Web Crawler should be available to large population of world.
5. Tradeoffs between different types of search services can be evaluated with the help of Meta Crawler. [7]

Laria V. et al observed that large number of search engines nowadays uses simple methods like indexing, collating and filtering. Semantic contents of the web document are not kept in mind. The authors quoted that in the process of simple indexing and filtering some of the valuable information of the web document like common links, connected links and cross document similarities are lost. The authors discuss the clustering techniques as one of the most effective technique used for Information Retrieval Process. In clustering techniques assumption is made that if any single document is relevant for user query then other documents belonging to particular cluster tend to be relevant for user query. Hence clusters can be built and they can also be used for offline queries. [8]

Selberg E.et al discussed the drawbacks of using single search engine for search services. They observed that there is huge difference in the list of references that are given as output for user query as well as in their positions in the output. Moreover the time duration taken by each search engine to respond to user query also differs drastically. These differences are due to the fact all the search engine uses different algorithms to index/rank the web information. These indexes may be indexed at different time also. Some are freshly indexed and some are stale references. Among the other drawbacks for engines is their user interfaces. All of them give different type of filters and search phrases. Sometimes the filters provided by search engine might not be sufficient for the end user. The authors created software programmed robotic Web crawler named Meta Crawler Softbot to overcome above limitations of a single web search service. Meta Crawler Softbot had following features:

1. It takes as input user search query
2. Formulate it properly
3. Parse and collates the results returned
4. Analyses and Elimination of duplication

Amongst the other salient features of MetaCrawler Softbot are intelligent user interface, faster processing capability and dynamic to the changing environment. [9]

3. TERRORIST META CRAWLER ARCHITECTURE

In this paper we propose an architectural model of Terrorist Meta Crawler which will collect terrorist attacks related information from the open source like internet. The most recent studies in the field of Meta Crawler design focuses on collating the data from different search engines. A Meta Crawler specialized in collecting Terrorist related information from the internet is focus of study in this paper. Moreover the current Meta Crawler designs use simple duplication removal methods as Filtration technique which might prove insufficient for spamming effected search engines but this Terrorist Meta Crawler architectural design uses more intelligent Filtration techniques. The following figure shows the architectural design of Terrorist Meta Crawler.

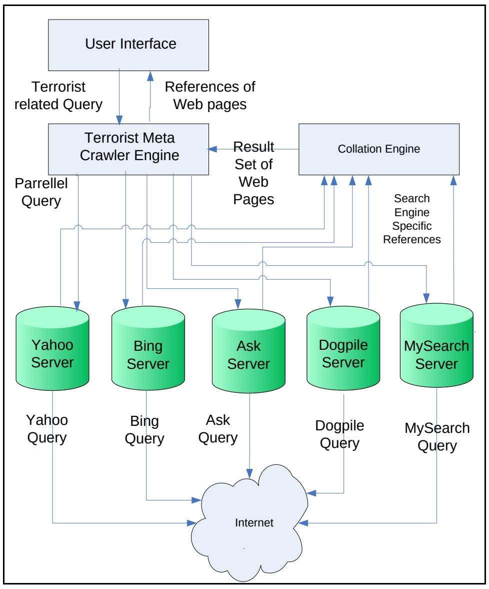

Fig 1: Terrorist Meta Crawler Architecture Diagram

The architectural design of Terrorist Meta Crawler consist of three main components

1. User Interface
2. Terrorist Meta Crawler Engine
3. Collation Engine

3.1 User Interface

It is an interface between end user and Terrorist Meta Crawler. User interface allows end user to enter terrorist related query that is to be searched on five different search engines. Hence user interface gives a uniform layer of interaction with variety of search engines. User entered query can be any structured English phrase so that user might find interface user friendly.

3.2 Terrorist Meta Crawler Engine

This part of Terrorist Meta Crawler is responsible formatting proper query that can run on different search engines. Terrorist Meta Crawler engine takes the structured English user query from the user interface and converts it into formatted query which can be fired successfully on search engines like Yahoo Search, Bing, Ask, Dogpile and My Search. Hence Terrorist Meta Crawler engine requires sophisticated logic. After formatting query properly these queries are assigned to different threads. Each thread is associated to a search engine so that they may run parallel on different search engines. When collation engine returns the different outputs from different search engines Terrorist Meta Crawler engine filters and stores them in appropriate database. Terrorist Meta Crawler engine uses novel Minimum Rank Removal technique for filtration process. In this technique a rank is assigned to each web page stored in database where rank is the number of times that particular web page appeared in search result. In simple duplication removal method sometimes the relevant links are removed. Hence by assigning ranks Terrorist Meta Crawler engine assigns relevance to web page. If any web page appears in search result of any particular search engine due to spamming problem that web page will automatically not appear in the search result of any other search engines and hence will not be allocated high rank by Terrorist Meta Crawler engine. This technique is effective in removing spamming effect on the performance of the Terrorist Meta Crawler

3.3 Collation Engine

It takes as input the formatted query from Terrorist Meta Crawler engine and gives set of web pages as an output. Collation Engine collects the results from different search engines and creates a singleton set of relevant references. This set of relevant references is returned to Terrorist Meta Crawler engine which appropriately filters it and stores the result in database.

3.4 Advantages of the architectural design of the Terrorist Meta Crawler

User friendly: Any application which is developed for end user must be user friendly. Terrorist Meta Crawler gives end user an interface which is very much similar to any search engine user interface which allows the user to type their queries phrase-wise. The user query can be in structured English language. Hence any naïve person can also use this application. Users have to simply type their query and click on search button, rest of the work will be handled by Terrorist Meta Crawler engine. User interface of Terrorist Meta Crawler is user friendly to that extent that naïve user will not feel different using Terrorist Meta Crawler from any popular search engine. This is one of the biggest advantages of Terrorist Meta Crawler.

Multithreaded: Many Meta Crawler application available in internet faces performance efficiency issues. Terrorist Meta Crawler uses multithreaded architecture for running user query on multiple search engines as well as for downloading the relevant web pages. Multithreaded design of Terrorist Meta Crawler uplifts the performance efficiency. Moreover initial response time of Terrorist Meta Crawler also decreases as one search query running on one search engine does not wait for other query running on other search engine. Hence the multithreaded architecture decreases the average waiting time.

Dynamic: When user fires any query, the phrases are searched on multiple search engines and the relevant web pages are downloaded and stored in database. This process will be repeated for every user query. Hence the system dynamically stores web pages for every end user query. Moreover each downloaded web pages is assigned a rank depending on number of times it appears in output result set. Rank of any web page indicates the relevance of that particular page in the output set. The rank changes when user fires other query. Hence rank values are assigned dynamically.

Scalability: Nowadays the internet users and internet data are growing rapidly. Hence it is necessary that Terrorist Meta Crawler must be capable of handling this rapid change. Multithreaded architecture of Terrorist Meta Crawler engine will be able to scale with the growth of internet. Hence there will be no need to change hardware to cope with the situation.

4. CONCLUSION

This paper focuses on architectural design of Terrorist Meta Crawler. This crawler is customized to collect the terrorist networks related information from multiple search engines. Terrorist Meta Crawler is divided into three main components like User Interface, Terrorist Meta Crawler Engine and Collation Engine. Terrorist Meta Crawler Engine will be responsible for transforming end user query to multiple search engine specific queries. Moreover this crawler dilutes the effect of spamming on the results of search services by using Minimum Rank Removal Technique. This filtration technique eliminates the spamming results to influence the effectiveness of search results. This paper shows several the advantages of Terrorist Meta Crawler like user friendly, dynamic and scalable.

5. REFERENCES

[1] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in 7th World Wide Conference(WWW7), Brisbane, 1998.

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

[3] Etzioni, Oren. "Moving up the information food chain: Deploying softbots on the world wide web." AI magazine 18.2 (1997): 11.

[4] R. D. Gaharwar, D. B. Shah, and G.K.S. Gaharwar, "Proposed Architecture for Terrorist Web Miner," International Journal of Computer Applications, vol. 128, no. 9, pp. 18-20, October 2015.

[5] Gyongi, Zoltan, and Hector Garcia-Molina, "Spam: It's not just for inboxes anymore.", Computer 38.10 (2005): 28-34.

[6] Yang, X., & Zhang, M. (2000). Necessary Constraints for Fusion Algorithms in Meta Search Engine Systems. In Proc. of the Int'l Conf. on Intelligent Technologies (pp. 409-416)

[7] Selberg, E., & Etzioni, O. (1995, December). Multi-service search and comparison using the MetaCrawler. In Proceedings of the Fourth Int'l WWW Conference, Boston.

[8] Laria, Víctor González, Richard Griffiths, and Graham Winstanley. "Application of a clustering algorithm to recover topic content in an unstructured text-based environment."

[9] Selberg, E., & Etzioni, O. (1997). The MetaCrawler architecture for resource aggregation on the Web. IEEE expert, 12(1), 11-14.