А.И. Мартышкин

АЛГОРИТМЫ ПОИСКА ДЛЯ СБОРА ДАННЫХ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ

Россия, г. Пенза, Пензенский государственный технологический университет

Modern social networks provide the data that many stakeholders need. But to get this data, you need to effectively organize the process of extracting it from social networks. One of the problems is the speed limits that the program must take into account to get the data. To overcome these limitations, one possible way is to use multiple credentials for data collection. However, the user is often unable to register too many credentials to perform the crawl, as this is controlled by the social networks themselves.

The article describes search algorithms for collecting data from social networks: adaptive search algorithms are proposed, namely: an algorithm for identifying new keywords and an algorithm for contextual monitoring, together with a mathematical model. These algorithms have been tested, which significantly expand the possibilities of data collection.

Введение. Современные социальные сети предоставляют данные, необходимые многим заинтересованным сторонам. Но для получения этих данных необходимо эффективно организовать процесс извлечения из социальных сетей. Одна из проблем – ограничения скорости, которые должна учитывать программа (поисковый робот, краулер) для получения данных. Чтобы преодолеть эти ограничения, одним из возможных способов является использование нескольких учетных данных для сбора данных. Адаптивный сбор данных – автоматическое уточнение запросов пользователя к социальной сети во времени на основе выделения информации, относящейся к контексту. Под контекстом подразумевается набор ключевых слов, пользователей, групп, постов, предоставленный пользователем. Адаптивный сбор данных решает проблему, когда у пользователя нет всей релевантной полной информации необходимой для сбора данных за счет решения следующих подзадач [1]:

- определения новых ключевых слов, связанных контекстом
- мониторинга связанных с контекстом групп, пользователей, постов

Алгоритм определения новых ключевых слов. Для реализации цели, необходимо ввести в архитектуру краулера дополнительные компоненты (рисунок 1):

- Распределенная очередь сообщений
- Распределенный фреймворк обработки данных

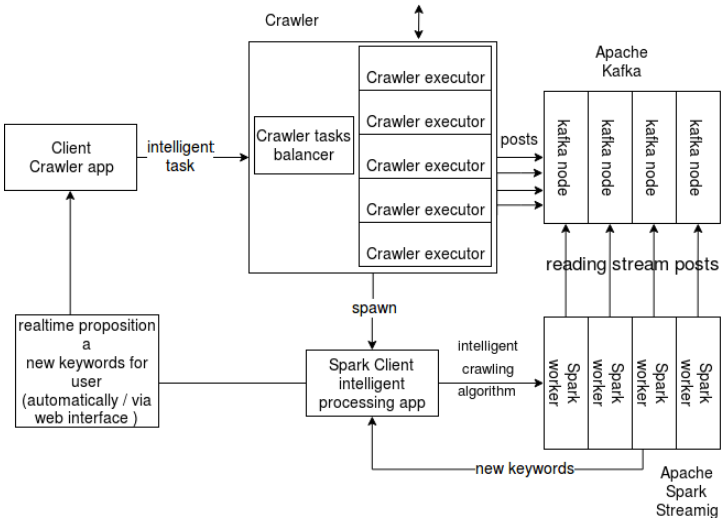

Рисунок 1 – Архитектура краулера

Распределенная очередь сообщений Apache Kafka позволяет параллельно обрабатывать данные пользователя в реальном времени распределенным фреймворком обработки данных Apache Spark Streaming [Spark Streaming – Spark 2.1.0 Documentation] [2, 3] с помощью следующего алгоритма (рисунок 2).

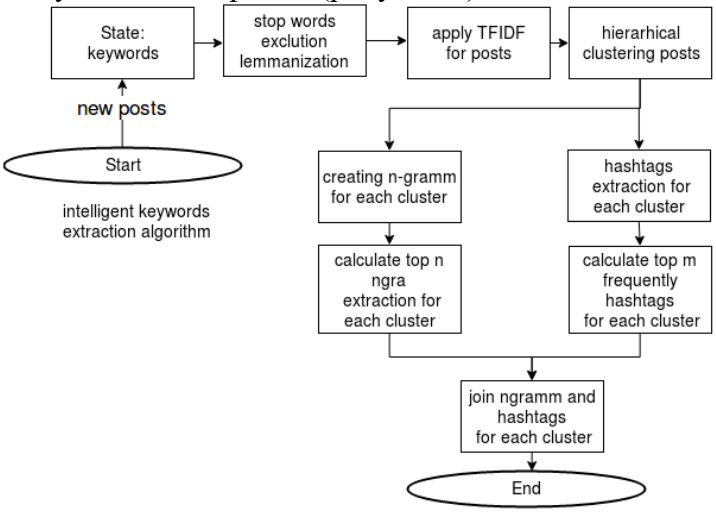

Рисунок 2 – Алгоритм определения ключевых слов

Алгоритм является потоковым алгоритмом данных с сохранением состояний. Состояние включает в себя все множество ключевых слов. На вход алгоритма подаются новые посты, затем из текста поста удаляются стоп-слова (предлоги и не несущие информации слова) происходит процесс лемманизации (приведение слова в нормальную форму). После чего, с помощью алгоритма TFIDF для каждого поста составляется future вектор и запускается процесс иерархической кластеризации для определения постов с похожей темой. Внутри каждого кластера извлекаются hash-теги и н- граммы (множество сочетаний слов). Затем, для каждого кластера постов происходит подсчет и сортировка наиболее часто встречающихся ключевых слов. В результате работы алгоритма новые ключевые слова предлагаются пользователю через веб интерфейс или автоматически попадают в приложения для дальнейшего использования.

Был проведен эксперимент: в социальной сети ВКонтакте [4] краулером собраны посты за 03.04.2021 с 14 до 17 часов по ключевому слову "спб". Интеллектуальный алгоритм предложил ключевые слова, заключенные в таблице 1.

Таблица 1 — Результат работы адаптивного алгоритма.

Ключевое слово	**Частота встречаемости**
#spb	4036
#питер	903
#санктпетербург	494
Санкт Петербург	494
наращивание ресниц	354
салон красоты	93

Алгоритм контекстного мониторинга. Пользователь может предоставить идентификаторы записей (постов), пользователей, групп, которые необходимо мониторить, либо же они могут появиться в ходе сбора данных как наиболее популярные и интересные для пользователя. Алгоритм контекстного мониторинга решает проблему остановки сбора данных на основе разработанной математической

модели, которая позволяет определить активность и актуальность сущности, находящейся под мониторингом.

Периодический процесс сбора данных по ключевым словам $T = \{k_1, ...k_z\}$ порождает создание множества постов $posts_i$ в интервале времени t_i

$$posts_i = crawling(OSN, t_i, T) = \{< pid, t_{i,} postScore >\}, \quad (1)$$

где pid – идентификатор поста, $postScore$ – функция счета поста, отражающая активность и актуальность поста: $postScore_{pid,i} = \alpha_1 \cdot cS_{pid,i} + \alpha_2 \cdot rS_{pid,i} + \alpha_3 \cdot lS_{pid,i}$.

Содержит 3 основные компоненты: функцию оценки активности комментариев, репостов и лайков поста. Коэффициент α введен для «взвешивания» компонент и позволяют более точно настроить модель. Поскольку функции оценки активности комментариев, репостов и лайков высчитываются схожим образом, а мониторинг можно производить независимо, будет рассмотрена только функция оценки активности комментариев.

Вводятся понятия перемещения (накопления), скорости и ускорения, которые помогают вычислить динамику процесса обновления информации. Сначала рассчитывается скорость $CV_{pid,i}$ как разница в количестве комментариев за определенный интервал во времени.

$$CV_{pid,i} = \frac{commentsCount_{pid,i} - commentsCount_{pid,i-1}}{t_i - t_{i-1}}, \quad (2)$$

$$CD_{pid,i} = \sum_{k=0}^{i-1} CV_{pid,k}, \quad (3)$$

$$CA_{pid,i} = CD_{pid,i} - CD_{pid,i-1}. \quad (4)$$

Функция оценки активности комментариев состоит из 3х взвешенных компонент накопления, скорости и ускорения:

$$cS_{pid,i} = \beta_1 \cdot CD_{pid,i-1} + \beta_2 \cdot CV_{pid,i-1} + \beta_3 \cdot CA_{pid,i-1}, \quad (5)$$

Остановка мониторинга происходит если функция счета достигла некоторого критического значения и определяется на основе выражения:

$$\begin{cases} true, cS_{pid,i} < \max(cS) \cdot \gamma \\ false \end{cases}, \quad (6)$$

Проведена апробация модели. Результаты активности можно увидеть ниже (рисунок 3).

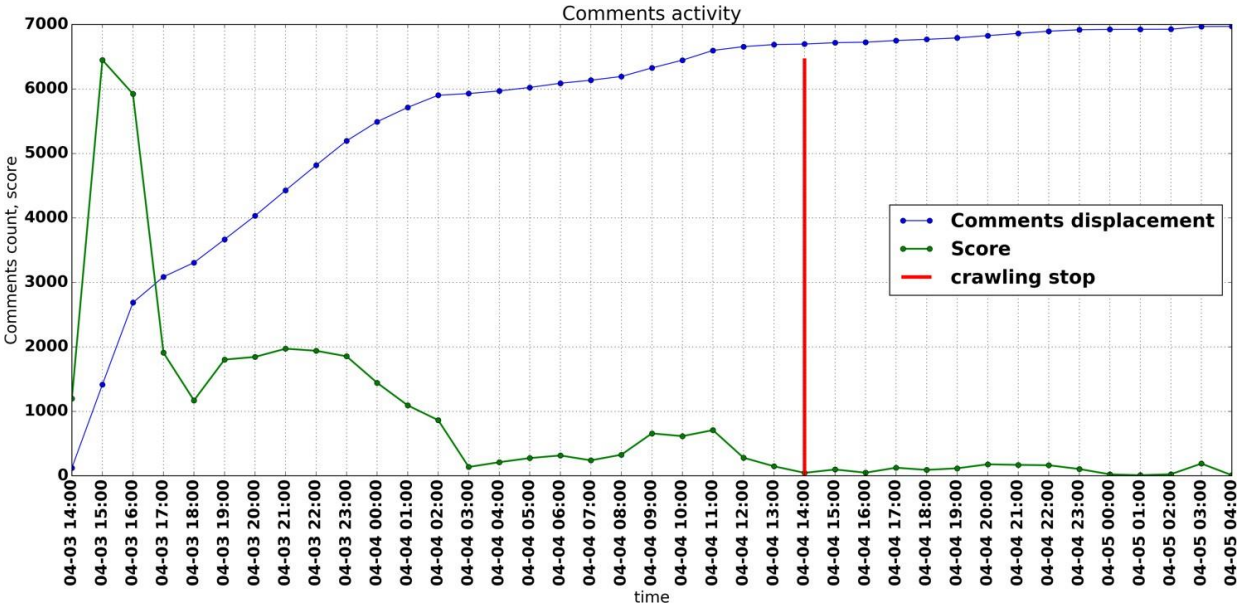

Рисунок 3 – Активность комментирования популярных записей в указанных группах

Можно выделить зоны активности: с 14 до 17 часов наблюдалась повышенная активность, затем следует спад активности и последующая зона остановки мониторинга. Следует отметить своевременность остановки мониторинга, так как количество новых данных пренебрежимо мало по сравнению с собранным (1%).

Выводы. В статье предложены алгоритмы адаптивного поиска, а именно: алгоритм определения новых ключевых слов и алгоритм контекстного мониторинга, вместе с математической моделью. Проведена апробация указанных алгоритмов, которые существенно расширяют возможности сбора данных.

1. Интеллектуальный анализ данных социальных сетей – YouTube [Электронный ресурс]. URL: https://www.youtube.com/watch?v=CGfIsVd5goQ (дата обращения: 28.03.2021).
2. GridGain Software Documentation [Электронный ресурс]. URL: https://www.gridgain.com/resources/documentation (дата обращения: 28.03.2021).
3. Github apache/samza [Электронный ресурс]. URL: https://github.com/apache/samza (дата обращения: 02.04.2021).
4. VK API Документация [Электронный ресурс]. URL: https://vk.com/dev/manuals (дата обращения: 28.03.2021).