International Journal on Recent and Innovation Trends in Computing and Communication
Volume: 2 Issue: 10

ISSN: 2321-8169
3222 – 3225

Social Media Aggregator
Using a focused Crawler and a Web & Android UI

Vivek Patani.
Computer department,
K. J. Somaiya College of Engineering,
Mumbai, India.
vivek.patani@somaiya.edu

Rhythm Shah.
Computer department,
K. J. Somaiya College of Engineering,
Mumbai, India.
rhythm.shah@somaiya.edu

Vruksha Shah.
Computer department,
K. J. Somaiya College of Engineering,
Mumbai, India.
vruksha.shah@somaiya.edu

Riya Patni.
Computer department,
K. J. Somaiya College of Engineering,
Mumbai, India.
riya.patni@somaiya.edu

Nirmala Shinde.
Computer department,
K. J. Somaiya College of Engineering,
Mumbai, India.
nirmala.shinde@somaiya.edu

***Abstract*—**today we live in a digital age, with almost each and every one of us having at least one social media account. With things as trivial as what one had for breakfast to the happenings of the world, everything is discussed about on these social media accounts. This paper intends to suggest using this social media as a policing solution, as in contrast to other traditional media such as television and print media, social media offers velocity, veracity, variety, real-time and a large volume of information. A focused web crawler is used to crawl the internet and to stick to the relevant and required topics only. This crawler is integrated with a database to store the information and the information is projected on an android application as well as a web application for the user's perusal.

***Keywords*—** *aggregator, crawler, focused, social media*

*****

I. INTRODUCTION

The aggregation of relevant information is becoming difficult by the day, as the internet keeps on growing at an astronomical rate. As a result keeping tab of the relevant data and related trends is becoming a tedious task. The solution to this problem is the Social Media Aggregator which collects the relevant data and the related trends in a single portal and subsequently displays it. The data is collected using a focused crawler, following which this data is filtered according to the requirements and stored into a database. The data is then displayed on a Web UI (User Interface) as well as an Android UI, for ease of access.

A. *Motivation*

The data is collected using a focused crawler, following which this data, As the internet consists of myriad amounts of data, to keep tab of trending information from different sources is a tedious and difficult task and hence the motivation to an aggregator which would provide information from various sources, but without testing the authenticity of the information but mostly emphasizing on the relevancy of the information.

B. *Need for the aggregator*

To provide an easy access to the ongoing trend related to a certain domain. This can be better explained by an example, such as; Police can use this in an advantageous way by keeping a tab on the commoners though and can curb flow of wrong information as they are alerted with trends in the society. Sometimes, police can also understand what and how people think and can provide better service by taking into consideration the information provided by the aggregator.

C. *Structure of the paper*

The structure of the paper is based on the flow of technology used and integrated. Primarily the technology and its aspect are discussed which is then followed by explaining what we want to achieve through the social media aggregator. In detail explanation of the aggregator is listed. Then a brief outline of the proposal and its shortcomings are listed.

II. WORKING

In this section, we discuss the working of the aggregator by integrating the underlying technologies.

A. *Working amd architecture of the web crawler*

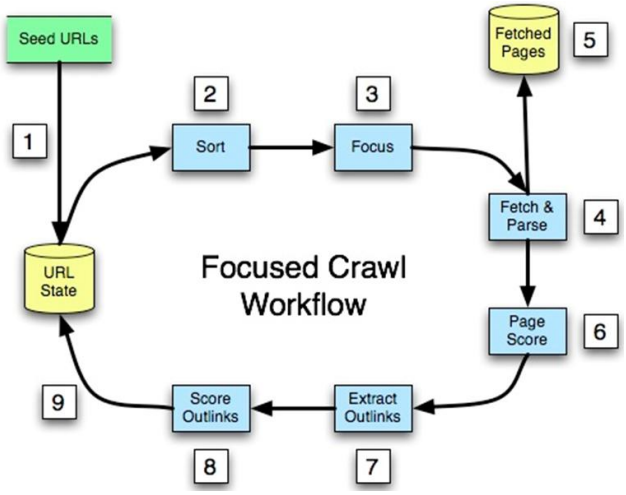

Figure 1: Architecture of Crawler

[5]Initially, the URL State database is loaded with a set of URLs. These URLs can be a broad set of domains with the

highest traffic or the result from some selective searches against some other index or manually handpicked URLs that point tot specific high quality pages. Once the URL state database is loaded, the first loop of the crawler begins. The prime step of all the loops is to extract all the unprocessed URLs and sort them according to their link score. Next is the critical step of deciding which how many URLs to process further in this loop. The fewer the URLs, the tighter the focus of the crawler. The selection can be based on a minimum link score, a fixed percentage of all URLs, a maximum count or a cutoff score that represents the transition point (elbow) in a power curve.

Now, having the set of accepted URLs the fetching process begins which entails all of the usual steps required for polite & efficient fetching, such as robots.txt processing. Pages that are fetched are normally stored in the fetched pages database and are then parsed. The content of the parsed page is given to the page scorer, which returns a value representing how closely the page matches the focus of the crawl. Normally this is a value from 0.0 to 1.0, with higher scores being better. Once the page has been scored, each out-link found in the parse page is extracted and the score for the concerned page is divided among all of its out-links. Finally, the URL State database is updated with the results of fetch attempts (succeeded, failed), all newly discovered URLs are added, and any existing URLs get their link score increased by all matching out-links that were extracted during this loop. At this point the focused crawl can terminate, if sufficient pages of high enough quality (score) have been found, or the next loop can begin. In this manner the crawl proceeds in a depth-first manner, focusing on areas of the web graph where the most high scoring pages are found.

III. Goal Of Aggregator

The Social Media Aggregator uses the Web Crawler to collect information from the internet. When the Web Crawler is first executed it requires the following - 1.) Seed URLs & 2.) Keyword to be searched and tracked. The Crawler provides the user with two kinds of information - 1.) The trending topics & 2.) The sensitive (searched) words with their frequency. The list of sensitive words is already stored prior to running the crawler and is provided by the client as per their requirements. For e.g.:- If a certain company would like to know how much and what is discussed about them and their products in the media and social networks, they would provide the list of sensitive words as their products and Seed URLs as the news portals and social media websites.

A. Proposal & Integration of the Technology

Firstly the java file is executed in order to initiate the process of crawling. Crawler now goes through the various seed URLs and searches for the keywords provided. The data is then inserted into the database through executing SQL Queries which are incorporated in the java file. The log-in credentials of the database are required by the crawler, in order to store the data dynamically. A certain time period is defined for the crawler to run periodically. Generally, one run of the aggregation process on an average consumes 2 hours.

After the crawler completes its iteration and stores the data into the database, the data then needs to be accessed by the Web database and the Android SQLite. The refresh rate of the Android as well as the Web UI needs to be defined; so as to regularly update new data to the intended user, also the user has the option to manually refresh the page.

The proposal of integrating various technologies will result into an outcome as:

- The Word list (Searched or sensitive) and the corresponding frequency of the words.
- Trending topics to a related domain as specified by the user.

IV. Design and Implementation Issues

A. Spam – There are millions of users on social websites such as Facebook, twitter, tumblr etc. and thus have their fair share of spammers. Social spam is unwanted spam content appearing on social networks and any website with user-generated content (comments, chat, etc.).It can be manifested in many ways, including bulk messages, profanity, insults, hate speech, malicious links, fraudulent reviews, fake friends, and personally identifiable information. Hence for its use as a policing solution the aggregator should be able to identify these spam users. For this a duplicate detection algorithm called LSH-with-filtering can be used. It treats the pairs of tweets whose minhash similarities are larger than the threshold as duplicated ones.[7]

B. Duplicate Pages – When the focused crawler crawls the internet it looks for relevant pages and then stores them in a database. If there are duplicate pages the crawler will store them both, hereby increasing the index storage space and he computation cost. The presence of near duplicate web pages thus plays an important role in this performance degradation while integrating data from heterogeneous sources. By introducing efficient methods to detect and remove such documents from the Web not only decreases the computation time but also increases the relevancy of search results. One solution for this is finding these near duplicate pages using minimum weight overlapping method. [8] It uses a TDW matrix based algorithm having three phases- rendering, filtering and verification, which receives an input web page and a threshold in its first phase, prefix filtering and positional filtering to reduce the size of record set in the second phase and returns an optimal set of near duplicate web pages in the verification phase by using Minimum Weight Overlapping (MWO) method.

C. Relevancy of a Page – When deciding whether a page is relevant enough to be stored or not, the crawler checks out the score of the page. However the way these pages are scored is one of the most important aspects of any focused crawler. There are many algorithms available for scoring pages such as genetic algorithm, page rank algorithm, naïve bayes classification algorithm and so on. One such method is classification of links using decision tree induction and neural network classifiers to improve the performance of focused

crawler.[9] Depending upon the requirements and the environment the best fit should be chosen to improve the efficacy of the crawler.

V. TECHNOLOGICAL ASPECTS

This section of the paper discusses the various underlying technologies and their usage with respect to the project at hand.

A. *List of Technologies Used* (though not a complete list but sufficient),

- SQL (Structured Query Language): To create, delete, update entries into the Database.
- Java: To design the Web Crawler.
- MySQL (Server): To store the Data collected by Crawler (Database).
- Eclipse (Java IDE / XML / SQLite): To design the Android Application.
- HTML/CSS/PHP: To design the Web UI and obtain statistics.

B. *Database Storage Aspect*

The web crawler keeps on downloading data and for the storage of this data, a database is essential. The Android UI will make use of the **SQLite** Database - connected to the master **MySQL** which stores data received from the Web Crawler. Queries written will insert the data into the database after applying the necessary filters. The security of data is maintained by SQLAuth as if the data is not protected, it is possible that it could become worthless through ad hoc data manipulation and the data is inadvertently or maliciously modified with incorrect values or deleted entirely. The crawler requires an average of ____GB/MB/KB/B of data. The storage of data follows a particular pattern to maintain a consistency in the Database. Serial Number is maintained as the primary key for accessing records from the Database. We select MySQL as it is open source and also holds good efficiency while input and output of data.

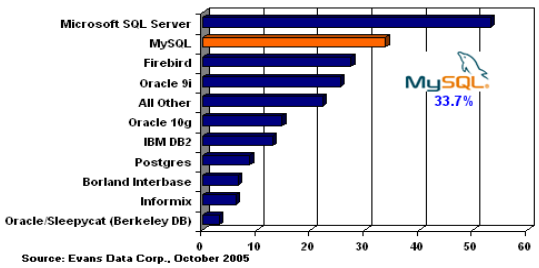

Figure 2: Comparison Chart of Databases

C. *Information Collection Aspect*

The Aggregation of information is based on the concept of a **Web Focused Crawler**. The Web Crawler is designed in the **Java** programming language and majorly consists of commands that can extract information from **HTML** tags. These are helpful in terms of collection of data and also assessing the relevancy of data and filtering the relevant data. The crawler crawls the seed URLs as provided by the programmer and runs up to 'n' levels deep depending upon the requirement. Constraints are applied in order to limit the crawling and to enhance the efficiency, as otherwise it would consume excessive amounts of time and data. The information collected is pertinent to the keywords stored in order to focus our crawling on the related topics and trends only. The frequency of the words is calculated as well and this is taken into consideration while determining the importance of the word and ranking the word on the **trending list**. This list provides us with the **sensitive trending words**. The crawler will for the most part crawl renowned news blogs and various Social Networking websites.

D. *Security Aspects*

Login- ID and password are used in order to gain access to the services. The password is bundled with the **md5 password encryption algorithm**. The android UI is similar as it tends to authenticate user using the same algorithm in order to maintain security of the information. The master database is also secured by the **SQLAuth**. To maintain consistency and exclusive access per user, it is suggested that the mobile number be used as Login - ID and those with exceptional cases should contact the DBA(Database Administrator).

E. *Information Viewing Aspect*

The information that is stored in the database can be viewed by the user in two ways, one is by viewing through the web browser (IE, Mozilla Firefox, Google Chrome or Safari) after logging in to the account & the second method is through the Android application after authorization. There are two different tabs on the Web as well as Android UI to see the list of trends to a related domain and the other tab to view the sensitive word list with frequencies.

VI. RELATED WORK

Due to its astronomical proliferation, social media is receiving a lot of attention now-a-days. The wealth of information it provides can be used for a wide range of application. As such there are number of people working on how best to use the social media. Among the most extensive work is the monitoring of the social network and the analysis of the retrieved information. Semenov, Veijalainen and Boukhanovsky proposed a Generic Architecture for a Social Network Monitoring and Analysis System that consists of three main modules, the crawler, the repository and the analyzer[10]. The first module can be adapted to crawl different sites based on ontology describing the structure of the site. The repository stores the crawled and analyzed persistent data using efficient data structures. It can be implemented using special purpose graph databases and/or object-relational database. The analyzer hosts modules that can be used for various graph and multimedia contents analysis tasks. The results can be again stored to the repository, and so on. All modules can be run concurrently.

Like us Papadopoulos & Kompatsiaris are working on Social Multimedia Crawling for Mining and Search[11] as social multimedia can be leveraged for a wide range of applications, but mining and search systems require innovative crawling solutions to meet both technical and policy-related obstacles.

Also Zhang & Nasraoui have put forth a Profile-Based Focused Crawler for Social Media-Sharing Websites[12] which treat users' profiles as ranking criteria for guiding the crawling process. It divides a user's profile into two parts, an internal part, which comes from the user's own contribution, and an external part, which comes from the user's social contacts. In order to efficiently and effectively extract data from a social media-sharing website for focused crawling, a path string based page-classification method was first developed for identifying list pages, detail pages and profile pages.

VII. ACKNOWLEDGMENT

We would like to thank our guide, Prof. Swati Mali for taking out the time to guide us throughout the project and for helping us as and when required and the K. J. Somaiya College of Engineering, Vidyavihar for providing us with the infrastructure and a platform to help us research and develop such a project.

VIII. REFERENCES

[1] Allen Heydon and Mark Najork, "Mercator: A Scalable, Extensible Web Crawler", Compaq Systems Research Center, 130 Lytton Ave, Palo Alto, CA 94301, 2001.

[2] Francis Crimmins, "Web Crawler Review", Journal of Information Science, Sep.2001. Twitter is the new police scanner http://www.popsci.com/technology/article/2013-04/twitter-is-the-new-police-scanner,2013.

[3] Shi Zhou, Ingemar Cox, Vaclav Petricek, "Characterising Web Site Link Structure", Dept. of Computer Science, University College London, UK, IEEE 2007.

[4] May, 2014: New technical report: "Online Social Media and Police in India: Behavior, Perceptions, Challenges". Authors: Niharika and PK.

[5] http://www.scaleunlimited.com/about/focused-crawler/

[6] Semantic Focused Crawling for Retrieving E-Commerce Information by Huang Wei, Zhang Liyi, Zhang Jidong and Zhu Mingzhu; http://www.researchgate.net/publication/42804620_Semantic_Focused_Crawling_for_Retrieving_E-Commerce_Information

[7] Qunyan Zhang; Haixin Ma; Weining Qian; Aoying Zhou, "Duplicate Detection for Identifying Social Spam in Microblogs," *Big Data (BigData Congress), 2013 IEEE International Congress on* , vol., no., pp.141,148, June 27 2013-July 2 2013

[8] Midhun Mathew, Shine N. Das, "An Efficient Approach for Finding Near Duplicate Web Pages Using Minimum Weight Overlapping Method", *ITNG*, 2012, 2012 Ninth International Conference on Information Technology: New Generations (ITNG), 2012 Ninth International Conference on Information Technology: New Generations (ITNG) 2012, pp. 121-126.

[9] Goyal, D.; Kalra, M., "A novel prediction method of relevancy for focused crawling in topic specific search," *Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on* , vol., no., pp.257,262, 12-13 July 2014

[10] Semenov, A; Veijalainen, J.; Boukhanovsky, A, "A Generic Architecture for a Social Network Monitoring and Analysis System," *Network-Based Information Systems (NBiS), 2011 14th International Conference on* , vol., no., pp.178,185, 7-9 Sept. 2011

[11] Papadopoulos, S.; Kompatsiaris, Y., "Social Multimedia Crawling for Mining and Search," *Computer* , vol.47, no.5, pp.84,87, May 2014

[12] Zhang, Zhiyong; Nasraoui, O., "Profile-Based Focused Crawler for Social Media-Sharing Websites," *Tools with Artificial Intelligence, 2008. ICTAI '08. 20th IEEE International Conference on* , vol.1, no., pp.317,324, 3-5 Nov. 2008