УДК 519.237.8 DOI 10.25559/SITITO.2017.3.439

Отраднов К.К., Жуков Д.О., Новикова О.А.

Московский технологический университет (МИРЭА), г. Москва, Россия

МОДЕЛЬ КЛАСТЕРИЗАЦИИ СЛАБОСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ

Аннотация

В статье предложена модель кластеризации коллекций новостных текстовых сообщений, а также соответствующий алгоритм кластеризации «bubble trap». Суть предлагаемого подхода заключается в разделения всего векторного пространства текстовых документов на оболочки смысловых кластеров с минимальными ограничениями на признаки отбора таким образом, что объем смыслового кластера и положение его центра остаются неизменными в процессе добавления в него новых векторов, а критерием соотнесения является заданная постоянная метрика точности.

Ключевые слова

Модель кластеризации; алгоритм кластеризации; алгоритм bubble trap; обработка текстовых данных; интеллектуальный анализ текстов.

Otradnov K.K., Zhukov D.O., Novikova O.A.

Moscow Technological University (MIREA), Moscow, Russia

CLUSTERING MODEL OF LOW-STRUCTURED TEXT DATA

Abstract

The article proposes a clustering model for collections of news text messages, as well as the corresponding bubble trap clustering algorithm. The essence of the proposed approach is to divide the entire vector space of text documents into shells of semantic clusters with minimal restrictions on the selection criteria in such a way that the volume of the semantic cluster and the position of its center remain unchanged in the process of adding new vectors to it, and the criterion of affiliation is a given constant accuracy metric.

Keywords

Clustering model; clustering algorithm; bubble trap algorithm; processing of text data; text mining.

Введение

Анализ неструктурированных или слабоструктурированных естественно – языковых текстовых данных, является очень важной и актуальной задачей с точки зрения поиска скрытых закономерностей в больших объемах информации, создания прогностических моделей различных процессов (в том числе социальных и экономических) и извлечения знаний.

При анализе больших объемов неструктурированных или слабоструктурированных текстовых данных и поиске в них скрытых закономерностей возникает проблема отбора признаков существенной для анализа информации. С одной стороны, необходимо, чтобы выполнение поиска происходило за приемлемое время, а с другой стороны необходимо, чтобы существенные данные не были потеряны. Отбор большого количества разнообразных признаков приводит к увеличению вычислительной сложности модели, и как следствие к необходимости использования значительных вычислительных ресурсов и затрат времени. Однако отбрасывание признаков, кажущихся несущественными, или проявляющимися на уровне шума, может привести к потере информации, значимой для поиска скрытых закономерностей. Таким образом, априори является неизвестным, какая часть имеющихся данных может понадобиться, а какая – нет. Определение связи шумовых признаков с целевой переменной решаемой задаче всегда является проблемой.

Следует отметить, что достаточно часто используемый для ускорения вычислений на практике метод, когда новые признаки формируются на основе старых, в данном случае является ненадежным, поскольку скрытые закономерности могут быть не линейным образом связаны с имеющимися данными и не являться их простой суперпозицией. Вместе с тем можно отметить, что формирование новых признаков на основе старых позволяет уменьшить их число для использования в моделях информационного поиска и приводит к ускорению выполнения модели.

При работе с большими объемами неструктурированных или слабоструктурированных естественно-языковых текстовых данных возникает необходимость осуществлять операции информационного поиска (в том числе поиска скрытых закономерностей), извлечения знаний и классификации, что требует создания соответствующих моделей для автоматизации этих процессов. Следует отметить, что в данном случае нет физически измеряемых величин, с которыми можно проводить математические операции и строить формальные модели.

Тексты состоят из универсальных лингвистических единиц (переменных), которыми являются слова и словосочетания. Лингвистические переменные сами по себе не имеют числовых или логических значений, поэтому построение математических моделей без предварительной формализации текстов не представляется возможным.

Основой формализации анализа естественно-языковых текстовых данных на основе лексического подхода с целью информационного поиска и классификации является статистическая семантическая гипотеза (statistical semantics hypothesis), согласно которой статистические зависимости употребления слов и терминов могут быть использованы для поиска заложенного в текст смысла [1]. В соответствии с гипотезой о скрытых связях, слова, термины и словосочетания, которые встречаются в похожих документах, стремятся иметь близкую семантическую зависимость [2].

Статистическая семантическая гипотеза позволяет формализовывать представление текстовые документы в виде векторной модели. Главная идея векторной модели семантики (vector space model, VSM) заключается в возможности представления каждого текстового документа в качестве точки в многомерном пространстве (вектора в векторном пространстве, размерности $\boldsymbol{R^M}$). Близко лежащие друг к другу точки соответствуют семантически схожим документам.

Для векторного представления текстов необходим словарь терминов (модель «bag of terms»), который может создаваться, как в процессе обработки документов, так, и взят готовый. Следует отметить, что использование для анализа заложенного в текст смысла, словаря, состоящего только из единичных слов, является недостаточным. Для повышения точности анализа текстов (более глубокого сохранения смысла) необходимо добавлять в словарь словосочетания (би-, три -, три-, тетра и т.д. граммы), более точно отображающие смысл описываемого объекта. Однако увеличение размера словаря приводит к увеличению времени обработки текстов, информационного анализа и поиска.

Суть модели векторизации текстов заключается в том, что если общее количество имеющихся различных уникальных терминов равно M (термы могут быть пронумерованы от 1 до M), а число текстовых документов равно N (документы могут быть пронумерованы от 1 до N), то данному документу (обозначим его $\boldsymbol{X_i}$, i – принимает значения от 1 до N) можно поставить в соответствие вектор $\boldsymbol{X_i} = \{x_{1,i}, x_{2,i}, \cdots x_{k,i}, \cdots x_{M,i}\}$, где первый элемент характеризует частоту вхождений в документ первого термина, второй – второго, третий – третьего и т.д. Каждый элемент вектора $x_{k,i}$ может прининать положительное значение. Все $\boldsymbol{X_i}$ образуют матрицу (размера N на M , где N – число документов, а M – размер словаря терминов): термин – документ (столбцами матрицы являются вектора $\boldsymbol{X_i}$, а строки определяют частоты вхождения отдельных терминов в документы коллекции):

$$\begin{vmatrix} x_{1,1} & x_{1,2} & \cdot & x_{1,i} & \cdot & x_{1,N} \\ x_{2,1} & x_{2,2} & \cdot & x_{2,i} & \cdot & x_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{j,1} & x_{i,2} & \cdot & x_{k,i} & \cdot & x_{i,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{M,1} & x_{M,2} & \cdot & x_{M,i} & \cdot & x_{M,N} \end{vmatrix}$$

Отметим, что данная матрица является сильно разреженной, т.е. содержит большое число нулей. Это связано с тем, что размер текстов по числу слов может быть существенно меньше, чем размер словаря.

Векторная модель документа может быть использована для решения таких задач информационного поиска, как кластеризации документов.

Кластеризацию документов можно определить, как процесс обнаружения в их коллекции групп, имеющих примерно одинаковое смысловое значение (однородных классов в произвольной предметной области). Обратим внимание на две проблемы. Во-первых, заранее неизвестны кластеры, которые можно отнести к истинным. Во- вторых, неизвестно число истинных кластеров, на которые предстоит разделить коллекцию.

При кластеризации документов необходимо, чтобы попадающие в один кластер документы имели смысловую схожесть, которую можно задать численно следующими способами:

• значение близости между любыми двумя элементами векторов, принадлежащих кластеру, должно быть не меньше определённого порога;

• значение близости между любым документом множества и центроидом кластера должно быть не меньше определённого порога, а центроид кластера рассчитывается как среднее арифметическое между всеми документами кластера.

В отличие от классификации, при кластеризации кластеры формируются на основе схожести описаний документов, и никакие характеристики этих групп не задаются заранее.

Следует отметить, что задача смысловой кластеризации документов является одной из ключевых в анализе больших объемов неструктурированных или слабоструктурированных естественно – языковых текстовых данных, и информационном поиске. В частности, кластеризация является единственным решением задачи, если нет точного представления о составе и структуры данных.

При использовании кластеризации на основе лексического подхода возникает ряд проблем, связанных с: выбором первичных кластеров, зависимостью качества кластеризации от длины текста, определением общего количества кластеров, отсутствие связи между близкими по смыслу текстами, в которых используется разная лексика и т.п. Для преодоления проблемы, связанной с различием в лексике необходимо использовать методы, основанные не только на основе лексического сходства, а еще и на основе семантической смежности или ассоциативности.

Современное состояние исследований в области кластеризации слабоструктурированных текстовых данных

На сегодняшний момент существует множество методов кластеризации документов. Например, Custom Search Folders, Latent Semantic Analysis/Indexing (LSA/LSI); Suffix Tree Clustering (STC); Single Link, Complete Link, Group Average; Scatter/Gather, K-means, Concept Indexing (CI); Self-Organizing Maps (SOM) и пр. Однако, несмотря на достаточно серьёзные успехи в этой области, остаются нерешёнными ряд насущных задач кластеризации, таких как выбор первоначального расположения ядер кластеров, масштабирование и управление размерностью данных; повышение производительности вычислений и ускорение процессов кластеризации, повышение точности кластеризации и др. Работы целого ряда исследователей посвящены актуальным вопросам кластерного анализа слабоструктурированных текстов. Предлагаемые авторами таких работ модели кластеризации весьма разнообразны и зачастую представляют собой гибридное решение на основе существующих алгоритмов.

Так, в работе [3] рассматривается алгоритм «искусственной пчелиной колонии» (ABC), который представляет собой метод стохастической оптимизации. Авторами предлагается гибридный алгоритм (MABCKM), основанный на модифицированных алгоритмах ABC и K-Means. Решения, создаваемые модифицированными ABC, рассматриваются как начальные решения для алгоритма K-Means. Производительность предлагаемого алгоритма оценивается в терминах разных параметров на шести стандартных наборах данных из алгоритма UCI Machine Learning Repository и сравнивается с алгоритмами ABC и K-Means. Экспериментальные результаты подтверждают превосходство алгоритма MABCKM для кластеризации текстовых данных и показывают, что предложенный алгоритм способен избегать локальных оптимумов и находить лучшие значения целевых функций с гораздо более низким стандартным отклонением по сравнению с другими двумя алгоритмами. Предлагаемый алгоритм также превосходит другие методы с точки зрения некоторых показателей (F-меры, ARI, CH и DB) и обеспечивает наилучшее ранжирование среди трех методов. Результаты показывают, что модифицированный алгоритм дает лучшие кластерные разделы и, естественно, приводит к выводу, что MABCKM является жизнеспособным методом кластеризации данных. Но предлагаемый метод нуждается в улучшении для выполнения автоматической кластеризации без какого-либо предварительного знания количества кластеров.

В работе [4] предлагаются три алгоритма кластеризации текстовых документов: с выбором функций, с функцией весовой схемы и динамического уменьшения размеров. В процессе кластеризации текстовые документы разделяются на несколько когерентных кластеров в соответствии с тщательно подобранными информативными особенностями, используя правильную функцию оценки, которая обычно зависит от временной частоты. Информативные функции в каждом документе выбираются с использованием

методов выбора объектов. Генетический алгоритм (GA), алгоритм гармонического поиска (HS) и алгоритм оптимизации (PSO) являются наиболее успешными методами выбора признаков, установленными с использованием новой схемы взвешивания, а именно: весовой коэффициент длины (LFW), который зависит от частоты и появления признаков в других документах. Также в статье предлагается новый метод уменьшения динамических размеров (DDR) для уменьшения числа функций, используемых в кластеризации, и, таким образом, для повышения эффективности алгоритмов. Алгоритм K-Means используется для кластеризации набора текстовых документов на основе терминов (или признаков), полученных динамическим сокращением. Оцениваются семь тестовых наборов текстовых данных разного размера и сложности. Анализ результатов показывает, что оптимизация с весом длины и динамическим уменьшением дает оптимальные результаты для почти всех тестируемых наборов данных.

В [5] представлен высоко масштабируемый алгоритм расширенной кластеризации, основанный на использовании n-грамм для уменьшения высокой размерности и получения высококачественных кластеров тестовых документов. Также в статье приведён сравнительный анализ, показывающий, что для образцов текстовых наборов данных с удалением стоп-слов, предлагаемый алгоритм работает лучше, чем без удаления стоп-слов.

Авторы статьи [6] изложили методы улучшения работы системы по извлечению событий из социальных данных. В ранних работах, например, в [7], сообщения обрабатывались индивидуально, что генерировало множество бессмысленных событий из-за недостающих деталей, разбросанных в миллионах текстовых сегментов. Кроме того, было проанализировано значительное число ненужных текстов, которые увеличили время обработки и снизили производительность системы. В данной работе была предложена кластеризация для группирования семантически связанных текстовых сегментов, фильтрации шума, уменьшения объема данных для обработки и продвижения только соответствующих текстовых сегментов в конвейер для извлечения информации. Алгоритм кластеризации был перенесён в инфраструктуру обработки потоков, под названием Storm, чтобы создать решение для кластеризации потоков и масштабировать до постоянно растущих объемов данных.

В работе [8] подчёркивается важность совместной кластеризации, особенно при рассмотрении разреженных высокоразмерных данных. В статье представлена новая генеративная модель Sparse Poisson Latent Block Model (SPLBM), основанная на распределении Пуассона, которое возникает для таблиц, таких как матрица документов-термов. Авторы утверждают, что SPLBM имеет два больших преимущества. Во-первых, это строгая статистическая модель. Во-вторых, алгоритм был разработан с нуля, чтобы справиться с проблемами разреженности данных. Как следствие, помимо поиска однородных блоков, как и других доступных алгоритмов, он также отфильтровывает однородные, но шумные из-за разреженности данные. Эксперименты по различным наборам данных различного размера и структуры показывают, что алгоритм, основанный на SPLBM, явно превосходит современные алгоритмы. В частности, представленный здесь алгоритм на основе SPLBM преуспевает в получении естественной кластерной структуры сложных несбалансированных наборов данных, которые другие известные алгоритмы не могут эффективно обрабатывать.

В статье [9] представлен новый метод автоматического обнаружения и отслеживания тем новостей из новостного архива мультимодального ТВ. Авторы предлагают тему и график (MT-AOG) для совместного представления текстовых и визуальных элементов новостей и их скрытых тематических структур. MT-AOG использует контекстно-зависимую грамматику, которая может описывать иерархический состав тем новостей по семантическим элементам, касающимся вовлеченных людей, связанных с ними мест и того, что произошло, и моделировать контекстуальные отношения между элементами в иерархии. В статье описано как с помощью процесса кластеризации, который объединяет истории о близких событиях, были выделены новостные темы. Сначала новостные документы кластеризуются, затем обнаруженные темы непрерывно отслеживаются и обновляются входящими потоками новостей. Генерируются траектории темы, чтобы показать, как темы появляются, эволюционируют и исчезают с течением времени.

Авторы [10] рассматривают способы кластеризации документов по темам. Они провели ряд экспериментов на разных реальных и искусственных наборах данных, таких как NEWS 20, Reuters, электронные письма, исследовательские работы по различным темам. Используется алгоритм TFI-DF с Fuzzy K-Means с

иерархическим алгоритмом. Изначально эксперименты были выполнены на небольшом наборе новостных данных. Далее, наилучший алгоритм был применён к расширенному набору данных. Наряду с различными кластерами соответствующих документов представлен приведенный коэффициент, энтропия и тренд F-меры, чтобы показать поведение алгоритма для каждого набора данных.

В статье [11] предлагается новый параллельный дизайн недавно появившейся эвристики для неконтролируемой и жесткой кластеризации. Исходный метод находит кластеры путем разбиения категориальных больших наборов данных в соответствии с линейной математической моделью, известной как реляционный анализ. Авторы улучшили этот метод для обработки кластеризации документов с итеративным использованием модели программирования MapReduce. Эффективность предлагаемого метода оценивается с использованием набора данных BBC, предоставленного сайтом BBC News. И эксперименты показали, что этот метод дает результаты качества с низкой вычислительной стоимостью.

В работе [12] была рассмотрена проблема роста цифровых новостей, хранящихся в базах данных и необходимости их моделирования, для того, чтобы их было легче понять и получить возможность извлечь из них важную информацию. Для упрощения обработки информации в базе данных требуется модель и конкретный метод кластеризации новостей, основанный на близости и характеристиках цифровых новостей. Авторы предлагают использовать модели и методы базы данных графов, алгоритм кластеризации графов MCL (алгоритм кластеризации Маркова), который может упростить обработку информации, идентифицируя характеристики каждой вершины на графике. В процессе идентификации кластеров графика цифровые новостные документы будут храниться в одной вершине, чтобы быть связанными с другой вершиной в общей категории новостей. Процесс расширения и раздувания матрицы является основным процессом кластеризации цифровых новостей, которые были преобразованы в модели базы данных графов. Между тем, процесс раздувания направлен на укрепление сильного края и ослабление слабого края. Так что при кластеризации цифровых новостей процесс раздувания матрицы очень влияет на время выполнения алгоритма MCL, а процесс раздувания матрицы влияет на формирование количества кластеров.

В статье [13] авторы рассматривают технологии качественного поиска документа в качестве ответа на любой поисковый запрос и анализируют, сравнивают различные алгоритмы кластеризации документов, используемых другими исследователями.

В работе [14] решается так называемая «новая проблема пользователя». Она заключается в необходимости получить некоторые данные о новом пользователе, чтобы начать делать персонализированные предложения. Авторы пытаются решить новую проблему пользователей с помощью уникальной персонализированной стратегии и изучают возможное улучшение результатов кластеризации документов и, в частности, кластеризацию новостных статей из Интернета при использовании текстовых n-грамм во время фазы извлечения ключевых слов. В статье представляется и оценивается автоматизированный подход, который объединяет кластеризацию новостных статей, полученных из Интернета, используя n-граммы. Затем этот метод сравнивается с методом «мешок слов», которое ранее был использован авторами в качестве алгоритма кластеризации (W-kmeans). Эксперименты показывают, что путем точной настройки весовых параметров между ключевым словом и n-граммами, а также самого значения n можно добиться значительного улучшения в отношении показателей результатов кластеризации.

Авторы работы [15] предлагает идею применения кластеризации документов на основе алгоритма Feature Based Clustering (FBC), т.к. считают, что будет проще использовать огромное количество существующих документов, если они объединены в несколько тем. FBC для кластеризации функций последовательных данных использует алгоритм K-Means. Особенности текстового документа могут быть представлены как последовательность слов. Для обработки в виде последовательных данных функции должны быть извлечены из коллекции неструктурированных текстовых документов. Поэтому авторами ставятся задачи предварительной обработки данных, чтобы предоставить для кластеризации соответствующие формы документов. В статье рассматривается два типа последовательного шаблона с использованием простой формы: частая последовательность слов (FWS) и максимальная частая последовательность (MFS). Оба типа подходят для текстовых данных. Разница заключается в применении принципа максимума в MFS. Следовательно, сумма MFS из

текстового документа будет меньше суммы его FWS. В этом исследовании авторы выбирают максимальные частотные последовательности (MFS) как представление функции и предлагают структуру для проведения FBC с использованием MFS в качестве функций. Структура тестируется на наборе данных кластеров, который является подмножеством текстовых данных группы новостей. Результат показывает, что на точность результата кластеризации влияет значение параметра, набор данных и количество целевого кластера.

В работе [16] иллюстрируется система поиска новостей, базовый алгоритм, ориентированный на нахождение новостных событий. Благодаря этой системе новостные статьи группируются вокруг одного центрального новостного события или события и его вспомогательных событий. Представленный алгоритм может помочь создать новые события и их компактные ярлыки в качестве исходных документов для процесса кластеризации. Система сконфигурирована для создания кластеров новостных событий верхнего уровня на основе текстовой метки, снабженной редакцией, известной как «линия слизняка», и создания кластеров на основе алгоритма, ориентированных на подтемы. Система использует алгоритм агломерационной кластеризации для сбора и структурирования документов в различные результирующие наборы. Решения о том, следует ли объединять связанные документы или кластеры, производятся в соответствии со сравнением подобия доказательств, полученных из двух разных источников: один – на основе цифровой подписи на основе неструктурированного текста в документе, другой – на основании наличия тегов с именованным объектом, которые были назначены документу с помощью тегатора, в данном случае движком Calais Thomson Reuters.

Авторы статьи [17] рассматривают проблему коротких текстовых сообщений на основе новостных документов, акцентируя внимание на смысле названия новостей. Короткие текстовые сообщения, которые сопровождаются метками времени и обычно краткими событиями, используя только несколько слов, отличаются от других более длинных текстовых документов, таких как веб-страницы, новостные статьи, блоги, технические документы и книги. Например, несколько слов повторяются в тех же заголовках новостей, поэтому частота слова не столь важна в коротком текстовом корпусе, как в более длинном текстовом корпусе. В статье предложен способ поиска кластеров объектов с использованием алгоритма совместного использования (FTCCT) для автоматического создания тем из короткого текстового корпуса и разработки алгоритма эволюции событий Evolution Mining (EEM) для обнаружения актуальных событий и их эволюций (т. к. степень популярности события со временем меняется). В FTCCT термин (одно слово или многословная фраза) относится только к одной теме в корпусе. Авторами были проведены и описаны эксперименты по новостям из 157 стран в течение 4 месяцев (с июля по октябрь 2013 года), которые показали, что метод FTCCT (объединяющий FTCCT и EEM) обеспечивает гораздо более высокое качество содержания и описания слов события, чем метод LDA (объединение LDA и EEM) для анализа потоков короткого текста. Метод также визуализирует эволюцию актуальных событий. Достоинством данного метода является возможность предсказывать и объяснять некоторые события. Обнаружившиеся всемирные события развили некоторые интересные корреляции наступления всемирных событий. Например, экстремальные погодные явления, происходящие в разных местах – тайфун в Гонконге и на Филиппинах, следовали за ураганом и штормовым наводнением в Мексике в сентябре 2013 года.

В работе [18] уделено внимание выбору алгоритмов кластеризации текстовых поисковых запросов. Основными критериями отбора стали: возможность работы с большим объёмом информации и вычислительная сложность. В данной работе использовались метод K-Means и DBSCAN.

В статье [19] сделан обзор существующих подходов к решению задачи кластерного анализа. Рассматриваются новые разработки в области кластерного анализа, основанные на привлечении ансамблей алгоритмов и логических моделей. Описываются преимущества таких алгоритмов. Ансамблевые методы позволяют значительно повысить устойчивость группировочных решений. Логические модели позволяют группировать разнотипные данные, а также давать объяснение результатов анализа на языке логических высказываний. Формулируются перспективные направления развития кластерного анализа.

Модели анализа слабоструктурированной текстовой информации созхданые на основе стохастичеких моделей описаны в работах [20-22].

Анализ перечисленных и прочих работ по тематике кластерного анализа слабоструктурированных текстов

свидетельствует о том, что несмотря на значительные результаты в повышении эффективности кластеризации, многие актуальные вопросы, затронутые нами в начале обзора, по-прежнему остаются вынесенными за скобки.

Постановка задачи для разработки модели кластеризации слабоструктурированных текстовых данных

Большинство используемых способов кластеризации базируются на определении расстояний между векторами и между кластерами. Вектора относят к определённым кластерам таким образом, чтобы расстояние между векторами, принадлежащими одному кластеру, было минимальным.

Построение моделей кластеризации текстовых документов основывается на методологии отбора существенных признаков для решения поставленной задачи информационного поиска. Не существует модели кластеризации, которая одновременно имела бы очень небольшое время выполнения, обладала бы абсолютной точностью и обрабатывала бы большие объемы неструктурированных или слабоструктурированных данных. Кроме того, следует отметить, что набор признаков необходимых для анализа скрытых закономерностей в больших объемах данных априори является неизвестным, поэтому при построении модели кластеризации и информационного поиска необходимо основываться на минимальном числе допущений, при максимальном сохранении разнообразных признаков, используемых для кластеризации. Однако следует учесть, что время выполнения модели при этом должны оставаться приемлемым, а реализация вычислений не требовать значительных аппаратных ресурсов.

Существующие модели кластеризации вне зависимости от того являются они лексическими или семантическими основываются на том, что исходя из задач информационного поиска для решения которых они предназначены, производится отбор значимых признаков. Далее всё множество документов векторизуется (с использованием терминов, или ассоциативно-семантических классов или т.д.) и в соответствии со значимыми признаками и заданной метрикой точности (или, например, заданным числом возможных кластеров и т.д.) кластеризуется по смысловым группам. Для каждого кластера определяется его центр и вектор задающий положение центра (центроид). При появлении новых документов, их вектора сравниваются по заданной метрике с центроидом, и по результатам сравнения добавляются в тот или иной кластер, или создается новый кластер. При таком подходе добавление каждого нового вектора изменяет положение центра кластера, и каждого из ранее вошедших в него векторов по отношению к нему. В конечном итоге это может привести к размыванию первоначального смыслового значения и необходимости повторной перекластеризации всего множества векторов, или использовании пост кластеризации, при обработке блоков документов. В существующих моделях, кластеры могут динамически изменяться не только по числу векторов, но и по объему.

Предлагаемая нами модель кластеризации (bubble trap) основана на других принципах. В соответствии с которыми, объем смыслового кластера и положение его центра (центроид), не должны изменяться в процессе добавления в него новых векторов. Критерием соотнесения является заданная постоянная метрика точности, на которую должно отличаться расстояние между данным вектором и центроидом оболочки кластера, для вхождения вектора в кластер.

Главной задачей при создании такой модели кластеризации является методика разделения всего векторного пространства текстовых документов на оболочки смысловых кластеров, по возможности с минимальными ограничениями на признаки отбора (чтобы не потерять скрытые закономерности, поведение которых имеет характер шума), что необходимо для создания эффективной модели анализа неструктурированных или слабоструктурированных текстовых данных.

Иными словами, при построении модели необходимо разделить всё информационное пространство размерности R^M на отдельные плотно упакованные области (назовем их space box), которые не изменяются в процессе кластеризации. Положение центра определяет смысловое значение всех векторов, попадающих в данную область, и не должно изменяться (даже незначительно) в процессе кластеризации. Разделение должно удовлетворять следующим требованиям. Если, длина вектора задающего положение центра данной области равна некоторой величине Y, то все вектора текстовых документов, положение которых отличается от положения центра не более чем на величину ξY (где ξ заданная точность кластеризации) принадлежат данной смысловой области.

При построении модели необходимо решить две основные задачи. Во-первых, разработать

методику разделения информационного пространства на смысловые области (space box). Во-вторых, разработать методику отнесения вектора конкретной области.

Перечислим основные требования к модели:

1. Исходными данными модели являются: словарь значимых терминов (содержащий некоторое число объектов), коллекция документов (содержащая некоторое произвольное число объектов, каждый из которых содержит некоторое произвольное число значимых терминов), заданная точность отнесения документа к определенной смысловой группе. Метрикой точности является заданная величина отклонения данного вектора от центра кластера;

2. Положение центра и объем каждого кластера не должны изменяться при добавлении новых векторов. Изменяться может только число векторов внутри кластера, что не приводит к изменению первоначально заданного смыслового значения, а значит появлению или исчезновению анализируемой скрытой закономерности;

3. Выполнение модели не должно требовать сверх больших аппаратных ресурсов и иметь в зависимости от объема данных приемлемое время их обработки.

Предлагаемая модель кластеризации

Для кластеризации коллекции текстовых документов по смысловым группам (определение однородных классов в произвольной проблемной области) после создания матрицы терм-документ мы предлагаем следующую модель. Воспользовавшись классическим определением расстояния между двумя точками в пространстве любой размерности, определяем длины всех N – векторов:

$$l_1 = \sqrt{x_{1,1}^2 + x_{2,1}^2 + x_{3,1}^2 + \cdots + x_{i,1}^2 + \cdots + x_{M,1}^2}$$

$$l_2 = \sqrt{x_{1,2}^2 + x_{2,2}^2 + x_{3,2}^2 + \cdots + x_{i,2}^2 + \cdots + x_{M,2}^2}$$

$$l_j = \sqrt{x_{1,j}^2 + x_{2,j}^2 + x_{3,j}^2 + \cdots + x_{i,j}^2 + \cdots + x_{M,j}^2}$$

$$l_N = \sqrt{x_{1,N}^2 + x_{2,N}^2 + x_{3,N}^2 + \cdots + x_{i,1}^2 + \cdots + x_{M,N}^2}$$

Далее выбираем вектор с максимальной длиной $l_k^{(max)}$, и вектор с минимальной длиной $l_p^{(min)}$. Величина $l_p^{(min)}$ определяет радиус гиперсферы в пространстве $\boldsymbol{R^M}$, внутри которой не будет находиться ни одного вектора, т.к. их длин будут больше $l_p^{(min)}$. Все вектора будут лежать внутри гипершарового слоя r, для которого будет выполняться условие $l_p^{(min)} \leq r \leq l_k^{(max)}$.

При использовании TF-IDF меры длина каждого из векторов будет одинакова (равна 1) и их координаты находятся не внутри гипершара, а попадают на поверхность гиперсферы. В этом случае деления шара на слои не требуется.

В случае, когда длины векторов имеют разные значения, то для кластеризации документов можно сформулировать следующую задачу: на сколько областей можно разделить гипершаровой слой r таким образом, чтобы выполнялись следующие условия:

- документы, попавшие в один кластер, имели бы примерно одинаковое смысловое значение;
- при очень большом числе документов (например, $N \to \infty$) все кластеры имели различимую границу, т.е. не пересекались бы между собой и не образовывали бы единый кластер.

Начнем с определения критерия одинаковости смысловых значений векторов. Возьмём некоторый вектор (любой из их множества), имеющий длину $l_p^{(min)}$, точка соответствующая концу этого вектора будет иметь координаты $(x_{1,p}, x_{2,p}, x_{3,p}, \cdots x_{M,p})$ и принадлежать определенному смысловому кластеру, вектора которого близки к данному вектору. Следует отметить, что в пространстве гипершара существует множество таких векторов, а множество точек, определяющих их концы будет образовывать гиперсферу радиуса $l_p^{(min)}$.

В качестве критерия «близости» можно выбрать условие, что расстояние от центроида кластера C (на рисунке 1 обозначен как ξl_c) до любого другого вектора (точки соответствующей его концу) входящего в данный кластер (в том числе и вектора l_p), не превышала бы некоторую заданную величину ξ от величины вектора l_c.

Подобный подход позволяет предположить, что при большом и плотном заполнении кластера векторами, его форма должна будет стремиться к гипершару радиуса ξl_c в пространстве размерности $\boldsymbol{R^M}$. А оболочки этих кластеров должны образовать субслой находящийся от центра координат на расстоянии $l_c = l_p^{min}/(1-\xi)$ (т.к. $l_c = l_p^{min} + \xi l_c$, см. рисунок 1).

Поскольку радиус кластера связан с величиной вектора центроида, а каждый из последующих векторов центроидов будет иметь длину больше, чем предыдущий, то при переходе к следующим субслоям, радиус кластеров должен увеличиваться.

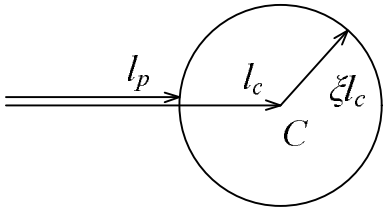

Рисунок 1 – Графическая интерпретация критерия близости векторов из выбранного кластера к его центроиду (вектору задающему положение центра данного кластера)

Используя описанный подход, разделим гипершаровой слой на «субслои» 1, 2 ,3 ,4, … , *K*, как это показано на рисунке 2, от $l_p^{(min)}$ до $l_k^{(max)}$.

При этом каждый следующий субслой будет несколько больше по ширине, чем предыдущий. Оболочки кластеров внутри каждого из субслоев показаны в виде сфер с увеличивающимся от субслоя к субслою размером. Еще раз отметим, что заполнение субслоев оболочками кластеров может быть максимально компактным. Однако, какие-то из кластеров, могут остаться пустыми (или содержать малое количество векторов, показаны на рисунке 2 белыми), а какие-то заполненными векторами (показаны на рисунке 2 серым цветом).

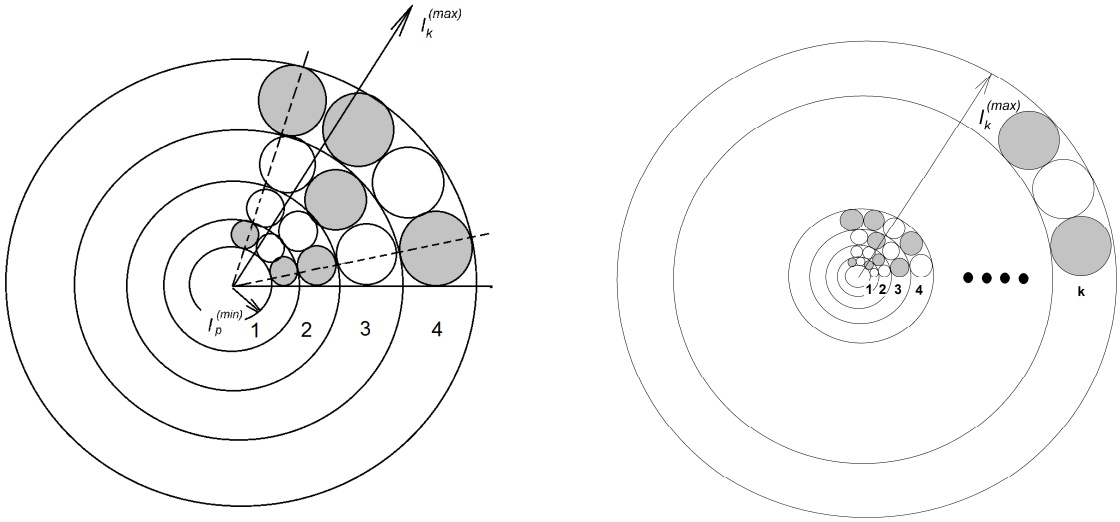

Рисунок 2 – Разделение гипершарового слоя на субслои с плотно упакованными кластерами

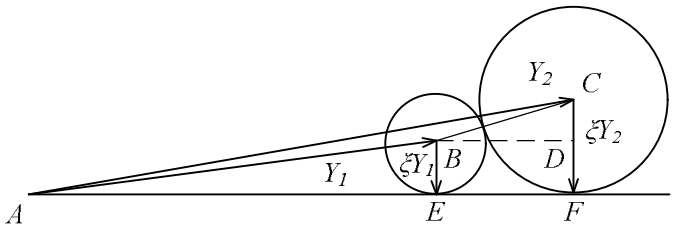

Рисунок 3 – Взаимное расположение соседних кластеров

Изучение топологии распределения пустых кластеров в информационном пространстве с точки зрения анализа само по себе является интересной задачей, однако это выходит за рамки данной работы.

Оболочки всех кластеров первого субслоя будут находиться от центра системы координат на расстоянии $Y_1=l_p^{min}/(1-\xi)$, второго на расстоянии Y_2, третьего на расстоянии Y_3 и т.д. до тех пор, пока не выполниться условие $Y_n \le l_k^{max}/(1+\xi)$. Определим величину Y_2 (радиус слоя). Обратим внимание на рисунок 2, на котором можно заметить, что оболочки кластеров различных субслоев могут быть выстроены в одну линию (проходящую через их центры, показана на левой части рисунка 2 пунктиром), либо несколько смещены относительно осевой линии (показана на левой части рисунка 2 пунктиром). Однако для определения Y_2 это не играет роли, т.к. при сдвиге оболочек кластеров внутри своих субслоев их центры будут по-прежнему оставаться на расстоянии Y_n от начала координат. Для нахождения Y_2 воспользуемся рисунком 3.

Из рисунка 3 видно, что

$$Y_2^2 = |AC|^2 = |AF|^2 + |CF|^2 = (|AE| + |EF|)^2 + |CF|^2 =$$

$$= \left(\sqrt{|AB|^2 - |BE|^2} + \sqrt{|BC|^2 - |CD|^2}\right)^2 + |CF|^2$$

$$|AB| = Y_1 = \frac{l_p^{(min)}}{1-\xi}$$
$$|BE| = \xi Y_1$$
$$|BC| = \xi Y_1 + \xi Y_2 = \xi(Y_1 + Y_2)$$
$$|CF| = \xi Y_2$$
$$|CD| = |CF| - |BE| = \xi Y_2 - \xi Y_1 = \xi(Y_2 - Y_1)$$
$$Y_2^2 = \left(\sqrt{|AB|^2 - |BE|^2} + \sqrt{|BC|^2 - |CD|^2}\right)^2 + |CF|^2 =$$
$$= \left(\sqrt{Y_1^2 - \xi^2 Y_1^2} + \sqrt{\xi^2(Y_1 + Y_2)^2 - \xi^2(Y_2 - Y_1)^2}\right)^2 +$$
$$+\xi^2 Y_2^2 = \left(Y_1\sqrt{1-\xi^2} + 2\xi\sqrt{Y_1 Y_2}\right)^2 + \xi^2 Y_2^2 =$$
$$= (1-\xi^2)Y_1^2 + 4\xi^2 Y_1 Y_2 +$$
$$+4\,\xi Y_1\sqrt{Y_1 Y_2(1-\xi^2)} + \xi^2 Y_2^2$$

или

$$Y_2^2(1-\xi^2) - (1-\xi^2)Y_1^2 - 4\xi^2 Y_1 Y_2 -$$
$$-4\,\xi Y_1\sqrt{Y_1 Y_2(1-\xi^2)} = 0 \qquad (1)$$

Таким образом, Y_2 может быть найдено из решения уравнения (1) с учетом того, что Y_2 должно быть больше Y_1, и любые Y_n не являются комплексными. Легко заметить из рисунка 1, что Y_3 может быть найдено заменой Y_1 в формуле (1) на Y_2, а Y_2 на Y_3, после нахождения Y_2 по формуле (1).

$$Y_3^2(1-\xi^2) - (1-\xi^2)Y_2^2 - 4\xi^2 Y_2 Y_3 -$$
$$4\,\xi Y_2\sqrt{Y_2 Y_3(1-\xi^2)} = 0 \qquad (2)$$

В общем случае получаем:

$$Y_n^2(1-\xi^2) - (1-\xi^2)Y_{n-1}^2 - 4\xi^2 Y_{n-1} Y_n -$$
$$4\,\xi Y_{n-1}\sqrt{Y_{n-1} Y_n(1-\xi^2)} = 0, \qquad (3)$$

где n=2, 3, 4, …, k, а для любого Y_n должно выполняться условие $Y_n \le l_k^{max}/(1+\xi)$.

Если $Y_{n+1} > l_k^{max}/(1+\xi)$, а $Y_n \le l_k^{max}/(1+\xi)$, то ширина последнего слоя (с номером n+1) определяется условием: $\{l_k^{max} - Y_n(1+\xi)\}$.

Следует отметить, что упаковка оболочек кластеров может содержать разреженные области между слоями может и не являться самой плотной. Наличие разреженных областей снижает общую плотность упаковки оболочек кластеров в гипершаре, но вместе с тем не снижает точности соотнесения векторов между субслоями, поскольку в разреженных областях выполняется условие

$$|BC| > \xi Y_1 + \xi Y_2 = \xi(Y_1 + Y_2)$$

и кроме того не сказывается на дальнейшем распределении векторов между оболочками кластеров внутри своего субслоя. Таким образом, плотное расположение дает нижнюю границу обеспечения точности распределения текстов по смысловым группам.

Рассмотренный подход позволяет найти число субслоев K в пространстве $\boldsymbol{R^M}$, чтобы с заданной точностью (1-ξ) распределить всё множество текстовых документов M по смысловым значениям.

Описанная методика позволяет дать определение класса, согласно которому понятие класса совпадает с понятием субслоя.

После распределения векторов по отдельным субслоям, рассмотрим задачу отнесения вектора к конкретному кластеру внутри субслоя.

Из соображений удобства желательно, чтобы покрытие слоя оболочками кластеров было равномерным. Учитывая, что часть векторов слоя может попадать в пространство вне оболочек, то для того, чтобы число таких не соотнесенных векторов было минимальным, необходимо стремиться, чтобы упаковка оболочек внутри слоя была максимально плотной. Далее необходимо определить координаты узлов сети (центров оболочек кластеров). Затем для каждого вектора слоя необходимо попарно вычислить метрики его расстояний от узлов сети. Величина каждой метрики должна быть меньше или равна величине ξY_n. Удовлетворяющие данной метрике вектора относятся к выбранному кластеру.

Для соотнесения векторов субслоя определенному кластеру необходимо найти координаты центров оболочек кластеров в пространстве $\boldsymbol{R^M}$.

Для удобства вычислений перейдем в сферическую систему координат.

В n-мерном пространстве, сферические координаты определяются следующим образом:

$$\begin{cases} x_M = Y\cos\theta_1 \\ \cdot \\ \cdot \\ x_i = Y\cos\theta_{i-1}\prod_{q=i}^{M-1}\sin\theta_q, \\ \text{при } i = 2,3,4,\cdots M-1, \\ \cdot \\ \cdot \\ x_1 = Y\sin\theta_1\sin\theta_2\sin\theta_3 \cdots\cdots \sin\theta_{M-1} \end{cases} \qquad (4)$$

где Y – сферический радиус ($Y \ge 0$), $\theta_1, \theta_2, \ldots, \theta_{n-1}$ – сферические углы, которые в общем случае изменяются в пределах: $0 \le \theta_1 < 2\pi$, $0 \le \theta_i \le \pi$ при i= 2, 3, …, M-1.

Учитывая, что все координаты x_i являются положительными: $0 \le \theta_1 < \pi/2$, $0 \le \theta_i \le \pi/2$ при i= 2, 3, …, M-1. Вектора находятся не во всем гипершаре, а только в его части.

Учитывая изотропность пространства и равнозначность осей можно предположить, что для обеспечения равномерности покрытия угловые координаты θ_i должны изменяться с одинаковым дискретным шагом δ_i от 0 до $\pi/2$.

Выберем субслой с номером n, и определим $\delta_{1,1}$ – шаг по углу θ_1 (θ_1 – угол между осью M и вектором, определяющим положение центра кластера) для цепочки кластеров с номером 1. Поясним нахождение величины $\delta_{1,1}$. На рисунке 5 показаны ось M и произвольно выбранная ось F,

индексы 1,2, 3 и т.д. одновременно обозначают цепочки кластеров и их проекции на плоскость OMF.

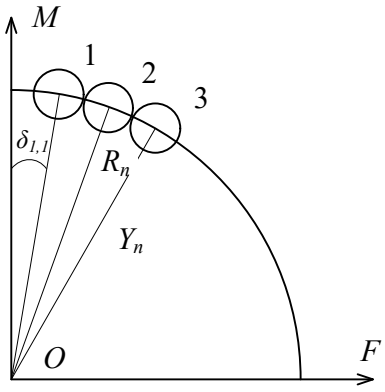

Рисунок 5 – Проекции цепочек кластеров на дугу между произвольными осями M и F

Выберем значение всех углов θ_i равными нулю, и получим следующий набор координат:

$$\begin{cases} x_{M,0} = Y_n \\ \cdot \\ \cdot \\ x_{i,0} = 0 \\ \cdot \\ \cdot \\ \cdot \\ x_{1,0} = 0 \end{cases}$$

Точка с данными координатами (обозначим её A) лежит на оси M.

Далее рассмотрим ближайшую к данной точке цепочку кластеров, у которой относительно угол между осью M и вектором задающим центр первой осевой линии $\theta_1=\delta_{1,1}$ (см. рисунок 7), а угловые координаты относительно остальных осей равны 0.

$$\begin{cases} x_{M,1} = Y_n cos\delta_{1,1} \\ x_{\mathrm{M}-1,1} = Y_n sin\delta_{1,1} \\ \cdot \\ x_{i,1} = 0 \\ \cdot \\ \cdot \\ \cdot \\ x_{1,1} = 0 \end{cases} \quad (5)$$

Далее запишем уравнение для метрики, определяющей расстояние между вектором, задающим осевую линию первой цепочки кластеров и точкой A, лежащей на оси M:

$$\begin{gathered} \left(x_{1,1} - x_{1,0}\right)^2 + \left(x_{2,1} - x_{2,0}\right)^2 + \cdots\cdots \\ + \left(x_{i,1} - x_{i,0}\right)^2 + \cdots\cdots + \left(x_{M,1} - x_{M,0}\right)^2 = \{2\xi Y_n\}^2 \\ Y_n^2 sin^2\delta_{1,1} + \left(Y_n cos\delta_{1,1} - Y_n\right)^2 = \{2\xi Y_n\}^2 \\ 2Y_n^2(1 - cos\delta_{1,1}) = \{2\xi Y_n\}^2 \\ cos\delta_{1,1} = 1 - \frac{\{2\xi Y_n\}^2}{2Y_n^2} = 1 - 2\xi^2 \end{gathered} \quad (6)$$

Угол θ_1 будет изменяться с дискретным шагом $\delta_{1,1}$ от 0 до π/2. Общее число цепочек кластеров будет равно $K_M=\pi/2\delta_{1,1}$ с округлением до минимального целого значения.

Определим $\delta_{1,2}$ – шаг по углу $\theta_{2,1}$ ($\theta_{2,1}$ – угол для оси M-1) для цепочки кластеров с номером 1. Выберем значение всех углов θ_i, (кроме $\theta_1=\delta_{1,1}$ и $\theta_{2,1}=\delta_{1,2}$) равными нулю, и получим следующий набор координат:

$$\begin{cases} x_{M,2} = Y_n cos\delta_{1,1} \\ x_{\mathrm{M}-1,2} = Y_n sin\delta_{1,1} cos\delta_{1,2} \\ x_{\mathrm{M}-2,2} = Y_n sin\delta_{1,1} sin\delta_{1,2} \\ \cdot \\ \cdot \\ x_{i,2} = 0 \\ \cdot \\ \cdot \\ \cdot \\ x_{1,2} = 0 \end{cases} \quad (7)$$

Отметим, что данный набор координат должен определять положение внутри цепочки кластера, являющимся ближайшим к кластеру, положение центра которого задается системой уравнений (5).

Далее запишем уравнение для метрики, определяющей расстояние между вектором, задающим центр кластера, с координатами, заданными системой уравнений (7) и вектором, задающим центр кластера, с координатами, заданными системой уравнений (5):

$$\begin{gathered} \left(x_{1,2} - x_{1,1}\right)^2 + \left(x_{2,2} - x_{2,1}\right)^2 + \cdots\cdots \\ + \left(x_{i,2} - x_{i,1}\right)^2 + \cdots\cdots + \left(x_{M,2} - x_{M,1}\right)^2 = \{2\xi Y_n\}^2 \\ \left(Y_n sin\delta_{1,1} sin\delta_{1,2}\right)^2 \\ + \left(Y_n sin\delta_{1,1} cos\delta_{n,2} - Y_n sin\delta_{1,1}\right)^2 \\ + \left(Y_n cos\delta_{1,1} - Y_n cos\delta_{1,1}\right)^2 \\ = \{2\xi Y_n\}^2 \\ Y_n^2 sin^2\delta_{1,1} sin^2\delta_{1,2} + Y_n^2 sin^2\delta_{1,1}\left(cos\delta_{1,2} - 1\right)^2 \\ = \{2\xi Y_n\}^2 \\ 2Y_n^2 sin^2\delta_{1,1}\{1 - cos\delta_{1,2}\} = \{2\xi Y_n\}^2 \\ cos\delta_{1,2} = 1 - \frac{\{2\xi Y_n\}^2}{2Y_n^2 sin^2\delta_{1,1}} = 1 - \frac{2\xi^2}{sin^2\delta_{1,1}} \end{gathered} \quad (8)$$

Угол $\theta_{2,1}$ будет изменяться с дискретным шагом $\delta_{1,2}$ от 0 до π/2. Общее число кластеров будет равно $K_{M\text{-}1,1}=\pi/2\delta_{1,2}$ с округлением до минимального целого значения.

Далее необходимо взять угол $\theta_1=2\delta_{1,1}$, и определить $\delta_{2,2}$ – шаг по углу $\theta_{2,2}$ ($\theta_{2,2}$ – угол для оси M-1) для цепочки кластеров с номером 2. Выберем значение всех углов θ_i, (кроме $\theta_1=2\delta_{1,1}$ и $\theta_{2,2}=\delta_{2,2}$) равными нулю, и получим следующий набор координат:

$$\begin{cases} x_{M,3} = Y_n cos2\delta_{1,1} \\ x_{\mathrm{M}-1,3} = Y_n sin2\delta_{1,1} cos\delta_{2,2} \\ x_{\mathrm{M}-2,3} = Y_n sin2\delta_{1,1} sin\delta_{2,2} \\ \cdot \\ \cdot \\ x_{i,3} = 0 \\ \cdot \\ \cdot \\ \cdot \\ x_{1,3} = 0 \end{cases} \quad (9)$$

Далее запишем уравнение для метрики, определяющей расстояние между вектором, задающим центр кластера, с координатами, заданными системой уравнений (9) и вектором, задающим центр кластера, с координатами заданными системой уравнений (10):

$$\begin{cases} x_{M,4} = Y_n cos2\delta_{1,1} \\ x_{\mathrm{M}-1,4} = Y_n sin2\delta_{1,1} \\ \cdot \\ x_{i,4} = 0 \\ \cdot \\ \cdot \\ \cdot \\ x_{1,4} = 0 \end{cases} \quad (10)$$

$$(x_{1,3} - x_{1,4})^2 + (x_{2,3} - x_{2,4})^2 + \cdots\cdots$$
$$+(x_{i,3} - x_{i,4})^2 + \cdots\cdots + (x_{M,3} - x_{M,4})^2 = \{2\xi Y_n\}^2$$
$$(Y_n sin2\delta_{1,1} sin\delta_{2,2})^2$$
$$+ (Y_n sin2\delta_{1,1} cos\delta_{2,2}$$
$$- Y_n sin2\delta_{1,1})^2 +$$
$$+(Y_n cos2\delta_{1,1} - Y_n cos2\delta_{1,1})^2 = \{2\xi Y_n\}^2$$
$$Y_n^2 sin^2 2\delta_{1,1} sin^2\delta_{2,2} + Y_n^2 sin^2 2\delta_{1,1} cos^2\delta_{2,2} -$$
$$-2Y_n^2 sin^2 2\delta_{1,1} cos\delta_{2,2} + Y_n^2 sin^2 2\delta_{1,1} = \{2\xi Y_n\}^2$$
$$2Y_n^2 sin^2 2\delta_{1,1}\{1 - cos\delta_{2,2}\} = \{2\xi Y_n\}^2$$
$$cos\delta_{2,2} = 1 - \frac{\{2\xi Y_n\}^2}{2Y_n^2 sin^2 2\delta_{1,1}} = 1 - \frac{2\xi^2}{sin^2 2\delta_{1,1}} \quad (11)$$

Угол $\theta_{2,2}$ будет изменяться с дискретным шагом $\delta_{2,2}$ от 0 до π/2. Общее число кластеров будет равно $K_{M\text{-}1,2}$=π/2$\delta_{2,2}$ с округлением до минимального целого значения.

Аналогичным образом берем угол θ_1=3$\delta_{1,1}$, и определить $\delta_{2,3}$ – шаг по углу $\theta_{2,2}$ ($\theta_{2,3}$ – угол для оси M-1) для цепочки кластеров с номером 3.

$$cos\delta_{2,3} = 1 - \frac{\{2\xi Y_n\}^2}{2Y_n^2 sin^2 3\delta_{1,1}} = 1 - \frac{2\xi^2}{sin^2 3\delta_{1,1}} \quad (12)$$

И т.д. получаем, что для оси M шаг по углу составляет $\delta_{1,1}$, а число цепочек K_M=π/2$\delta_{1,1}$ (с округлением до наименьшего целого). Для первой цепочки кластеров относительно оси M-1 угловой шаг составит $\delta_{1,2}$, для второй цепочки кластеров $\delta_{2,2}$, для третей цепочки кластеров $\delta_{2,3}$, и т.д. до цепочки K_M:

$$cos\delta_{1,1} = 1 - 2\xi^2$$
$$cos\delta_{1,2} = 1 - \frac{2\xi^2}{sin^2\delta_{1,1}}$$
$$cos\delta_{2,2} = 1 - \frac{2\xi^2}{sin^2 2\delta_{1,1}}$$
$$cos\delta_{3,2} = 1 - \frac{2\xi^2}{sin^2 3\delta_{1,1}}$$
$$cos\delta_{K_M,2} = 1 - \frac{2\xi^2}{sin^2 K_M\delta_{1,1}}$$

Далее переходим к оси M-2. Определим шаг по углу θ_3 (θ_3 – угол для оси M-2). Здесь должна быть реализована следующая процедура. Выбираем угол $\delta_{1,1}$, затем берем угол $\delta_{1,2}$ и определяем углы $\delta_{1,3}$, $\delta_{2,3}$, и т.д. как это было проделано ранее для углов $\delta_{1,2}$, $\delta_{2,2}$ и т.д. при условии, что углы $\theta_{3,1}$, $\theta_{3,2}$, $\theta_{3,3}$, и т.д. дискретно изменяются с шагом $\delta_{1,3}$, $\delta_{2,3}$, и т.д. от 0 до π/2. Далее выбираем угол $\delta_{1,1}$ и угол $\delta_{2,2}$ и проделываем описанную выше процедуру. Далее выбираем угол $\delta_{1,1}$ и угол $\delta_{3,2}$ и проделываем описанную выше процедуру до $\delta_{K_M,2}$, пока не будет достигнут угол π/2. Потом берем угол $2\delta_{1,1}$ и повторяем все вновь, потом берем $2\delta_{1,1}$ и т.д. до тех пор, пока θ_1 не достигнет π/2.

Выберем значение всех углов θ_i, (кроме θ_1=$\delta_{1,1}$, θ_2=$\delta_{1,2}$ и θ_3=$\delta_{1,3}$) равными нулю, и получим следующий набор координат:

$$\begin{cases} x_{M,5} = Y_n cos\delta_{1,1} \\ x_{\mathrm{M}-1,5} = Y_n sin\delta_{1,1} cos\delta_{1,2} \\ x_{\mathrm{M}-2,5} = Y_n sin\delta_{1,1} sin\delta_{1,2} cos\delta_{1,3} \\ x_{\mathrm{M}-3,5} = Y_n sin\delta_{1,1} sin\delta_{1,2} sin\delta_{1,3} \\ \cdot \\ x_{i,5} = 0 \\ \cdot \\ \cdot \\ \cdot \\ x_{1,5} = 0 \end{cases} \quad (13)$$

Далее запишем уравнение для метрики, определяющей расстояние между вектором, задающим центр кластера, с координатами, заданными системой уравнений (13) и вектором, задающим центр ближайшего по цепочке кластера, с координатами, заданными системой уравнений (7):

$$(x_{1,5} - x_{1,2})^2 + (x_{2,5} - x_{2,2})^2 + \cdots\cdots$$
$$+(x_{i,5} - x_{i,2})^2 + \cdots\cdots + (x_{M,5} - x_{M,2})^2 = \{2\xi Y_n\}^2$$
$$(Y_n sin\delta_{1,1} sin\delta_{1,2} sin\delta_{1,3})^2 +$$
$$+(Y_n sin\delta_{1,1} sin\delta_{1,2} cos\delta_{1,3} - Y_n sin\delta_{1,1} sin\delta_{1,2})^2 +$$
$$+(Y_n sin\delta_{1,1} cos\delta_{1,2} - Y_n sin\delta_{1,1} cos\delta_{1,2})^2 +$$
$$+(Y_n cos\delta_{1,1} - Y_n cos\delta_{1,1})^2 = \{2\xi Y_n\}^2$$
$$(Y_n sin\delta_{1,1} sin\delta_{1,2} sin\delta_{1,3})^2 +$$
$$+(Y_n sin\delta_{1,1} sin\delta_{1,2} cos\delta_{1,3} - Y_n sin\delta_{1,1} sin\delta_{1,2})^2$$
$$= \{2\xi Y_n\}^2$$
$$Y_n^2 sin^2\delta_{1,1} sin^2\delta_{1,2}\left\{sin^2\delta_{1,3} + (cos\delta_{1,3} - 1)^2\right\}$$
$$= \{2\xi Y_n\}^2$$
$$2Y_n^2 sin^2\delta_{1,1} sin^2\delta_{1,2}(1 - cos\delta_{1,3}) = \{2\xi Y_n\}^2$$
$$cos\delta_{1,3} = 1 - \frac{\{2\xi Y_n\}^2}{2Y_n^2 sin^2\delta_{1,1} sin^2\delta_{1,2}} = 1 -$$
$$\frac{2\xi^2}{sin^2\delta_{1,1} sin^2\delta_{1,2}} \quad (14)$$

Используя описанный ранее подход, рекурсивно можно найти угловые координаты всех оболочек кластеров для данного субслоя.

Очевидно, что число оболочек во всем субслое будет настолько велико, что нахождение координат их центров является технически неразрешимой задачей. Однако отметим, что её

решение для кластеризации векторов внутри слоя не требуется.

Рассмотрим текстовый документ *j* состоящий из w_j значимых терминов. Соответственно из всего множества R^M, отличными от 0 будет только w_j координат, для которых можно записать систему уравнений (4):

$$\begin{cases} x_M = Y cos\theta_1 \\ \cdot \\ \cdot \\ x_i = Y cos\theta_{i-1} \prod_{q=i}^{M-1} sin\theta_q, \text{ при } i = 2,3,4,\cdots\cdots M-1 \\ \cdot \\ \cdot \\ \cdot \\ x_1 = Y sin\theta_1 sin\theta_2 sin\theta_3 \cdots\cdots sin\theta_{M-1} \end{cases}$$

В которой останутся только строки соответствующие отличным от нуля x_i. Для того чтобы рассматриваемый вектор, принадлежал определенному кластеру, необходимо чтобы выполнялось условие:

$$(x_{1,c} - x_{1,v})^2 + (x_{2,c} - x_{2,v})^2 + \cdots + (x_{i,c} - x_{i,v})^2 + \cdots + (x_{M,c} - x_{M,v})^2 \le \{\xi Y_n\}^2,$$

где $x_{i,c}$ – координаты центра кластера, а $x_{i,v}$ – соответствующие координаты вектора.

Используя, для угловых координат, полученные ранее рекурсивные соотношения и систему уравнений (4) можно восстановить по ним соответствующие координаты $x_{i,c}$ центров кластеров. Критерием точности нахождения координат является соотношение: $|x_{i,v} - x_{i,c}| \le \Delta$, где Δ – величина ошибки (по сути, каждая из $x_{i,c}$ подгоняется к $x_{i,v}$, таким образом, чтобы рассчитываемая с использованием формул (4) величина отличалась от соответствующей координаты вектора не более чем на Δ). Величину Δ можно оценить используя уравнение:

$$(\Delta)^2 + (\Delta)^2 + \cdots + (\Delta)^2 + \cdots + (\Delta)^2 \le \{\xi Y_n\}^2$$

$$w_i(\Delta)^2 \le \{\xi Y_n\}^2$$

$$\Delta \le \frac{\xi Y_n}{\sqrt{w_i}}$$

Далее зная к какому кластеру принадлежит данный вектор, можно кластеризовать все документы по смысловым группам с заданной ошибкой δ.

Пусть документ имеет значимый термин, соответствующий в рассматриваемом пространстве координате X_1. Тогда используя условие:

$x_{i,v} - \frac{\xi Y_n}{\sqrt{w_i}} \le x_{i,c} \le x_{i,v} + \frac{\xi Y_n}{\sqrt{w_i}}$ подбираем θ_1, используя для этого его наборы угловых координат δ. Если X_1=0, то $cos\theta_1 = 0$, а $sin\theta_1 = 1$. При $cos\theta_1 = 0$, координата кластера по оси 1 выбирается из множества: $\theta_1 = \delta_{1,1}, 2\delta_{1,1}, 3\delta_{1,1}, \cdots\cdots\cdots\cdots, n_{\theta_1}\delta_{1,1}$, где $\delta_{1,1}$ – шаг по углу θ_1 (θ_1 – угол по оси 1). Отметим, что необходимо найти $n_{max}\delta_{1,1} \approx \pi/2$ или $n_{max} \approx \pi/2\delta_{1,1}$ с округлением до минимального целого числа. Если $cos\theta_1 \neq 0$, то θ_1 сразу выбирается из $\theta_1 = \delta_{1,1}, 2\delta_{1,1}, 3\delta_{1,1}, \cdots\cdots\cdots\cdots, n_{\theta_1}\delta_{1,1}$.

Переходим к определению θ_2. Зная $cos\theta_1$ (подобранный или найденный из значения координаты X_1), находим величину $sin\theta_1$, а затем величину $cos\theta_2 = \frac{x_2}{Y sin\theta_1}$. Далее используя условие: $x_{i,v} - \frac{\xi Y_n}{\sqrt{w_i}} \le x_{i,c} \le x_{i,v} + \frac{\xi Y_n}{\sqrt{w_i}}$ подбираем θ_2, используя для этого его наборы угловых координат δ для θ_2, но уже с известным $\theta_1 = n_{\theta_1}\delta_{1,1}$ (с ограничением для θ_1), что существенно уменьшает число вариантов для угловых наборов θ_2. Отметим, что предварительного расчета всех угловых наборов для всех углов θ не требуется, т.к. все значения можно рассчитывать при отнесении вектора к кластеру. Используя для этого описанную ранее рекурсивную методику и полученные с её помощью в общем виде для любых θ и δ математические выражения, что существенно уменьшает объем вычислений и увеличивает скорость кластеризации.

Далее аналогичным образом определяем все остальные координаты.

$\theta_1^{(vector)} = arccos\left|\frac{x_1^{(vector)} \pm \frac{\xi * Y}{\sqrt{w_i}}}{Y}\right|$, при j=1, $x_1^{(vector)}$ – координата вектора по оси 1;

$\delta_1 = \arccos(1 - 2 * \xi^2)$, при j=1;

$n_{\theta_1} = min\left\{\frac{\theta_1^{(vector)}}{\delta_1}\right\}_{integer}$, (выбирается целое n_{θ_1}, для которого наименьшая погрешность при округлении);

$\theta_1^{(cluster)} = n_{\theta_1} * \delta_1 \le \frac{\pi}{2}$, где $\theta_1^{(cluster)}$ – угловая координата по оси 1 центра кластера, которому принадлежит данный вектор;

$$\theta_j^{(vector)} = arccos\left|\frac{x_j^{(vector)} \pm \frac{\xi * Y}{\sqrt{w_i}}}{Y * \prod_{j-1}^{M-1} sin\left\{n_{\theta_{j-1}} * \delta_{j-1}\right\}}\right|,$$

при $j = 2,3,4,\cdots\cdots M$, $x_j^{(vector)}$ – координата данного вектора по оси j;

$$\delta_j = arccos\left\{1 - \frac{2 * \xi^2}{\prod_{j-1}^{M-1} sin^2\left\{n_{\theta_{j-1}} * \delta_{j-1}\right\}}\right\},$$

при $j = 2,3,4,\cdots\cdots M$;

$n_{\theta_j} = min\left\{\frac{\theta_j^{(vector)}}{\delta_j}\right\}_{integer}$, (выбирается целое n_{θ_j}, для которого наименьшая погрешность при округлении);

$\theta_j^{(cluster)} = n_{\theta_j} * \delta_j \leq \frac{\pi}{2}$, где $\theta_j^{(cluster)}$ – угловые координаты по осям j=1,2,3, М центра кластера, которому принадлежит рассматриваемый вектор.

Заключение

Обобщая изложенное выше, представим осуществление кластеризации слабоструктурированных текстов на основе разработанной модели в виде следующей последовательности действий (алгоритм кластеризации «bubble trap»):

1. Обрабатываем с помощью методов математической лингвистики (разметка, нормализация и т.д.) и словаря терминов содержащего слова и n – граммы всю коллекцию текстовых документов, и создаем матрицу документ – термин (с использованием TF-IDF метрик или абсолютных частот вхождения терминов);

2. Для повышения точности информационного анализа можно учесть смысловую близости различных семантических единиц, и перейти от матрицы документ – термин к матрице документ – семантический класс. Для этого можно использовать следующую методику:

a) создаем обучающую выборку документов, и на её множестве определяем ассоциативно – семантические классы для часто употребляемых общих терминов (например, используя алгоритм Word2Vec). В один класс попадают имеющие разную семантику, но обладающие схожим смысловым значением термины (слова, словосочетания и т.д.);

b) заменяем термины на семантические классы. При этом первоначальная размерность информационного пространства уменьшается, за счет того, что отдельные термины частично объединяются в классы. Те термины, которые не объединились, образуют классы, каждый из которых включает всего один элемент;

c) рассчитываем частотные характеристики вхождения классов в документы (используя либо TF-IDF метрику, либо абсолютные частоты вхождения объектов семантических классов в документы);

3. Определяем длины всех N – векторов коллекции документов. Выбираем вектор с максимальной длиной $l_k^{(max)}$, и вектор с минимальной длиной $l_p^{(min)}$. До момента выбора минимального и максимального векторов данная операция может осуществляться параллельно;

4. Задаем величину критерия ξ смыслового совпадения документов. И решая уравнение: $Y_n^2(1-\xi^2) - (1-\xi^2)Y_{n-1}^2 - 4\xi^2 Y_{n-1}Y_n - 4\xi Y_{n-1}\sqrt{Y_{n-1}Y_n(1-\xi^2)} = 0$ (где n=2, 3, 4, …, k) определяем радиусы (Y_n) слоев оболочек кластеров в информационном пространстве (учитывая, что $Y_1 = l_p^{(min)}/(1-\xi)$). Для любого Y_n должно выполняться условие $Y_n \leq l_k^{(max)}/(1+\xi)$, n – номер соответствующего слоя. Если $Y_{n+1} > l_k^{max}/(1+\xi)$, а $Y_n \leq l_k^{max}/(1+\xi)$, то ширина последнего слоя (с номером n+1) определяется условием: $\{l_k^{max} - Y_n(1+\xi)\}$;

5. Сортируем все вектора по слоям (классам), используя следующее условие:

$$Y_n(1-\xi) \leq l_j \leq Y_n(1+\xi);$$

6. Используя описанные ранее рекурсивные методику определения координат оболочек кластеров при их покрытии субслоя и распределения векторов внутри классов, определяем угловые координаты кластера (наборы δ для углов θ), к которому относится данный вектор. Данная операция может осуществляться параллельно не только по обработки каждого субслоя, но и по обработке каждого из векторов, отнесенного к тому или иному субслою;

7. Зная координаты центров кластеров, к которым относятся вектора, проводим кластеризацию документов по смысловым группам;

8. Извлекаем знания из текстов документов образовавших кластер, проводим их аннотирование и определяем смысловую тематику данного кластера.

Как видно из описания, данный подход не требует предварительного расчета значений координат оболочек всех кластеров. Отметим, что при нахождении координат на каждом следующем этапе используются результаты предыдущего этапа. Координата принадлежности вектора определенному кластеру находится в процессе обработки самого вектора и не связана с обработкой других векторов, что существенно уменьшает объем вычислений и сокращает время обработки данных. Кроме того, поскольку обработка векторов производится независимо, это позволяет значительно распараллелить вычисления и получить выигрыш в производительности.

Благодарности

Работа выполнена за счет финансирования Министерством образования и науки Российской Федерации конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов, номер проекта 28.2635.2017/ПЧ, наименование

«Разработка моделей стохастической самоорганизации слабоструктурированной информации и реализации памяти при прогнозировании новостных событий на основе массивов естественно-языковых текстов».

Acknowledgments

The work is executed due to financing by the Ministry of Education and Science of the Russian Federation as the competitive part of government tasks of higher education and scientific organizations for the implementation the initiative scientific projects; the number of the project – 28.2635.2017 / IF named «Development of models of stochastic self-semistructured information and implementation of memory in predicting the news events array-based natural language texts».

Литература

1. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press, 2007. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008, 544 p.
2. Turney, P. D. The latent relation mapping engine: Algorithm and experiments. // Journal of Artificial Intelligence Research, 33, 2008, P. 615–655.
3. Kumar, A., Kumar, D., Jarial, S.K. A novel hybrid K-means and artificial bee colony algorithm approach for data clustering. // Decision Science Letters . – 2018, Vol. 7, Issue 1, P. 65-76.
4. Abualigah, L.M., Khader, A.T., Al-Betar, M.A., Alomari, O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. / Expert Systems with Applications. – 2017, 84, P. 24-36.
5. Kanimozhi, K.V., Venkatesan, M. A novel map-reduce based augmented clustering algorithm for big text datasets. // Advances in Intelligent Systems and Computing. – 2018, Vol. 542, P. 427-436.
6. Jenhani, F., Gouider, M.S., Said, L.B. Social stream clustering to improve events extraction. // Smart Innovation, Systems and Technologies. – 2018, Vol. 73, P. 319-329.
7. Jenhani, F., Gouider, M.S., Ben Said, L.A hybrid approach for drug abuse events extraction from Twitter (2016). 20Th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (ICKIIES 2016), York, United Kingdom. -2016.
8. Ailem, M., Role, F., Nadif, M. Sparse Poisson Latent Block Model for Document Clustering. / IEEE Transactions on Knowledge and Data Engineering. – 2017, №29 (7), P. 1563-1576.
9. Li, W., Joo, J., Qi, H., Zhu, S.-C. Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph. // IEEE Transactions on Multimedia. – 2017, Vol. 19, Issue 2, 19(2), P. 367-381.
10. Bafna, P., Pramod, D., Vaidya, A. Document clustering: TF-IDF approach. // International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016. – 2016, P. 61-66.
11. Lamari, Y., Slaoui, S.C. Parallel document clustering using iterative mapreduce. 2016 International Conference on Big Data and Advanced Wireless Technologies, BDAW 2016; Blagoevgrad; Bulgaria; 10 November 2016 to 11 November 2016. // ACM International Conference Proceeding Series.
12. Al-Fath, A.M.U., Saleh, W.K.R., Sa'Adah, S. Implementation of MCL algorithm in clustering digital news with graph representation. // 4th International Conference on Information and Communication Technology, ICoICT 2016; Bandung; Indonesia; 25 May 2016 до 27 May 2016.
13. Patil, H., Thakur, R.S. Document clustering: A summarized survey (Book Chapter) // Pattern and Data Analysis in Healthcare Settings. – 22 July 2016, P. 264-281.
14. Le, T.M.V., Lauw, H.W. Semantic visualization with neighborhood graph regularization. / Journal of Artificial Intelligence Research. – 2016, Vol. 55, P. 1091-1133.
15. Rahmawati, D., Putri Saptawati, G.A., Widyani, Y. Document clustering using sequential pattern (SP): Maximal frequent sequences (MFS) as SP representation. // Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015. – 2016, P. 98-102.
16. Conrad, J.G., Bender, M. Semi-supervised events clustering in news retrieval. // CEUR Workshop Proceedings. – 2016, P. 21-26.
17. Huang, G., He, J., Zhang, Y., (...), You, Y., Cao, J. Mining streams of short text for analysis of world-wide event evolutions. / World Wide Web . -2015, 18(5), P. 1201-1217.
18. О. А. Безверхий, С. Г. Самохвалова. Кластеризация большого объёма текстовых поисковых запросов. // «Ученые заметки ТОГУ». – 2016, Том 7, № 3.
19. А. В. Потемкин, А. Ю. Бородащенко. Алгоритм динамической кластеризации сообщений средств массовой информации сети Интернет по сюжетным линиям. // «Human Progress». –2016, Том 2, № 8.
20. Lesko, S.A.. Zhukov, D.O Trends, self-similarity, and forecasting of news events in the information domain, its structure and director. Proceedings – 2015 IEEE International Conference on Big Data Intelligence and Computing, DataCom 2015, **DOI:** 10.1109/SmartCity.2015.178
21. D.O. Zhukov, S.A. Lesko. Stochastic self-organisation of poorly structured data and memory realisation in an information domain when designing news events forecasting models. The 2nd IEEE International Conference on Big Data Intelligence and Computing. 2016, August 8-12, Auckland, New Zealand, **DOI:** 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.153
22. Sigov, A., Zhukov, D., Novikova, O. Modelling of memory realization processes and the implementation of information self-organization in forecasting the new's events using arrays of natural language texts. Proceeding the 1st International Scientific Conference Convergent Cognitive Information Technologies, Convergent 2016; Moscow; Russian Federation; 25 November 2016 through 26 November 2016; Code 125487, CEUR Workshop Proceedings Volume 1763, 2016, pp. 42-55

References

1. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press, 2007. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008, 544 p.
2. Turney, P. D. The latent relation mapping engine: Algorithm and experiments. // Journal of Artificial Intelligence Research, 33, 2008, P. 615–655.
3. Kumar, A., Kumar, D., Jarial, S.K. A novel hybrid K-means and artificial bee colony algorithm approach for data clustering. // Decision

Science Letters . – 2018, Vol. 7, Issue 1, P. 65-76.
4. Abualigah, L.M., Khader, A.T., Al-Betar, M.A., Alomari, O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. / Expert Systems with Applications. – 2017, 84, P. 24-36.
5. Kanimozhi, K.V., Venkatesan, M. A novel map-reduce based augmented clustering algorithm for big text datasets. // Advances in Intelligent Systems and Computing. – 2018, Vol. 542, P. 427-436.
6. Jenhani, F., Gouider, M.S., Said, L.B. Social stream clustering to improve events extraction. // Smart Innovation, Systems and Technologies. – 2018, Vol. 73, P. 319-329.
7. Jenhani, F., Gouider, M.S., Ben Said, L.A hybrid approach for drug abuse events extraction from Twitter (2016). 20Th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (ICKIIES 2016), York, United Kingdom. -2016.
8. Ailem, M., Role, F., Nadif, M. Sparse Poisson Latent Block Model for Document Clustering. / IEEE Transactions on Knowledge and Data Engineering. – 2017, №29 (7), P. 1563-1576.
9. Li, W., Joo, J., Qi, H., Zhu, S.-C. Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph. // IEEE Transactions on Multimedia. – 2017, Vol. 19, Issue 2, 19(2), P. 367-381.
10. Bafna, P., Pramod, D., Vaidya, A. Document clustering: TF-IDF approach. // International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016. – 2016, P. 61-66.
11. Lamari, Y., Slaoui, S.C. Parallel document clustering using iterative mapreduce. 2016 International Conference on Big Data and Advanced Wireless Technologies, BDAW 2016; Blagoevgrad; Bulgaria; 10 November 2016 to 11 November 2016. // ACM International Conference Proceeding Series.
12. Al-Fath, A.M.U., Saleh, W.K.R., Sa'Adah, S. Implementation of MCL algorithm in clustering digital news with graph representation. // 4th International Conference on Information and Communication Technology, ICoICT 2016; Bandung; Indonesia; 25 May 2016 до 27 May 2016.
13. Patil, H., Thakur, R.S. Document clustering: A summarized survey (Book Chapter) // Pattern and Data Analysis in Healthcare Settings. – 22 July 2016, P. 264-281.
14. Le, T.M.V., Lauw, H.W. Semantic visualization with neighborhood graph regularization. / Journal of Artificial Intelligence Research. – 2016, Vol. 55, P. 1091-1133.
15. Rahmawati, D., Putri Saptawati, G.A., Widyani, Y. Document clustering using sequential pattern (SP): Maximal frequent sequences (MFS) as SP representation. // Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015. – 2016, P. 98-102.
16. Conrad, J.G., Bender, M. Semi-supervised events clustering in news retrieval. // CEUR Workshop Proceedings. – 2016, P. 21-26.
17. Huang, G., He, J., Zhang, Y., (...), You, Y., Cao, J. Mining streams of short text for analysis of world-wide event evolutions. / World Wide Web . -2015, 18(5), P. 1201-1217.
18. O. A. Bezverhij, S. G. Samohvalova. Klasterizacija bol'shogo objoma tekstovyh poiskovyh zaprosov. // «Uchenye zametki TOGU». – 2016, Tom 7, № 3.
19. V. Potemkin, A. Ju. Borodashhenko. Algoritm dinamicheskoj klasterizacii soobshhenij sredstv massovoj informacii seti Internet po sjuzhetnym linijam. // «Human Progress». –2016, Tom 2, № 8.
20. Lesko, S.A. Zhukov, D.O Trends, self-similarity, and forecasting of news events in the information domain, its structure and director. Proceedings – 2015 IEEE International Conference on Big Data Intelligence and Computing, DataCom 2015, **DOI:** 10.1109/SmartCity.2015.178
21. D.O. Zhukov, S.A. Lesko. Stochastic self-organisation of poorly structured data and memory realisation in an information domain when designing news events forecasting models. The 2nd IEEE International Conference on Big Data Intelligence and Computing. 2016, August 8-12, Auckland, New Zealand, **DOI:** 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.153
22. Sigov, A., Zhukov, D., Novikova, O. Modelling of memory realization processes and the implementation of information self-organization in forecasting the new's events using arrays of natural language texts. Proceeding the 1st International Scientific Conference Convergent Cognitive Information Technologies, Convergent 2016; Moscow; Russian Federation; 25 November 2016 through 26 November 2016; Code 125487, CEUR Workshop Proceedings Volume 1763, 2016, pp. 42-55

Поступила: 29.09.2017

Об авторах:

Отраднов Константин Константинович, соискатель, старший преподаватель кафедры автоматизированных систем управления Института комплексной безопасности и специального приборостроения, Московский технологический университет (МИРЭА), const.otradnov@yandex.ru

Жуков Дмитрий Олегович, доктор технических наук, профессор, заместитель директора по научной работе института комплексной безопасности и специального приборостроения, Московский технологический университет (МИРЭА), zhukovdm@yandex.ru

Новикова Ольга Александровна, соискатель, заместитель заведующего аспирантурой, Московский технологический университет (МИРЭА), ol-novikova@bk.ru

Note on the authors:

Otradnov Konstantin K., Applicant, Senior Lecturer, Moscow Technological University (MIREA), const.otradnov@yandex.ru

Zhukov Dmitry O., Doctor of Technical Science, Professor, Deputy Director on scientific work, Moscow Technological University (MIREA), zhukovdm@yandex.ru

Novikova Olga A., Applicant, Deputy Head of the graduate school, Moscow Technological University (MIREA), ol-novikova@bk.ru