Ontology based Multi-Threaded Crawler for Collecting Twitter Data Streams

Dr. M. Arun Manicka Raja,
Associate Professor,
Department of Computer Science and Engineering,
RMK College of Engineering and Technology, Chennai.
arunmanickarajam@gmail.com

Ms.T.Sumitha,
Assistant Professor,
Department of Computer Science and Engineering,
RMK Engineering College, Chennai.
Sumitharmk90@gmail.com

Dr. N.R. Rejin Paul,
Associate Professor,
Department of Computer Science and Engineering,
RMK College of Engineering and Technology, Chennai.
nrrejinpaul@gmail.com

Abstract—
Twitter is one of the most prominent social media applications. Millions of users post lot of valuable information in twitter. These information is potential for many other organizations to know about their product or services. An average of 500 million tweets are generated in a day around the world. Analyzing and extracting fruitful information from these vast data is a great challenge for anyone who is interested to perform data analytics about a product or service. It is essential to collect the generated data quickly for decision making on the intelligent recommendation systems. Many research works have concentrated on collecting and analyzing the twitter data for various purposes. But only very few works have been carried out regarding the twitter data collection. Twitter data collection is as important as analyzing the twitter data. In this work, a multi-threaded tweet crawler has been proposed to collect all the tweets related to event or person or product. In an average, approximately 1 million tweets are generated about the popular topics such as politics, sports, weather, finance, education, electronic gadgets etc. It is not easy to collect all these 1 million tweets by the traditional crawler to collect. Most of the typical crawlers do not collect entire tweets related to any specific event. The proposed multi-threaded crawler creates the number of threads dynamically per the increased availability of tweets.

Keywords—crawler, tweets, ontology, multi-threaded, search, crawl rate, tweet categorization.
DOINUMBER:10.48047/NQ.2022.20.19.NQ99377 **NEUROQUANTOLOGY2022;20(19):4142-4148**

I. INTRODUCTION

The advantage of social media websites like twitter helps the users to detect various events or product trend or popularity of a person in a short time by collecting the generated tweets. The tweets are highly associated with spatial and time related information, it is easy to make a statistical analysis of the tweets. It is important to use the meaningful data for performing the analysis. Large amount of data is generated in twitter by millions of users from various regions of the world on different topics. Analyzing these huge collections of data and making a better analytical result is an essential task. The main job of this twitter data analytics is the data collection.

The twitter data must be collected with respect to covering all the data pertaining to any topic or person or event. Since the is generated very rapidly, there are millions of tweets are available for any popular topic. If the trending topic or the popular person or trending event needs to be identified daily, then it is very important to collect the data generated related to any specific event. It means that the data collection needs an efficient crawler to collect the tweets in the speed of its generation.

Crawler is a software agent for collecting the web data. Tweet crawler is the twitter data collection agent focusing on collecting the tweets about various topics. Typical tweet crawler involves the use of keywords for searching the tweet content, commonly knowns as keyword based tweet crawler. Keyword based tweet crawler is appropriate for collecting the tweets containing the specific keywords. But these typical keywords based tweet crawler failed to collect most of the similar tweets related or associated with those keywords.

In such situation, it is vital to concentrate on a crawler which would focus on all the tweets associated with a specific event. Ontology is the knowledge representation of any specific domain. In many data collection agents, ontologies are used as the key representation for domain knowledge. Ontologies related these specific domain helps to collect the data in a very focused way for increasing the tweet collection. Though ontology based tweet crawlers are participating as key element in data collection, it is essential to increase the number of tweet collection instances. Thus, process threads are used for making the instances of these tweet collection agents. Ontologies along with these multithreaded tweet crawlers highly increases the tweet collection process.

The paper has been organized in the following manner. In section 2, the related works of the ontology based multi-threaded tweet crawler is discussed. In section 3, the architecture of the ontology based multi-threaded tweet crawler is explained. In section 4, the results are discussed by comparing the result of ontology based multi-threaded crawler with the typical keyword based tweet crawler. Section 5 includes the conclusion and future enhancements of the ontology based multi-threaded tweet crawler.

II. Related Works

The need for analyzing the tweets increases every day. It leads to the design of the tweet crawler. In the literature, many works have concentrated on analyzing the tweets for finding the trending topic. But only very few works have concentrated on the tweet crawlers. It is necessary to create a focused event crawling to collect the web data regarding many important events. This is mainly needed, when an important event is occurred everyone is trying to locate the most trending information. To achieve this, there is no specific systematic way of collecting the information. The intelligent event focused crawling is designed to effectively collect the data from the web. The event model that helps to capture the key events related information for providing the model state for the data crawling. The similarity between the events is measured to the relevance of them to help the prediction easy. The event model based method performs well for focused crawling of the relevant web pages from the web. [1]

The vast social media data resides in the databases contain the high-quality information. The hidden information is also contained in the web databases wherein it is difficult for the traditional web crawlers. The query generation techniques help to generate appropriate queries for extracting the information from the web. It is also important that only domain specific content can be gathered by navigating through the highly relevant websites. The crawler has been designed in such a way that it only crawls the relevant portion of data from the entire web [2].

There is always an increasing demand for the effective crawling of data from the web. Parallel crawlers are used for crawling the data simultaneously. The URLs related the specific contents are given the parallel crawling process. The crawling process threads are concurrently executed to collect the data [3]. The high-performance web crawler runs in a distributed and decentralized manner. It helps to scrap the twitter or any social media data. The crawler is working as a scalable and adaptable to various situations. [4]

The continuous evolution of the online social networking applications leads to large amount of data generation. It is very important to consistently crawl the data at all the times. The fast crawler is designed to gather the up-to-date information very precisely. The node discovery strategy of the random walk along with the backtrack gives the promising results of fastest crawling. The random search crawler is implemented for the purpose crawling large number of tweets [5]. Tweet crawling also helps to improve the news result among the web data. in traditional systems, recent news has been provided based on the click through rate. It is the comparison between the clicks and the views of content. But in such cases, recent news URL has only very few clicks. Thus, it is not possible to retrieve the less click through news. Due to this, tweets

are used for recent news search for providing efficient information [6]. Microblogs are created for sharing the information about social causes and many other problems. There is need for the development of the tools for effectively collecting the ever-updated micro blogs. In addition, it is important keep tracking of the most attracting event that is evolving among the user communities. It is a big challenge to detect the event as an emerging one, before it becomes a trending topic. The tracking of these evolving contents must be monitored continuously [7]. The collaborative filtering recommendation is proposed to microblog prediction and recommendation. The relationships between the tag, microblogs and the users are constructed. [8]

Models provides the mechanisms to improve system performance. Various methods are applied for modeling and controlling on micro-blog crawler. With the rapid development of social studies and social network, millions of people present or comment or share their opinions on the platform every day, and thus, produce or spread their opinions and sentiments on different topics. The microblog has been an effective platform to know or mine social opinions. To do so, crawling the relevant microblog data is necessary. But it is hard for a traditional web crawler to crawl micro-blog data as usual, as by using Web 2.0 techniques such as AJAX, the micro-blog data is dynamically generated rapidly. As most microblogs' official platforms cannot offer some suitable tools or RPC interface to collect the big data effectively and efficiently, an algorithm is provided on modeling and controlling on micro-blog data crawler based on simulating browsers' behaviors. This needs to analyze the simulated browsers' behaviors to obtain the requesting URLs to simulate and parse and analyze the sending URL requests according to the order of data sequence. [9]

Researchers have capitalized on microblogging services, such as Twitter, for detecting and monitoring real world events. Existing approaches have based their conclusions on data collected by monitoring a set of pre-defined keywords. It is shown that the manner of data collection risks losing a significant amount of relevant information. It is proposed an adaptive crawling model that detects emerging popular hashtags, and monitors them to retrieve greater amounts of highly associated data for events of interest. The model analyzes the traffic patterns of the hashtags collected from the live stream to update subsequent collection queries. To evaluate this adaptive crawling model, it has been applied it to a dataset collected during the 2012 London Olympic Games. The analysis shows that adaptive crawling based on the proposed Refined Keyword Adaptation algorithm collects a more comprehensive dataset than pre-defined keyword crawling, while only introducing a minimum amount of noise. [10]

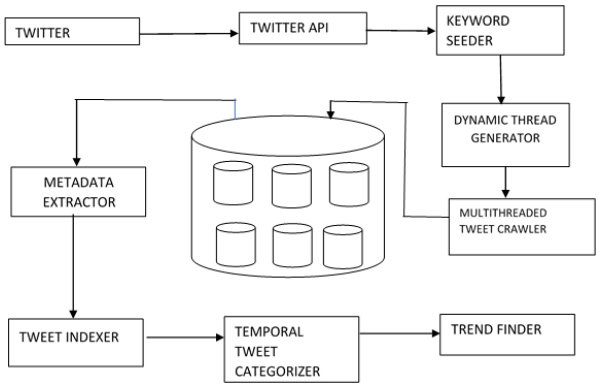
These days' social networks have attracted people to express and share their interests. It is aimed to monitor public opinions and other valuable discoveries by using the data collected from social network website. A distributed web crawler framework called SWORM is used, which runs on the Raspberry Pi for fetching the micro-blog data and overwhelms the traditional web crawlers on efficiency, scale, scalability and cost. The framework can easily be extended per the specific needs of the user with the help of some simple python scripts. A model for micro-blog network is used to confirm what and how the crawler will crawl from social website. Some crawler has been used in this framework on the Raspberry Pi and stored the obtained resources in Shared MongoDB which is a category of NoSQL. The distributed framework greatly improves the efficiency and accuracy for collecting data. [11]

With the development of the Web, social media, the SINA Microblog for instance, is gaining wider popularity. This indicates most users' attitudes towards emerging technologies that happened recently. To obtain the data on social media, the traditional Web crawler is only able to obtain part of the information by collecting data in a non-login status because the crawler is not provided with the capabilities to log into the system. To enable information sharing among users, the SINA Microblog Open Platform offers a number APIs in which certain limitations such as those on the number of user requests of the microblog server, still exist. In this paper, a method that simulates user login to collect Microblog data related to hot topics on the SINA platform is proposed. What's more, the method overcomes the limitations of the SINA Microblog APIs, thus leading to getting more data from the system. In addition, the typical user behaviors, through the collected data, is found when they interact with social media [12]. MicroBlog is an effective vehicle for the network public opinion, and plays an important role in

dissemination of the public opinion. A crawler which consisted of user crawling and contents crawling used to crawl MicroBlog is designed. The crawler used protocol-driven strategy, event-driven strategy and template extraction methods to achieve the successful extraction and data storage. It shows that the crawler has an efficiency and integrity of information collection compared with the crawler BFS. A more flexible crawler is needed with the more complexity of DOM Tree [13].

Despite all the works carried out regarding the tweet crawler, there is a scope for developing the ontology based multi-threaded tweet crawler to collect the tweets efficiently. The multi-threaded tweet crawler collects the tweets based on the term mentioned in the ontology. The following section describes the architecture of the multi-threaded tweet crawler.

III. MULTI-THREADED TWEET CRAWLER ARCHITECTURE



A. Twitter

The architecture of the multi-threaded tweet crawler is depicted in fig 1 with its various components. Twitter is the most commonly used social networking website. It provides the facility to user to share the micro blog contents instantly to other social networking users. The short micro blog content is known as tweet. It allows maximum 140 characters for generating the message. The social networking users are posting tweets about their opinions on the products purchased or services consumed or political views and knowledge sharing using URLs.

B. Twitter API

The twitter provides two kinds of APIs for collecting the tweets from the twitter website. The APIs are streaming API and REST API. These APIs have certain time interval limit for collecting the tweets from any twitter user account. Once the users have created the twitter account, it allows to create the authentication and secret keys for connecting with twitter API. Using these user credentials, the users get connected with the twitter website.

C. Keyword Seeder

Twitter contains lot of large volume on various topics. When it is required to collect the tweets from the twitter, it is necessary to provide the input keywords which are mentioned in the tweets by the users who posted it. The keyword seeder provides the related domain specific keywords to the tweet crawler as input. As per the input domain keywords, the domain ontology is constructed. Further the domain ontology is invoked to identify the similar words of each word present in the ontology tree and then those similar identified keywords are given as further inputs to the crawler.

D. Dynamic thread generator

The thread generator helps to create the number of threads required as per the number of elements present in the ontology tree at each level. Whenever the ontology tree is processed for extracting the search terms mentioned in the branches, at each level the threads are generated. These threads are generated dynamically per the input terms.

E. Multi-threaded tweet crawler

The crawler helps to retrieve the tweets from the twitter website. Multiple threads are created and each thread is achieving individual results. Combinedly all these thread results are invoked to make the huge collection of results as the final one. The thread manager is used for managing the multiple threads used in the tweet crawler. The tweet crawling status of each thread is monitored by the thread manager.

F. Tweet repository

Each thread works as an individual tweet collector agent and the tweets are collected individually and at the same time simultaneously collected by the thread crawler. These collected tweets are stored as a repository. The overall repository is maintained for managing all the tweets collected by the individual crawler.

G. Meta-data extractor

The metadata extractor extracts the information about each of the tweets posted by the individual user. The metadata includes the date and time, location, user id, user replies etc.

H. Tweet indexer

The tweets are indexed based on the date and time the tweets were created. The indexed tweets later help to categorize the tweets and identify the top categories which were discussed most prominently by most of the users.

I. Temporal tweet categorizer

The tweets are categorized based on the time attribute. It is easy to categorize the tweets based on the tweet time. In addition, these categorization helps to segregate the tweets with respect to the product type or service type.

J. Trend finder

The trend of the product or person or political party or any event is identifiable based on the categorized tweets. The tweets which often reaches certain threshold limit is described as the most trending category of the day or hour.

IV. RESULTS AND DISCUSSION

Twitter API is used for effectively streaming the required tweets for various topics. Tweets are the 140 characters of short information. These are generated by millions of users around the clock from different regions of the world. Many text-based retrieval techniques are used for variety of purposes. Data collection is the vital task for these kinds of information retrieval processes. Twitter provides few ways for the technical community to collect the tweets for many analytical tasks. These tweets are the core information for many decision-making processes. Thus, it is important to decide the domain for which the tweets must be collected. Many electronic gadgets are being prominently used by everyone. Tweets related to these electronic gadgets include potential information. These information is very crucial in decision making processes related to various businesses. Tweets are collected mainly to know whether the product or service or event or person in currently trending or not. The twitter provides ways for collecting the tweets using twitter API. The tweets returned by the twitter are in the form of semi-readable or semi-understandable form. The necessary information is extracted from these formats to perform the needed analytical tasks.

It is possible to collect all the tweets containing the related keywords with the help of the twitter API. If the keyword is general, then it is not possible to collect all the tweets containing the search term. The proposed solution for this problem is to make the search more specific b combining many keywords. It is found that for every moment more than 500 million tweets are created per day. Thus, any search keywords leading to a better crawler will at least collect 1 million tweets in a day. That is 10,00,000 tweets must be collected within 24 hours. It means that for every hour more than 40,000 tweets must be collected. It is meant that for every minute averagely 666 tweets must be collected. But many of the typical tweet collecting programs failed to achieve result. In other words, many research works have not concentrated on collecting all these possible tweets.

In this work, a multi-threaded tweet crawler has been created for collecting all the available tweets with the possible ways. Initially, the input search keywords are seeded to the crawler. The tweet collection begins with these initial search keywords. The number of threads which are to be used in the beginning is decided based on these initial keywords. The domain ontology has been created for the electronic gadget (laptop) domain. When the input terms are given to the search crawlers, the tweets are identified with the matching of contents. At every level of the ontology tree, the number of crawlers are increased and tweet collection is also proportionally increased. This is continued till it reaches the leaf nodes of the ontology.

When each key term is inputted to the crawler, the collected tweets are keep on storing in a repository. These all collected tweets are contained in a common tweet repository. Each tweet contains various attributes or meta information about the tweets. These meta data are helpful in indexing the tweets based on anyone attribute.

Table 1. Crawled Tweet Information

	Keyword-based crawler	Ontology-based multi-threaded crawler
No. of tweets	1,24,000	9,26,000
Valid tweets	1,02,000	8,97,000
Crawling time	24 hours	24 hours

Table 1 shows the crawled tweet information using both keyword-based crawler and ontology based multi-threaded tweet crawler.

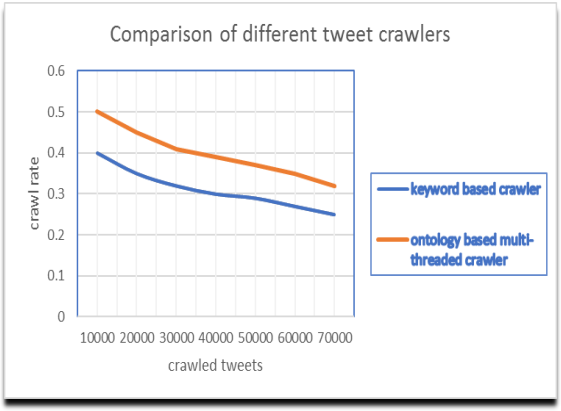


Fig 2. Comparison of different tweet crawlers

It is shown in the figure 2 that the keyword based crawler is compared with the ontology based multi-threaded crawler. It is observed that the performance of the keyword based tweet crawler is lesser than the ontology based multi-threaded crawler. In the graph, it is very clear that the performance decreases when the size of the crawled tweet increases. In case of ontology based multi-threaded tweet crawler, though the size of the crawled tweets decreases, the overall performance is comparatively better than that of keyword based tweet crawler.

V. CONCLUSION

Due to the widespread usage of social networking applications and many social media websites, the data is generated uninterruptedly. These led to the space for collecting those data and performing analysis. Data collection is the ultimate task for performing the analysis. Tweet crawler is required to be designed as a self-adaptive one for effectively crawling the dynamically generating tweets instantly without omitting any data. Thus, in this work an ontology based multi-threaded tweet crawler has been proposed and experimentally evaluated with the normal keyword based tweet crawler and the comparative results have concluded that the performance of multi-threaded crawler is far better than the typical tweet crawler. In future, the ontology based multi-threaded crawler may be enhanced to traverse with various hybrid ontology to generate multiple level threads to effectively crawl the tweets generated on various topics.

REFERENCES

1) Farag, M.M.G., Lee, S. & Fox, “Focused crawler for events”, International journal on Digital Libraries, Springer, 2017, vol. 16, no. 4, pp. 1-17.
2) Liakos, P., Ntoulas, A., Labrinidis, A., “Focused crawling for the hidden web”, World Wide Web, Springer, 2016, vol.19, no.4, pp.605-631
3) S. Gupta and K. K. Bhatia, "CrawlPart: Creating Crawl Partitions in Parallel Crawlers," International Symposium on Computational and Business Intelligence, 2013, pp. 137-142.
4) A. I. Vasile, B. Pävăloiu and P. Dan Cristea, "Building a specialized high performance web crawler," 20th International Conference on Systems, Signals and Image Processing (IWSSIP), 2013, pp. 183-186.
5) A. Saroop and A. Karnik, "Crawlers for social networks & structural analysis of Twitter," IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application, 2011, pp. 1-8.
6) A. K. Santoso and G. A. P. Saptawati, "Using Twitter data to improve news results on search engine," International Conference on Data and Software Engineering (ICODSE), 2014, pp. 1-5
7) Huang, J., Peng, M., Wang, H., A probabilistic method for emerging topic tracking in Microblog strea m, World Wide Web, Springer, vol.20, no.2, pp.325-350.
8) Yuan, Z., Huang, C., Sun, X., A microblog recommendation algorithm based on social tagging and a temporal interest evolution model, Journal of Frontiers of Information Technology & Electronic Engineering, Springer, vol.16, no.7, pp.532-540.
9) Kai Gao, Er-Liang Zhou, Steven Grover, “Applied Methods and Techniques for Modeling and Control on Micro-Blog Data Crawler”, Applied methods and techniques for mechatronic systems, vol.452, Springer Lecture Notes in Control and Information Sciences,pp.171-188.
10) X. Wang, L. Tokarchuk, F. Cuadrado and S. Poslad, "Exploiting hashtags for adaptive microblog crawling," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 2013, pp. 311-315.
11) J. Xia, W. Wan, R. Liu, G. Chen and Q. Feng, "Distributed web crawling: A framework for

crawling of micro-blog data," International Conference on Smart and Sustainable City and Big Data (ICSSC), Shanghai, 2015, pp. 62-68.

12) Hernandez, Julio, Heidy M. Marin-Castro, and Miguel Morales-Sandoval. 2020. "A Semantic Focused Web Crawler Based on a Knowledge Representation Schema" Applied Sciences 10, no. 11: 3837. https://doi.org/10.3390/app10113837

13) D. Shen, H. Wang, J. Cao, P. Li and Z. Jiang, "The Design and Implement of High Efficient Incremental Microblogging Crawler," Fourth International Conference on Multimedia Information Networking and Security, 2012, pp. 537-540.