электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.93'1

Реализация и исследование методов автоматической кластеризации текстовых документов с помощью методов машинного обучения

***Лыфенко Н.Д.**, аспирант*
Россия, 125993, г. Москва, Российский государственный гуманитарный университет, «Кафедра математики, логики и интеллектуальных систем в гуманитарной сфере»

Научный руководитель: Бобков А.В., к.т.н., доцент
Россия, 125993, г. Москва, Российский государственный гуманитарный университет
LyfenkoNick@yandex.ru

В настоящее время в связи с увеличением текстовой информации в сети Интернет возникает потребность в ее структурировании, тем самым повышается интерес к такой области искусственного интеллекта и компьютерной лингвистики, как автоматическая обработка естественного языка (*Natural Language Processing*). Поэтому задача автоматической кластеризации документов (разбиения информации на группы), решаемая в данном исследовании, представляется актуальной. Цель данного исследования заключается в реализации и исследовании методов автоматической кластеризации документов.

Под кластеризацией будем понимать разбиение массива информации на группы (кластеры) такие, что внутри кластера буду находиться наиболее близкие объекты. А расстояние между кластерами будет максимально.

Существуют различные методы кластеризации данных: *иерархические* (строят систему вложенных разбиения объектов на кластеры), *плоские* (строят одно разбиение), *нечеткие* (объект может принадлежать только одному кластеру) и *четкие* (объект относится к кластеру с некоторой вероятностью). В данной статье рассматриваются иерархические четкие алгоритмы кластеризации: плотностный *DBSCAN (Density-based spatial clustering of applications with noise)* и итеративный *к-средних,* определяющий количество кластеров во время работы с помощью метода аномального кластера.

Основная схема работы программы представлена на рис.1 и сводится к трем этапам: *получение, обработка данных* и *анализ результатов*. Необходимо получить вектор признаков с помощью лингвистических методов и применить алгоритмы кластеризации.

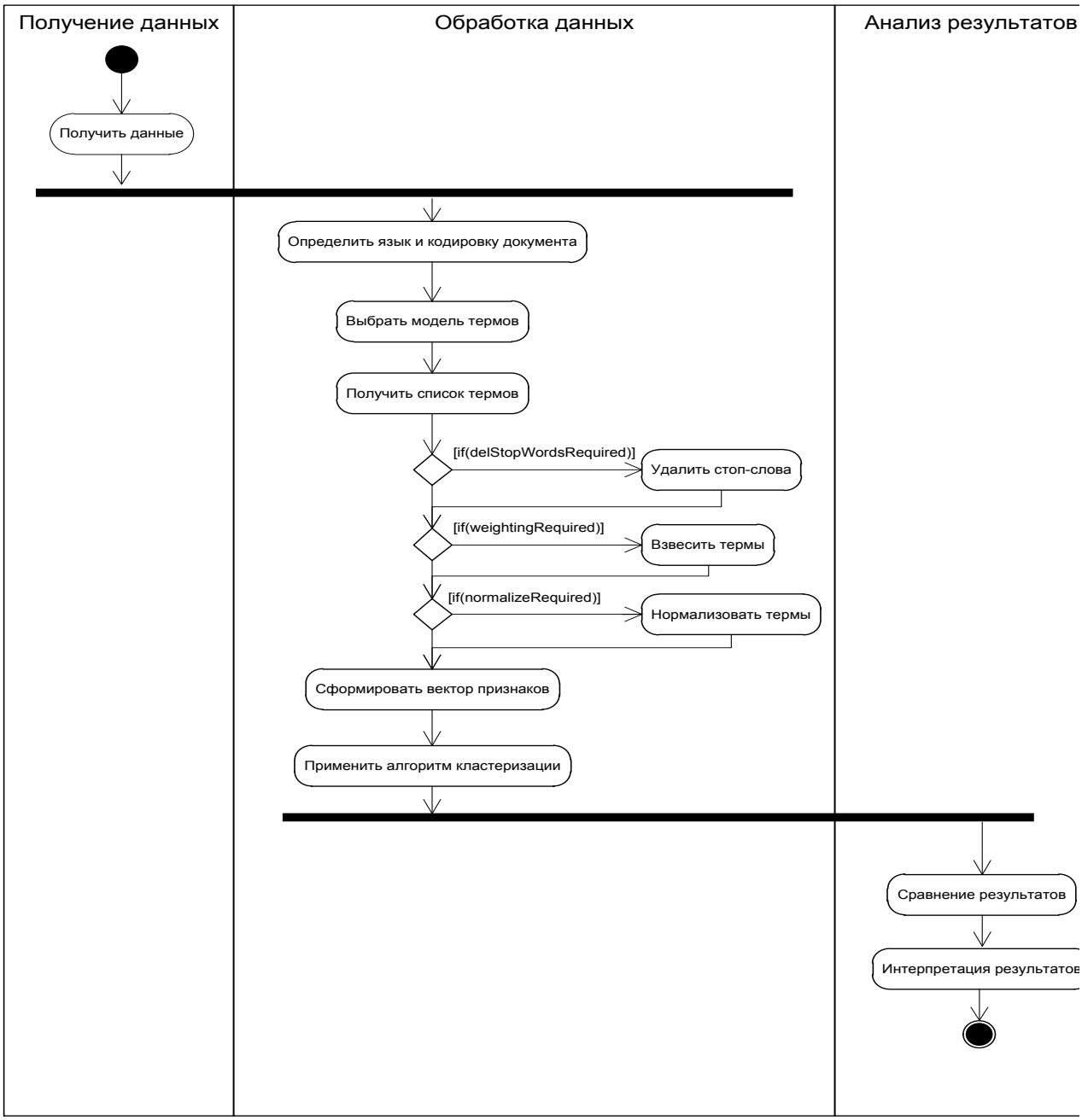

Рис. 1. Структурная схема работы программы автоматической кластеризации текстовых документов

Под первым этапом получения данных подразумевается конвертация документов в текстовый файл, содержащий только значимую текстовую информацию. Так зачастую объектами интереса являются различные новостные статьи, расположенные в сети Интернет. Очевидно, что теги в html–документе не несут семантической нагрузки и их можно не учитывать. Но обычно удаление тегов не приводит к безошибочному получению «чистого» текста из-за наличия большого числа рекламы, новостного мусора, комментариев, которые вносят большой шум в обучающую выборку. Поэтому был

предложен и реализован механизм более точного извлечения чистого текста с учетом тегов, который повысил качество получения текстовых данных из html–документов.

Поддержка, по крайней мере, двух и более естественных языков приводит к тому, что нужно иметь несколько списков стоп-слов и механизмов нормализации термов, поэтому перед кластеризацией текстовые данные нужно предварительно правильно обработать. Вследствие чего первым шагам определяется язык и кодовая страница документа. Используется статистический анализ для формирования гистограммы классов (язык, кодировка) и евклидово расстояние, косинус угла между текстами (векторами), как метрику близости. Каждый документ, представленный соответствующим вектором, сравнивается с вектором, характеризующим класс (кодировку или язык), и выбирается наиболее похожий.

Для работы с текстовыми документами обычно используют векторное представление (*Vector Space Model*), т. е. отображают текст в вектор. Данную процедуру можно осуществить несколькими способами. Наиболее популярными в области интеллектуальной обработки текста (*Text Mining*) являются: *мешок слов* (*Bag of Words*) и *учет взаимного положения слов*.

Мы используем модель мешка слов, под которыми понимаются *n*-граммы, т. е. словосочетания длины не более 3, т. к. данное представление хорошо зарекомендовало себя в задачах автоматической классификации [1,2] и использует простую математическую модель, которую можно эффективно реализовать. Получение списка термов (tokenization) проводится путем выделения термов, между которым есть разделитель (например, пробел). Под этапом взвешивания подразумевается приписывание большего значения некоторым термам. Для подсчета значимости слова в тексте (веса) применяется модель *tf-idf*.

$$W_i = tf_i * idf_i$$

Где W_i - вес i-го терма, tf_i - частота встречаемости *i*-го терма в данном документе (term frequency), $idf_i = Log\frac{N}{n}$ - логарифм отношения количества всех документов в коллекции к количеству документов, в которых встречается *i*-ый терм (inverse document frequency). Вес признака определяется тем, что чем больше локальная частота терма в документе (term frequency) и больше «редкость» (то есть чем реже он встречается в других документах) терма в коллекции (inversed document frequency), тем выше вес данного документа по отношению к терму. Большой вес в TF-IDF получат термы с высокой частотой в пределах конкретного документа и с низкой частотой употребления терма в других документах.

Зачастую получение всех термов из текста приводит к очень большой размерности

вектора признаков, и некоторые слова вносят шум в выборку. Поэтому стоп-слова, несущие в себе небольшую семантическую нагрузку, имеет смысл не учитывать. Чаще всего это — служебные части речи (союзы, предлоги) и наречия (например, вводные слова).

Отображение нескольких словоформ в одну лексему, также снижает размерность пространства и увеличивает качество классификации. Например, словоформы *столы, столе* отображаются в одну лексему *стол.* Данный процесс называется лемматизация, который существует наряду со стеммингом (отображения нескольких словоформ в одну основу). Это два варианта нормализации терма, которые используются в задачах интеллектуальной обработки текста.

В данной работе рассмотрены: классический метод *к-средних*, с процедурой выявления аномального кластера [3] для определения количества кластеров во время работы и плотностный алгоритм кластеризации — DBSCAN, распознающий кластеры не только сферической формы в отличие от метода *к-средних*.

В ходе исследования были проведены эксперименты, сравнивающие методы кластеризации данных (к-средних и DBSCAN) на предмет количества сформированных кластеров и их структуру. Данные были взяты со страниц пользователей из социальной сети «Вконтакте» (vk.com). Под документом (вектором) подразумевается отдельное сообщение на странице. Для анализа было взято 50000 документов суммарным объемом ~14 МБ.

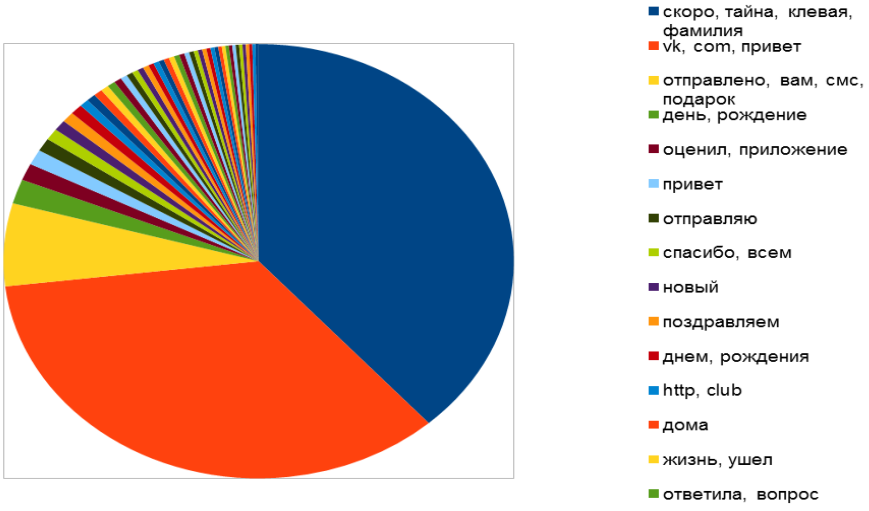

Рис 2. Результаты кластеризации с помощью метода *DBSCAN*

На рис. 2 представлен результат работы алгоритма *DBSCAN*. В легенде описаны ключевые слова для соответствующего кластера. С помощью алгоритма *DBSCAN* найдены два больших шумных кластера (19193, 17381 документов, соответственно). Остальные 46

кластеров имеют небольшой размер (<1000), но очень плотные, т.е. документы внутри кластера очень близки и в основном содержат дубликаты.

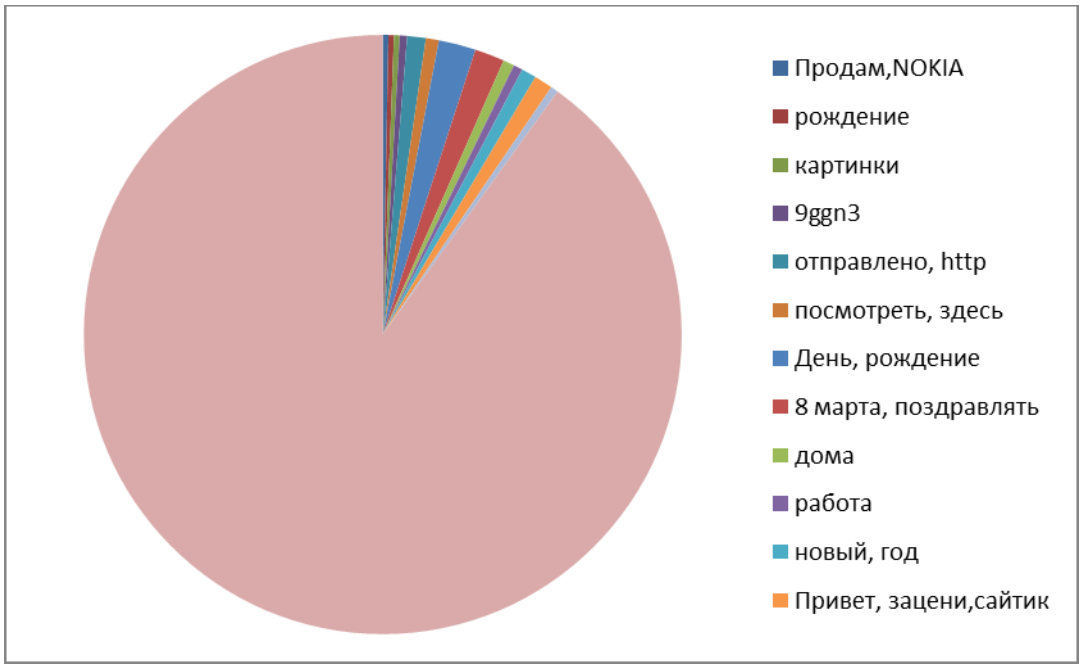

Рис. 3. Результаты кластеризации с помощью метода *к-средних*

На рис. 3 предоставлены результаты работы алгоритма *к-средних* на 1000 документов. Также получается «мусорный» кластер (85% документов) и несколько информативных.

По результатам сравнения двух методов можно сказать, что самыми большими и информативными оказались кластеры с метками «день рождение» и всевозможные статусы пользователей (на работе, дома и пр.)

Таким образом, в данной работе были описаны основные этапы обработки текстовых документов и представлены результаты экспериментов по сравнению методов автоматической кластеризации текстовых документов на естественном языке. Рассмотрен метод плотностной кластеризации *DBSCAN* и итеративный к-средних. Можно сделать вывод, что последний метод находит большее количество кластеров, по сравнению с плотностным на используемых данных. Документы кластера для метода *к-средних* содержат минимальное количество дубликатов, чего нельзя сказать о методе DBSCAN.

Результаты данного исследования планируется использовать в задачах структурирования информации новостного потока. В дальнейшем имеет смысл сравнить методы по метрикам информационного поиска (точность, полнота, f-мера) и провести больше экспериментов на различных источниках информации (новостные ленты,

аналитические статьи, публикации в блогах, форумах и социальных сетях).

Список литературы

1. Náther P. N-gram based Text Categorization, Diploma thesis. Institute of Informatics, Comenius University, Bratislava. 2005. 119 p.
2. Cavnar W.B., Trenkle J.M. N-Gram-Based Text Categorization // The Third Symposium on Document Analysis and Information Retrieval: proceedings. Las Vegas: University of Nevada, 1994. P. 366-376.
3. Amorim R.C., Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering // Pattern Recognition. 2012. No. 45(3). P. 1061–1075.