Разработка алгоритма выявления сообщений о политических событиях

Соловьёв Дмитрий Андреевич
студент 1 курса магистратуры
Московского технологического университета
г. Москва
E-mail: solovey7887@mail.ru

Ключевые слова: Кластеризация, Алгоритмы поиска, Политические события, Информационные сообщения.

Задачи по разграничению сообщений, которые могут быть связаны с тем или иным политическим событием в мире, возможно, решить с использованием кластеризации. Стоит отметить, что задача по выделению искомой группы однотипных сообщений по конкретным сюжетам может иметь ряд отличительных особенностей.

Конкретные алгоритмы кластеризации должны иметь заранее выделенный объем информации. Они не предоставляют шанса добавить новый элемент поиска, без проверки всех уже выбранных кластеров обрабатываемой информации.

Принцип действия алгоритмов состоит в том, что количество кластеров, на которые необходимо разделить всю информацию, заранее оговорено. Это верно для решения многих задач, но для кластеризации информационных сообщений это не имеет значения.

Каждая поставленная задача обработки выделенной информации имеет высокую размерность. Это говорит о том, что большая часть способов кластеризации однозначно работает с информацией, которая задана как пространственный вектор Rn.[1] Анализ выделенных данных текстовой информации осуществляется как сопоставление имеющихся признаков с функцией-индикатором выбранного слова. То есть вся размерность поставленной задачи позиционируется суммарным количеством выбранных определений. В связи с этим величина размерности исследуемого пространства колеблется в пределах от 100 тысяч до 1 миллиона измерений, для нахождения нужной информации.[1] При этом вектор признаков сообщения имеет только малую часть ненулевых значений функций-индикаторов. Алгоритмы кластеризации данных текстовой информации приведены во множестве источников. Но самыми популярными для работы являются три вида алгоритмов:

- «k средних»;
- Иерархический;
- Scatter-Gather.

Каждый из них имеет свои минусы при использовании методики анализа сообщений. Они не приспособлены для работы в режиме онлайн, а производят обработку тех данных, которые система в данный момент выделила во всем потоке информации. То есть для работы с сообщениями это не совсем приемлемо, так как новые, которые только появились, сообщения не попадают в обработку.

Например, алгоритм типа «k средних» имеет предварительно заданное условие — заранее фиксированную величину обрабатываемых кластеров. Второй алгоритм — иерархический, требует циклическую обработку выбранной информации.

Исходя из выше указанных алгоритмов, предлагается использовать пошаговый алгоритм анализа имеющихся сообщений, который представлен ниже на рисунке 1.

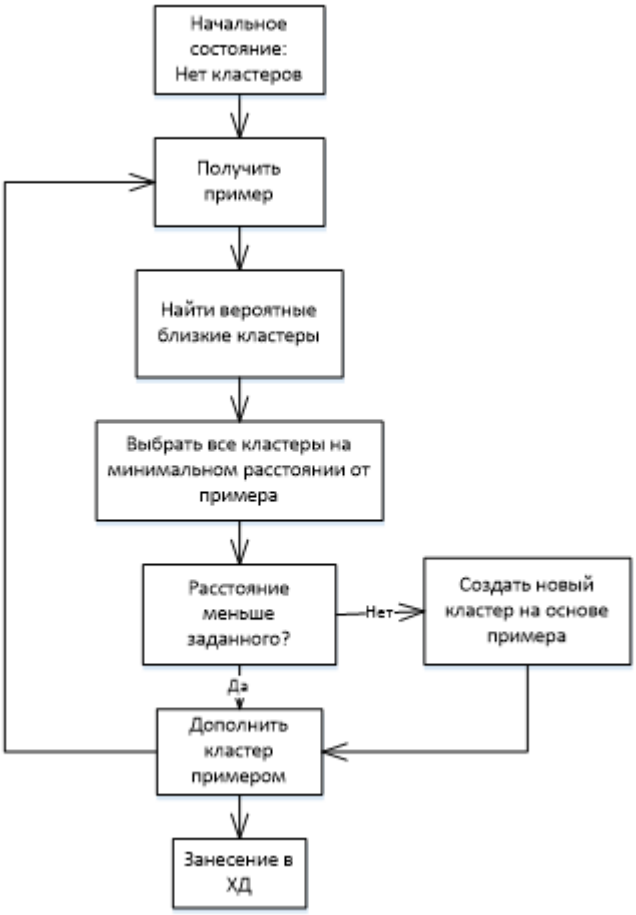

Рисунок 1 Алгоритм кластеризации.

Суть использования такой методики состоит в привлечении дополнительных параметров, которые можно извлечь из необходимого текстового сообщения, используя для этого метод глубокого лингвистического анализа. Главным параметром при этом служит признак. Затем идет метрика расстояния между сопоставленными точками в пространстве вектора Rn. Используемые метрики для работы: косинусное расстояние, евклидово расстояние, метрики на основе множеств.

За основу исследований берутся поверхностные показатели конкретных текстовых сообщений, которые довольно часто применяются при методике информационного поиска:

о совокупность слов;

о вес, рассчитанный по tf-idf;

о точное время и дата составления сообщения.

Дополнение базовой модели выражается в процедурах всестороннего анализа выбранных текстов на основе синтаксического анализа нужного сообщения. Анализатором служит гибридный алгоритм[2], который имеет в своей основе лингвистические понятия, представленные в виде правил и статистический анализ (робастность).

Результатом синтаксического анализа предложения является так называемое дерево синтаксической зависимости слов, которые включает в себя поисковая фраза.

Из этого всего можно сделать заключение о том, что кластеризация имеет не только поверхностные и линейные определения текста, а еще берет во внимание глубинные нелинейные. Для более точного показа синтаксических характеристик используется тройная схема «источник-

отношение-приемник». Источником и приемником выступают словоформы, имеющиеся части речи и леммы.

Для минимизации емкости ресурса главного алгоритма применяется процедура хеширования. Она используется взамен таблицы признаков и координат вектора. Это позволяет не хранить в памяти объем таблицы и варьировать величиной пространства признаков.

Список используемых источников

1. Feldman R., Sanger J. The Text Mining Handbook, Cambridge University Press, 2007.
2. Казенников А.О., Куракин Д.В., Трифонов Н.И. Гибридный алгоритм синтаксического разбора для системы анализа новостных потоков, Информатизация образования и науки № 1(13) 2012, 90-97.