Метод кластеризации текстов, основанный на попарной близости термов, характеризующих тексты, и его сравнение с метрическими методами кластеризации

М. Киселев
Megaputer Intelligence Ltd.
mkiselev@megaputer.ru

Аннотация

Настоящее исследование посвящено разработке нового метода автоматической кластеризации массивов текстов, основанного на представлении текстов в виде наборов *ключевых термов* (различающихся по количеству и составу для разных текстов), а не как точек единого для всех текстов метрического пространства, как в большинстве существующих алгоритмов кластеризации. При этом мера близости текстов основывается на попарной близости термов, характеризующих тексты. Близость термов, в свою очередь, определяется на основе их близости в некотором тезаурусе. Вследствие отсутствия на данный момент общеязыкового русскоязычного тезауруса важную роль в данном исследовании играла разработка (полу)автоматических методов построения тезаурусов с помощью матрицы совместной встречаемости лексем, рассчитанной для большого текстового корпуса. При проведении сравнительного анализа результатов кластеризации использовалась как математическая оценка качества кластеризации, так и ручная оценка качества смысловой пометки найденных кластеров.

1. Введение

Задача автоматической (unsupervised) кластеризации набора текстов [1] при ее актуальности для разнообразных практических приложений остается до сих пор далеко не решенной. Кроме проблем, общих для всех задач кластеризации, дополнительная трудность кластеризации текстов определяется необходимостью индикации смысла найденного кластера – так как обычно результаты кластеризации непосредственно интерпретируются человеком, он или она должны понимать основания для выделения кластера и отнесения к нему того или иного текста. Наиболее распространенный подход к решению этой проблемы состоит в использовании представления текста в виде (обычно нормированного) вектора признаков и евклидовой меры близости между текстами (косинуса угла между векторами), поэтому этот класс методов кластеризации мы будем далее называть метрическими методами. Главный упор в них делается на применении подходящего метода понижения размерности пространства признаков, а находимые кластеры помечаются наиболее отличающимися от 0 координатами центроидов. Наиболее популярные из методов понижения размерности, не требующих специальных лингвистических ресурсов, это отбор наиболее информативных размерностей в соответствии с некоторым критерием (например, суммой tfidf [4,6]) и использование первых n главных размерностей в латентном семантическом анализе [8].

Представление всех текстов изучаемого корпуса как точек единого евклидова пространства, предоставляя естественную меру близости текстов, может содержать в себе в то же время существенное ограничение, особенно явное в случае очень разнообразного по составу корпуса текстов. В этом случае многие из слов, хорошо представляющих смысл текста, будут наличествовать в слишком малом проценте текстов, чтобы быть выбранными или войти в состав редуцированного числа размерностей пространства кластеризации. Очевидно, что это обстоятельство может сделать результаты кластеризации неудовлетворительными.

В данном исследовании в качестве альтернативы или дополнения к евклидовой мере близости текстов предлагается мера, основанная на близости множеств признаков (термов), характеризующих тексты (вообще говоря, разных для разных текстов), – далее они будут называться *ключевыми термами*. Причем, эта мера основана не только на величине пересечения множеств (что редко приводит к качественным результатам), но и на попарной близости отдельных термов из множеств, характеризующих разные тексты. Близость термов должна рассчитываться на основе информации, внешней по отношению к решаемой задаче кластеризации, например, иерархического тезауруса типа WordNet [9] или матрицы совместной встречаемости лексем, рассчитанной на большом корпусе текстов (желательно относящемся к той же области, что и кластеризуемый корпус).

Учитывая специфику данных, предоставленных компанией Яндекс для этого исследования, было решено ограничить наше рассмотрение только русскоязычными текстовыми корпусами. При этом автор столкнулся с трудностью, вызванной отсутствием доступного общеязыкового русскоязычного тезауруса. Для решения этой проблемы мной была разработана методика автоматического создания на основе вышеупомянутой матрицы тезаурусоподобной древовидной структуры (называемой далее *автотезаурусом*), оказавшейся, как мы увидим, вполне пригодной для использования в предлагаемой кластеризационной процедуре. Кроме того, автотезаурус может рассматриваться в качестве исходной точки для ручной модификации с целью создания полноценного тезауруса (что,

впрочем, может составить предмет отдельного исследования). Тем самым, разработанная процедура кластеризации стала обладать дополнительным преимуществом – независимостью от какого-либо создаваемого вручную лингвистического ресурса (кроме программы морфологического анализа).

Имея построенный автотезаурус, мы получили естественную меру близости термов – расстояние по графу автотезауруса между его узлами, соответствующими рассматриваемым термам.

Для сравнения предлагаемого метода кластеризации с метрическими методами были использованы несколько текстовых корпусов с сильно различающимися свойствами: случайная выборка страниц русского Интернета, узкотематическая выборка страниц русского Интернета (страницы, содержащие фамилию Столыпин, из числа страниц, предоставленных Яндексом для данного исследования), тексты новостей за одну неделю, заголовки текстов новостей за одну неделю, статьи уголовного кодекса РФ. Эти текстовые корпуса были выбраны, во-первых, потому, что они представляют разные и наиболее типичные случаи применения текстовой кластеризации, а также потому, что все они снабжены смысловыми классификаторами (иногда многоуровневыми), что позволяет применить числовую меру качества смысловой кластеризации текстов – шенноновскую меру взаимной информации между идентификатором кластера документа и его классификатором. Кроме того, оценивалась адекватность смысловой пометки кластеров, найденных сравниваемыми методами. Эта оценка производилась вручную и имела качественный характер.

2. Метрические методы кластеризации текстов

Как уже говорилось во введении, метрические методы кластеризации предполагают представление каждого текста из анализируемого корпуса как точки в некотором евклидовом пространстве (или вектора). Такой подход дает возможность применить большой спектр существующих процедур кластерного анализа числовых данных. При этом мера близости текстов, необходимая для реализации этих методов, есть просто величина, обратная евклидову расстоянию между соответствующими точками этого пространства, либо косинус угла между соответствующими векторами.

В самом простом варианте каждая размерность этого пространства соответствует некоторому слову или точнее лексеме, включающей все его словоформы (или как вариант – все однокоренные слова, приведенные с помощью процедуры *стемминга* к одной корневой основе). При этом обычно не учитываются часто встречающиеся несмысловые слова, входящие в *стоп-лист*, слова, встречающиеся в малом (1–2) количестве текстов, а также слова, не относящиеся к основным смысловым частям речи (существительное, глагол, прилагательное, наречие). Далее мы будем употреблять для такой сущности, соответствующей одному измерению, название *терм*. Значение по каждой оси вычисляется в разных методах по-разному. Это может быть бинарное значение (0/1), индицирующее присутствие данного терма в данном тексте, либо частота этого терма (отношение количества этого терма к количеству всех слов или термов в тексте), либо более сложная величина, обозначаемая tfidf (term frequency * inverse document frequency – частота терма * обратная частота по документам). Так как в большинстве методов кластеризации применяется именно эта величина, приведем формулу для ее вычисления относительно терма t и документа d:

$$tfidf(d,t) = tf(d,t)\log\left(\frac{N}{df(t)}\right) \qquad (1)$$

Здесь $tf(d,t)$ – частота терма t в документе d, N – число документов в корпусе, $df(t)$ – количество документов, в которых встречается терм t. Обычно также вектора, соответствующие текстам, нормируются на единицу, – т.е. тексты можно рассматривать как точки на гиперсфере с единичным радиусом.

Первая трудность, с которой сталкиваются методы этого типа, – это очень большая размерность получающегося пространства. Анализ показывает, что, в дополнение к вычислительной сложности, кластеризация в пространстве очень высокой размерности редко бывает эффективной, так как каждая точка имеет очень близкие величины расстояния до остальных точек [2]. Поэтому применяются различные методы уменьшения размерности пространства кластеризации. Они разбиваются на две следующие группы.

2.1. Отбор наиболее информативных размерностей

Для каждой размерности вычисляется некоторая величина, характеризующая ее предполагаемую ценность для кластеризации, после чего оставляются n размерностей с максимальным значением этой величины. Например, в методах, где документы представляются их значениями tfidf, в качестве такой величины используется сумма tfidf данного терма по всем документам корпуса [6]. Метрическая кластеризация с применением данного метода сокращения размерности будет использоваться нами в качестве первого метода для сравнения с предлагаемой в данной работе неметрической кластеризацией. Будем обозначать ее и относящиеся к ней результаты как MSEL.

2.2. Использование синтетических размерностей

В этом методе на основе нескольких исходных размерностей вычисляется производная величина, которая соответствует одной размерности нового евклидового пространства. То есть происходит проекция точек исходного евклидового пространства на некоторую его гиперплоскость. Например, при наличии тезауруса можно слить размерности, соответствующие термам-синонимам (или синонимам и гипонимам) в одну новую размерность [6]. Как уже говорилось, этот метод не применим в нашем случае вследствие отсутствия нужного тезауру-

са. Другой механизм построения производных размерностей, эффективность которого была продемонстрирована во многих приложениях, основывается на латентном семантическом анализе [8]. Этот метод, будучи обобщением анализа основных компонент (principal component analysis), основан на так называемом разложении на сингулярные значения (singular value decomposition) матрицы tfidf. В результате такого разложения получаются новые размерности, являющиеся линейными комбинациями изначальных размерностей. Каждой новой размерности соответствует некоторое число, отражающее величину вклада этой размерности в воссоздание исходной матрицы tfidf. В данном методе оставляются n новых композитных размерностей с максимальными значениями этих чисел. Данный метод также используется в нашем исследовании в качестве референтного. Обозначим его как MLSA.

Проблема избыточных размерностей пространства кластеризации является не единственной проблемой метрических методов. Очевидно, что при достаточном относительном разнообразии текстов в корпусе, в том случае, если большинство термов, характеризующих содержание текстов, представлены в небольшом проценте текстов, понижение размерности вряд ли даст хороший результат, так как для образования отдельных кластеров будут важны разные наборы размерностей. Проекция всего пространства кластеризации на одну гиперплоскость просто приведет к потере важной для кластеризации информации.

3. Предлагаемая альтернатива: мера близости текстов, основанная на близости характеризующих их термов

Идея предлагаемого метода состоит в том, чтобы использовать для кластеризации весь набор термов, характеризующих каждый текст в корпусе, но при этом использовать внешнюю по отношению к решаемой задаче кластеризации информацию о смысловой близости, похожести термов. При этом тексты не представляются как точки некоторого евклидового пространства, что не вызывает трудностей, так как существует большой класс алгоритмов кластеризации, которые требуют для своей работы только возможности определения меры близости между объектами кластеризации. Такой подход к определению смысловой близости текстов будет обозначаться как KTPB (key term proximity based).

3.1. Ключевые термы

Прежде всего, определим формально, что мы понимаем под множеством термов, характеризующих тексты, или, как мы далее будем их называть, множество ключевых термов. Как и во многих других подходах, мы рассматриваем в качестве ключевых те термы, частота которых в данном тексте существенно превышает некоторую среднюю частоту. В данном исследовании рассматривались два варианта определения этой средней частоты. В первом варианте в качестве средней принимается частота терма по всему кластеризуемому корпусу. Во втором – частота терма по некоторому очень большому корпусу, представляющему самые разные тексты данного языка, либо, может быть, их некое тематическое подмножество. Однако в последствие было решено не использовать первый метод для кластеризации, так как на многих кластеризуемых корпусах с высокой тематической однородностью текстов он приводил к тому, что для значительной части текстов в корпусе множество ключевых термов оказывалось пустым.

Коль скоро мы знаем среднюю частоту $f(t)$ терма t, мы можем оценить существенность ее превышения $p(d,t)$ в данном тексте d. Пусть текст включает $n(d)$ термов, из них $n(d,t)$ термов t. Тогда в качестве меры значимости превышения частоты терма в данном тексте можно взять вероятность того, что, сделав $n(d)$ испытаний с априорной вероятностью успеха $f(t)$, мы получим $n(d,t)$ или более успехов. Как известно, эта вероятность есть кумулятивная вероятность биномиального распределения с параметрами $f(t)$ и $n(d)$. Чем меньше $p(d,t)$, тем значимее терм t характеризует документ d. Для всех ключевых термов должен выполняться критерий

$$p(d,t) < \frac{0.03}{Nn(d)} \qquad (2)$$

Здесь сделана поправка часто выбираемого порога вероятности при проверке статистических гипотез 0.03 на количество независимых гипотез, которое грубо оценено как количество документов, умноженное на количество термов в данном документе. Не все термы, удовлетворяющие критерию (2), входят во множество ключевых термов данного текста. Как будет описано ниже, процедура кластеризации использует тезаурус, вводящий на множестве термов отношения гипонимии и гипернимии. Мы будем требовать, чтобы никакие два терма из множества ключевых термов одного текста не были связаны отношением гипернимии. Это достигается следующей процедурой прореживания: мы располагаем термы по паденью их значимости и затем, отбирая по одному термы из этого списка во множество ключевых термов, удаляем из списка все гипонимы и гипернимы отбираемого терма. Эта процедура повторяется, пока список не станет пуст.

В процессе проведения данного исследования встречались ситуации (в случае большого корпуса текстов, состоящих из нескольких слов), когда для значительного количества текстов критерий (2) не выполнялся ни для какого терма. В этом случае в качестве ключевых термов брались все лексемы текста.

Скажем несколько слов о том, как во втором методе определялась $f(t)$. Для ее определения была взята случайная выборка текстов из русского Интернета объемом около 30 MB. Для того, чтобы исключить эффекты случайных тематических и жанровых флуктуаций из каждого текста были выкинуты лексемы, ключевые для этого текста (определенные по первому методу). После чего частоты определялись только по неключевым лексемам. В случае, если

лексема не встречалась среди этих неключевых лексем, значение $f(t)$ бралось для нее равным $1/(3n)$, где n – общее количество неключевых лексем в этом корпусе.

Отметим что, при использовании ключевых термов для описания текстов нет необходимости пользоваться стоп-листами, так как маловероятно, что частые несмысловые слова окажутся среди ключевых.

3.2. Определение близости текстов

Итак, для каждого документа d мы знаем его множество ключевых термов $\mathbf{K}(d)$. Предположим также, что для каждой пары термов t_1 и t_2 мы можем определить меру их близости $P(t_1, t_2)$ (это будет предметом дальнейшего рассмотрения). Тогда близость документов d_1 и d_2 выражается следующей формулой:

$$P(d_1,d_2)=\frac{\overline{P}(d_1,d_2)+\overline{P}(d_2,d_1)}{n(d_1)+n(d_2)}, \qquad (3)$$

где

$$\overline{P}(d_1,d_2)=\sum_{t\in\mathbf{K}(d_1)} n(d_1,t)\max_{s\in\mathbf{K}(d_2)} P(t,s). \qquad (4)$$

Отметим, что, если мы не располагаем информацией о близости термов, а можем лишь сравнивать их на равенство, то мера (3) превращается в

$$P_1(d_1,d_2)=\frac{\sum_{t\in\mathbf{K}(d_1)\cap\mathbf{K}(d_2)}(n(d_1,t)+n(d_2,t))}{n(d_1)+n(d_2)}, \qquad (5)$$

то есть просто в меру пересечения множеств ключевых термов d_1 и d_2.

Смысловая близость термов могла бы быть достаточно естественно определена при наличии тезауруса на основе расстояния по графу, представляющему этот тезаурус, от одного терма до другого. Как уже отмечалось, необходимого русскоязычного тезауруса автору найти не удалось. О ручном его создании в рамках отпущенных на проект ресурсов речь идти не могла. Поэтому было решено разработать методику автоматического создания подобной тезаурусу древовидной структуры (далее называемой автотезаурусом). Как мы увидим в следующем разделе, основой для ее создания послужила матрица частот совместного появления термов в текстах, что явилось отражением первоначальной идеи данного проекта о том, что мера близости термов может основываться на этой матрице.

4. Построение автотезауруса

Поскольку ручное построение тезаурусов требует существенных ресурсов, неоднократно предлагались различные автоматические методы, позволяющие создавать, так сказать, заготовку тезауруса, ручное доведение которой до готового уровня потребовало бы гораздо меньших усилий. Теоретическим основанием практически всех этих методов является так называемая *распределительная гипотеза* [5], предполагающая, что семантически подобные термы встречаются, как правило, в похожем лингвистическом контексте. Точное понимание термина «контекст» в этой связи сильно разнятся у разных исследователей. Например, в [3] в это понятие включается не только набор лексем, находящихся вблизи от рассматриваемых термов, но также и типы синтаксических структур, в которые эти термы входят. В данном исследовании пришлось ограничиться чисто позиционными признаками для определения контекста (контекст = файл или контекст = абзац), так как не было возможности проводить синтаксический анализ больших текстовых массивов.

В работе [7] используется еще одно, более сильное, допущение, что термы, соответствующие более частным концептам (гипонимам) встречаются в подмножестве контекстов, в которых встречаются термы – их обобщения (гипернимы). В нашем исследовании мы несколько ослабили это допущение, постулировав, что частота встречаемости гипернимов в текстах больше, чем у любого их гипонима. Ручная проверка на нескольких десятках русскоязычных пар гипоним-гипернимов показала, что в подавляющем большинстве случаев это предположение соответствует действительности. По аналогии с предыдущим допущением назовем его *частотной гипотезой*.

Основой для построения автотезауруса в нашем случае послужила матрица совместной встречаемости лексем, которая строится по большому текстовому корпусу, представительному для всего языка или какой-нибудь прикладной области. Значения ячеек этой матрицы $p(l_1,l_2)$ отражают статистическую значимость взаимозависимости появления лексем l_1 и l_2 в одних и тех же текстах. Если появление этих лексем не зависимо друг от друга, то вероятность встретить обе лексемы в некотором одном тексте равна произведению вероятностей встретить в нем каждую лексему по отдельности: $df(l_1)df(l_2)/N^2$. Тем самым, если в рассматриваемом корпусе данные лексемы встречаются в $df(l_1,l_2)$ документов, в качестве меры значимости связи этих лексем можно взять вероятность того, что, сделав N испытаний с априорной вероятностью успеха $df(l_1)df(l_2)/N^2$, мы получим $df(l_1,l_2)$ или более успехов, т.е. кумулятивную вероятность биномиального распределения с параметрами $df(l_1)df(l_2)/N^2$ и N. Чем меньше $p(l_1,l_2)$, тем значимее связь l_1 и l_2, и следовательно, согласно распределительной гипотезе, тем ближе они по смыслу.

Значения ячеек матрицы $p(l_1,l_2)$ используются в качестве меры расстояния между лексемами, что позволяет запустить на множестве лексем процедуру бинарной иерархической агломеративной кластеризации. Каждый кластер представляет собой одну или несколько лексем. Расстояние между двумя кластерами есть минимальное расстояние между входящими в них лексемами (на парах лексем из разных кластеров). Перед началом процедуры кластеризации каждая лексема есть один кластер. На каждом шаге процедуры кластеризации два кластера с наименьшим расстоянием между ними сливаются в один. При этом каждый кластер «помнит», какие два кластера его образовали. Процедура продолжается пока есть кластеры, расстояние между которыми меньше 0,03, деленное на количество лексем в квадрате, что соответствует обычно выбираемому порогу вероятности при проверке ста-

тистических гипотез с поправкой на количество независимых гипотез. В результате этой процедуры образуется набор бинарных деревьев, листьями которых являются лексемы. Нетерминальные узлы деревьев могут рассматриваться как концепты, соответствующие синонимическим группам и понятиям-гипернимам в тезаурусах, построенных вручную. В соответствии с частотной гипотезой мы используем для пометки этих псевдо-гипернимов небольшое количество их самых частых (имеющих самую большую $f(t)$) лексем-гипонимов. Я использовал для этого не одну самую частую лексему, а несколько наиболее частых лексем, чтобы учесть возможную неточность частотной гипотезы. А именно, если взять для пометки три самых частых лексем, то это вряд ли скажется на понятности пометки, но существенно уменьшит вероятность упустить слово, наиболее точно соответствующее помечаемому псевдо-гиперниму.

В данном исследовании был использован автотезаурус, построенный на основе той же выборки из русского Интернета, что была использована для вычисления средних частот лексем. Он получился состоящим из 398 отдельных деревьев с общим числом нетерминальных узлов 2390. Детальное изучение его содержимого показало на удивление высокое соответствие его структуры представлениям о семантической связи слов, основанным на здравом смысле, блестяще подтверждая распределительную и частотную гипотезы. Для иллюстрации приведем три случайно выбранных поддерева, которые демонстрируют совершенно различные структурные свойства автотезауруса в разных его частях.

Первый фрагмент представляет собой линейную структуру, где на каждом уровне к растущему кластеру присоединяются по одной все более и более далекие от его центра лексемы, в следующем порядке: аренда, квартира, сдавать, недвижимость, комната, офис, кухня, новостройка, комна (?), комнатный. При этом пометка соответствующих псевдо-гипонимов происходит следующим образом: [квартира, аренда] -> [квартира, аренда, сдавать] -> [квартира, недвижимость, аренда] -> [квартира, недвижимость, аренда] -> [офис, квартира, недвижимость]. Добавление четырех последних лексем не меняет пометку вышележащих узлов. Этот комплекс лексем хорошо представляется словом «недвижимость», которое и вошло в пометку. Видно, что в отличие от отношения гипернимии в обычном понимании, у нас это отношение может связывать разные части речи.

Второй фрагмент представляют собой простую ассоциацию слов бурение и буровой и, вероятно, не был бы построен, если бы слова подвергались не только приведению к нормальной форме, но и стеммингу (распознаванию однокоренных слов). Но в распоряжении автора не было русскоязычного стеммера – его роль как раз и сыграл автотезаурус.

Наконец, третий фрагмент дает пример сбалансированного дерева, демонстрирующего также весьма разумную смысловую пометку псевдо-гипернимов (рис. 1). Иногда объединение лексем на низких уровнях иерархии, как, например, для лексем спаниель и мастиф, является в большой степени результатом случайности.

Часто в построенном автотезаурусе близкими оказываются лексемы, не связанные, строго говоря, какими-то просто формализуемыми семантическими отношениями, но обладающие выраженной интуитивной смысловой близостью – как, например, «щенок» и «порода».

В завершение этого раздела точно определим понятия терма и близости термов, используемые в нашем подходе. Терм соответствует у нас узлу автотезауруса – т. е. лексеме или «концепту», включающему несколько лексем. При этом частота концепта в каком-либо тексте определяется как сумма частот входящих в него лексем. Близость термов определяется как $P(t_1, t_2) = 1/(1 + \frac{1}{2}D(t_1, t_2))$, где $D(t_1, t_2)$ – расстояние между термами в графе автотезауруса, $D(t_1, t_2) = \infty$ для термов из разных деревьев. Таким образом, определив понятия терма, расстояния между термами и вычисляемого на его основе расстояния между текстами, мы можем формально определить нашу процедуру кластеризации, что мы и сделаем в следующем разделе.

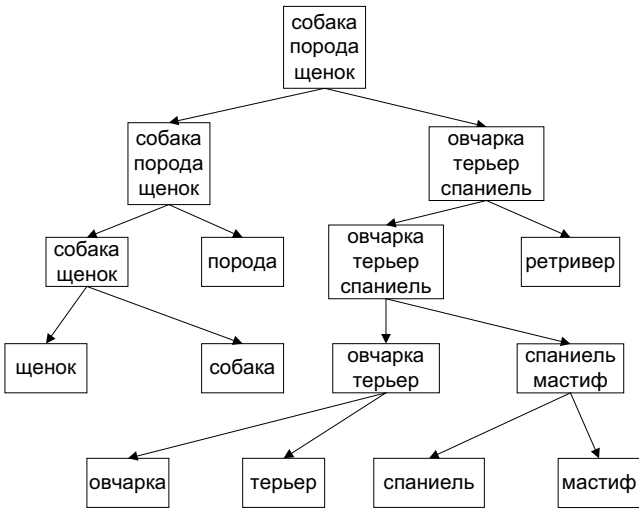

Рис. 1. Фрагмент автотезауруса, построенного на выборке из русского Интернета

5. Процедура кластеризации

Мы используем простую аггломеративную процедуру кластеризации. В начале каждый кластер представлен одним текстом. На каждом шаге этой процедуры пара кластеров, близость между которыми максимальна, сливаются в один. При определении близости кластеров все тексты, составляющие кластер, сливаются в один текст, так что мера близости кластеров определяется как близость их совокупных текстов. Процесс слияния кластеров останавливается по достижении заданного количества кластеров, либо при выполнении более сложного критерия, который мы рассмотрим в следующем разделе.

Для целей сравнения мы выбрали два точно таких же алгоритма, но использующие метрику евклидова пространства для определения близости между кластерами (где размерности этого пространства определялись в соответствии с методами MSEL и MLSA, соответственно, – см. раздел 2). Так как нас

в данном исследовании интересовало сравнение метрического и неметрического подходов к определению близости текстов, мы во всех сравниваемых случаях использовали один и тот же базовый алгоритм кластеризации. Критерии этого сравнения и использованные для него корпуса текстов будут обсуждаться в следующем разделе.

6. Критерии сравнения методов кластеризации и используемые для этого текстовые корпуса

Как уже говорилось во введении, процедура кластеризации текстов должна обладать двумя свойствами. Во-первых, она должна группировать тексты в соответствии с их смысловой близостью, и, во-вторых, должна предоставлять средства для связывания найденных кластеров с некоторыми понятными для человека пометками, передающими смысловое отличие документов в данном кластере от остальных документов. В нашем исследовании качество кластеризации оценивалось с точки зрения обоих этих критериев. Однако, если первый критерий допускает (хотя и с оговорками – см. ниже) достаточно хорошую формализацию и может быть оценен численно, то второй может быть, по-видимому, оценен только качественно на основе мнения экспертов.

Для оценивания соответствия полученных кластеров внутренней смысловой структуре текстового корпуса был применен следующий подход. Решено было использовать для этой цели лишь наборы текстов, снабженные верным a priori смысловым классификатором. Для оценки соответствия структуры полученных кластеров и смысловых классов, на которые разбит анализируемый корпус, применялась мера взаимной информации Шеннона. Эта часто используемая оценка дает прямой ответ на вопрос, сколько информации об одном параметре (в нашем случае – идентификаторе смыслового класса) содержится в другом параметре (идентификаторе кластера). Если количество текстов из смыслового класса i отнесенных к кластеру j обозначим как $m(i, j)$, то значение взаимной информации будет определяться как

$$IG = \sum_i \sum_j q(i,j) \log q(i,j) - \sum_i q_1(i) \log q_1(i) - \\ - \sum_j q_2(j) \log q_2(j), \quad (6)$$

где

$$q(i,j) = \frac{m(i,j)}{N}, q_1(i) = \sum_j q(i,j), q_2(j) = \sum_i q(i,j).$$

Видно, что, если отнесение текста к кластеру и смысловому классу независимы друг от друга, то $q(i, j) \approx q_1(i)q_2(j)$ и $IG \approx 0$. Если документы из каждого кластера принадлежат только одному смысловому классу, то IG равно максимально возможному значению, равному энтропии распределения текстов по кластерам:

$$IG_{max} = -\sum_i q_2(i) \log q_2(i) \quad (7)$$

Конечно, можно представить себе ситуации, когда такая оценка окажется весьма спорной. Например, мы хотим запустить такой тест на корпусе отзывов клиентов о качестве гостиничных услуг, которые заранее отнесены к смысловым классам «положительные», «нейтральные» и «отрицательные» оценки. Однако вполне может быть, что процедура кластеризации разобьет отзывы на замечания о качестве питания, постельного белья и предупредительности персонала, что также будет адекватным смысловым разбиением, но плохим с точки зрения значения IG. Тем не менее, можно надеяться, что такая оценка будет разумной в большинстве случаев при еще одном условии. А именно, энтропия распределения текстов по кластерам должна не слишком отличаться от энтропии распределения текстов по смысловым классам. Действительно, если кластеров очень много и они очень мелкие (предельный случай – каждый документ – один кластер), то IG в любом случае будет близка к IG_{max}. Если же кластеров гораздо меньше, чем смысловых классов, то невозможно будет достичь ситуации, когда документы из каждого кластера принадлежат только одному смысловому классу и IG будет всегда сильно меньше IG_{max}. С учетом этого обстоятельства в качестве критерия для остановки алгоритма кластеризации было выбрано не количество кластеров, а энтропия распределения текстов по кластерам. Алгоритм (во всех его изучаемых вариантах) останавливается, как только эта энтропия становится меньше энтропии распределения текстов по смысловым классам.

Таким образом, пригодные для целей тестирования текстовые корпуса должны содержать тексты, заранее распределенные по смысловым классам. Кроме того, целью исследования было проведение сравнения метода KTPB с метрической кластеризацией на разнообразных по своим свойствам текстовых корпусах, представляющих в то же время наиболее характерные области применения текстовой кластеризации. Эти два соображения наряду с возможностью использования для целей данного исследования данных, предоставляемых компанией Яндекс, определили следующий выбор текстовых корпусов:

1. Корпус Wl. Случайная выборка 965 веб-сраниц русского Интернета общим объемом 4.2 MB. На этой выборке заданы рубрики многоуровневого тематического рубрикатора Яндекса. В этом корпусе было оставлено 36 высокоуровневых рубрик, соответствующих 36 смысловым классам. Это выборка имитирует, например, ситуацию, когда надо кластеризовать результаты низкоспецифичного запроса к какому-либо текстовому хранилищу.

2. Корпус Wh. Та же выборка, но с более подробной разбивкой по смысловым классам (более низкие уровни рубрикатора) – 59 классов.

3. Корпус Ql. Набор страниц из выборки русского Интернета, предоставленной Яндексом, которые содержат фамилию Столыпин. 56 текстов общим объемом 1.7MB. Тематические классы опять же определяются рубрикатором Яндекс – 7 классов. Это соответствует задаче кластеризации результатов небольшого по объему специфичного запроса, содержащих несколько тематических линий.

4. Корпус Qh. Тот же корпус, 12 классов.

5. Корпус N. Выборка текстов новостей, из архива, предоставленного компанией Яндекс («обычная неделя»). 430kB в 295 текстах, распределенных по одноуровневому классификатору тем – 16 классов.

6. Корпус H. Заголовки новостей из того же самого архива. Общий объем – 135kB. Количество текстов – 2020. Количество смысловых классов – 431. Такие параметры задач кластеризации характерны для многих областей применения из области бизнеса и, в частности, торговли, например, для анализа кратких описаний причин недовольства клиентов тем или иным сервисом.

7. Корпус Cl. Статьи уголовного кодекса РФ, посвященные отдельным видам преступлений. 276 текстов, 270kB. В качестве смысловых классов служат шесть разделов этой части УК. Эта задача может служить прообразом проблемы автоматического смыслового структурирования узкотематического отраслевого текстового репозитория.

8. Корпус Ch. Статьи УК, но классифицированные по 19 его главам.

Таким образом, сравнение методов кластеризации проводилось в весьма разнообразных условиях. Кроме того, в каждом случае делалась попытка понять, какие свойства изучаемого метода в сочетании с конкретными параметрами кластеризуемого корпуса приводили к его относительному успеху или неуспеху. Эти результаты представлены ниже.

7. Результаты сравнения

В первую очередь было проведено сравнение между собой метрических методов MSEL и MLSA для разного количества оставляемых размерностей. Оказалось, что MSEL ни в одном из проделанных экспериментов не дал результат, лучший, чем MLSA, что, во-первых, согласуется с имеющимися в литературе указаниями на очень высокую эффективность последнего, а во-вторых, является косвенным свидетельством адекватности выбранного метода оценки качества кластеризации.

Зависимость результатов этих методов от размерности пространства кластеризации представлена на рис. 2.

Интересно, что результаты, получаемые с использованием метода MLSA, весьма несильно варьируются для размерностей пространства кластеризации, изменяющихся почти на два порядка (для MSEL эта зависимость была гораздо сильнее). Устойчивость по отношению к неправильному выбору настроечных параметров, безусловно, надо отнести к достоинствам метода. Отметим, что для метода KTPB эта проблема вообще не стоит, так как у него нет настроечных параметров. В качестве окончательного значения размерности, выбранного для цели сравнения методов MLSA и KTPB, мной было принято 300, что является оптимальным по всей совокупности проведенных экспериментов.

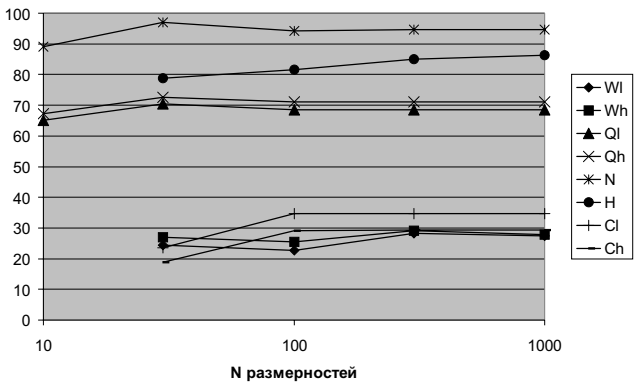

Рис. 2. Результаты кластеризации с помощью метода MLSA

Результаты этого сравнения представлены в табл. 1.

Таблица 1

Результаты кластеризации для методов MLSA и KTPB

Эксперимент	MLSA	KTPB
Wl	28.27	35.49
Wh	29.04	33.79
Ql	68.53	34.06
Qh	71.12	37.16
N	94.63	52.76
H	84.91	86.1
Cl	34.79	23.61
Ch	29.31	23.03

Как видно из этой таблицы, в трех экспериментах из восьми метод KTPB показал лучшие результаты, хотя в нескольких экспериментах его результаты были существенно хуже, чем у MLSA. Как и ожидалось, преимущества KTPB проявились там, где корпус состоял из очень разнообразных текстов, либо включал много мелких текстов. И в том и в другом случае наличие большого количества смысловых слов, встречающихся лишь в малом проценте текстов, препятствует правильному определению близости текстов без учета смысловой близости самих этих слов.

Теперь рассмотрим второй критерий сравнения методов кластеризации – адекватность смысловой пометки найденных кластеров. В метрических методах основной характеристикой найденных кластеров являются координаты их центроидов. Поскольку размерность пространства кластеризации велика, для более-менее лаконичной характеристики берется некоторое количество наиболее отличающихся от 0 координат. Так как каждой размерности соответствует некоторый терм, то соответствующий набор термов и используется для пометки данного кластера. Несколько более сложная ситуация имеет место в методе MLSA, так как в этом случае размерности пространства кластеризации являются линейной комбинацией размерностей, соответствующих отдельным термам. Поскольку вклад каждого терма в эту линейную комбинацию определяется соответствующим числовым коэффициентом, то разумно помечать каждую размерность пространства кластеризации MLSA некоторым количеством термов с наибольшими коэффициентами. В данном исследо-

вании я брал для пометки три размерности, каждая из которых характеризуется тремя наиболее значимыми для нее лексемами. Итого получается девять характеризующих кластер слов, что, как представляется, является разумным компромиссом между краткостью и полнотой описания кластера.

В методе KTPB ситуация более проста. Так как каждый кластер характеризуется множеством его ключевых термов, то надо просто из них выбрать подмножество, наиболее полно описывающее все входящие в него тексты, а именно, те термы, которые являются ключевыми для самого большого количества текстов данного кластера. Учитывая, что каждый терм описывается от одного до трех слов, то для того, чтобы общая длина описания кластера в сравниваемых методах была приблизительно одинакова, для пометки кластеров берется переменное количество термов, так чтобы общее число слов было максимально близко к девяти.

Так как кластеров во всех проведенных экспериментах было найдено несколько десятков, для сравнения в каждом случае используются три самых многочисленных кластера. Результаты этого сравнения представлены в табл. 2. В колонке KTPB слова в квадратных скобках соответствуют описанию нетерминальных узлов автотезауруса.

Обсудим результаты по каждому из корпусов.

W. Для этого корпуса результаты обоих методов не особенно хороши. MLSA выделил два больших кластера, относящихся к сфере бизнеса, но чем они различаются, непонятно. Третий кластер соответствует, вероятно, развлечениям. KTPB собрал все страницы, соответствующие бизнесу, в один большой кластер (на мой взгляд, более правильно). При этом можно понять, что там собраны тексты про работу (в разном понимании), торговлю (в частности, Интернет-магазины) и строительство/технологию. Так как тексты в этом корпусе очень разнообразны, то для пометки кластера выбраны узлы автотезауруса очень высокого уровня, в то время как проведенный ручной анализ показывает, что качество смысловой пометки узлов дерева автотезауруса ухудшается при приближении к корню (например [мочь; это; работа]), что, в общем-то, неудивительно. Хотя даже и в этом случае общий смысл остается понятен. Второй кластер явился результатом технической проблемы, связанной с неправильной интерпретацией значительного количества юникодовских страниц программой подготовки текстовых массивов. Было решено не выкидывать эти тексты из анализа, так как в реальной практике всякого рода испорченные и содержащие «мусор» тексты встречаются довольно часто. Метод KTPB поместил их все в один кластер, пометив его часто встречающимися в них тройками букв «рес» и «рер», что на мой взгляд, правильно. Наконец смысл третьего, более узкотематического кластера, совершенно ясен – это тексты про домашних животных.

Q. На этом корпусе KTPB дает существенно лучшие результаты. Чем отличаются между собой кластеры, найденные MLSA, из их описания совершенно непонятно. KTPB нашел один большой кластер, так или иначе связанный с известным русским реформатором начала XX века и его влиянием на политику и государственное устройство России. Второй кластер содержит особую группу текстов, представляющих в этой связи публицистику и другие материалы «национал-патриотов» в Интернете. Наконец, третий – это Интернет-публикации журнала «Полития» (отсюда – загадочное «политие» – результат нормализации этого несловарного слова программой морфологического анализа).

N. Опять результаты KTPB значительно превосходят MLSA, который только во втором кластере собрал тематически близкие тексты и понятно их пометил. Первый и третий кластеры смешивают несколько тем, что хорошо видно из их пометки. Обратим внимание, что для данного корпуса синтетические размерности пространства кластеризации MLSA очень хорошо соответствуют имеющимся тематическим линиям. Проблема с пометкой здесь вызвана самим метрическим подходом к кластеризации. KTPB только в самом многочисленном кластере смешивает несколько тематических линий, «оправдывая» это, как видно из пометки этого кластера, тем, что все эти темы касаются власти, президентов и пр. Два остальных кластера включают тематически однородные тексты и помечены абсолютно адекватно.

H. На этом корпусе MLSA дал малопонятные результаты. Например, чем различаются кластеры 2 и 3, сказать очень трудно. Самый большой кластер смешивает несколько тем, и не понятно, что их объединяет. Результаты KTPB гораздо более разумны. Первый кластер – про главную новость того периода, выборы в Алтайском крае. Второй – про разного рода экономические новости. Смысл третьего кластера менее понятен, но видно, что одна из его составляющих – известия про проходившую в то время неделю безопасности дорожного движения.

C. Результаты одинаково неудовлетворительны у обоих методов. Проблема здесь заключается в том, что они оба получили два больших тематически неоднородных кластера (относительно – так как сам корпус очень однороден) плюс много мелких узкоспецифичных кластеров. Различие между этими большими кластерами неясно ни в том, ни в другом методе.

Таким образом, для всех рассмотренных корпусов KTPB продемонстрировал более адекватную (или в одном случае – равно неудовлетворительную) смысловую пометку найденных кластеров, чем MLSA.

В заключение отметим сильное отличие результатов сравнения рассматриваемых методов с точки зрения количественного и качественного критериев. По-видимому, это означает, что в случаях, когда более важна интерпретация человеком результатов кластеризации, (или в случае тематически разнообразных корпусов) предпочтительно использовать предлагаемый в работе метод. В противном случае более оправдано использование метрических методов. Возможно также, что, как говорилось в разделе 6, критерий IG не совсем точно отражает качество кластеризации в каких-то из поставленных экспериментов.

Таблица 2

Смысловая пометка трех самых многочисленных кластеров в методах MLSA и KTPB

Корпус	MLSA	KTPB
W	[компания каталог услуга] [пользователь руб корзина] [квартира руб знакомство]	[мочь;это;работа] [контакт;лист;прайс] [двигатель; жидкость; арматура]
	[руб корзина мебель] [руб компания игра] [скачивать банк файл]	[рес; рер]
	[мара кинотеатр тело] [кинотеатр конференция отправлять] [клуб санкт оборудование]	сайт [собака; порода; щенок] радиодиагностик животное шнауцер поводок
Q	[политие это они] [библиотека борис еврей] [михаил власть писание]	[политический; партия] [российский; государственный; закон] [история; наука; научный]
	[политие семинар опубликовывать] [семинар еврейский еврей] [семинар рабочий практический]	книга борис еврей россия миф башилов масонство уило пирс
	[политие семинар опубликовывать] [давление медный филиал] [студент михаил давление]	политие [политический; партия] поршневый федерализм осень тоталитаризм общественный постсоветский
N	[грузия аджария тбилиси] [паксас литва сейм] [сутягин путин грузия]	[область; президент; район] покушение [грузия; грузинский; тбилиси] ингушетие [выбор; комиссия; кандидат]
	[ингушетие покушение мурат] [ингушетие покушение мурат] [ингушетие край избиратель]	масхадов [чечня;чеченский; ингушетие] турлай сдаваться шаа кадыр Аслан
	[масхадов чечня турлай] [ингушетие покушение мурат] [лебедев ходорковский менатеп]	призыв служба военный призывник призывный альтернативный комиссариат заявление
H	[ингушетие россия покушение] [новое тюмень уренгой] [россия сша покушение]	референтура калимулин евдоким край сельхозналог губернатор путин алтайский назначать
	[евро чечня масхадов] [евро газпром москва] [человек россия погибать]	евро рубль пенсия проц млрд инфляция тыс росфнефть бугурусланнефть
	[чечня масхадов сдаваться] [евро чечня масхадов] [челябинский масхадов новый]	азербайджаный поезд безопасность движение неделя дорожный дорога экономический уральский
C	[занимать определять размер] [применение неосторожность смерть] [занимать определять должность]	лишение срок наказывать осужденный лет размер [плата;заработный] деяние
	[занимать определять размер] [крупный применение миллион] [занимать определять применение]	[российский; государственный; закон] насильственный иной
	[занимать определять размер] [занимать определять применение] [наркотический вещество психотропный]	Самоубийство

8. Заключение

В данном исследовании предложен новый метод кластеризации текстов, основанный на их представлении в виде наборов ключевых термов и мере их близости, вычисляемой на основе смысловой близости их ключевых термов. Проводилось сравнение предлагаемого метода с метрическими методами кластеризации, представляющими тексты как точки единого евклидового пространства. Результаты сравнения подтверждают предположение о том, что предлагаемый метод должен иметь преимущество в случае сильного тематического разнообразия анализируемого корпуса либо малого размера отдельных текстов. Кроме того, он приводит к гораздо более понятной и точной смысловой пометке найденных кластеров. Важным побочным результатом данной работы, заслуживающим отдельного исследования, стала разработка метода автоматического создания напоминающей тезаурус структуры на базе матрицы совместной встречаемости лексем, построенной на большом текстовом корпусе. Эта структура не только дает основу для достаточно адекватной оценки смысловой близости термов, что необходимо для предлагаемого метода кластеризации, но и может быть удобна как отправная точка для ручного создания настоящего общеязыкового или тематического тезауруса.

9. Благодарности

Автор благодарит компанию Яндекс за поддержку данного исследования.

10. Литература

[1] *Berry M. W.* Survey of Text Mining, Springer – 2003.
[2] *Beyer K., Goldstein J., Ramakrishnan R., & Shaft U.* When is 'nearest neighbor' meaningful. // Proc. of ICDT-1999, Jerusalem, Israel, – P. 217-235.
[3] *Bloehdorn S., and Cimiano Ph., & Hotho A.* Learning Ontologies to Improve Text Clustering and Classification. // Proc of GFKL. 2005. Mode of access: http://www.kde.cs.uni-kassel.de/hotho/pub/2005/bloehdorn05learning_gfkl_final.pdf
[4] *Efron M., Marchionini G. & Zhiang J.* «Implications of the Recursive Representation problem for Automatic Concept Identification in On-line Governmental Information» // ASIST SIG-CR Workshop. – 2003 Mode of access: http://ils.unc.edu/govstat/papers/efronASISpaper.pdf
[5] *Harris Z.* Mathematical Structures of Language. – 1968.
[6] *Hotho A., Maedche A. & Staab S.* Ontology-based text clustering. // Proceedings of the IJCAI-2001 Workshop «Text Learning: Beyond Supervision», Seattle, USA.- 2001 – Mode of access: http://www.aifb.uni-karlsruhe.de/WBS/Publ/2001/hothoetal.pdf
[7] *Kashyap V., Ramakrishnan C., Thomas C. & Sheth A.* TaxaMiner: An Experimental Framework for Automated Taxonomy Bootstrapping. // International Journal of Web and Grid Services, Special Issue on Semantic Web and Mining Reasoning. – 2005 – Mode of access: http://lsdis.cs.uga.edu/~cartic/publications/TaxaMinerIJGWS.pdf
[8] *Landauer T. K., Foltz P. W. & Laham D.* Introduction to Latent Semantic Analysis. // Discourse Processes. 1998. Vol. 25. P. 259–284.
[9] Miller G. WordNet: A lexical database for English. // CACM. 1995. Vol. 38. No. 11. P. 39–41.

Text clustering procedure based on pair-wise proximity of key terms and its comparison with metric clustering methods

Mikhail Kiselev

This work is devoted to development of a new method for automated text clustering based on representation of text documents as sets of their *key terms* which differ in size and contents in contrast with majority of existing methods representing texts as points of a Euclidean space. In this method proximity measure of two texts is calculated on the basis of pair-wise proximities of their key terms. Proximity of two terms in its turn is determined from distance between them in a tree representing certain ontology. Lack of available all-language Russian ontology (at present) made it necessary to develop methods for (semi)automated ontology construction on the basis of lexeme co-occurrence matrix created for large text corpus. Comparative analysis of explored clustering procedures included numeric clustering quality estimation as well as manual evaluation of cluster semantic tagging.