FACULDADE ENGENHARIA
Departamento de Informática

SNCrawler V0.2
Proposta de Projeto

Orientador: Sebastião Pais(sebastiao@di.ubi.pt)

Objetives

Crawling social networks is a relatively new concept, one which needs to be performed carefully for ensuring compliance with data privacy norms of various social media sites. Most of these sites control the access of information to crawlers (as well as individuals) via APIs, though publicly available, non-sensitive information can also be extracted otherwise.

A crawler in a social network is a program that starts with a list of the user's page to visit. As the crawler visits these pages, it identifies all the friend pages embedded in the current page and then follows those just identified pages to discover more new pages. This process continues until some criteria are met.

However, there are many challenges when crawling Online Social Networks. For example, the complexity of today's web technologies (e.g., JavaScript and Ajax) makes it challenging when interpreting the content of a page. And, the access restrictions of most Online Social Network services (e.g., login requirements, limited view, API query limits) impose difficulty to crawl the network with sufficient amount of samples. On the other hand, privacy control policies do not allow a crawler to access the entire online social network.

The task of extracting and analyzing data from online social networks has attracted the interest of researchers. The most popular social network, Facebook, naturally gets the most attention from researchers, who measured some large-scale network properties of the Facebook graph through sampling, crawling, and other methods to collect network data.

The objective of this project is to continue to design and implement a crawler for the most popular Online Social Networks and provide practical recommendations to tackle the challenges during the crawling process.

One of the objectives of this project is to continue to design and implementation of a crawler for Online Social Networks, which is an automated program that systematically collects data on Online Social Networks. The data collected can be used to create news approaches for sentiment analysis. The second objective is to provide practical recommendations to tackle the challenges during the crawling process.

Workplan

T1 Preliminary Investigation and Initial Requirements Specification;

T2 Preliminary Research on the various approaches for extracting and analyzing data from online social networks;

T3 Research, Conceptualization and Experimental of the Public Platform for access at information;

T4 Research, Conceptualization and Experimental Development;

T5 Integration, Testing and Evaluation;

T1 Report Writing.

Academic Prerequisites

Interest about Artificial Intelligence and Natural Language Processing.

Assessment elements to deliver.

Source code and documentation of all code development; Project report.

Expected Results

* Open Cloud Computing Platform;

* Project report.

Contacts

Sebastião Pais (sebastiao@di.ubi.pt)