DOI: 10.34185/1991-7848.itmm.2020.01.031

КЛАСТЕРИЗАЦИЯ ИНФОРМАЦИОННЫХ НОВОСТНЫХ СООБЩЕНИЙ НА СОБЫТИЙНЫЕ ГРУППЫ

Горобец Д.В. к.т.н.

Институт технической механики Национальной академии наук Украины и Государственного космического агентства Украины

Аннотация. В работе рассмотрены вопросы обработки информационных сообщений. Выделение среди сообщений новостей, их классификацией по тематикам, формирование в группы новостных сюжетов, ранжирование новостных сюжетов по важности. Предложенный алгоритм основан на формировании для сообщений множества слов-маркеров и сравнением данных множеств, принадлежащих разным сообщениям между собой.

Ключевые слова: ТЕКСТ, ИНФОРМАЦИЯ, НОВОСТИ, КЛАСТЕРИЗАЦИЯ, КЛАССИФИКАЦИЯ.

В настоящее время большой актуальностью пользуются алгоритмы кластеризации больших массивов данных, которые генерируются социальными сетями, новостными и тематическими сайтами. Такие алгоритмы используются при решении задач более высокого уровня. Например, отслеживание причин социальных волнений на основе информационных сообщений в постах социальных сетей; выявление спроса на продукты и услуги, которых еще нет на рынке или еще не имеют должного распространения. Если рассматривать глобальную экономику, то применение таких алгоритмов позволяет по политическим заявлениям и фактическим показателям государств, производить анализ социальных и экономических изменений в конкретных регионах мира с последующим прогнозированием угроз и потенциальных возможностей для инвестиционной и деловой деятельности.

В данной работе рассматривалась задача анализа множества новостных сообщений с целью их кластеризации по тематике и схожести информационных поводов.

По условию задачи [1] дано множество файлов, содержащих текстовые сообщения с указанием источника информации, даты публикации, ее название и непосредственно само текстовое сообщение. Необходимо произвести

выделение информационных сообщений на русском и английских языках; произвести идентификацию сообщения, как новости; произвести классификацию новостей за тематикой; выполнить группировку новостей в новостные сюжеты; произвести ранжирование новостей по важности. При работе алгоритма запрещается обращаться к источникам данных, находящихся в интернете.

Решение подзадачи по идентификации языка, на котором написан текст, показало, что языков содержащих символы на латинице и кириллице довольно большое количество. Наиболее просто оказалось выделение английского языка из группы языков, содержащих латиницу. Выделение сводилось к отслеживанию отсутствия дифтонгов и умлаутов. Анализ языков с кириллистическими алфавитами показал, что наличие символа «ы» указывает на то, что данный алфавит принадлежит одному из трех языков: русскому, белорусскому или молдавскому. Русский алфавит из этих трех языков определяло условие отсутствия символов «і», «ў» (символы белорусского алфавита) и «ӂ» (символ молдавского алфавита).

Решение оставшихся подзадач приведены на примере русского языка, так как все полученные выводы справедливы так же и для английского языка.

Выделение новостных сообщений из множества данных возможно либо по наличию официального стиля в подаче информации, либо по источнику информационного сообщения. В первом случае составляется словарь словосочетаний характерных новостным сообщениям, например как «по сообщению», «согласно информации» и т.д. В другом случае производится предварительное составление списка достоверных источников информации экспертами, а затем проверка на принадлежность источника рассматриваемого сообщения данному списку. Сам по себе официальный стиль не является достаточным условием для идентификации информации, как новости, так как по сути можно формировать дезинформационные потоки данных в больших количествах, что соответственно будет проводить к существенному влиянию на конечный результат обработки данных. Поэтому действенным направлением остается использование списков достоверных источников информации.

Идентификация тематики производится составлением списков слов, присущих основным категориям, как «социум», «спорт», «технологии» и т.д

(для примера см. Таблицу). При решении данной задачи словари по тематикам создавались вручную и имели небольшой, но качественный объем слов. Если есть в наличии уже отсортированные новости по тематическим группам, то возможно создание таких списков на основе поиска часто встречающихся слов и отсеивания тех слов, которые присутствуют в новостях всех тематических групп.

Таблица – Тематические словари

Социум	Спорт	Технологии	Развлечения
ООН	болельщик	смарт	сериал
закон	спорт	процессор	кино
власть	побед	производительност	фестивал
импичмент	поражени	технолог	картин
междунар	тренер	сервер	портрет
сенат	тренировк	интернет	пейзаж
вотум	титул	разработч	музыка
переворот	инструктор	устройств	джаз
суд	чемпион	инженер	фильм
санкци	игр	конструкц	продюсер
канслер	игрок	оружие	режисер
президент	претендент	прибор	актер
правительств	форвард	патч	актрис
государств	судья	стартап	автор
министерство	лидер	startup	телеведущ
министр	соревн	систем	поэт
аннексия	кибер	мобильн	спектакл
парламент	игра	гаджет	кинематограф
палата	трек	ноут	концерт
совет	гоночн	планшет	книг
депутат	гонка	девайс	певец
чиновник	матч	портативн	певиц
чиновница	турнир	многофункц	публик
человек	команд	прогресс	звезд
МИД	счет	аккумулят	театр
…	…	…	…

Для сравнения новостей на предмет идентичности одному событию сначала формировался информационный «образ» текста. В него входит три списка слов, состоящих из слов, встречающихся в тексте. Список

нарицательных слов: слова, которые начинаются с заглавной буквы или аббревиатуры. Список характерных слов, которые образуют множество длинных слов, которые по статистике реже употребляются. И список слов из тематических словарей.

Для возможности корректного сравнения слов, последние символы слов опускаются, так как там находятся суффиксы и окончания. Принадлежность новостей к одному новостному событию определяется превышение отношения количества общих слов-маркеров к размеру меньшей статьи некоторого значения, выбранного экспериментальным путем. Варьирование этого экспериментально выбранного значения определяет узконаправленность событийной группы новостей. При этом учитывается, что новости, принадлежащие одной событийной группе должны иметь вхождение в ограниченный временной диапазон.

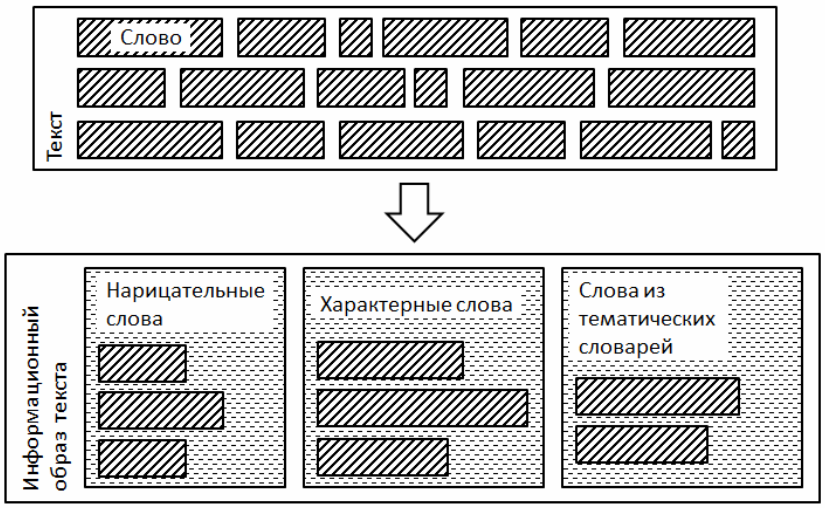

Рисунок – Формирование информационного образа текста

Сортировка сюжетных новостей по важности хорошо определяется количеством источников, которые представляют данный сюжет. Т.е. после выделения группы новостей в один информационный сюжет, ранжирование сюжетов производилось сортировкой по количеству новостей входящих в сюжет.

Необходимо отметить, что, несмотря на хорошую эффективность алгоритма, сам алгоритм не предполагает анализа смыслового содержания в тексте. Поэтому представленный алгоритм подвержен ошибочным срабатываниям при подаче некорректных или преднамеренно искаженных

данных. Результаты работы алгоритма на реальных данных можно найти по ссылке [2].

Литература

1. Data Clustering Contest: Round 1 // сайт Developer Challenges / Telegram. URL: https://contest.com/docs/data_clustering (дата обращения: 21.02.2020)
2. Результат роботы алгоритма на тестовых данных // сайт Developer Challenges / Data Clustering Contest / Telegram. URL: https://entry1178-dcround1.usercontent.dev (дата обращения: 21.02.2020)

CLUSTERING NEWS FEEDS FOR EVENT GROUPS

Horobets Dmytro

Abstract. The paper considers the processing of information messages. Highlighting news reports, their classification by theme, forming news stories in groups of news, ranking news stories by importance. The proposed algorithm is based on the formation of a set of marker words for messages and a comparison of these sets belonging to different messages among themselves.

Keywords: TEXT, INFORMATION, NEWS, CLUSTERIZATION, CLASSIFICATION.

References

1. Data Clustering Contest: Round 1 // site of Developer Challenges / Telegram. URL: https://contest.com/docs/data_clustering (access date: 21.02.2020)
2. The result of the algorithm’s functioning on test data // site of Developer Challenges / Data Clustering Contest / Telegram. URL: https://entry1178-dcround1.usercontent.dev (access date: 21.02.2020)