Procedia Computer Science

Procedia Computer Science 101, 2016, Pages 115 – 124

YSC 2016. 5th International Young Scientist Conference on Computational Science

Multitenant approach to crawling of online social networks

Nikolay Butakov,
Maxim Petrov,
Anton Radice
ITMO University, Saint-Petersburg, Russia
alipoov.nb@gmail.com, djvipmax@gmail.com, antonradice@gmail.com

Abstract

The importance of online social networks (OSN) and their data leads to the need to collect this data for different purposes. Restrictions imposed by various OSNs prevents obtaining this data in the required volume and time. Sharing credentials by many users in combination with different user needs and their request types can solve this problem, but in its own turn requires a new approach to organize such sharing efficiently and fairly among users. One of the most critical characteristics is throughput. In order for throughput to be fairly provided to users, sophisticated load balancing methods in addition to crawler architecture that has to manage multiple credentials and users must be developed. This work proposes a new approach that deals with the aforementioned issues and can improve characteristics of throughput for multiple users.

Keywords: Online Social Networks, Crawling, Multitenant, Social Networks, Load Balancing.

1 Introduction

Nowadays, modern OSN provide data which are needed by many stakeholders. However, to obtain this data, one needs to efficiently organize crawling process. One of the problems is the rate limits the crawler must take into account to obtain data. To overcome these limits, one possible approach is to use several credentials to collect data, but a user often cannot register too many credentials to perform crawling because it is restricted by the OSN itself. Additionally, at any given time, there may be different people who want to collect data who posses some set of credentials which are significant in their sum. Their need for credentials also varies in time, giving a user the chance to share their credentials with others when they don't need it and take more credentials when they require it, thus increasing overall performance. The second option especially may be the case when different users have different workloads in terms of the type of queries they are making, which is the case for Twitter and Instagram because of their restrictions on particular methods calls.

Peer-review under responsibility of organizing committee of the scientific committee of the
5th International Young Scientist Conference on Computational Science

doi:10.1016/j.procs.2016.11.015

In order to use such an ability, a new approach is needed to be able to manage credentials efficiently and fairly among different users. The crawler must control access to individual credentials and perform balancing among requests of different users that are dispatched to a set of credentials.

Fairness of such balancing has to be ensured in a way that each user gets access to credentials and throughput of their requests is close as possible to the throughput of others. The last can be expressed as a max min fairness.

In this work, a new approach is proposed to deal with the organization of crawling multiple online social networks (OSNs) for multiple users, using multiple credentials and taking into account different requests types and restrictions on them. The approach consists of an architecture of a multitenant crawler and workload balancing algorithm to control fairness of dispatching requests and throughput characteristics of multiple users.

This paper contributes with: (a) an architecture of a multitenant crawler that assumes organization of crawling leveraging multiple credentials; (b) a modification of the round robin algorithm to perform load balancing for multiple users, multiple credentials and multiple types of queries; (c) an experimental study that investigates efficiency of the proposed approach.

2 Related works

Currently, several classes of research aimed towards crawling support and automation can be found.

(Psallidas et al., 2013) proposed a general model of OSN that can be applied to collect data from multiple OSNs. The authors also introduce a special framework which contains multiple workers that exploits a set of credentials to collect data. In contrast to our approach, their solution is dedicated to be used by a single user and the problem of credentials sharing is not stated and discussed.

(**Valkanas et al., 2014**) proposed a method of hybrid crawling which combines two different ways of crawling - streaming and searching - to crawl Twitter. Also, the author introduced the idea of a crawl flow - a description of crawling operations to get the desired data. However, this work doesn't highlight working with multiple users or management of accounts used for crawling.

There is a huge amount of studies (**L. Lopes et al., 2009, W. Galuba et al., 2010, A. Pak and P. Paroubek, 2010, Xiong et al. 2012, Li et al. 2012, Wang et al. 2012**) dedicated to the development of different crawlers that are designed for specific tasks such as study of URLs propagation, epidemic events in several countries and sentiment analysis. Despite the fact that these crawlers assume the use of multiple accounts, they are designed for a single user and the problem of workload balancing among multiple users aren't assumed to be solved.

The framework proposed in (**Bošnjak et al., 2012**)) allows one to build a platform that can be used by multiple users to monitor and gather information of a pre-defined set of users, which can be expanded during the monitoring. Also, the framework is a distributed system and assumes using multiple accounts to retrieve information. However, there is no way to add applications or workload of multiples users and this framework does not support other social networks than Twitter.

BlogoScope system (**Bansal & Koudas, 2007**) is able to give the information for users on the request from various OSNs. Such request describes information that needs to be searched for in the text of posts, e.g. topic or discussion. The closest analogue is a search engine like Google. The approach is used to serve multiple users, but workload balancing is not a point discussed by the authors at all.

All the mentioned works above present solutions for different applications which can be useful only in specific situations for a single user, but there is still no crawling system which can work with multiple OSNs and collect data using multiple credentials to handle requests which simultaneously come from multiple users.