



Московский государственный университет имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра Системного Программирования

Лазарев Владимир Александрович

Исследование методов OSINT для поиска информации о человеке

Курсовая работа

Научный руководитель:
к.ф.-м.н.

Турдаков Денис Юрьевич

Научный консультант:

Яцков Александр Константинович

Москва, 2021

Аннотация

Исследование методов OSINT для поиска информации о человеке

Лазарев Владимир Александрович

Данная работа посвящена исследованию и разработке методов OSINT для поиска информации о человеке. Данная курсовая содержит описание реализованных методологий и повествует о созданных приемах извлечения информации.

В ходе работы были изучены и представлены существующие различные методы как по способу взаимодействия с сервисами: извлечение данных с web-страницы и посредством скрытого или открытого api; так и по типу сервиса: поисковый агрегатор и социальные сети.

Содержание

1	Введение	4
2	Постановка задачи	5
3	Обзор существующих решений	7
3.1	Поиск данных в поисковых сервисах	7
3.1.1	Google Dorks (Google Hacking)	7
3.1.2	Carrot2	8
3.1.3	Yippy	9
3.2	Поиск данных в социальных сетях	10
3.2.1	Maltego	10
3.2.2	ITools	11
3.2.3	FindThatLead	12
3.3	Универсальные приложения	13
3.3.1	Виток OSINT	13
3.3.2	Palantir	13
3.4	Выводы	13
4	Исследование и построение решения задачи	15
5	Описание практической части	16
5.1	Описание выбранного инструментария	16
5.1.1	Архитектура работы сборщиков в поисковых сервисах	16
5.1.2	Архитектура работы сборщиков в социальной сети LinkedIn	17
6	Заключение	19
	Список литературы	20

1 Введение

В разделе 1 сформулирована постановка задачи. В разделе 2 приведен анализ существующих решений методов поиска, сбора и анализа информации из открытых источников. В разделе 3 описано исследование и построение решения задачи. В разделе 4 приведено описание практической части курсовой работы. В конце документа сформулировано заключение.

2 Постановка задачи

Целью данной курсовой работы является исследование и разработка методов OSINT для поиска информации о человеке. Для решения задачи, ее можно разбить на несколько подзадач: сбор информации при помощи поисковых сервисов, сбор информации с помощью социальных сетей. В свою очередь каждую из подзадач также можно поделить на следующие части: определение структуры web-страницы и извлечение данных непосредственно из страницы, поиск более быстрого доступа к информации посредством открытого или закрытого аri.

В итоге для достижения поставленной цели необходимо решить следующие задачи:

- Поиск данных в поисковых сервисах:
 - Провести анализ литературы и существующих решений для извлечения данных из поисковых систем;
 - Разработать методы поиска и сбора информации из поисковых систем:
 - * Проанализировать структуру web-страниц поискового сервиса;
 - * Реализовать метод поиска и извлечения информации при помощи атрибутов web-страницы;
 - * Провести исследование о возможности получения данных из ресурса посредством открытого или закрытого аri;
 - * Если аri реализовано на стороне сервиса, то реализовать метод поиска и сбора посредством аri;
 - Получить тестовые данные от реализованных методов и провести анализ, исследование полученной информации;
- Поиск данных в социальных сетях:
 - Провести анализ литературы и существующих решений для извлечения данных из социальных сетей;
 - Разработать методы поиска и сбора информации из социальных сетей:
 - * Проанализировать структуру web-страниц социальных сетей;

- * Реализовать метод поиска и извлечения информации при помощи атрибутов web-страницы;
 - * Провести исследование о возможности получения данных из ресурса посредством открытого или закрытого API;
 - * Если API реализовано на стороне соц. сети, то реализовать метод поиска и сбора посредством API;
- Получить тестовые данные от реализованных методов и провести анализ, исследование полученной информации;

3 Обзор существующих решений

3.1 Поиск данных в поисковых сервисах

3.1.1 Google Dorks (Google Hacking)

Google Dorks¹ - это по сути та же самая поисковая система от Google. Отличие заключается только в том, что обычный пользователь вбивает типовые запросы а-ля "Какая погода в Москве? то Google Dorks позволяет использовать специальные запросы для получения конкретной информации. Google Dorks имеет множество операторов, которые можно использовать для составления очень гибких и точных запросов [1]. По факту, это запросы, с помощью которых можно проверить безопасность того или иного сайта, найти IP-адреса сервисов, камер. Весьма эффективна для поиска документации по ключевым словам, а также поиску людей с помощью тех же самых Google Dorks Queries.

Плюсы данной системы:

- быстрый и объемный поиск по ключевым словам.

Из недостатков системы можно определить следующее:

- составленный запрос выдаст перечень ссылок в интерфейсе поисковой системы, а не сами данные;
- перед использованием необходимо изучить синтаксис запросов;
- нет накопления собранной информации, нельзя отслеживать изменения (дельты);
- нет построения графа зависимостей объекта.

¹<https://www.google.com/>

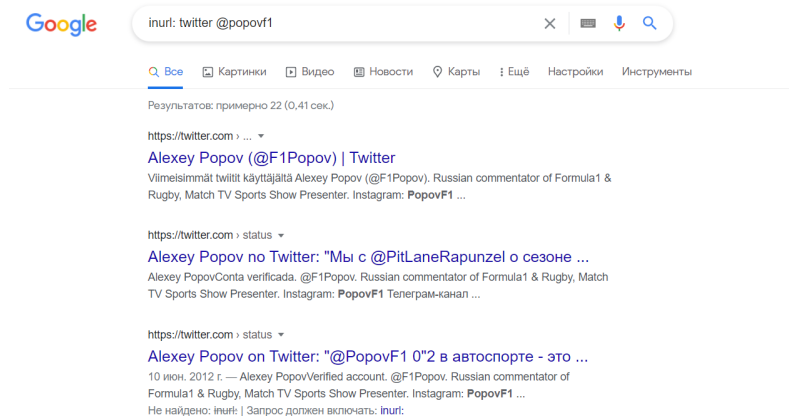


Рис. 1: Пример использования GDQ для поиска человека.

3.1.2 Carrot2

Carrot2 - движок кластеризации результатов поисковых запросов с открытым исходным кодом. Carrot2 может самостоятельно группировать по категориям найденные документы или данные. Работает в свою очередь как обычный поисковик, то есть по указанному ключевому слову возвращает некоторое множество ссылок, затем которые группируются по категориям [2].

Преимущества:

- быстрый и обширный поиск по ключевым словам;
- автоматическая группировка данных в соответствии с категориями;
- наличие удобного интерфейса с возможностью просмотра древовидной карты и круговидной диаграммы.

Недостатки:

- как и в случае с Google Dorks, Carrot2 возвращает нам перечень ссылок на источники данных, а не сами данные непосредственно;
- невозможно произвести точечный поиск файлов и данных, как это реализовано в Google Dorks. Как следствие - большое количество лишней информации.

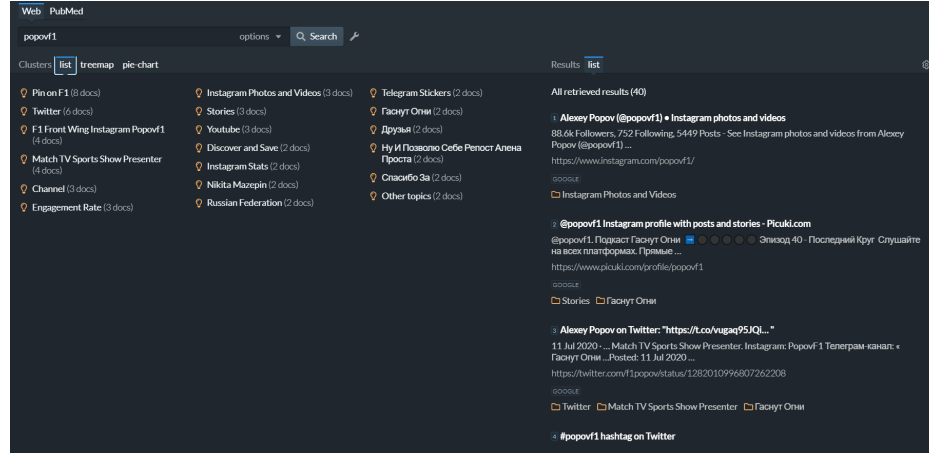


Рис. 2: Пример использования Carrot2 с разбиением результатов на группы.

3.1.3 Yippy

Yippy² - это метапоисковый движок, который группирует результаты поиска на категориям в группы. Наделен обширным функционалом: позволяет искать по ключевым словам новости, вакансии, правительственную информацию и блоги. Также позволяет вручную настраивать источники данных для собственного уникального метапоиска. [3]

Преимущества:

- группирует данные по тематическим категориям;
- есть возможность поиска не только ссылок в web-пространстве, но и непосредственно новостей, изображений и видео;

Недостатки:

- сервис недоступен на территории РФ;
- нет поддержки GDQ.

²<http://yippy.com/>

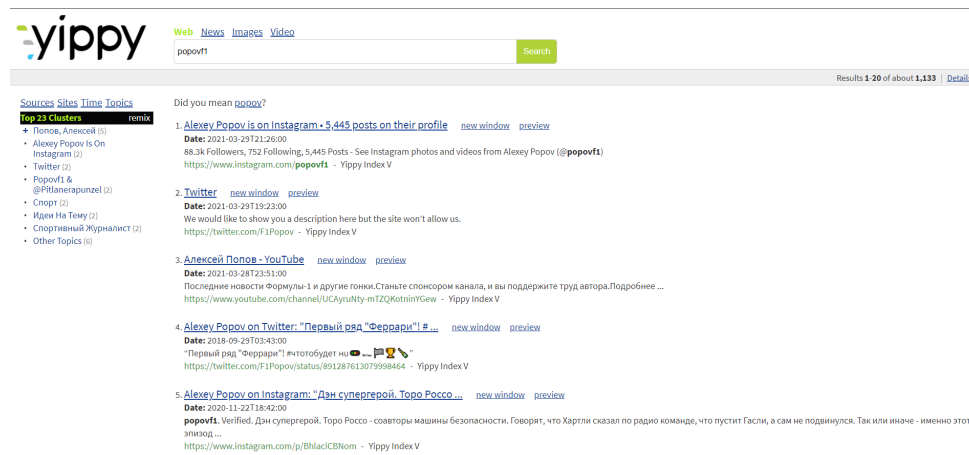


Рис. 3: Пример использования Yippy.

3.2 Поиск данных в социальных сетях

3.2.1 Maltego

Maltego³ - это комплексное решение с множеством поддерживаемых источников информации. Представляет из себя не движок, способный просто находить ссылки и группировать их, а проводит полноценный поиск и анализ данных, выстраивает деревья взаимосвязей. Например, может показать все активные адреса электронной почты заданного пользователя. [4]

Преимущества:

- выстраивание связей между объектами поиска, которыми могут быть как человек, так и группа лиц, компании, веб-сайты, организации и тому подобное;
- user-friendly интерфейс;
- возможность сохранения данных на стороне клиента с помощью СУБД;
- обладает гибкими настройками;
- является ПО с открытым исходным кодом, базовая версия которой поставляется абсолютно бесплатно в Kali Linux.

Недостатки:

³<https://www.maltego.com/>

- для доступа ко всем возможностям программы необходимо оплачивать лицензию.

3.2.2 ITools

iTools⁴ - это некий агрегатор всех инструментов, перечисленных выше. Имеет возможности искать по ключевым словам людей и организаций во многих популярных современных социальных сетях. Для каждого из подключенного метода поиска имеет свои настройки.

Преимущества:

- большой перечень источников информации с настройками для каждого из них.

Недостатки:

- нет никакой аналитики и сбора данных, просто поиск и ничего более;
- нет возможности запустить сбор по всем источникам одновременно;
- данные не собираются, не хранятся. Как следствие для полноценного использования необходимо будет писать ПО поверх данного сервиса;

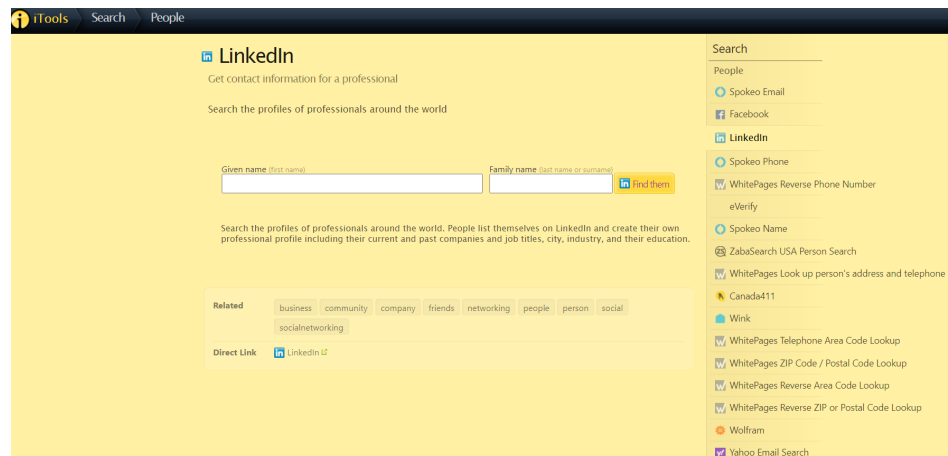


Рис. 4: Интерфейс агрегатора iTools.

⁴<http://itools.com/search/people-search>

3.2.3 FindThatLead

FindThatLead⁵ - это онлайн-сервис, позволяющий осуществлять поиск e-mail адресов и страниц пользователей в социальных сетях LinkedIn и Twitter. Обладает возможностью проверять валидность найденного адреса электронной почты. Главным отличием является то, что можно установить данное ПО как расширение браузера Chrome.

Преимущества:

- лаконичный и понятный интерфейс, наличие расширения для браузера;
- поиск e-mail адресов по профилю в социальных сетях.

Недостатки:

- анализ данных можно совершить только вручную;
- малое количество собираемой информации;
- не подходит для комплексного и обширного анализа сущностей.

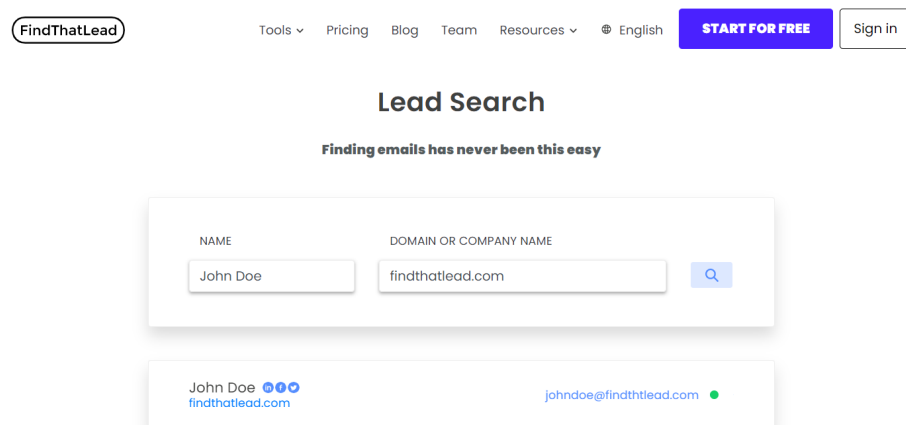


Рис. 5: Интерфейс FindThatLead.

⁵<https://findthatlead.com/en>

3.3 Универсальные приложения

3.3.1 Виток OSINT

Виток OSINT⁶ - это отечественное решение для спецслужб, позволяет собирать информацию с помощью поисковых сервисов, так и анализируя данные социальных сетей. Строит деревья зависимостей между объектами поиска, которыми могут быть: человек, организация, событие. Имеет индексацию и дедупликацию данных, в следствие чего система не перегружена излишками данных и повышает производительность. Вся информация также имеет привязку к географическому положению, что позволяет более наглядно воспринимать собранные и проанализированные ПО данные.

Главным и единственным недостатком является приватность системы, программы нет в свободном доступе и оценить ее возможности вживую не представляется возможным.

3.3.2 Palantir

Palantir⁷ - это зарубежное решение для спецслужб, делающее ставку прежде всего на безопасность собранной информации, удобную и развернутую подачу последней. Присутствует возможность как просто получать информацию из социальных сетей и прочих открытых источников, так и наблюдать за видеопотоком с камер наблюдений. Имеет визуализацию на карте мира.

Главным и единственным недостатком является приватность системы, программы нет в свободном доступе и оценить ее возможности вживую не представляется возможным.

3.4 Выводы

В результате исследования существующих методов сбора информации были выделены два подхода: поиск с помощью поисковых сервисов; поиск внутри социальных сетей. Однако большинство решений, которые производили поиск через поисковые сервисы, зачастую не могли предоставить полноценный сбор и анализ данных, которые можно было бы в последствии загрузить в СУБД для отображения в каком-либо интерфейсе.

⁶<https://norsi-trans.ru/catalog/vitok-osint/>

⁷<https://www.palantir.com/solutions/intelligence/>

Пожалуй, это главный недостаток приложений с таким подходом. Второй путь, поиск внутри соц сетей - зачастую реализован только в коммерческих проектах, и проверить объем извлекаемых данных невозможно.

4 Исследование и построение решения задачи

С целью исследования и разработки своих собственных OSINT методов сбора информации о человеке с помощью поисковых сервисов и социальных сетей предстоит решить следующие задачи:

- поисковые сервисы:

1. Определить структуру поискового сайта. В качестве таких сайтов возьмем следующие ресурсы:
 - DuckDuckGo;
 - Google;
 - Yandex;
 - Yahoo.
2. Извлечение найденных ссылок по заданному ключевому слову.
3. Сбор информации с сайтов по отобранным ссылкам.
4. Для случая с Google попробовать Google Search API: определить шаблон GET-запроса, структуру возвращаемых данных.

- социальные сети:

1. Определить структуру сайта социальной сети. Будем работать над социальной сетью LinkedIn.
2. Реализовать поиск и сбор данных пользователей и организаций посредством веб-краулинга сайта.
3. Реализовать сбор данных пользователей и организаций посредством закрытого API LinkedIn. Для этого потребуется:
 - реализовать вход систему через закрытое API посредством GET и POST запросов;
 - определить шаблон GET-запроса для получения данных по указанным ключевым словам, структуру возвращаемых данных.

- реализовать все указанные выше подзадачи в систему сбора данных.

5 Описание практической части

5.1 Описание выбранного инструментария

Работа была написана на языке Python, основной фреймворк для сбора данных - Scrapy, так как эта библиотека позволяет гибко настраивать параметры запросов, их обработку, генерацию cookie-файлов [5]. В качестве базы данных выступает MongoDB, поскольку она хранит данные в формате JSON-подобных документов [6].

Поскольку поиск в поисковых сервисах и поиск в социальной сети LinkedIn отличается по концепции и настройке пауков Scrapy, то они были выделены в 2 различных проекта.

5.1.1 Архитектура работы сборщиков в поисковых сервисах

Диаграмма классов приведена на рис. 7. (А рисунок то где!!!)



Рис. 6: Диаграмма классов сборщик в поисковых сервисах.

Система включает следующие 9 классов:

- FetchSpider - позволяет собирать все документы, изображения и html-код страницы;
- AbstractSearchSpider - содержит общие метода генерации запросов, обхода страниц и сбора данных с них;

- DuckDuckGoSearchSpider - реализует конструктор запуска сборщика для поискового сервиса DuckDuckGo и несколько специфичных констант, таких как шаблон url с query и CSS-селектор найденных ссылок;
- GoogleSearchSpider - реализует конструктор запуска сборщика для поискового сервиса Google и несколько специфичных констант, таких как шаблон url с query и CSS-селектор найденных ссылок;
- YahooSearch - реализует конструктор запуска сборщика для поискового сервиса Yahoo и несколько специфичных констант, таких как шаблон url с query и CSS-селектор найденных ссылок;
- YandexSearch - реализует конструктор запуска сборщика для поискового сервиса Yandex и несколько специфичных констант, таких как шаблон url с query и CSS-селектор найденных ссылок, настройки прокси;
- GoogleSearchApiSpider - реализует сборщик для поискового сервиса Google, который будет производить сбор с помощью Google API Search;
- GoogleAPICredentialsDownloaderMiddleware - данный класс производит неким проводником между Scrapy Engine и GoogleSearchApiSpider, в нем идет выбор API-ключа по стратегии "выбери тот ключ, у которого осталось наибольшее количество запросов" и обработка 429 ошибки (случай, когда API-ключ неожиданно превысил лимит использований и его необходимо признать невалидным, и запустить запрос с новым ключом);
- SplashFilesPipeline - выкачивает все файлы, которые были получены в ходе сбора, если отображенная ссылка была ссылкой не на html-страницу.

5.1.2 Архитектура работы сборщиков в социальной сети LinkedIn

Диаграмма классов приведена на рис. 6. (А рисунок то где!!!)



Рис. 7: Диаграмма классов сборщик в социальной сети LinkedIn.

Система включает следующие n классов:

- поиск и сбор с помощью навигации по атрибутам html-кода страницы и извлечение информации из атрибутов:

—

- поиск и сбор с помощью закрытого LinkedIn API:

—

6 Заключение

В данной работе были исследованы методы OSINT для поиска информации о человеке. Её решение было разбито на следующие задачи:

- Поиск данных в поисковых сервисах:
 - Провести анализ литературы и существующих решений для извлечения данных из поисковых систем;
 - Разработать методы поиска и сбора информации из поисковых систем.
- Поиск данных в социальных сетях:
 - Провести анализ литературы и существующих решений для извлечения данных из социальных сетей;
 - Разработать методы поиска и сбора информации из социальных сетей.

Список литературы

- [1] *ru.wikipedia.org*. Google hacking. — 2020. — Ноябрь. https://ru.wikipedia.org/wiki/Google_hacking.
- [2] *en.wikipedia.org*. Carrot2. — 2021. — Март. <https://en.wikipedia.org/wiki/Carrot2>.
- [3] *en.wikipedia.org*. Yippy. — 2021. — Февраль. <https://en.wikipedia.org/wiki/Yippy>.
- [4] *Опанюк, Игорь*. Maltego. Нароет все. — 2009. — October. <https://habr.com/ru/post/73306/>.
- [5] *Kouzis-Loukas, Dimitrios*. Learning Scrapy: Learn the art of efficient web scraping and crawling with Python / Dimitrios Kouzis-Loukas. — Packt Publishing, 2016. — Pp. 198–210.
- [6] MongoDB in Action, Second Edition / Kyle Banker, Peter Bakkum, Shaun Verch et al.; Ed. by Mihalis Tsoukalos. — Manning Publications, 2016. — Pp. 75–97.
- [7] *Ольга, Дзюба*. OSINT: что это, кому он нужен, какие методы сбора и типы информации использует? — 2020. — Август. <https://yushchuk.livejournal.com/1451268.html>.
- [8] *Карев, Антон*. SHODAN: САМЫЙ СТРАШНЫЙ ПОИСКОВИК ИНТЕРНЕТА. — 2018. <http://samag.ru/archive/article/3714>.
- [9] *Шагаев, Иван*. Поисковая система Shodan не то, чем кажется. — 2018. — Май. https://www.anti-malware.ru/analytics/Threats_Analysis/Shodan.
- [10] *kali.tools*. theHarvester. <https://kali.tools/?p=2286#:~:text=theHarvester>.
- [11] <https://www.spiderfoot.net/>. SpiderFoot: OSINT Automation. — 2019. — Сентябрь. https://ai-news.ru/2019/09/spiderfoot_osint_automation.html#:~:text=SpiderFoot.
- [12] *geocreepy*. Creepy. <https://www.geocreepy.com>.

- [13] <https://jivoi.github.io/>. Awesome OSINT. — 2021. <https://github.com/jivoi/awesome-osint>.
- [14] *Kozhuh*. Что такое Google Dorks? <https://spy-soft.net/gugl-dorki/>.
- [15] *Goossens, Michel*. The L^AT_EX Companion / Michel Goossens, Frank Mittelbach, Alexander Samarin. — Reading, Massachusetts: Addison-Wesley, 1993.