Двухшаговый метод объединения новостей в сюжеты

^{1,2} К.А. Скорняков, ORCID: 0000-0002-8218-4258 <kirill.skorniakov@ispras.ru>

^{1,2} А.С. Ласкина, 0000-0003-0878-7023 <laskina.as@ispras.ru>

^{1,3} Д.Ю. Турдаков, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

¹ Институт системного программирования им. В.П. Иванникова РАН,
109004, Россия, г. Москва, ул. А. Солженицына, д. 25.

² Московский физико-технический институт
141701, Россия, Москва, Керченская, д.1 А, корп. 1

³ Московский государственный университет имени М.В. Ломоносова,
119991, Россия, Москва, Ленинские горы, д. 1.

Аннотация. Работа посвящена разработке метода выделения сюжетов в новостях на русском языке. Сюжетом мы считаем группу новостей про одно событие реального мира. Предлагается двухэтапная схема кластеризации, при которой результаты первого «грубого» шага уточняются с помощью бинарного классификатора на парах новостей. В рамках работы создан размеченный на принадлежность сюжетам корпус новостей на русском языке, доступный для скачивания. На этом наборе данных показывается, что предложенный метод превосходит существующие решения по основным внешним метрикам кластеризации.

Ключевые слова: выделение сюжетов; кластеризация; новости

Для цитирования: Скорняков К. А., Ласкина А. С., Турдаков Д. Ю. Двухшаговый метод объединения новостей в сюжеты. Труды ИСП РАН, том 32, вып. 4, 2020 г., стр. 165–174. DOI: 10.15514/ISPRAS-2020-32(4)-12

Благодарности: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта No18-07-01059.

Two Step Method for Grouping News with Similar Topics

^{1,2} K.A. Skorniakov, ORCID: 0000-0002-8218-4258 <kirill.skorniakov@ispras.ru>

^{1,2} A.S. Laskina, 0000-0003-0878-7023 <laskina.as@ispras.ru>

^{1,3} D.Yu. Turdakov, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

¹ Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia

² Moscow Institute of Physics and Technology, MIPT,
Kerchenskaya, Moscow, 141701, Russian Federation

³ Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia

Abstract. Amount of news is rapidly growing up in recent years. People cannot handle them effectively. This is the main reason why automatic methods of news stream analysis have become an important part of modern science. The paper is devoted to the part of the news stream analysis which is called “event detection”. “Event” is a group of news dedicated to one real-world event. We study news from Russian news agencies. We consider this task as clusterization on news and compare algorithms by external clusterization metrics. The paper

introduces a novel approach to detect events at news in Russian language. We propose a two-staged clustering method. It comprises “rough” clustering algorithm at the first stage and clarifying classifier at the second stage. At the first stage, a combination of shingles method and naive named entity based clusterization is used. Also we present a labeled dataset of news event detection based on «Yandex News» service. This manually labeled dataset can be used to estimate event detection methods performance. Empirical evaluation on these corpora proved the effectiveness of the proposed method for event detection at news texts.

Keywords: event detection; clustering; news

For citation: Skorniakov K.A., Laskina A.S., Turdakov D.Yu. Two Step Method for Grouping News with Similar Topics. Trudy ISP RAN/Proc. ISP RAS, vol. 32, issue 4, 2020. pp. 165–174 (in Russian). DOI: 10.15514/ISPRAS-2020-32(4)-12

Acknowledgements. This work was supported by a grant from the Russian Foundation For Basic Research №18-07-01059

1. Введение

Объем новостных изданий и количество новостей в день растет с каждым годом. Это усложняет задачи мониторинга актуальной информации и анализа новостных потоков. Поэтому автоматизация этих процессов интересует исследователей последние 20 лет ([1–6]). Статья посвящена одной из подзадач анализа новостного потока – построению системы выделения сюжетов. Сюжетом мы будем называть группу новостей про одно событие реального мира ([1]). Мы будем рассматривать новости от русскоязычных новостных изданий, без дополнительных ограничений на их тематику. Такие тексты посвящены какому-либо событию в реальном мире (в отличие от постов в социальных сетях/блогах), написаны профессионалами с соблюдением правил орфографии и пунктуации. Мы рассматриваем задачу выделения сюжетов как задачу кластеризации и сравниваем ее с существующими решениями, используя внешние метрики кластеризации. Такая система может использоваться как отдельная сущность (новостной агрегатор), так и как вспомогательный модуль, позволяющий проводить анализ распространения информации на уровне сюжетов, а не отдельных новостей.

На наш взгляд, большинство работ про выделение сюжетов имеют один общий недостаток – недостаточно строго оценочную часть. В данной работе мы исправим этот недочет и покажем, что разработанная система превосходит существующие решения относительно основных метрик внешней оценки кластеризации для новостей на русском языке.

2. Методы выделения сюжетов

Методы выделения сюжетов различаются по наличию заранее заданных событий реального мира. Например, А.-М. Попеску (А.-М. Popescu) и М. Пеннакиотти (М. Pennacchiotti) [4] считают, что сюжеты связаны с заданным набором знаменитостей. В своей работе мы считаем, что у нас нет никакой априорной информации о связях новостей.

Методы можно разделить на 2 группы: без использования учителя и комбинированные методы (использующие методы обучения с учителем вместе с методами без учителя).

2.1 Выделение сюжетов без учителя

К группе методов обучения без учителя относятся работы, в которых авторы решают задачу выделения сюжетов как задачу кластеризации. В этой группе можно выделить два основных направления – потоковая и «групповая» кластеризации. При потоковой обработке каждое сообщение проверяется на принадлежность одному из существующих кластеров-сюжетов. При «групповой» на массиве сообщений запускается алгоритм кластеризации и полученные кластеры считаются сюжетами.

Типичным примером потоковой кластеризации является работа Дж. Шанкарараяна (J. Sankaranarayanan) и др. [7]. Авторы предлагают поддерживать список актуальных кластеров и проверять каждую входящую новость на принадлежность одному из них.

Р. Лонг (R. Long) и др. в [8] предложили метод «групповой» кластеризации, основанный на группировке новостей по «тематическим словам» (topical words).

В работе С. Петровича (S. Petrović) и др. [5] была предложена модификация алгоритма Local Sensitive Hashing (LSH, стандартный метод для «групповой» кластеризации) для выявления сюжетов в потоке сообщений Twitter.

2.2 Комбинированные методы

К комбинированным методам относятся методы, использующие классификацию «поверх» кластеризации для повышения качества ее работы. Т. Сакаки (T. Sakaki) и др. [9] использовали классификацию сообщений поверх поисковика Twitter, чтобы оставить сообщения, посвященным землетрясениям. Х. Беккер (H. Becker) и др. [6] предлагают объединять алгоритм потоковой кластеризации с классификатором, определяющим относятся ли выявленные кластеры к сюжетам или это набор случайных сообщений. В работе Дж. Г. Гонрада (J.G. Conrad) и М. Бендера (M. Bender) [10] авторы используют двухэтапную кластеризацию, где на втором этапе они уточняют грубые кластеры с помощью извлеченных векторов именованных сущностей.

2.3 Признаки

Для построения качественного алгоритма машинного обучения очень важен выбор подходящего признакового пространства. Рассмотрим подробнее, какие варианты признаков используют для решения задачи выделения сюжетов другие исследователи. Многие методы ([2, 3, 11]) выделения сюжетов основаны на представлении документа в виде «мешка слов» (bag of words) и его вариациях: мешок самых популярных слов, мешок именованных сущностей и т.д.

Часть методов использует различные виды графов (термин-текст, текст-текст и т.д.) для кластеризации, например, Т. Хуа (T. Hua) и др. в работе [12].

Другой популярный подход – использование тематического моделирования. При таком подходе сначала фиксируется число тем t , каждый текст представляется в виде вектора размера t . Элементы этого вектора – вероятности принадлежности текста к соответствующей теме. Так в [13], [14] для кластеризации сообщений в сюжеты используется популярный метод тематического моделирования – латентное размещение Дирихле (LDA) и его модификации.

2.4 Наборы данных

В значительном числе статей используется Twitter как основной источник для данных и тестирования. Для этого есть несколько причин: популярность в англоязычном Интернет, удобное API для скачивания, наличие хештегов (по ним удобно группировать сообщения).

К сожалению, они не подходят нам по нескольким причинам.

- Основная причина в том, что исследователи не выкладывают оригиналы текстов твитов (из-за правил Twitter¹). В редких случаях доступны id твитов, но их технически сложно получить из-за ограничений API Twitter, при этом твиты заблокированных/удаленных пользователей получить и вовсе невозможно.
- Твитты – это короткие тексты, в основном на английском языке, со своим языком. Такие тексты содержат сленг, опечатки. Тексты новостных изданий, напротив, средней длины,

¹ <https://developer.twitter.com/en/developer-terms/policy>

написаны с соблюдением правил языка, без использования узкоспециализированной лексики.

Однако сотрудники исследовательского подразделения Thomson Reuters (крупная медиакомпания) Дж. Г. Гонрад (J.G. Conrad) и М. Бендер (M. Bender) в работе [10] используют англоязычный набор новостей от Reuters для решения задачи объединения в сюжеты новостей Thomson Reuters. Также некоторые исследователи ([2, 3]) используют набор данных TDT², содержащий в себе распределение англоязычных и арабских новостей по сюжетам. Набор данных закрыт и предоставляется за отдельную плату.

2.5 Варианты постановки задачи

Определение сюжета – «набор новостей/текстов про одно событие реального мира» – достаточно общее, поэтому в различных работах можно встретить сужение этого определения на конкретную область. Кроме того, есть различные вариации выделения сюжетов.

По теме сюжетов работы делятся на:

- сюжеты о знаменитостях ([4]);
- сюжеты о землетрясениях ([9]);
- сюжеты в новостях из различных областей (политика, спорт, экономика и т.д., [2, 3, 5, 6, 8, 10, 11]);
- сюжеты, привязанные к гео-координатам/конкретному городу ([7, 12 – 14]).

По типу выделения сюжетов:

- определение новых сюжетов в новостном потоке (new event detection, [1 – 3, 5, 7, 14]);
- для фиксированного сюжета определение его продолжения в новостном потоке (event tracking, [1, 8, 11]);
- группировка всех новостей в сюжеты ([4, 6, 10, 12, 13]).

В нашей работе нас интересуют сюжеты о новостях из различных областей жизни с группировкой всех новостей в сюжеты.

2.6 Кандидаты на сравнение

Как видно из обзора, «ландшафт» методов выделения сюжетов достаточно обширен. Различаются как варианты постановки задачи, так и используемые в решениях признаки.

Ближе всего к нашей постановке находится подход, описанный в [10].

- Сюжет – группа новостей про событие реального мира
- В качестве признаков доступны только текст, заголовок и время публикации новости. Не используются специфичные для платформы признаки (гео-локация, хештеги, упоминания пользователей и т.д.).
- Используется кластеризация на группе новостей.

3. Двухэтапная кластеризация

Для сохранения баланса между скоростью и качеством работы мы используем двухэтапный комбинированный метод. На быстром, но «грубом» этапе формируются блоки из кандидатов на принадлежность одному сюжету. На следующем этапе происходит уточнение полученного разбиения с помощью классификации на парах объектов.

В качестве предварительной кластеризации мы использовали комбинацию двух методов:

- Local Sensitive Hashing (LSH) метод для сравнения текстов -- алгоритм «шинглов» ([15]).

² <https://catalog.lde.upenn.edu/LDC2005T16>

- «наивную» кластеризацию по именованным сущностям: в один кластер попадают новости, рассказывающие про одни именованные сущности.

За ними следует попарная классификация объектов каждого кластера с помощью алгоритмов машинного обучения с учителем.

Общая схема работы приведена на рис. 1.

Итоговый алгоритм выглядит следующим образом:

1. предобработка текста;
2. применение алгоритма шинглов; в один кластер попадают новости, имеющие хотя бы один общий супершингл (мы использовали 84 хэш-функции и супершинглы размером 4); подробнее см. в 3.1;
3. выделение именованных сущностей в тексте;
4. группировка текстов, имеющих общие именованные сущности; подробнее см. в 3.1;
5. объединение результатов 2 и 4 шагов; подробнее см. в разделе 3.2;
6. уточняющая классификация, подробнее см в 3.3;
7. объединение новостей в сюжеты после предыдущего шага.

Рис. 1. Общая схема
Fig. 1. General scheme

3.1 «Грубая» кластеризация

Перед кластеризацией проводится предобработка данных: удаление html тегов и других артефактов сборщика данных, удаление стоп-слов и пунктуации, лемматизация слов.

«Грубая» кластеризация состоит из объединения двух методов: шинглов и «наивной» кластеризации. Методы запускаются независимо.

Для методов шинглов мы использовали 84 хэш-функции, супершинглы размером 4, шинглы размером 1. Мы относили новости к одному сюжету, если у них совпал один супершингл.

При «наивной» кластеризации новости принадлежат одному сюжету, если имеют общую именованную сущность за исключением «шумовых сущностей». «Шумовыми» мы считаем такие именованные сущности, которые встречаются в большой доле новостей в корпусе. Для их отсека мы использовали пороговое значение такой доли в 0.8.

3.2 Объединение сюжетов

В процессе кластеризации одна и та же новость может попасть в несколько кластеров.

Мы считаем, что каждая новость должна принадлежать не более, чем одному сюжету и отношение «новости принадлежат одному сюжету» должно быть транзитивным, поэтому такие кластеры нужно объединить в один сюжет.

Для решения этой проблемы мы предлагаем построить граф связности новостей, где вершины – это новости, а ребра – отношение «принадлежат одному сюжету». Для этого достаточно пройти по всем парам внутри каждого кластера, тем самым получив список ребер такого графа.

Тогда для формирования итогового разбиения на сюжеты нам нужно найти все компоненты связности в таком графе. Мы использовали реализацию метода Д. Пирса (D.J. Pearce) [16]. Таким образом мы группируем новости в сюжеты так, что отношение «новости принадлежат одному сюжету» транзитивно.

3.3 Уточняющая классификация

Следующим шагом идет уточнение полученного разбиения новостей.

Для каждого блока мы будем образовывать всевозможные пары текстов, принадлежащих данному блоку.

Далее для каждой пары новостных текстов определим, принадлежат ли они одному сюжету или нет, т.е. решим задачу бинарной классификации, отмечая каждую пару метками «0» или «1».

Для этого мы используем логистическую регрессию на следующих признаках:

- коэффициент Жаккара (P. Jaccard) [17], Дайса (L.R. Dice) [18] между «мешком слов»;
- косинус между tf-idf представлениями документов;
- расстояние Левенштейна [19] относительно слов и букв;
- нормализованное расстояние сжатия ([20]).

После уточняющей классификации ребра, получившие метку «0», удаляются из графа и повторяется шаг 5, описанный в 3.2.

4. Эксперименты

4.1 Данные

В качестве данных для экспериментов мы использовали тексты новостей и информацию о разбиении новостей на сюжеты с сервиса «Яндекс Новости»³. Данные собирались с помощью краулера на основе фреймворка scrapy, аналогично работе [21].

```
{
  '_id': {'$oid': '5cddb5ba73619bf7289fcbd47'},
  'title': 'Ученый КФУ: Из-за сильной магнитной бури татарстанцы могут увидеть полярное сияние',
  'url': 'https://www.tatar-inform.ru/news/2019/05/14/650928/?utm_source=yxnews&utm...',
  'story_url': 'https://news.yandex.ru/story/Na_Zemle_zafiksirovana_moshhnejshaya_magnitnaya...',
  'text': 'Мощная магнитная буря, которая началась сегодня примерно в 6 часов утра по московскому времени, может вызвать полярное сияние, сбои в работе систем связи...',
  'agency': 'Татар-информ',
  'category': 'Общество|society',
  'platform': 'yandex',
  'type': 'news',
  'datetime': '1557878460'
}
```

Рис. 2. Пример новостной статьи
Fig. 2. News example

³ <https://yandex.ru/news/>

Скачанные данные представляют собой json-объект с информацией о названии, тексте, времени размещения, идентификаторе сюжета и прочих служебных полях для новостей. Пример приведен на рис. 2.

4.2 Формирование обучающей выборки

Для использования уточняющей классификации нам нужно построить набор пар новостей с метками «1» (принадлежат одному сюжету) и «0» (принадлежат разным сюжетам).

Для формирования положительных примеров мы использовали разбиение новостей на сюжеты в «Яндекс Новостях» (поле *story_url*) -- пара получает метку «1», если принадлежит одному сюжету в «Яндекс Новостях».

Основная проблема при создании подобных обучающих наборов из пар объектов – генерация отрицательных примеров. Стандартные подходы ([22]) предполагают отбор пар текстов с высоким коэффициентом схожести, но размеченные человеком как «0».

Однако наши эксперименты показали (мы не описываем эти предварительные эксперименты в этой статье), что в нашей задаче эффективней строить выборку иначе.

Так как основная цель нашего классификатора – исправлять ошибки кластеризации, то обучающая выборка формировалась по следующему алгоритму.

- 1. Получаем набор кластеров из алгоритма предварительной кластеризации.
- 2. Для каждого кластера строим все возможные пары новостей из этого кластера.
- 3. Пары, в которых новости принадлежат одному сюжету в «Яндекс Новостях» получают метку «1», остальные метку «0».

Такой подход к созданию негативных примеров показал себя лучше, чем отбор новостей из разных сюжетов с коэффициентом Жаккара выше порога.

4.3 Формирование тестовой выборки

Новости на «Яндекс Новостях» объединяются в сюжеты с помощью автоматических алгоритмов, поэтому в таком разбиении бывают ошибки.

Для исправления этих ошибок при построении тестовой выборки использовалась ручная разметка.

Идеальный вариант разметки новостей на сюжеты – попросить аннотаторов из массива новостей выделить группы сюжетов. К сожалению, он не исполним в реальности, т.к. требует огромной работы аннотаторов уже на сотнях новостей. Поэтому мы просили аннотаторов не группировать новости в сюжеты «с нуля», а исправлять предварительное разбиение, полученное в «Яндекс Новостях».

Процесс разметки был устроен следующим образом.

- Каждому аннотатору показывались по очереди новости из сюжетов «Яндекс Новостей».
- Для каждого сюжета аннотатору нужно было исключить новости, не подходящие к основному событию сюжета.
- Параллельно аннотатор отмечал номера сюжетов, описывающих одно событие.

Такая разметка позволяет построить такой набор кластеризованных новостей, в котором новости принадлежат одному кластеру, только когда они принадлежат одному сюжету (за счет шага 2). И такой набор кластеров, что новости из разных кластеров принадлежат разным сюжетам (шаг 3).

В результате разметки получился корпус из 547 новостей, размеченных на принадлежность сюжетам, доступный здесь⁴. Разметка проводилась двумя аннотаторами, коэффициент согласия каппа Кохена (J. Cohen) [23]: 0.99, что говорит о высокой степени уверенности

⁴ http://talisman.ispras.ru/wp-content/uploads/2020/09/news_events.json_gz

разметчиков. В спорных случаях привлекался третий аннотатор, и решение принималось большинством голосов.

В результате разметки из 51 исходного кластера получилось 70 новых (большая часть новых – кластеры размера 1, которые попали в изначальные кластеры по ошибке).

4.4 Метрики качества

Задача выделения сюжетов – это кластеризация, для которой известны метки настоящих кластеров на тесте. Поэтому мы будем использовать внешние метрики качества кластеризации ([24]), такие как: Adjusted Mutual Information (AMI), homogeneity, completeness, v-measure.

Дадим качественные характеристик этих метрик. Гомогенность (Homogeneity) максимальна, если все объекты в выделенных кластерах принадлежат одному сюжету, полнота (Completeness) максимальна, если все новости из сюжета попали в один кластер. V-мера (v-measure) равна среднему гармоническому гомогенности и полноты.

Скорректированная взаимная информация (AMI) основана на энтропии пар объектов, в зависимости от их попадания в разные кластеры. Принимает значение 1, когда выделенные кластеры совпадают с сюжетами, значение 0, когда разбиение случайно.

4.5 Сравнение с существующими работами

Основную сложность при тестировании систем выделения сюжетов представляет сравнение с существующими методами. Многие из них используют специфичные для конкретной площадки признаки (хештеги в Twitter), закрытые наборы данных. Код подавляющего большинства методов нельзя найти в открытом доступе. Также разные авторы по-разному ставят саму задачу выделения сюжетов (см. подраздел 2.5), что делает сравнение между такими работами некорректным.

Мы будем сравнивать наш метод с работой [10], по нескольким причинам:

- авторы группируют все новости в сюжеты;
- авторы рассматривают новости из различных областей;
- авторы не используют признаки, специфичные для конкретной площадки;
- авторы также используют комбинированный метод.

Введем следующие обозначения: Shingles – кластеризация на основе шинглов, clf – бинарный классификатор на втором шаге, Naive – «наивная» кластеризация, SSEC – Semi-Supervised Events Clustering [10].

Замеры проводились на 547 новостях, размеченных на принадлежность сюжетам.

Табл. 1. Сравнение методов выделения сюжетов

Table 1. Comparison of plot extraction methods

Метод	AMI	Homogeneity	Completeness	V-measure
Shingles	0.45	0.98	0.74	0.85
Naive	0.14	0.26	0.79	0.39
Shingles + Naive	0.12	0.22	0.80	0.34
Shingles + Naive + clf	0.81	0.91	0.89	0.90
SSEC	0.42	0.98	0.73	0.84

Из таблицы видно, что предложенный метод превосходит по трем из четырех метрик существующие решения. Это достигается за счет комбинации высокой гомогенности (Homogeneity) метода шинглов с полнотой (Completeness) «наивной» кластеризации.

Как мы видим, эта комбинация не работает без уточняющей классификации, т.к. из-за «наивной» кластеризации сильно падает гомогенность выделения сюжетов. Гомогенность

максимальна, если все объекты в выделенных кластерах принадлежат одному сюжету, полнота максимальна, если все новости из сюжета попали в один кластер.

5. Заключение

В ходе работы был разработан метод выделения сюжетов, превосходящий по качеству существующие решения. Была предложена двухэтапная схема, объединяющая методы кластеризации и классификации на парах новостей. Показано, что второй этап классификации пар позволяет комбинировать различные методы кластеризации, достигая нужного баланса между гомогенностью и полнотой. Был предложен способ генерации данных для качественного обучения классификатора для уточняющей классификации.

Также мы попытались исправить пробел в методике оценки качества методов выделения сюжетов и построили корпус новостей, позволяющий оценивать точность и полноту методов. Размеченные данные позволили нам оценить различные методы выделения сюжета, используя внешние метрики кластеризации.

В дальнейшем мы планируем расширить тестовый корпус и проверить применимость такого подхода на менее формальных текстах из социальных сетей. Также интерес представляет совместное выделение сюжетов в текстах новостных изданий и постах/комментариях обычных пользователей.

Список литературы / References

- [1]. J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194-218.
- [2]. T. Brants, F. Chen, and A. Farahat. A system for new event detection. In Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 330-337.
- [3]. G. Kumaran and J. Allan. Text classification and named entities for new event detection. In Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pages 297-304.
- [4]. A.-M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In Proc. of the 19th ACM International Conference on Information and Knowledge Management, 2010, pp. 1873-1876.
- [5]. S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In Proc. of the Annual Conference of the North American Chapter of the Association for Computational linguistics, 2010, pp. 181-189.
- [6]. H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: real-world event identification on twitter. In Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 438-441.
- [7]. J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In Proc. of the 17th ACM Sigspatial International Conference on Advances in Geographic Information Systems, 2009, pp. 42-51.
- [8]. R. Long, H. Wang, Y. Chen, O. Jin, and Y. Yu. Towards effective event detection, tracking and summarization on microblog data. Lecture Notes in Computer Science, vol. 6897, 2011, pp. 652-663.
- [9]. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proc. of the 19th International Conference on World Wide web, 2010, pp. 851-860.
- [10]. J.G. Conrad and M. Bender. Semi-supervised events clustering in news retrieval. In Proc. of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval, 2016, pp. 21–26.
- [11]. M. Mohd. Named entity patterns across news domains. In Proc. of the 1st BCS IRSG Conference on Future Directions in Information Access, 2007, 5 p.
- [12]. T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan. Sted: semi-supervised targeted-interest event detection in twitter. In Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1466-1469.
- [13]. K. N. Vavliakis, F. A. Tzima, and P. A. Mitkas. Event detection via lda for the mediaeval 2012 sed task. In Proc. of the Multimedia Benchmark Workshop, 2012, 2 p.

- [14]. X. Zhou and L. Chen. Event detection over twitter social media streams. The VLDB journal, vol. 23, no. 3, 2014, pp. 381-400.
- [15]. A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. Computer Networks and ISDNs, vol. 29, no. 8-13, 1997, pp. 1157-1166.
- [16]. D. J. Pearce. An improved algorithm for finding the strongly connected components of a directed graph. Victoria University, Wellington, NZ, Tech. Rep, 2005.
- [17]. P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles, vol. 37. 1901, pp. 547–579 (in French).
- [18]. L. R. Dice. Measures of the amount of ecologic association between species. Ecology, vol. 26, no. 3, 1945, pp. 297-302.
- [19]. В.И. Левенштейн. Двоичные коды, способные исправлять удаления, вставки и обращения. Доклады Академии Наук СССР, том 163, no. 4, 1966 г., стр. 845-848. / V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady, vol. 10, no. 8, 1966, pp. 707-710.
- [20]. R. Cilibrasi and P. M. Vitányi. Clustering by compression. IEEE Transactions on Information Theory, vol. 51, no. 4, 2005, pp. 1523–1545.
- [21]. А.К. Яцков, М.И. Варламов, Д.Ю. Турдаков. Сбор и извлечение данных с веб-сайтов СМИ. Программирование, том 44, no. 5, 2018 г., стр. 68-80 / A.K. Yatskov, M.I. Varlamov, and D.Yu. Turdakov. Extraction of data from mass media web sites. Programming and Computer Software, vol. 44, no.5, 2018, pp. 344-352.
- [22]. E. Pronoza, E. Yagunova, and A. Pronoza. Construction of a russian paraphrase corpus: unsupervised paraphrase extraction. In Proc. of the Russian Summer School in Information Retrieval, 2015, pp. 146–157.
- [23]. J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, vol. 20, no. 1, 1960, pp. 37–46.
- [24]. П.А. Пархоменко, А.А. Григорьев, Н.А. Астраханцев. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, том 29, вып. 2, 2017. DOI: 10.15514/ISPRAS-2017-29(2)-6 / P.A. Parhomenko, A.A. Grigorev, N.A. Astrakhantsev. A survey and an experimental comparison of methods for text clustering: application to scientific articles. Trudy ISP RAN/Proc. ISP RAS, 2017, vol.29, issue 2, pp.161-200 (in Russian).

Информация об авторах / Information about the authors

Кирилл Андреевич СКОРНЯКОВ – аспирант. Научные интересы: анализ нормативных документов, обработка графов и методы получения их векторных представлений небольшой размерности, методы адаптации существующих алгоритмов к новому домену, распределенные алгоритмы машинного обучения, обнаружение дубликатов текстов, обработка текстов на естественном языке.

Kirill Andreevich SKORNYAKOV – postgraduate student. Research interests: analysis of regulatory documents, graph embedding, domain adaptation, transfer learning, distributed machine learning algorithms, detection of duplicate texts, natural language processing.

Анна Сергеевна ЛАСКИНА – студентка магистратуры. Научные интересы: обработка текстов на естественном языке, машинное обучение.

Anna Sergeevna LASKINA – Master's student. Research interests: natural language processing, machine learning.

Денис Юрьевич ТУРДАКОВ – к.ф.-м.н., заведующий отделом «Информационные системы» ИСП РАН, доцент МГУ. Сфера научных интересов: машинное обучение, интеллектуальный анализ данных, извлечение информации, обработка естественного языка, сложные сети, анализ социальных сетей, большие данные.

Denis Yurievich TURDAKOV – Ph.D. head of the Information Systems Department at ISP RAS, associated professor at MSU. Research interests: machine learning, data mining, information extraction, natural language processing, complex networks, social network analysis, big data.