Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики[*].

Киселев М. В. Пивоваров В. С. Шмулевич М. М.
Компания Megaputer Intelligence

Данная работа посвящена автоматической смысловой кластеризации текстов и ее применению к анализу динамики тематического состава потока новостей. Проанализированы существующие методики кластеризации, и показано, что ни одна из них не обладает полным набором качеств, необходимых для успешного решения этой задачи. С целью преодоления этих трудностей предложен новый метод, названный островной кластеризацией, который основан на статистической мере корреляции встречаемости в текстах термов, характеризующихся значимым превышением их частот над средним уровнем. Показано, что он успешно решает проблемы плоской и иерархической кластеризации новостей, а также отслеживания динамики тем новостного потока.

This paper is devoted to automated clustering of document sets and its application to analysis of electronic news topic structure dynamics. The existing clustering algorithms are considered and it is shown that none of them obeys the full set of requirements necessary for successful solution of this problem. In order to overcome these difficulties a novel method called *island clustering* is proposed. It is based on a statistical measure of term co-occurrence calculated only for the terms showing in some texts significant frequency excess over the average level. It is demonstrated that our method obtains high quality flat and hierarchical clustering of news and allows user to monitor qualitatively and quantitatively evolution of the news stream thematic structure.

[*] Данная работа поддерживалась компанией Яндекс (грант №102930).

1. Введение

Наблюдаемое на протяжении последних десятилетий лавинообразное увеличение числа и общего объема создаваемых и хранимых человечеством документов делает автоматическую кластеризацию, т.е. разбиение текстовых массивов на систему (возможно иерархических) подмножеств, помеченных какими-то их смысловыми описателями, одной из приоритетных задач, решаемых системами поддержки документооборота и другими информационными системами. Примеры частных случаев этой проблемы, интересных с точки зрения их применения в бизнесе и технологии, весьма многочисленны. Службы поддержки клиентов сталкиваются с задачей структурирования и анализа отзывов или жалоб клиентов с целью определения наиболее сильных источников их недовольства и раннего выявления новых причин недовольства. Аналитическим службам производственных компаний приходится кластеризовать отчеты о нештатных ситуациях для организации более эффективного управления производством, выявления опасных трендов повышения потенциальной аварийности. Характерной чертой этих и многих других примеров является то, что кластеризуемые документы представляют собой не фиксированный массив, а скорее поток поступающих текстов, что делает актуальной более сложную задачу отслеживания динамики картины кластеризации. К этому типу относится и представляющая особый интерес в контексте данного исследования задача автоматической кластеризации новостного потока. Корректное разбиение новостного потока в соответствии с его тематической структурой важно не только для провайдеров новостей как фактор, повышающий удобство и привлекательность этого сервиса для его потребителей, но и для конечных пользователей, например, использующих новостную ленту в ручном или полуавтоматическом режиме для принятия решений по биржевым операциям. При этом альтернатива автоматическим методам кластеризации, ручная кластеризация, производимая экспертами, во многих случаях становится все менее и менее привлекательной вследствие растущей относительной дороговизны их труда и, часто, субъективности получающейся кластеризации.

Отметим, что задача кластеризации корпуса текстов имеет много общего с задачей классификации текстов в заранее заданную пользователем систему категорий, заполненную заведомо правильно классифицированными текстами (называемыми обучающим массивом текстов). Однако имеется ряд особенностей, приводящих к разительному контрасту между обилием хорошо изученных и эф-

фективных методов классификации, многие из которых успешно применяются на практике, и гораздо более бедным набором кластеризационных алгоритмов с весьма ограниченной практической применимостью. Одной из главных причин такого положения является то обстоятельство, что, в отличие от задачи классификации, задача кластеризации текстов с трудом поддается формализации. В то время, как существуют объективные и точные методы, позволяющие оценить точность классификационных моделей (которая является главным показателем их качества), оценка адекватности разбиения текстов на кластеры, как правило, основывается на мнении экспертов и трудновыразима в виде какой-то одной численной характеристики. С этим обстоятельством связана и еще одна трудность, отсутствующая в случае решения задач классификации, – требование интерпретируемости результата, полученного процедурой кластеризации. Действительно, тогда как автоматически построенная на обучающем массиве текстов классификационная модель может быть сразу применена для классификации новых текстов, даже не будучи проинспектирована человеком, результаты работы кластеризатора предполагают, как правило, их понимание и интерпретацию пользователем. Это означает, что кластеры должны быть не только получены, - им еще должны быть присвоены некоторые метки, отражающие их семантику.

Еще одно требование к процедуре кластеризации становится особенно важным, если мы хотим рассматривать динамику распределения документов по кластерам во времени. Это требование можно нестрого сформулировать как статистический характер получаемых кластеризатором результатов. Действительно, если нас интересует исключительно данный набор текстов, который предполагается нерасширяемым, то мы можем искать в нем систему кластеров, не задаваясь вопросом, отражает ли эта картина кластеризации свойства какой-то большой совокупности текстов, частью которой является данный набор. Если же, в противоположность этому, мы рассматриваем изучаемый массив только как часть некой генеральной совокупности, то мы хотели бы, чтобы полученные кластеры характеризовали всю генеральную совокупность – в частности, чтобы кластеры, полученные на разных выборках из этой совокупности, были бы в какой-то степени подобны. Эти подходы к кластеризации можно назвать, соответственно, онтологическим и статистическим. Разницу между ними можно ярко проиллюстрировать на следующем примере. Если мы предъявим корпус, где каждый текст состоит из полностью случайного набора слов, кластеризаторам этих двух типов, то первый найдет в нем некоторые кластеры, отражающие

случайные близости каких-то документов, тогда как второй никаких кластеров обнаруживать не должен. Очевидно, что для того чтобы картина кластеризации обладала свойством преемственности, устойчивости при анализе документов из соседних временных срезов, кластеризационная процедура должна быть статистической по своей природе.

К этим требованиям можно добавить еще несколько требований практического характера: возможность опционального отнесения одного документа к нескольким кластерам (что, как показывает опыт, важно при кластеризации новостей), масштабируемость – т.е. не слишком сильная зависимость времени работы алгоритма от количества текстов (обычно грань между масштабируемыми и немасштабируемыми системами проводят где-то между логлинейной и квадратичной зависимостью), а также требование минимальности числа настроечных параметров алгоритма, что обеспечивает легкость его использования и универсальность.

Таким образом, еще раз кратко сформулируем свойства, какими, по нашему мнению, должна обладать процедура кластерного анализа для того, чтобы быть практически применимой к кластеризации больших массивов текстов вообще и к анализу динамики тематической структуры потока новостей в частности:

- интерпретируемость найденных кластеров в терминах смысла содержания относящихся к ним документов;
- статистическая значимость группирования текстов в кластеры;
- возможность отнесения документа более, чем к одному кластеру;
- не более, чем логлинейный рост времени работы кластеризатора с увеличением количества текстов;
- минимальная (а лучше вообще отсутствующая) настройка со стороны пользователя.

В следующих разделах настоящей работы мы рассмотрим существующие техники кластеризации текстов, проанализируем присущие им слабые стороны, препятствующие удовлетворению сформулированных выше требований, предложим оригинальную методику кластеризации текстов, названную нами «островная кластеризация», лишенную этих слабых сторон, а также продемонстрируем результаты применения нашего кластеризатора для смыслового группирования текстов и отслеживания динамики тематической структуры новостного потока.

2. Краткий обзор существующих техник кластеризации текстов.

При рассмотрении в самом общем виде методы кластеризации могут быть разбиты на два типа: представляющие тексты в виде векторов в многомерном пространстве признаков (и использующие метрику близости между векторами) и методы, основывающиеся на других представлениях кластеризуемых текстов. Первая группа методов представлена алгоритмами иерархической кластеризации Single/Complete/Average Link, неиерархическими алгоритмами K-Means/EM, а также большим числом других базирующихся на них методов. Примерами алгоритмов второй группы являются алгоритм Suffix Trie Clustering (STC) и алгоритм категоризации текстов системы PolyAnalyst.

В методах кластеризации, основывающихся на метрике близости, документ представляется в виде многомерного вектора в пространстве признаков. Подходы к формированию вектора признаков документа могут существенно различаться. В простейшем случае каждый признак соответствует наличию в тексте одной из словоформ, встречающейся в рассматриваемом наборе текстов. Величина компоненты вектора тоже может определяться по-разному: например, компонента может быть равна единице, если данный термин присутствует, и нулю в противоположном случае; может быть равна числу вхождений термина в документ (*вектор частот*, далее обозначаемый tf) или вычисляться по несколько более сложным формулам, учитывающим среднюю встречаемость словоформы по всему набору текстов или по внешнему корпусу текстов (т.н. *интегральная значимость*, часто обозначаемая tfidf). Вектор, отвечающий данному тексту, нормируется на единицу. Мерой близости между текстами считается скалярное произведение между соответствующими векторами.

Большинство алгоритмов кластеризации оперирует в качестве входных данных прямоугольной матрицей T, составленной из векторов документов, и симметричной квадратной матрицей $S=T\cdot T^t$, именуемой *матрицей близости*. Матрица T, составленная из векторов tfidf, обозначается TFIDF. Основной проблемой методов, основанных на такой матрице, оказывается слишком большая размерность пространства признаков, большая часть которых является избыточными и даже вредными. Например, при кластеризации текстов в данной предметной области термины, не относящиеся к этой области, могут маскировать сходство между документами.

Для уменьшения размерности пространства признаков могут применяться, в частности, следующие приемы:

a) Использование стоп-листов, содержащих списки «несмысловых» слов.

b) Использование методов лингвистики:

   Использование словарей и тезаурусов для группировки словоформ по нормальным формам и объединения нормальных форм в синонимические группы.

   Более развитый вариант того же метода может использовать семантическую сеть (см., например, [6]) и группировать термины с использованием отношений более сложного типа (например, голонимии и гипернимии). Некоторые кластеризационные алгоритмы используют в качестве элементов вектора tf не отдельные словоформы, а выделяемые именные или глагольные группы, имена собственные, устойчивые словосочетания [9]. В качестве вспомогательных, для уточнения величин компонент вектора tf могут применяться разнообразные методы разрешения омонимии [2, 10] и полисемии [13].

c) Использование лингво-статистических методов, использующих информацию об априорных вероятностях встречаемости слов для включения в вектор tf только статистически значимых терминов, например, определив величину доверительного интервала для частот, и включая только термины, частота которых в каком либо из текстов выходит за верхнюю границу интервала.

d) Использование алгебраических и вероятностных методов для разложения векторов в пространстве признаков по минимальному числу главных линейно-независимых компонент, с учетом корреляционных связей между терминами.

При использовании описанного представления текстов кластер оказывается набором векторов в пространстве признаков. Функция близости между кластерами выводится тем или иным образом из функции близости между векторами. Единственное ограничение – требование монотонности функции близости. Важной характеристикой кластера является положение его *центроида* (т.е., вектора центра масс, среднего арифметического векторов).

Наиболее популярными иерархическими алгоритмами, основанными на близости текстов в пространстве признаков, являются алгоритмы Single/Complete/Average Link [14]. Есть варианты этих алгоритмов, работающие как сверху-вниз (рассматривающие сначала все тексты как один кластер), так и снизу-вверх (начинающие работу с кластеров, состоящих из единственного документа). Результатом работы этих алгоритмов является дендрограмма (бинар-

ное дерево), связывающее все тексты. При заданном вручную числе кластеров (или предельной величине близости) делается соответствующее сечение бинарного дерева, дающее разбиение текстов на кластеры. Алгоритм кластеризации «снизу-вверх» локально-оптимален, и работает на объединение двух ближайших элементов. Расстояние между кластерами может определяться как:

a) Минимальное расстояние между парой объектов в соседних кластерах (для Single Link)
b) Максимальное расстояние между парой (для Complete Link)
c) Среднее расстояние между элементами двух кластеров (что соответствует скалярному произведению центроидов).

Метод Single Link работает быстрее, чем Complete Link ($O(N^2)$ против $O(N^3)$, где N – число документов), но построенные кластеры часто оказываются чрезмерно «вытянутыми». Метод Group Average представляет собой компромисс между Single и Complete Link по скорости/точности.

В контексте нашей задачи эти методы представляются малопригодными, так как они не удовлетворяют, в сущности, ни одному из сформулированных во введении требований. Смысл полученных кластеров трудно формализовать, так как кластер описывается главным образом положением его центроида в многомерном пространстве признаков. Процедура кластеризации не статистична. Документы относятся лишь к одному кластеру. Время работы не менее, чем квадратично. И кроме того, требуется заранее знать, сколько кластеров присутствует в тексте.

Несколько более подходящими представляются алгоритмы K-Means и EM (Expectation Maximization)[5]. Под этим названием объединяется обширная группа отличающихся высокой производительностью (порядка O(N)) неинкрементных методов плоской кластеризации. В простейшем случае (K-Means) требуется задание числа кластеров и начальных положений центроидов кластеров, после чего запускается итеративный процесс, стабилизирующий положение центроидов. На каждом шаге процесса документы приписываются к кластеру с ближайшим центроидом. После того, как все тексты распределены, вычисляется новое положение центроидов. Процесс останавливается, когда центроиды перестают перемещаться, или удовлетворяется критерий остановки. Существуют вариант этого алгоритма [4], в котором применяется вычисление положений затравочных центроидов при помощи методов Single/Average Link на случайном подмножестве данных. Размер случайной выборки в этом случае выбирается равным $\sqrt{kn}$, где k – желательное число кластеров, n – число документов. Этот вариант часто применяется

поисковыми серверами для кластеризации поисковых запросов. Применяются также методики, позволяющие оценивать необходимое число кластеров автоматически. Наиболее известной является Minimum Description Length (MDL) [3], применяющаяся в системе AUTOCLASS. Идея этого метода состоит в том, что для каждого кластера и для каждого объекта, входящего в кластер, некоторой процедурой генерируется строка описания. Предполагается, что оптимальное количество соответствует минимальной длине описания. K-Means, один из самых простых в реализации и, вероятно, самый распространенный из методов кластеризации, является одним из частных случаев общего метода EM. Метод EM оперирует вероятностной моделью отнесения документа к определенному кластеру. Предполагается, что существует k (по числу кластеров) скрытых независимых «генераторов», подчиняющихся многомерному закону нормального распределения. Вектора документов рассматриваются в качестве реализации многомерной случайной величины. Если параметры распределений известны, можно вычислить условную вероятность принадлежности вектора документа к одному из «генераторов».

Из этого описания видно, что и эти методы не решают кардинально многих проблем предыдущего семейства кластеризаторов. Их самая сильная сторона – это линейная зависимость времени их работы от количества документов.

Описанные выше техники кластеризации могут являться базой для построения более эффективных кластеризаторов, будучи скомбинированы с разнообразными методами уменьшения размерности пространства признаков. Эти методы понижают размерность пространства признаков за счет анализа корреляций терминов, позволяя также снимать проблемы синонимии и омонимии без использования тезауруса, так как фактически объединяют в один процесс вариант кластеризации на уровне терминов (которая используется для группировки родственных слов) и кластеризацию на уровне документов. Мы рассмотрим наиболее часто применяемые из этих методов.

В методе *латентного семантического анализа* LSA/LSI [7] учитываются связи элементов векторов tfidf (т.е. используется совместная встречаемость на уровне документа). Матрица TFIDF раскладывается по базису главных компонент. Размер базиса обычно задается вручную. После того, как вектора tfidf спроецированы на пространство меньшей размерности, применяются обычные методы кластеризации, такие как Single/Average Link или K-means/EM. Производительность метода LSI/LSA порядка $O(k^3N^2)$, где k – редуцированная размерность пространства признаков, N – количество

документов. Метод имеет также интересные применения для улучшения качества поиска.

Метод анализа основных компонент (Principal Component Analysis – PCA)[1] использует диагонализацию полной ковариационной матрицы терминов. Пространство уменьшенной размерности строится на собственных векторах ковариационной матрицы, соответствующих нескольким наибольшим собственным значениям. В связи с невысокой скоростью работы этот метод вряд ли может использоваться для кластеризации больших корпусов текстов.

В [8] предложен двухпроходный метод, называемый индексацией концептов (Concept Indexing), в котором размерность пространства признаков уменьшается за счет использования базиса, полученного из центроидов кластеров, найденных на первом проходе процедуры кластеризации. Число кластеров k, соответствующее размерности редуцированного пространства, задается вручную.

Наконец, для целей уменьшения размерности пространства признаков может быть использован тезаурус, например, один из самых известных - WordNet [6]. В этом тезаурусе все слова и словосочетания сведены в синонимические группы или *синсеты*. Одно слово может входить в несколько таких групп. Группы связываются между собой отношениями различных типов, зависящими от частей речи. Особый интерес имеют связывающие между собой группы существительных и группы глаголов отношения *гипернимии* (частное-общее), а также связывающие группы существительных отношения *голонимии* (часть-целое). При использовании тезауруса WordNet можно заменять словоформу на набор маркеров синонимических групп, к которым она принадлежит, а затем, используя выбранный тип отношений, увеличивать веса групп, прямо или косвенно связанных с выбранными. Компонентам вектора признаков соответствуют в этом случае не слова, а синонимические группы.

Перейдем теперь к методам, не использующим векторного представления текстов и метрики близости. Мы рассмотрим два из них: Suffix Trie Clustering (STC) [15] и алгоритм категоризации текстов системы PolyAnalyst [12].

Метод STC основан на использовании древовидных структур suffix trie, использующихся для поиска за время, пропорциональное длине поисковой строки. Время построения поисковой структуры растет пропорционально величине индексируемой базы. В методе STC строится структура *suffix trie* для последовательности идентификаторов, объединяющей все тексты. Как обычно, при построении могут использоваться идентификаторы словоформ, нормальных форм или (см. выше) синонимических групп. В каждом узле сохра-

няется информация о текстах, прошедших через этот узел. Внешние узлы соответствуют (как правило) отдельным словам, внутренние – часто повторяющимся словосочетаниям. Узлы, отвечающие отдельным словам или словосочетаниям, рассматриваются как затравки для кластеров. Затравки объединяются, если содержат достаточный процент (порядка половины) общих текстов. Процесс останавливается, когда больше нет сильно пересекающихся кластеров. При этом каждый кластер помечается набором слов и словосочетаний, соответствующих вошедшим в него узлам первоначального дерева.

Этот метод в гораздо большей степени удовлетворяет нашим требованиям. К его недостаткам можно отнести отсутствие статистической верификации его результатов и значительное время работы при больших размерах первоначального дерева.

В завершение нашего обзора упомянем кластеризационный алгоритм, реализованный в системе анализа данных PolyAnalyst. Этот алгоритм основывается на двух статистических характеристиках распределения термов (нормальных форм, устойчивых словосочетаний и синсетов тезауруса WordNet) по документам. Одна из них – это мера аномальности частоты терма в тексте, вычисляемая на основе средней частоты этого терма по всему набору текстов. Она имеет значение вероятности обнаружить данное или большее количество термов в тексте при предположении о равномерности распределения терма по всем текстам. Вторая – это величина антикорреляции встречаемости пары термов в текстах. Идея алгоритма состоит в выборе набора термов, которые, с одной стороны, имеют аномально высокие частоты в большом количестве текстов, а с другой – показывают большие величины попарных антикорреляций. Каждому такому терму соответствует кластер содержащих его текстов. Алгоритм естественным образом иерархичен, так как эта процедура может быть повторена с каждым полученным на предыдущем шаге кластером, пока находится хотя бы одна пара значимо антикоррелирующих термов. Данный алгоритм предполагает неэксклюзивную кластеризацию – каждый документ может принадлежать нескольким кластерам (либо ни одному). Он имеет статистическую природу, линейное по количеству документов время работы и создает понятные описания кластеров. Однако данный алгоритм также имеет весьма существенный недостаток – малую устойчивость выделения множеств антикоррелирующих термов по отношению к вариации его настроечных параметров, которые достаточно многочисленны.

В следующем разделе мы опишем предлагаемый нами алгоритм кластеризации, который представляется лишенным этих не-

достатков, и удовлетворяет всем сформулированным выше требованиям.

3. Метод «островной кластеризации» текстов.

Предлагаемый нами метод, так же как и большинство методов кластеризации, упомянутых в предыдущем разделе, работает с текстами, рассматриваемыми как неупорядоченные наборы слов, отбрасывая всю информацию о положении слов в тексте и используя только их частоты. Точнее говоря, элементами этих наборов являются не слова, а нормальные формы слов, получаемые с помощью морфологического анализа (который, в принципе, может быть заменен стеммингом [11]), и устойчивые словосочетания. Последние выделяются в тексте на этапе предобработки на основе статистического алгоритма, находящего статистически маловероятные последовательности нормальных форм [12], - мы не будем на нем останавливаться, так как он не является принципиальной составляющей описываемого метода кластеризации.

Для сокращения времени работы алгоритма из этих наборов слов могут исключаться частые несмысловые слова (при наличии списка таковых), слова, не являющиеся существительными, или слова, чья частота в документе не превышает существенно частоты во всем кластеризуемом корпусе текстов. Также могут быть исключены слова, встречающиеся менее, чем в заранее заданном пороговом числе документов. Далее все слова и словосочетания, включаемые в эти наборы, будут обобщенно называться *термами*. Общее количество всех термов, включенных хотя бы в один набор, будет обозначаться N_{terms}, а общее количество документов - N_{docs}

Первая часть алгоритма состоит в построении так называемого *графа корреляций термов*. Этот граф задается матрицей парных корреляций булевых переменных a_{ip}, отражающих наличие терма i в документе p, так что связь между термами i и j считается существующей при достаточно сильной (большей пороговой) корреляции между переменными a_i и a_j. Степень корреляции между термами i и j определяется следующим образом. Пусть n – общее количество термов во всех документах, n_i – количество термов в документах, в которых встречается терм i. Обозначим общее число употреблений терма во всех текстах как N_j, а количество термов j в документах, содержащих терм i – как N_{ij}. Если принять гипотезу, что термы i и j распределены в документах независимо друг от друга, то вероятность того, что в документах, содержащих терм i, окажется N_{ij} или

более термов j – это вероятность получения не менее N_{ij} успехов в серии из N_j испытаний при вероятности успеха одного испытания, равной n_i/n, т.е. $P_B(N_{ij}, N_j, \frac{n_i}{n})$, где $P_B(n, N, p) = \sum_{i=n}^{N} b(i; N, p)$, b – биномиальное распределение.

Вероятность $p_{ij} = P_B(N_{ij}, N_j, \frac{n_i}{n})$ может быть принята в качестве меры корреляции между термами i и j, - чем она меньше, тем более коррелированы эти термы.

В программной реализации нашего алгоритма мы выбираем пороговое значение вероятности, определяющей достоверность связи межу термами, равным

$$p_c = \frac{0.03}{\max\left(N_{terms}^2, N_{docs}\right)}, \quad (1)$$

что позволяет избежать получения чрезмерного количества малозначимых связей в случае большого массива разнообразных текстов. Впрочем, эксперименты показывают, что результаты алгоритма мало зависят от этого параметра, так как он влияет на состав лишь самых малозначимых кластеров, тогда как связи между термами, составляющими основные кластеры, для сколько-нибудь значительных текстовых массивов с достаточно выраженной тематической структурой, характеризуются очень низкими значениями p_{ij}.

Перед тем как приступить к описанию второй части нашего алгоритма кластеризации, отметим, что величина p_{ij} все же не совсем подходит для описания силы связи термов i и j, в частности, потому что она, как легко видеть, не симметрична: $p_{ij} \neq p_{ji}$. Поэтому в нашем алгоритме в качестве меры корреляции термов берется $\tilde{p}_{ij} = \max(p_{ij}, p_{ji})$. То, что надо использовать максимум от этих двух вероятностей, можно продемонстрировать на следующем примере. Допустим, в массиве из N текстов одинакового размера терм i встречается лишь один раз и лишь в одном документе. Терм j тоже встречается лишь в этом же тексте, но его количество там достаточно велико. Интуитивно ясно, что связь между термами i и j не значима. Но если мы посмотрим на значения p_{ij} и p_{ji} в этом случае,

то увидим, что только p_{ji} достаточно велика ($= 1/N_{docs}$), чтобы сигнализировать о незначимости этой связи (см. (1)); $p_{ij} = \frac{1}{N_{docs}^{N_j}}$ может иметь весьма малую величину при большом количестве термов j. Операция взятия максимума от p_{ij} и p_{ji} устраняет эту проблему.

Вычисление матрицы $\widetilde{p}_{ij}$ является самым вычислительно-емким шагом алгоритма, однако отметим, что это однопроходная процедура – время ее работы линейно зависит от количества документов. Кроме того, на практике лишь небольшая доля всех значений $\widetilde{p}_{ij}$ оказывается ниже порогового значения вероятности, что позволяет эффективно оптимизировать ее вычисление (например, не вычисляя значения p_{ij} для пар термов с $N_{ij} \le \frac{n_i N_j}{n}$).

Подчеркнем, что процедура определения связей между термами носит по своему существу статистический характер, определяющий статистический характер и всей процедуры кластеризации. А именно, она основана на проверке статистической гипотезы о попарной независимости присутствия в документах термов. При этом уровень достоверности (1) корректируется на общее количество проверяемых гипотез ($\sim N_{terms}^2$). Это означает, что, например, для текстов, представляющих собой полностью случайные наборы слов, эта процедура с большой вероятностью не найдет ни одной связи между термами, а следовательно, и весь алгоритм кластеризации не найдет ни одного кластера.

Итак, мы имеем меру связи между термами $\widetilde{p}_{ij}$. Теперь обсудим, как на основе этой меры мы строим кластеры термов и соответствующие им кластеры документов, что составляет суть второй части описываемого алгоритма. Входными данными для процедуры кластеризации является тройки *связей* $< i, j, \widetilde{p}_{ij} >$, упорядоченные по возрастанию $\widetilde{p}_{ij}$ (т.е. в начале идут самые сильные связи). Каждый шаг алгоритма будет соответствовать очередной связи из этого набора связей. Алгоритм имеет один параметр T_S – минимальный размер кластера документов, при котором он перестает расти. Выбор его значений мы обсудим после описания алгоритма. Растущие в процессе выполнения алгоритма кластеры мы будем называть ост-

ровами. Каждый остров задается множеством его термов и множеством связей между ними. В алгоритме будут фигурировать множества растущих островов **G** и множество зафиксированных островов **F**. Еще в алгоритме будет использоваться так называемое *множество несвязываемых термов*, которое будет обозначаться как **PRO.** Оно будет включать термы, относящиеся к полностью сформированным островам, которые тем самым не могут быть связаны с новыми термами.

Наш алгоритм кластеризации существует в двух вариантах – мы будем называть их вариантами L и N. В варианте L множества **F**, **G** и **PRO** перед выполнением алгоритма предполагаются пустыми. В варианте N перед выполнением алгоритма множество **G** содержит N_{terms} островов, так что каждый остров включает лишь один терм, а множество **PRO** включает те термы, которые содержатся в T_S или более документах. Во время описания алгоритма будет использоваться обозначение Pop(C) – множество документов, относящихся к острову C. Оно определяется по-разному в вариантах L и N. В варианте L документ считается принадлежащим острову C, если он содержит пару термов, соответствующих хотя бы одной связи острова. В варианте N документ считается принадлежащим острову C, если он содержит хотя бы один терм, входящий в этот остров. Кроме того, в алгоритме часто будет использоваться операция *фиксации* острова C из множества **G**, состоящая в выполнении следующих трех шагов:

1. Остров C удаляется из множества **G**;
2. Остров C добавляется ко множеству **F**;
3. Все термы острова C добавляются ко множеству **PRO**.

Каждая итерация алгоритма состоит в рассмотрении очередной связи $< i, j, \widetilde{p}_{ij} >$ из списка связей с убывающей силой, пока все связи не будут исчерпаны, после чего алгоритм останавливается. Во время каждой итерации алгоритма выполняется следующая процедура:

Если $i \in \mathbf{PRO}$ и $j \in \mathbf{PRO}$,
    перейти к следующей связи.
Если $i \in \mathbf{PRO}$,
    если существует остров C, содержащий терм j,
        остров C фиксируется;
    иначе $\mathbf{PRO} \leftarrow \mathbf{PRO} \cup \{j\}$;

перейти к следующей связи.

Если $j \in \mathbf{PRO}$,

если существует остров С, содержащий терм *i*,

остров *C* фиксируется;

$\mathbf{PRO} \leftarrow \mathbf{PRO} \cup \{i\}$;

перейти к следующей связи.

Если ни *i,* ни *j* не принадлежат ни одному острову,

добавить к множеству **G** остров, состоящий из термов *i* и *j* и связи между ними;

перейти к следующей связи.

Если *i* и *j* принадлежат одному острову,

добавить к этому острову связь между термами *i* и *j*;

перейти к следующей связи.

Если *i* и *j* принадлежат двум разным островам *C* и *D*,

Если $|\text{Pop}(C)| > T_S$ или $|\text{Pop}(D)| > T_S$,

остров *C* фиксируется;

остров *D* фиксируется;

иначе

удалить острова *C* и *D* из **G** и добавить к **G** остров, являющийся объединением этих островов и связи между термами *i* и *j*;

перейти к следующей связи.

Если *i* принадлежит некоторому острову, а *j* не принадлежит ни одному острову,

добавить к этому острову связь между термами *i* и *j* и терм *j*;

перейти к следующей связи.

Если *j* принадлежит некоторому острову, а *i* не принадлежит ни одному острову,

добавить к этому острову связь между термами *i* и *j* и терм *i*;

перейти к следующей связи.

Когда исчерпаются все связи из списка, последним шагом алгоритма фиксируются все острова из множества **G**, если они там есть. Результатом алгоритма является множество **F**. Каждый остров *C* из этого списка соответствует найденному кластеру термов. Pop(*C*) – соответствующий ему кластер документов.

Кластеры термов, находимые этим алгоритмом, обладают следующим свойством. Как нетрудно видеть, все связи между термами, принадлежащие острову, сильнее любых других связей термов этого

острова с другими термами. Таким образом, острова соответствуют наборам выраженно ассоциированных в текстах понятий.

Чтобы закончить описание нашего алгоритма кластеризации, осталось сказать несколько слов о выборе параметра T_S. Этот параметр определяет размеры и количество находимых кластеров. Если установить его в 0, то в случае алгоритма N мы получим, что каждому терму соответствует кластер, а в случае алгоритма L, кластеров тоже будет много, и они будут, скорее всего, небольшими, так как будут заканчивать свой рост при соприкосновении с другими кластерами. Если сделать этот параметр равным N_{docs}, то кластерами будут служить все несвязанные между собой подграфы графа связей термов. Во многих случаях хорошая картина кластеризации получается при установке его в несколько раз меньшим N_{docs}. Так как этот параметр влияет лишь на вторую часть алгоритма, которая всегда работает очень быстро (так как вообще не зависит от размера исследуемого массива текстов), то аналитик, занимающийся кластеризацией, легко может подобрать величину T_S, соответствующую наиболее четкой и интерпретируемой кластеризации. Если состав текстов, подвергающихся кластеризации, остается более-менее однородным в разных кластеризуемых массивах (а это условие можно считать выполняющимся при анализе новостей, относящихся к последовательным временным интервалам), то выбор T_S для всех массивов должен быть сделан так, чтобы отношение T_S к N_{docs} было постоянно.

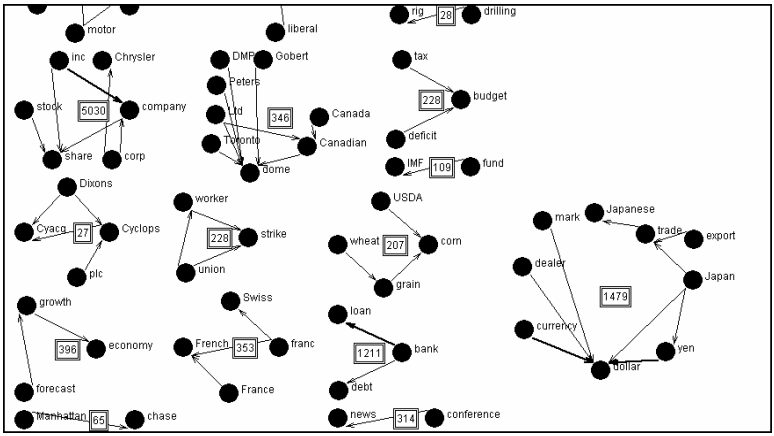
Описанием полученного кластера может служить перечень входящих в него термов и их связей. Обычно количество термов, составляющих остров, невелико, – практически никогда не превышает двух-трех десятков. Поэтому представление «смысла» найденного кластера часто бывает весьма прозрачно в отличие от описаний центроидов кластеров в многомерных пространствах признаков, получаемых большинством существующих кластеризаторов.

Легко видеть, что получаемое распределение документов по кластерам неэксклюзивно – один документ может находиться в нескольких кластерах или ни в каком кластере, если, скажем, он совсем не содержит смысловых слов. Учитывая масштабируемость нашего алгоритма (линейный рост времени его работы с увеличением числа документов) и наличие единственного существенного настроечного параметра T_S, мы можем прийти к заключению, что метод островной кластеризации вполне удовлетворяет всем сформулированным во введении требованиям. Как мы увидим в следующем разделе, это позволяет успешно применять технику островной кластеризации текстов к автоматическому группированию новостей

по тематическому признаку и для отслеживания динамики развития тем новостей как в смысле числа новостей, относящихся к разным темам, так и в смысле состава ассоциированных с данной темой понятий.

4. Островная кластеризация корпуса новостей: иерархия и динамика тем.

В процессе разработки и тестирования нашего алгоритма мы проверяли качество получаемых им результатов на нескольких больших корпусах текстов из корпоративных хранилищ, интернет-форумов и тематических текстовых хранилищ. Для целей данной работы мы продемонстрируем результаты, полученные алгоритмом островной кластеризации на публично доступном репозитории новостей, известном как Reuters-21578, который использовался как стандартный тестовый набор в многочисленных исследованиях по автоматической категоризации текстов[1]. Этот массив включает 21578 текстов новостей, помеченных соответствующими им топиками (так, что каждая новость может быть отнесена к нескольким топикам или ни к одному), и датой поступления новости. Количество топиков равно



[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

Рис. 1. Фрагмент общей картины кластеризации массива новостей Reuters-21578, полученной алгоритмом островной кластеризации.

135. В нашем исследовании мы отбросили слишком короткие новости (длиной менее 100 символов), так как таковые содержат главным образом нетекстовую информацию – биржевые котировки. После этого общее количество новостей составило 17545. В наших экспериментах мы пользовались реализацией алгоритма островной кластеризации в рамках системы автоматического анализа данных PolyAnalyst, производимой компанией Megaputer Intelligence.

В первом запуске нашего алгоритма мы применили его ко всему этому массиву текстов, задав порог роста кластеров равным 1000 документам, что должно было позволить нам выделить достаточно крупные подмножества текстов, объединенные общей тематикой. Был использован вариант L алгоритма, как более специфичный – дающий меньшее пересечение кластеров. В результате счета, который продолжался 9 минут на заранее проиндексированных текстах (компьютер - 3 GHz Pentium), получилось 54 кластера. Вся картина кластеризации весьма велика и не может быть здесь приведена, поэтому мы включили ее фрагмент (Рис. 1.), содержащий 3 наиболее многочисленных кластера, для того, чтобы дать общее представление о характере получаемых алгоритмом результатов. Каждый кластер, остров связанных термов, помечен количеством принадлежащих ему документов. Из рисунка видно, что разбиение на кластеры достаточно логично с человеческой точки зрения, тематика получающихся наборов документов в большинстве случаев вполне ясна. В результатах фигурируют как большие кластеры документов, объединенных довольно широкой тематикой, так и более мелкие более тематически-специфичные группы. К первым относятся 3 самых крупных кластера: биржевые корпоративные новости (и, в особенности, часто фигурирующая в них корпорация Крайслер) – 5030 текста; тексты о торговых взаимоотношениях Японии и США и их влиянии на курсы доллара и йены (1479 текста); неспецифичная группа текстов о банках и кредитных операциях (1211 текст). К более мелким специфичным кластерам относится, например, кластер с 27 документами о взаимоотношениях компаний Cyclops Corp, Cyacq Corp и Dixons Group PLC.

В пользу качества полученной кластеризации свидетельствует также то, что полученное нашим алгоритмом разбиение документов на группы находит хорошее соответствие системе маркирующих их топиков. О полном соответствии речь не может идти даже в идеале, так как топики сильно несбалансированны по количеству докумен-

тов, и много документов не помечены ни одним топиком, однако для большинства кластеров наблюдается явное доминирование одного из топиков, либо их некоторой комбинации. Например, из 5 самых крупных кластеров только 1 не имеет четко соответствующего ему набора топиков, как это показано на Таблице 1.

Таблица 1. Соответствие кластеров, найденных процедурой островной кластеризации, и топиков новостей, расставленных вручную.

Термы, характеризующие кластер	Число документов	Соответствующие топики	Число документов	Пересечение
share, inc, company, stock, corp, Chrysler	5029	earn, acq	4705	2054
currency, dollar, yen, dealer, mark, Japan, trade, export	1479	trade, money-fx, dlr	1158	661
loan, bank, debt	1211	N/A	N/A	N/A
cts, Shr, net, note	1180	earn	2506	1143
price, OPEC, production, quota, oil, gas, bpd	1097	crude, nat-gas	636	466

Для более тонкого анализа тематической структуры корпуса текстов весьма желательно, чтобы кластеризатор допускал иерархическую кластеризацию, не приводящую к ухудшению качества кластеризации с повышением тематической однородности текстов в более мелких кластерах. Произведенные нами эксперименты показывают, что алгоритм островной кластеризации успешно справляется с этой задачей. Например, фрагмент иерархической кластеризации самого большого кластера из Таблицы 1 показан на Рис. 2 (на каждом последующем уровне там берутся самые большие кластеры). Видно, что на самом верхнем уровне находятся все корпоративные новости, на среднем уровне они распределяются на подтипы (финансовые результаты компаний, биржевые сводки по их акциям, события в жизни компаний, такие как собрания акционеров или совета директоров), тогда как на нижнем уровне находятся новости по отдельным крупным компаниям или отраслям.

Наконец, обсудим, как алгоритм островной кластеризации позволяет проследить изменение тематической структуры новостного потока со временем. В рассматриваемом ниже примере мы будем выявлять достаточно крупномасштабную динамику тем новостей – временное окно, используемое для стратифицирования новостей по

времени, составляет примерно 2 недели. Так как плотность новостей по времени в рассматриваемом массиве сильно понижается в его

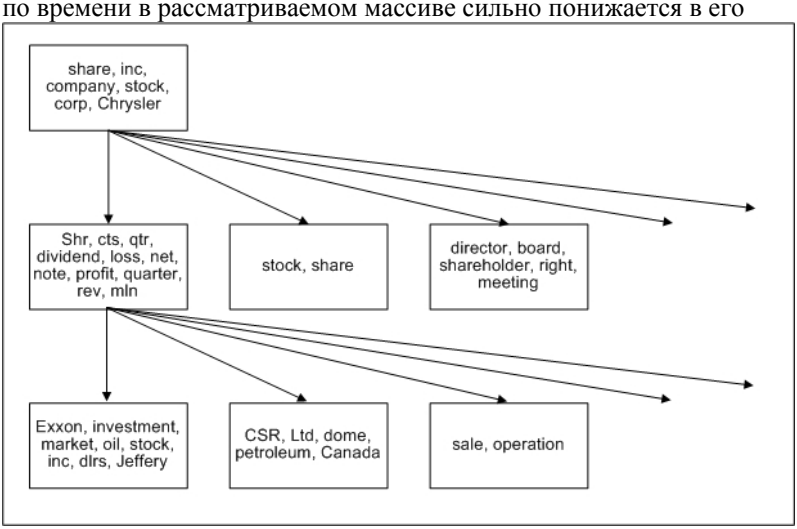


Рис. 2. Пример иерархической кластеризации, полученной алгоритмом островной кластеризации.

более позднем по времени конце, мы разбили все новости на 6 частей, упорядочив их по дате поступления, и взяли для нашего исследования первые 5 частей. В каждом из этих временных слоев мы провели кластеризацию новостей с помощью нашего алгоритма. Как и ожидалось, среди полученных кластеров были совершенно неизменные на протяжении всех 5 периодов, были кластеры с хорошо прослеживаемой преемственностью (в качестве кластеров, преемственных друг по отношению к другу мы брали кластеры, имеющие хотя бы 2 связи одинаковых термов), а также кластеры, специфичные для отдельных временных слоев. Эти типы являются отражением, соответственно, постоянных типовых новостей (например, формализованных биржевых сводок), новостей, содержащих актуальные долгое время темы, но относящиеся к разным конкретным событиям, и новостей об отдельных событиях, реакция новостных агентств на которые длится несколько дней. Нас главным образом интересовали два первых типа новостей. Далее новости, относящиеся к преемственным кластерам, мы будем называть тематическим потоком.

Предлагаемый тип кластеризации позволяет производить как качественный, так и количественный анализ тематических потоков. Количество новостей из разных тематических потоков является индикатором динамики интереса к ним и/или важности происходящих событий. Состав термов в кластерах отображает те или иные аспекты этого интереса или конкретику событий. Для иллюстрации применения наших методов мы взяли 5 выявленных тематических потоков, один из которых (соответствующий новостям мировой нефтегазовой индустрии) был весьма вариабельным в смысле состава термов, а остальные 4 практически не менялись. Это были:

1. финансовые новости по Западной Германии (термы `West, German, Germany, mark`);
2. известия о еврооблигациях (термы `denomination, concession, fee, Luxemburg`);
3. решения ЕС и других общеевропейских учреждений (термы `European, community`);
4. новости профсоюзов (термы `worker, union`).

Построенная диаграмма распределения количества новостей показана на Рис. 3. Там, где к стабильным кластерам добавляются какие-то специфические термы, они показаны на рисунке в виде подписей. Видно, например, как отдельно выявились забастовки моряков и шахтеров, вызвавшие повышение интереса к профсоюзной теме, а также обсуждения проблем экспорта-импорта сахара в ЕС.

Чтобы проиллюстрировать, как метод островной кластеризации позволяет изучить изменения тематической структуру внутри одного тематического потока, рассмотрим кластер, соответствующий новостям о нефтегазовой отрасли за 5 изученных периодов (Рис. 4). В структуре этого кластера прослеживаются отдельные события, влияющие на эту отрасль – землетрясение в Эквадоре (периоды 1 и

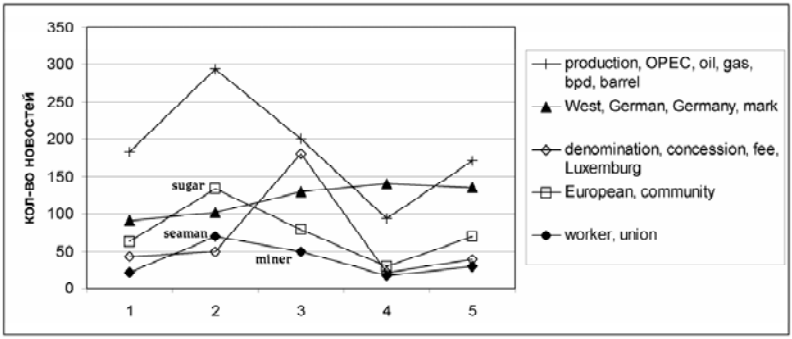


Рис. 3. Количество новостей по 5 постоянным темам (кластерам) в последовательных временных периодах

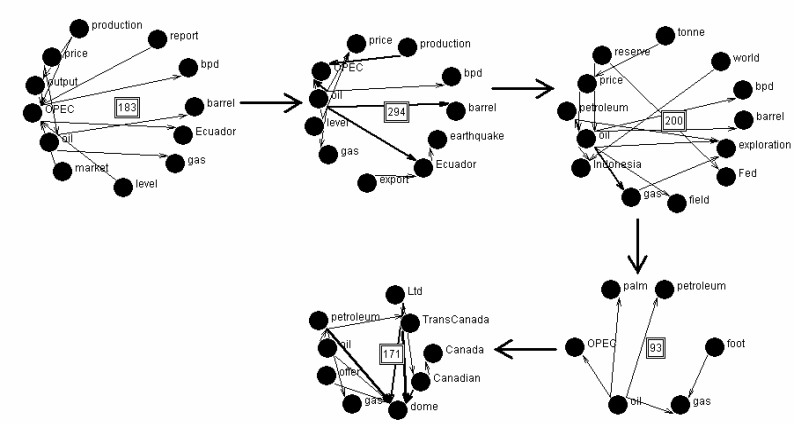


Рис. 4. Изменение состава термов, относящихся к кластеру «новости нефтегазовой отрасли», в последовательных временных периодах.

2), повышение добычи нефти в Индонезии (период 3), финансовые затруднения компании Dome Petroleum (период 5).

Таким образом, мы видим, как метод островной кластеризации позволяет как автоматически выявить тонкую тематическую структуру массива новостей и других документов, так и проследить ее развитие во времени. Так как критерий принадлежности документа к тому или иному кластеру легко формализуется, то это позволяет легко запрашивать соответствующие документы из хранилища на основе достаточно простых поисковых запросов.

5. Заключение.

В данной работе были сформулированы критерии, которым должна удовлетворять процедура кластеризации текстов для того, чтобы быть пригодной для автоматического разбиения массивов текстов и, в частности, новостных потоков на подмножества тематически близких документов, показа тематической структуры найденных подмножеств, а также прослеживания ее динамики во времени, что является актуальным для разнообразных приложений. С точки зре-

ния этих критериев были кратко проанализированы существующие методики кластеризации и показано, что ни одна из них не удовлетворяет этим критериям в полной мере. Для преодоления этих трудностей была предложена новая кластеризационная процедура, называемая методом островной кластеризации. Она не использует метрики близости документов (тем самым избегая квадратичной зависимости времени работы от числа документов, что характерно для многих методов, основанных на близости), но построена на основе статистических мер выраженности присутствия терма в тексте и корреляции встречаемости термов. Таким образом, удается придать нашей процедуре кластеризации статистический характер, позволяющий, в частности, получать сопоставимые результаты при кластеризации новостей, относящихся к соседним периодам времени. Кроме того, оказывается, что полученные такой процедурой кластеры имеют, как правило, весьма лаконичное и понятное пользователю описание в виде связей термов, характеризующих кластер.

Применение процедуры островной кластеризации было проиллюстрировано с использованием публично доступного массива новостей Reuters-21578 на примере трех типов задач: плоская кластеризация, иерархическая кластеризация и прослеживание динамики тематической структуры. Было показано, что метод островной кластеризации может успешно решать эти типы задач, давая в каждом случае описание полученных результатов в понятных человеку терминах.

Литература.

[1] Biber, D., Conrad, S., Reppen, R., Aitchison, J., '*Corpus Linguistics: Investigating Language Structure and Use*', Cambridge Univ. Press, 1998

[2] Brill, E., '*A simple rule-based part of speech tagger*', Proceedings of the Third Annual Conference on Applied Natural Language Processing, ACL, 1992, pp. 152-155

[3] Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D., '*AutoClass: A Bayesian classification system*', In Proc. of 5th Int. Conf. on Machine Learning, 1988, pp. 54-64

[4] Cutting, D. R., Pedersen, J. O., Karger, D., and Tukey, J. W., '*Scatter/gather: A cluster-based approach to browsing large document collections*', In Proceedings of 15th Annual ACM-SIGIR, 1992, pp. 318-329

[5] Dempster, A., Laird, N., Rubin, D., '*Maximum likelihood from incomplete data via EM algorithm*', J. Royal Stat. Society, Series B39, 1977, pp. 1-38

[6] Fellbaum, C. (editor), '*WordNet: An Electronic Lexical Database*', MIT Press, 2005
[7] Hofmann, T., '*Probabilistic Latent Semantic Indexing*', Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, 1999, pp. 50-57
[8] Karypis, G., Eui-Hong (Sam) Han, '*Concept Indexing A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization*', University of Minnesota, Department of Computer Science, Technical Report 00-016, 2000
[9] Kathleen, R., McKeown, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., Sable, C., Schiffman, B., Sigelman, S., '*Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster*', Proceedings of the Human Language Technology Conference, 2002
[10] Kupiec, J., '*Robust Part-of-speech tagging using a hidden Markov model*', Computer Speech and Language 6, 1992, pp. 225-242
[11] Lovins, J.B., '*Development of a stemming algorithm',* Mechanical Translation and Computation, 11(1-2), 1968, pp. 11-31
[12] PolyAnalyst data/text mining system. User manual. http://www.megaputer.com
[13] Ravin, Y. and Leacock, C. (editors), '*Polysemy: Theoretical and Computational Approaches*', New York: Oxford University Press, 2000
[14] van Rijsbergen, C. J., '*Information Retrieval*', London, 1979
[15] Ukkonen, E., '*On-line construction of suffix trees*', Algorithmica, 14(3), September 1995, pp. 249-260