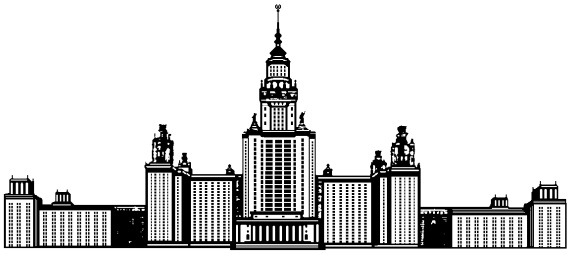


Московский государственный университет
имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Сендерович Никита Леонидович

Автоматизация кодирования открытых вопросов

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:
к.ф-м.н.
А.И. Майсурадзе

Москва, 2015

Оглавление

Глава 1

Введение

В современном мире для выяснения общественного мнения в самых разнообразных целях повсеместно используется анкетирование. Сбор статистической информации о взглядах различных групп населения по тем или иным политическим и социально-экономическим вопросам, относящимся к жизни страны, региона или города производится как государственными, так и независимыми исследовательскими компаниями с целью выявления существующих тенденций и прогнозирования. Руководство коммерческих предприятий может производить как опросы персонала с целью усовершенствования бизнес-процессов, так и опросы потребителей с целью выявления их запросов и увеличения продаж. Кроме того, собранные данные используются учёными соответствующих предметных областей в качестве фактического материала для исследований.

Вопросы анкеты могут предлагать респонденту различные способы ввода ответа, особенно широкое разнообразие можно встретить в интернет-опросах: выбор одного или нескольких из предложенных вариантов, ввод текста, ввод оценки, упорядочивание предложенных вариантов по степени некоторого свойства и т.д.

В целом, по степени свободы, предоставляемой респонденту, вопросы можно разбить на 3 группы:

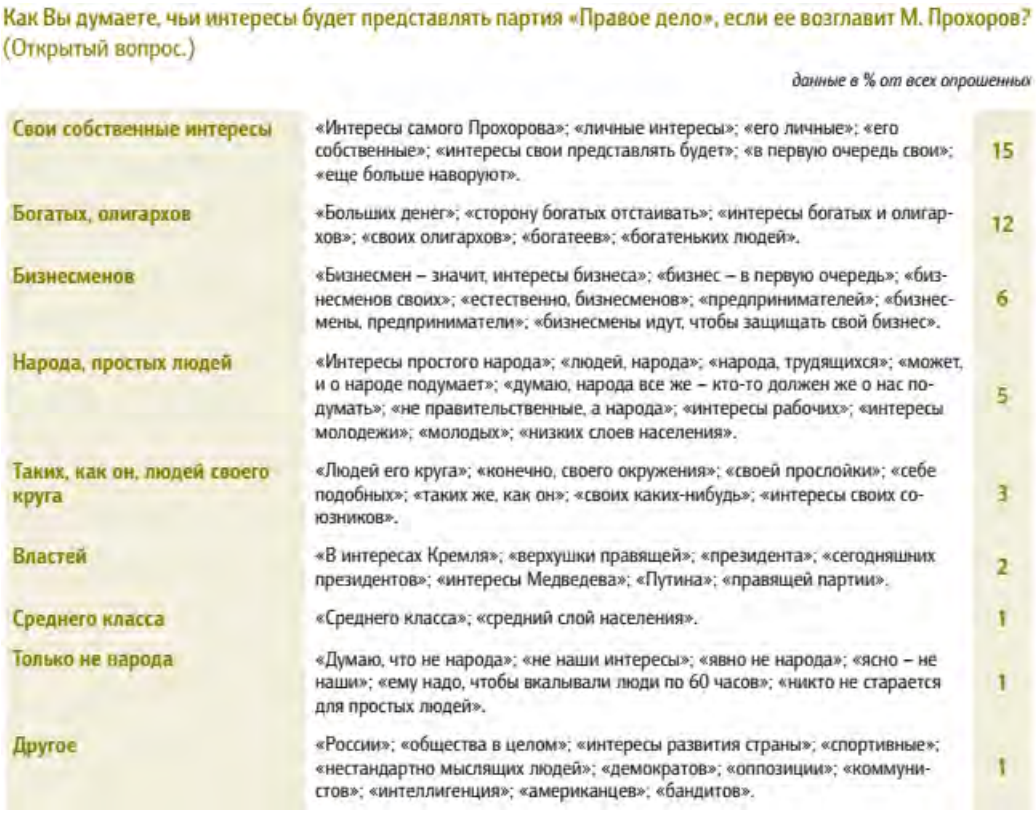
1. Закрытые вопросы — вопросы, для которых респондент выбирает один или несколько из предусмотренных составителем вариантов ответа. Этот тип вопросов включает вопросы с жёстко фиксированным форматом ответа: выбор из конечного числа данных вариантов, ввод числа и т.п.

2. Открытые вопросы — вопросы, на которые респонденту предлагается дать развёрнутый ответ своими словами.

3. Гибридные вопросы — вопросы, для которых респондент может выбрать один или несколько из предусмотренных вариантов ответа, может дать ответ своими словами, а может сделать и то, и другое.

Задача автоматизированного анализа ответов на закрытые вопросы представляется более простой, поскольку поступающие данные имеют известную структуру.

Более сложна задача анализа открытых вопросов, поскольку ответом на открытый вопрос, как правило, является текст на том или ином естественном языке



Как Вы думаете, чьи интересы будет представлять партия «Правое дело», если ее возглавит М. Прохоров?
(Открытый вопрос.)

данные в % от всех опрошенных

Свои собственные интересы	«Интересы самого Прохорова»; «личные интересы»; «его личные»; «его собственные»; «интересы свои представлять будет»; «в первую очередь свои»; «еще больше наворуют».	15
Богатых, олигархов	«Больших денег»; «сторону богатых отстаивать»; «интересы богатых и олигархов»; «своих олигархов»; «богатеев»; «богатеньких людей».	12
Бизнесменов	«Бизнесмен – значит, интересы бизнеса»; «бизнес – в первую очередь»; «бизнесменов своих»; «естественно, бизнесменов»; «предпринимателей»; «бизнесмены, предприниматели»; «бизнесмены идут, чтобы защищать свой бизнес».	6
Народа, простых людей	«Интересы простого народа»; «людей, народа»; «народа, трудящихся»; «может, и о народе подумает»; «думаю, народа все же – кто-то должен же о нас подумать»; «не правительственные, а народа»; «интересы рабочих»; «интересы молодежи»; «молодых»; «низких слоев населения».	5
Таких, как он, людей своего круга	«Людей его круга»; «конечно, своего окружения»; «своей прослойки»; «себе подобных»; «таких же, как он»; «своих каких-нибудь»; «интересы своих союзников».	3
Властей	«В интересах Кремля»; «верхушки правящей»; «президента»; «сегодняшних президентов»; «интересы Медведева»; «Путина»; «правящей партии».	2
Среднего класса	«Среднего класса»; «средний слой населения».	1
Только не народа	«Думаю, что не народа»; «не наши интересы»; «явно не народа»; «ясно – не наши»; «ему надо, чтобы вкалывали люди по 60 часов»; «никто не старается для простых людей».	1
Другое	«России»; «общества в целом»; «интересы развития страны»; «спортивные»; «нестандартно мыслящих людей»; «демократов»; «оппозиции»; «коммунистов»; «интеллигенция»; «американцев»; «бандитов».	1

Рис. 1.1: Пример результата анализа ответов на открытый вопрос

длиной от одного слова до нескольких абзацев. При этом ответы на открытые вопросы часто оказываются ценным активом, поскольку они полнее, чем закрытые, передают мнение респондента и могут содержать новые и важные для исследователей мысли ([16]).

При ручной обработке вопроса с развёрнутым ответом широко используется технология, позволяющая подготовить первичную социологическую информацию (тексты ответов) к последующей компьютерной обработке:

1. прочесть каждый из данных ответов и составить список основных встретившихся тем, идей и мнений;

2. отнести каждый ответ к одной или нескольким выделенным на предыдущем шаге категориям.

Этот процесс называется *кодированием открытых вопросов*, его сущность заключается в переводе качественных оценок, данных респондентами, в количественную форму, легче поддающуюся анализу. Пример результата работы аналитика изображен на рис. 1. Пример взят из социологического бюллетеня Фонда Общественное Мнение от 26 мая 2011 года. Видим, что все содержательные ответы (больше половины респондентов затруднились ответить) были разбиты кодировщиками на 8 групп, для каждой группы приведены наиболее характерные ответы.

У описанного метода кодирования две основных проблемы. Во-первых, крайне велика трудоёмкость описанной процедуры. Во-вторых, каждый из этапов процедуры содержит субъективную составляющую, поэтому унификация результатов

анализа требует от кодировщиков дополнительных усилий: координации работы, выработки детальных инструкций и т.п. Примеры используемых методов согласования действий аналитиков можно найти в работе [6]. Таким образом, издержки на ручной анализ открытых вопросов запретительно высоки, вследствие чего такие вопросы редко включаются в анкеты, а если и включаются, то при больших объёмах выборки собранные данные не получают адекватного анализа. В этих условиях остро встаёт вопрос об автоматизации описанного процесса.

Задача автоматического нахождения структуры тем, к которым относятся ответы на данный открытый вопрос, сложна, поскольку список этих тем заранее неизвестен. Эту задачу можно рассматривать как *задачу кластеризации* документов коллекции. Задача кластеризации предполагает выделение групп объектов и распределение имеющихся объектов по ним таким образом, что объекты, принадлежащие одной группе, схожи в большей степени, чем объекты, принадлежащие разным группам. Сами группы принято называть *кластерами*. В противоположность задачам классификации (в приложении к коллекциям текстов, эту задачу ещё называют задачей *категоризации*), не предопределен набор классов и не дано примеров, какой объект к какому классу относится. В случае коллекций текстов предполагается, что результатом кластеризации будет семантическое разбиение коллекции, поскольку выполнена *гипотеза компактности* ([17]), то есть гипотеза о лексическом сходстве текстов, относящихся к одинаковым темам.

В данной работе исследуются подходы к решению задачи кодирования открытых вопросов с помощью методов кластерного анализа. Она построена по следующему плану. В главе 2 рассматривается и анализируется ряд существующих моделей и алгоритмов, используемых для кластеризации коллекций текстов, в объёме, необходимом для дальнейшего исследования. В главе 3 даётся обзор основных свойств ультраметрик, рассматриваются существующие варианты применения теории ультраметрик для решения задачи кластеризации и доказывается теорема о взаимосвязи этих вариантов. В главе 4 описываются эксперименты по эмпирическому сравнению различных алгоритмов кластеризации коротких текстов. В главе 5 предлагается модель интерактивной кластеризации, проводится её анализ. Глава 6 содержит выводы, сделанные по итогам исследования.

Глава 2

Кластеризация коллекции коротких текстов

2.1 Варианты постановки задачи

Задача кластеризации коротких текстов встречается в многочисленных приложениях. Помимо рассматриваемой задачи кодирования открытых вопросов, к данному классу задач относится, к примеру, *задача анализа результатов, выдаваемых поисковой машиной по данному поисковому запросу*. Она состоит в следующем. Как правило, запрос пользователя, данный на вход поисковой машине содержит неопределённость, и результаты поиска содержат документы, относящиеся к разным темам. Для удобства пользователя необходимо выделить тематические категории и отнести найденные документы к соответствующим темам. При этом при принятии решения об отнесении результата к той или иной тематической категории может учитываться не всё содержимое документа, а только отрывок (англ. *snippet*), предоставляемый поисковой машиной. Отметим, что данная задача изначально предполагает автоматическое решение.

Рассмотрим следующую постановку задачи кластеризации текстов: на вход подаётся коллекция документов на естественном языке $D = \{d_1, d_2, \ldots, d_n\}$, необходимо найти такое наилучшее разбиение текстов на несколько непересекающихся групп, чтобы расстояния между текстами в одной группе были как можно меньше, а расстояния между текстами из разных групп — как можно больше.

Отметим, что поставленная в такой форме задача заведомо не имеет однозначного решения. Как правило, для того, чтобы формализовать постановку, вводятся понятия *модели текста* и *модели коллекции*, после чего на основании введённых объектов формулируется задача оптимизации, отражающая требования к искомой кластеризации. Решение поставленной задачи оптимизации и даёт наилучшее в каком-то смысле разбиение множества ответов на кластеры.

Разработано огромное число алгоритмов кластеризации текстов, основанных на различных математических моделях и идеях (подробный обзор даётся, например, в работе [2]). В данной главе производится необходимый для дальнейшего исследования обзор ряда подходов к автоматизированному анализу текстовых коллекций.

Дальнейшее изложение построено по следующему плану: в разделе 2.2 рассказано о существующих методах предварительной обработки текстов, затем в разде-

лах 2.3, 2.4 изложены классические подходы к кластеризации текстовых коллекций, а в разделе 2.5 рассказывается о проблемах, вызванных спецификой задачи обработки коротких текстов и существующих методах их решения.

2.2 Обработка текста на естественном языке

Будем предполагать, что на вход подаётся коллекция текстов на естественном языке. Прежде чем документ будет подан на вход тому или иному алгоритму кластеризации, он, как правило, подвергается следующим этапам предварительной обработки:

- Сегментация — разбиение текста на отдельные предложения.
- Токенизация — разбиение каждого из предложений на отдельные слова — термины.
- Нормализация — приведение каждого термина к выделенной *нормальной форме*.

Отметим, что если задачи сегментации и токенизации могут быть решены простыми техническими методами, то задача нормализации является нетривиальной задачей распознавания и обладает существенной сложностью. Решение этой задачи необходимо для того, чтобы различные формы одного и того же слова не рассматривались как различные термины. Её можно формулировать как задачу разбиения множества всех встречающихся в языке слов на классы эквивалентности, где к каждому классу относятся формы одного и того же слова. Для решения этой задачи разработаны различные методы, которые будут далее кратко описаны.

Один класс методов основан на преобразовании суффиксов слова в соответствии с системой правил, разработанных для конкретного естественного языка. Примерами могут служить стеммеры Портера [29], Ловинса [24] (англ. *rule-based stemmers*). Другой класс методов основан на вычислении статистики встречаемости последовательностей букв и слов в текстах (англ. *statistical stemmers*). При этом распространённые последовательности букв, встречающиеся в начале и в конце слова, признаются приставками и суффиксами соответственно и удаляются [26]. Также существуют стеммеры, основанные на построении скрытых марковских моделей [27]. Ещё ряд методов нормализации в рамках статистического подхода предполагает построение кластеризации слов на основе расстояний между терминами, построенными с учётом совместной встречаемости слов в дополнительном корпусе текстов [34], [25]. Более подробно методы стемминга освещены в обзоре [19].

Принципиально иной подход к нормализации предполагает использование словарей, в которых задано разбиение всех словоформ на группы родственных. В этом случае главная проблема состоит в разрешении неоднозначности при определении нормальной формы для данной словоформы — должна решаться задача *снятия морфологической омонимии*. Например, для слова «три», встречающегося в тексте, не ясно, является ли оно глаголом в повелительном наклонении (и тогда

его начальной формой является слово «тереть») или же это числительное в начальной форме. Для решения данной задачи может быть использована скрытая марковская модель ([22]).

Также в ходе предварительной обработки для повышения качества работы дальнейших алгоритмов может производиться удаление стоп-слов — слов, которые не должны учитываться при анализе. К таким словам могут относиться наиболее частые слова языка, встречающиеся во всех документах в большом количестве и не дающие дополнительной информации о документе. Также может производится удаление низкочастотных терминов — принято считать, что такие слова также не информативны. Тем не менее, нужно иметь в виду, что в приложении к задаче анализа коротких текстов этот этап может привести к дополнительной потере ценной информации об ответах, содержащих такие слова.

2.3 Векторная модель текста

2.3.1 Описание модели

Классической моделью текста является *векторная модель*, или *модель мешка слов*. Пусть D — множество документов в коллекции, W — словарь всех терминов, встретившихся в коллекции, n_{dw} — сколько раз слово w встречается в документе d, n_d — число слов в документе $d \in D$. Каждый документ d коллекции представляется в виде вектора-столбца длины $|W|$: $d = [f_1^d, \ldots, f_{|W|}^d]^T$. При этом в качестве значения признака может быть:

- бинарная величина: $f_w^d = [n_{dw} > 0]$ — признак говорит о том, встречается ли данный термин в коллекции или нет
- частота использования слова в документе: $f_w^d = \dfrac{n_{dw}}{n_d}$
- величина TF-IDF: $f_w^d = \dfrac{n_{dw}}{n_d} \log \dfrac{|D|}{|\{s : n_{sw} > 0\}|}$ — величина, учитывающая как частоту использования слова в данном документе, так и частоту использования в других документах коллекции

Результатом такого представления является *матрица термины-документы*.

В задачах информационного поиска наиболее часто используется величина TF-IDF, поскольку она позволяет достичь высоких результатов на практике и хорошо исследована теоретически ([1]).

Отметим, что описанная модель текста не является единственной. Предложены модели, в которых тексты представляются как множества частых словосочетаний (англ. *frequent itemsets*) или как последовательности терминов. Для таких моделей также разработаны соответствующие методы кластеризации ([4], [11]).

2.3.2 Функции расстояния

После того как документы представлены в описанном выше виде, для того, чтобы к ним можно было применить любой из стандартных алгоритмов класте-

ризации, необходимо выбрать функцию расстояния между текстами. Рассмотрим некоторые функции, часто используемые в этом случае.

Пусть документы рассматриваются как множества терминов (случай бинарных признаков):

$$D_1 = \{u_1, u_2, \ldots, u_n\}, \quad D_2 = \{v_1, v_2, \ldots, v_m\} \tag{2.1}$$

В этом случае используются следующие расстояния:

- Расстояние Джаккарда:

$$d_J(D_1, D_2) = 1 - \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} \tag{2.2}$$

- Расстояние Дайса:

$$d_D(D_1, D_2) = 1 - \frac{2|D_1 \cap D_2|}{|D_1| + |D_2|} \tag{2.3}$$

В более распространённом случае представления документов в виде векторов из $\mathbb{R}^n$ наиболее популярны косинусное расстояние и косинусная мера близости:

$$d_C = 1 - \frac{(v_1, v_2)}{||v_1||||v_2||}, \qquad s_C = \frac{(v_1, v_2)}{||v_1||||v_2||} \tag{2.4}$$

Далее перейдём к рассмотрению алгоритмов кластеризации, опирающихся на данное представление коллекции текстов.

2.4 Традиционные подходы к кластеризации текстов

По структуре множества кластеров алгоритмы кластеризации могут порождать *плоскую кластеризацию* или *иерархическую кластеризацию*. Иерархическая кластеризация подразумевает наличие дерева вложенных кластеров. Построение такого дерева называется также задачей *таксономии*. Плоская кластеризация, в отличие от иерархической, не подразумевает вложенности кластеров, все они располагаются на одном уровне. Задачи построения плоской и иерархической кластеризации тесно связаны между собой. С одной стороны, алгоритмы построения плоской кластеризации могут быть использованы при построении иерархической кластеризации, с другой стороны, результат каждого шага иерархической кластеризации можно рассматривать как очередную плоскую кластеризацию. В целом же, преимущество иерархических алгоритмов перед плоскими состоит в том, что иерархическая кластеризация позволяет получить больше информации о выборке документов и даёт пользователю возможность рассматривать разные уровни тематической организации коллекции ([9]).

2.4.1 Агломеративные и дивизивные алгоритмы иерархической кластеризации

Классическими подходами к построению иерархической кластеризации являются *агломеративные* и *дивизивные* алгоритмы кластеризации. При построении

иерархической кластеризации с помощью агломеративных алгоритмов объекты постепенно объединяются во всё более крупные кластеры. Таким образом из конфигурации, когда каждый объект является отдельным кластером, получается один кластер, содержащий все объекты. При использовании дивизивных алгоритмов же, наоборот, из более крупных кластеров получаются более мелкие. При этом из одного кластера, содержащего все объекты выборки, получаются кластеры из отдельных объектов.

2.4.2 Агломеративные алгоритмы

При построении кластеризации с помощью агломеративного метода на каждом шаге производится выбор двух наиболее похожих кластеров для слияния — U и V. Главный вопрос состоит в том, каким образом пересчитывать расстояние между получившимся кластером $U \cup V$ и остальными кластерами. Для решения этой задачи нужна адекватная мера расстояния между кластерами. В работе Ланса и Уильямса [33] был предложен общий подход для вычисления межкластерных расстояний в процессе агломеративной кластеризации:

$$d(U \cup V, S) = \alpha_U d(U, S) + \alpha_V d(V, S) + \beta d(U, V) + \gamma |d(U, S) - d(V, S)| \tag{2.5}$$

Приведём несколько часто используемых расстояний, являющихся частными случаями формулы Ланса-Вильямса:

- Расстояние ближнего соседа:

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{2.6}$$

- Расстояние дальнего соседа:

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \tag{2.7}$$

- Расстояние между центрами:

$$d(C_i, C_j) = d^2\left(\frac{1}{|C_i|} \sum_{x \in C_i} x, \frac{1}{|C_j|} \sum_{y \in C_j} y\right) \tag{2.8}$$

- Среднее расстояние:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y) \tag{2.9}$$

- Расстояние Уорда:

$$d(C_i, C_j) = \frac{|C_i||C_j|}{|C_i| + |C_j|} d^2\left(\frac{1}{|C_i|} \sum_{x \in C_i} x, \frac{1}{|C_j|} \sum_{y \in C_j} y\right) \tag{2.10}$$

Каждое из этих расстояний определяет стратегию объединения кластеров, и, соответственно, порождает агломеративный алгоритм кластеризации. Каждое из них обладает собственным набором свойств, и нет единого мнения о том, какое из них является наиболее универсальным. По-видимому, применимость того или иного расстояния в каждой задаче следует устанавливать эмпирически. В связи с этим упомянем работы [36], [32], в которых производится исследование ряда агломеративных и дивизивных алгоритмов кластеризации текстов, и устанавливается, что лучшей для агломеративных методов метрикой является среднее расстояние.

2.4.3 Дивизивные алгоритмы

При построении дивизивных алгоритмов неизбежно возникает две задачи: задача выбора кластера для разбиения и задача построения оптимального разбиения выбранного кластера. Распространённым подходом к решению первой задачи является выбор максимального кластера. Однако как правило, обе задачи решаются в комплексе: тем или иным способом строится семейство разбиений каждого имеющегося кластера, затем среди всех выбирается наилучшее. Отметим, что задача поиска оптимального разбиения, как правило, имеет экспоненциальную сложность, поэтому решается приближённо.

Критерием оценки разбиений выступает максимизация межкластерного расстояния (примеры используемых функций уже были приведены) и минимизация внутрикластерных расстояний. Для оценки компактности расположения объектов внутри кластера могут использоваться следующие функционалы:

- Диаметр кластера:
$$d(C_s) = \max_{x,y \in C_s} d(x, y) \tag{2.11}$$

- Средний квадрат расстояния до центра кластера:
$$d(C_s) = \frac{1}{|C_s|} \sum_{x \in C_s} d^2(x, \frac{1}{|C_s|} \sum_{x \in C_s} x) \tag{2.12}$$

- Средний квадрат расстояния между элементами кластера:
$$d(C_s) = \frac{2}{|C_s|(|C_s| - 1)} \sum_{x,y \in C_s} d^2(x, y) \tag{2.13}$$

В работе [36] утверждается, что минимизация среднего расстояния до центра кластера по косинусной мере (2.4) позволяет добиться наилучших результатов для текстовых коллекций.

Для построения разбиений может использоваться тот или иной алгоритм плоской кластеризации. Высокие результаты при кластеризации текстов показывает известный метод K средних для $K = 2$, т.н. алгоритм *bisecting K-means* [32].

Другим подходом к построению разбиения, предложенным в [36], является следующий эвристический оптимизационный процесс. Изначально выбираются два случайных документа выборки, служащие в качестве центров для порождаемых кластеров, и все документы относятся к кластеру ближайшего из этих двух документов. Далее на каждом шаге берётся случайный объект выборки, относится к

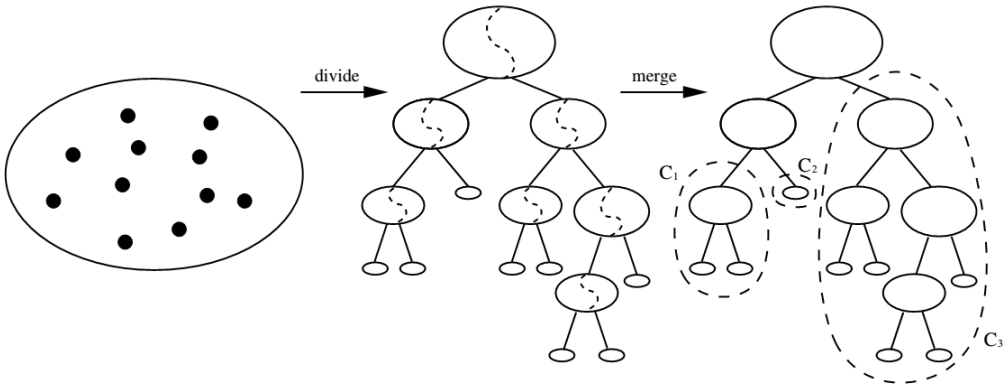


Рис. 2.1: Принцип работы гибридных алгоритмов

противоположному кластеру и происходит проверка, улучшится ли значение оптимизируемого функционала. В случае улучшения выбранный объект остаётся в новом кластере, в случае ухудшения процесс просто продолжается. Данный алгоритм имеет жадную природу и не гарантирует нахождение глобального оптимума функционала, однако процесс сходится к локальному минимуму.

Наконец, в работе [7] предлагается построение разбиений кластера путём перебора всевозможных пороговых значений для каждого из признаков, описывающих объекты выборки:

$$C \longrightarrow (C_l, C_r), \quad C_l = \{x : f_j(x) < c\}, \quad C_r = \{x : f_j(x) \geqslant c\} \tag{2.14}$$

2.4.4 Сравнение

Известно, что агломеративные алгоритмы менее вычислительно эффективны, что ограничивает их применение в приложениях. Однако традиционно в литературе считается, что агломеративные иерархические алгоритмы позволяют достичь более высокого качества кластеризации ([23]). Это мнение оспаривается в работах [36], [32]. В работе [32], например, доказывается, что алгоритм bisecting K-means превосходит агломеративные алгоритмы.

В ряде работ ([36], [8]) можно также встретить гибридные алгоритмы. Их идея состоит в том, чтобы сперва построить разбиение с помощью дивизивного алгоритма, а затем объединить промежуточные кластеры с помощью агломеративного алгоритма для получения более высококачественной кластеризации (рис. 2.4.4).

2.5 Проблема разреженности

Решение задачи кластеризации коротких текстов затруднено в связи с проблемой *разреженности исходных данных*. В длинных текстах, как правило, встречается большое количество слов, относящихся к основной теме документа, что позволяет, используя вероятностные методы и модели, выявить близкие по смыслу тексты. Для коллекций коротких текстов характерен недостаток статистиче-

ской информации о встречаемости слов, и недостаточное количество контекстной информации, общей для различных текстов. При этом особенно трудно найти адекватную меру сходства между текстами, что является главной трудностью при построении эффективного алгоритма кластеризации ([15]). С одной стороны, если рассматривать короткие тексты как документы, то окажутся неприменимыми стандартные методы вычисления сходства, основанные на общих для двух текстов словах (примеры таких функций расстояния приведены в разделе 2.3). С другой стороны, если рассматривать короткие тексты как слова и вычислять частоту совместной их встречаемости в корпусе, то она может оказаться близкой к нулю ([35]). Поэтому для решения проблемы разреженности применяются различные подходы, использующие вспомогательные данные.

В литературе описаны методы, использующие информацию двух видов:

1. семантические связи между терминами;
2. вспомогательную выборку релевантных «длинных» текстов для расширения контекста.

Рассмотрим далее соответствующие методы более подробно.

2.5.1 Семантическое сглаживание

При использовании векторной модели представление каждого документа содержит в себе лишь статистическую информацию о появлении терминов в коллекции. Для учёта информации о смысловых связях между терминами может быть использована техника *семантического сглаживания* ([31], [21]). Она состоит в следующем. Пусть P — неотрицательная симметричная матрица, где элемент $p_{ij} \in [0, 1]$ характеризует степень семантической близости терминов. Чем ближе термины по смыслу, тем больше соответствующее значение матрицы P; диагональные элементы равны единице. Тогда при вычислении близости документов друг к другу может быть использована формула:

$$s(d_i, d_j) = d_i^T P d_j. \tag{2.15}$$

Отметим, что мера близости 2.15 обобщает часто используемую косинусную меру близости, для которой $P = I$. Матрица P называется *матрицей семантической близости* (англ. *semantic proximity matrix*). В случае, если матрица P положительно определена, можно рассмотреть разложение Холецкого $P = L^T L$ (L — верхняя треугольная матрица), и трактовать меру близости 2.15 как скалярное произведение векторов Ld_i и Ld_j. При этом матрица L является матрицей линейного преобразования, переводящей вектора из исходного пространства в т.н. *семантическое пространство*. Таким образом, ядро сходства позволяет избежать изолированности терминов и тем самым произвести семантическое сглаживание при вычислении сходства между документами ([21]).

Вопрос о построении матрицы P решается по-разному. В работе [21] предлагается метод построения матрицы P по матрице документы-термины путём аналитического решения системы рекурсивных матричных уравнений относительно матрицы корреляций признаков и ядровой матрицы.

В работе [31] для построения P используется WordNet — семантическая база данных английских слов. Она представляет собой граф, в котором связаны между собой синонимы и пары гипоним-гипероним. Величина сходства двух терминов определяется обратной длиной пути между терминами в этом графе.

2.5.2 Расширение контекста

Рассмотрим теперь методы, опирающиеся на вспомогательные документы при решении задачи кластеризации коротких текстов.

В работе [30] для оценки семантического сходства между короткими текстами предлагается произвести процедуру расширения контекста. Для этого короткий текст подаётся на вход поисковой машине, после чего из первых n релевантных документов извлекаются наиболее частые термины, которыми частотами которых дополняется описание короткого текста. Данная идея получает развитие в работе [35], где для получения итоговой метрики применяются методы машинного обучения.

В работе [12] было показано, как вспомогательный массив текстов может быть использован для увеличения числа признаков для решения задачи категоризации текстов. В [3] было показано, что использование Wikipedia для извлечения дополнительных признаков и определения семантической близости между текстами позволяет улучшить качество кластеризации коротких текстов. В [15] для получения новых признаков по исходному короткому тексту используются одновременно Wikipedia и WordNet, что позволяет достичь более высокого результата по сравнению с использованием каждого из этих средств по отдельности и варианта без расширения признакового пространства.

Для решения задачи кластеризации коротких текстов также используются методы тематического моделирования. В работе [14] проводится исследование и сравнение различных схем обучения тематических моделей на сообщениях в Твиттере и демонстрируется, что более высокие результаты показывают модели, обученные на выборке из агрегированных текстов сообщений. В исследовании [18] предлагается тематическая модель DLDA (Dual LDA), позволяющая одновременно кластеризовать выборку вспомогательных документов и выборку коротких текстов, учитывая семантические взаимосвязи между ними.

Глава 3

Ультраметрики и задача кластеризации

В этой главе будет введено понятие ультраметрики и рассмотрены важнейшие свойства ультраметрик (раздел 3.1). Далее будет показано, что задача кластеризации может рассматриваться как задача построения ультраметрики в пространстве кластеризуемых объектов. В частности, при определённых ограничениях на коэффициенты формулы пересчёта (2.5) метод Ланса-Уильямса можно рассматривать как алгоритм превращения исходной функции расстояния между объектами в ультраметрику (раздел 3.2). В разделе 3.3 будут рассмотрены другие подходы к ультраметризации пространства объектов, основанные на решении оптимизационных задач в пространстве метрик. В 3.4 будет теоретически установлена связь между этими подходами к ультраметризации.

3.1 Основные свойства ультраметрик

Изложим необходимые известные факты из теории ультраметрических пространств.

Определение 1. *Рассмотрим непустое множество X. Функция двух аргументов $\rho : X \times X \longrightarrow \mathbb{R}$ называется* ***метрикой****, если выполнены следующие аксиомы:*

1. $\rho(x, y) = 0 \Leftrightarrow x = y \quad \forall x, y \in X,$
2. $\rho(x, y) = \rho(y, x) \quad \forall x, y \in X$
3. $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z) \quad \forall x, y, z \in X$

Если вместо первой аксиомы имеет место только условие $\rho(x, x) = 0 \quad \forall x \in X$, то ρ называется ***псевдометрикой****.*

Непосредственно из определения вытекает, что функция ρ неотрицательна:

$$0 = \rho(x, x) \leqslant \rho(x, y) + \rho(y, x) = 2\rho(x, y) \Longrightarrow \rho(x, y) \geqslant 0 \ \forall x, y \in X$$

Таким образом, функция ρ определяет расстояние между любыми двумя элементами множества X.

Определение 2. *Рассмотрим непустое множество X. Функция двух аргументов $\rho : X \times X \longrightarrow \mathbb{R}$ называется* ***ультраметрикой*** *(или неархимедовой метрикой), если она является метрикой и удовлетворяет усиленному неравенству треугольника:*

$$\rho(x,z) \leqslant \max\{\rho(x,y), \rho(y,z)\} \quad \forall x,y,z \in X$$

Если ρ является только псевдометрикой и удовлетворяет усиленному неравенству треугольника, то она называется ***псевдоультраметрикой****.*

Очевидно, что данное определение корректно, поскольку из усиленного неравенства треугольника вытекает неравенство треугольника в определении метрики, ибо для неотрицательных чисел имеет место неравенство $\max\{a,b\} \leqslant a+b$.

Нетрудно проверить, что примерами ультраметрик являются:

1. метрика пространства изолированных точек:

$$\rho(x,y) = [x \neq y]$$

2. метрика, порождённая p-адической нормой в пространстве рациональных чисел. Пусть p —простое число, тогда любое рациональное число r можно представить в виде $p^n\dfrac{a}{b}$, где a и b не делятся на p, и норма r определяется следующим образом: для $r \neq 0$ $|r|_p = p^{-n}$, $|0|_p = 0$.

Приведём некоторые важные свойства ультраметрик.

Утверждение 1. *Пусть ρ — ультраметрика, заданная на множестве X. Тогда для любых трёх элементов $x,y,z \in X$ среди трёх попарных расстояний $\rho(x,y)$, $\rho(x,z)$, $\rho(y,z)$ два расстояния равны и не меньше третьего.*

Доказательство. Докажем от противного, что максимальное из трёх попарных расстояний $\rho(x,y)$, $\rho(x,z)$, $\rho(y,z)$ не может строго превосходить оба других. В самом деле, если

$$\rho(x,z) > \rho(x,y),\ \rho(x,z) > \rho(y,z),$$

то это вступает в противоречие с усиленным неравенством треугольника:

$$\rho(x,z) \leqslant \max\{\rho(x,y), \rho(y,z)\}$$

Значит, из трёх расстояний максимальное значение принимают по крайней мере два. ■

Доказанный факт говорит о том, что любой треугольник с вершинами в элементах множества X является равнобедренным с бёдрами, превосходящими по длине основание. Это утверждение позволяет характеризовать ультраметрику с помощью *условия трёх точек*.

Определение 3. *Метрика ρ, заданная на множестве X удовлетворяет* ***условию трёх точек****, если любые три элемента $x,y,z \in X$ можно переименовать так, чтобы*

$$\rho(x,y) \leqslant \rho(x,z) = \rho(y,z)$$

Из определения вытекает очевидный факт.

Утверждение 2. *Метрика ρ является ультраметрикой $\Longleftrightarrow$ ρ удовлетворяет условию трёх точек.*

Ключевым свойством ультраметрики является следующий факт.

Утверждение 3. *Для любого $r > 0$ бинарное отношение на множестве X с ультраметрикой ρ, определяемое предикатом $[\rho(x, y) < r]$, является отношением эквивалентности.*

Доказательство. Рефлексивность и симметричность заданного отношения очевидны. Докажем транзитивность: пусть $\rho(x, y) < r$ и $\rho(y, z) < r$, тогда

$$\rho(x, z) \leqslant \max\{\rho(x, y), \rho(y, z)\} < r,$$

что завершает доказательство. ■

Из доказательства становится ясно, почему ультраметрики называют неархимедовыми метриками: за любое количество шагов любой величины, меньшей r по ультраметрике, невозможно выйти из шара радиуса r с центром в начальной точке.

Кроме того, этот факт говорит о том, что для любого значения порога с помощью ультраметрики пространство X разбивается классы эквивалентности, которые, как известно, либо не пересекаются, либо совпадают. Важная идея состоит в том, что эти классы можно рассматривать как кластеры, т.е. ультраметрическое пространство очень легко кластеризовать.

3.2 Ультраметрики и агломеративные алгоритмы кластеризации

В процессе работы агломеративного алгоритма кластеризации на каждом шаге t определяется, какие две ближайшие группы объектов нужно объединить друг с другом. Пусть на шаге t объединяются две группы, находящиеся на расстоянии M_t, а всего объектов в выборке n.

Определение 4. *Расстояние $d(C_i, C_j)$ называется* ***монотонным****, если*

$$R_1 \leqslant R_2 \leqslant \ldots \leqslant R_n$$

В случае, если расстояние, используемое в агломеративном алгоритме кластеризации, монотонно, то по результатам работы алгоритма может быть построена дендрограмма (рис. 3.1), на которой отображается процесс объединения кластеров. По оси абсцисс откладывается половина межкластерного расстояния, при котором происходит объединение, по оси ординат — все объекты выборки. Свойство монотонности гарантирует, что при любом порядке объединения объекты можно расположить на изображении так, что не произойдёт пересечения линий.

Определение 5. *Рассмотрим конечный связный взвешенный неориентированный граф $G = (V, E, W)$ с положительными (неотрицательными) весами рёбер. Определим для каждой пары вершин a, b расстояние $\rho(a, b)$ как длину кратчайшего пути между вершинами a и b в графе G. Тогда ρ называется* ***метрикой (псевдометрикой) кратчайших путей, порождённой графом*** *G.*

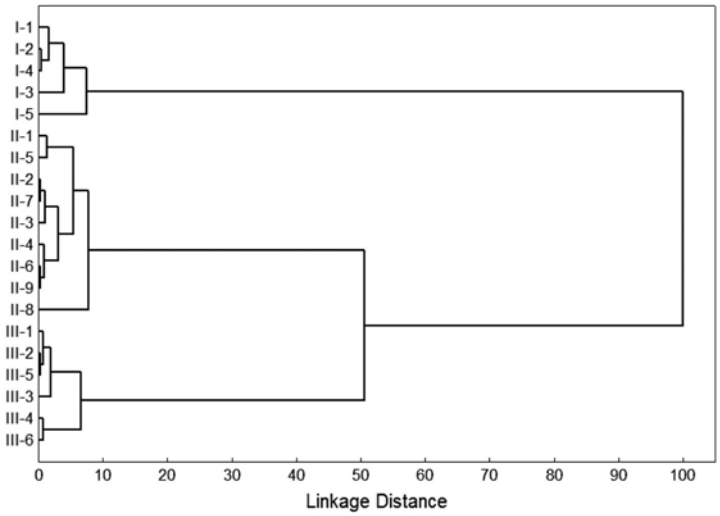


Рис. 3.1: Пример дендрограммы

Очевидно, что определённое выше расстояние действительно является метрикой (псевдометрикой).

Пусть исходные расстояния между разными объектами X были положительными. Построенную дендрограмму можно рассматривать как взвешенное дерево, где вершины — точки объединения групп объектов (вертикальные линии на рис. 3.1), рёбра — горизонтальные линии, веса на рёбрах — их длины в единицах горизонтальной оси. В дереве, как известно, существует единственный возможный путь между каждой парой вершин. Поэтому по построенному графу можно однозначно определить метрику, являющуюся сужением метрики кратчайших путей, порождённой деревом, на множество его листьев — объектов исходной выборки. По построению очевидно, что для листьев этого графа выполнено условие трёх точек. Из этого рассуждения вытекает следующая известная теорема ([20]).

Теорема 1. *Агломеративный алгоритм кластеризации объектов множества X с монотонной функцией расстояния порождает ультраметрику на этом множестве.*

Важным является вопрос о монотонности расстояний Ланса-Уильямса, заданных формулой (2.5). Ответ на неё даёт теорема Миллигана [28]:

Теорема 2. *Если выполняются следующие три условия, то кластеризация является монотонной:*

- $\alpha_U \geqslant 0, \alpha_V \geqslant 0$
- $\alpha_U + \alpha_V + \beta \geqslant 1$
- $\min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0$

Среди перечисленных в разделе 2.4.2 не удовлетворяет данному условию и не является мононтонной только функция расстояния между центрами 2.8.

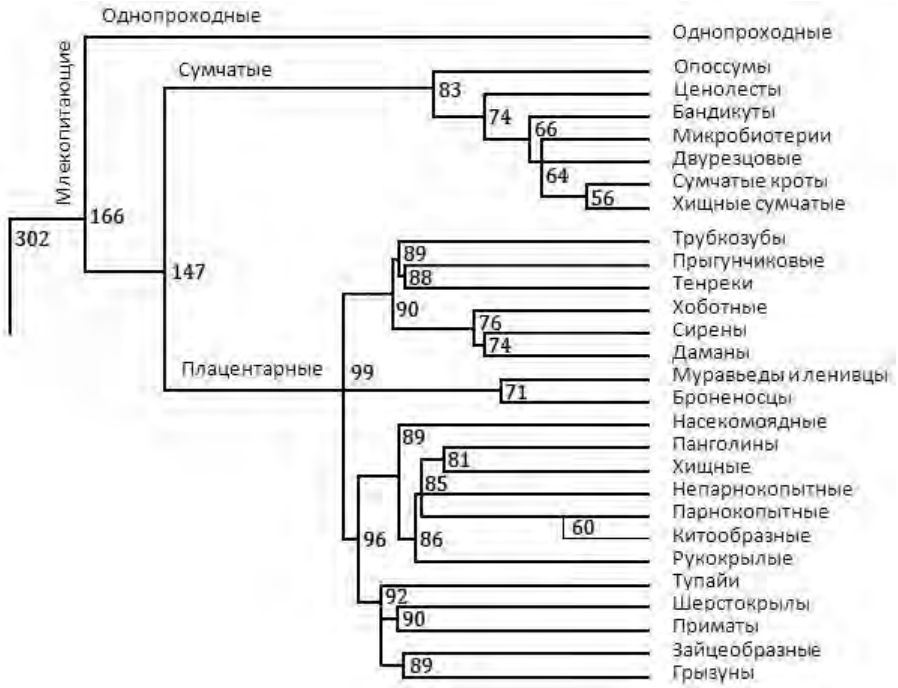


Рис. 3.2: Филогенетическое дерево современных отрядов млекопитающих. Цифры показывают ориентировочное время расхождения филогенетических групп (млн. лет)

Отметим, что каждому вертикальному сечению дендрограммы соответствует некое разбиение объектов на кластеры, т.е. плоская кластеризация. Нетрудно видеть, что по дендрограмме можно построить n различных плоских кластеризаций, где n равно числу объектов в выборке.

С точки зрения ультраметрики, каждая вертикальная прямая на дендрограмме определяет порог r для разбиения объектов на классы эквивалентности согласно утверждению 3. Таким образом, именно простота кластеризации пространств с ультраметрикой является теоретическим обоснованием для применения ряда агломеративных алгоритмов.

Наконец отметим, что использование дендрограмм и алгоритмов иерархической кластеризации является традиционным в биологических приложениях: для разнообразных процессов эволюции строятся филогенетические деревья, в которых расстояния имеют смысл времени, за счёт чего естественным образом удовлетворяют требованиям ультраметрики (см. рис.3.2).

3.3 Иные подходы к построению ультраметрик

В предыдущем разделе было показано, что классические агломеративные алгоритмы позволяют построить ультраметрику для набора объектов с известными расстояниями между ними, после чего элементарно может быть произведена их кластеризация.

Рассмотрим ряд других способов построения ультраметрик, связанных с исходной метрикой и при условии дополнительных ограничений, которые можно затем

использовать для решения задачи кластеризации. Приведённые результаты более подробно изложены в статьях [10] и [39].

3.3.1 Задача построения субдоминантной псевдоультраметрики

Определение 6. *На пространстве псевдометрик $\mathcal{F}$, заданных на непустом множестве X определим частичный порядок $\preccurlyeq$ следующим образом:*

$$d_1 \preccurlyeq d_2 \Longleftrightarrow \forall x, y \in X \quad d_1(x, y) \leqslant d_2(x, y)$$

Пусть задан конечный связный неориентированный граф $G = (V, E, W)$ с неотрицательными весами рёбер.

Определение 7. *Если в пространстве псевдометрик $\mathcal{F}$ на множестве V существует наибольший элемент ρ_W множества всех псевдоультраметрик, удовлетворяющих условию*

$$\rho(u, v) \leqslant W(\{u, v\}) \quad \forall \{u, v\} \in E,$$

то этот элемент называется ***субдоминантной псевдоультраметрикой для веса*** W***.***

Приведём без доказательства следующие известные факты ([10]):

Утверждение 4. *Пусть $\mathcal{P}_{x,y}$ — совокупность всевозможных путей в графе G, соединяющих вершины x и y. Тогда функция*

$$\rho_W(x, y) = \begin{cases} 0, & x = y \\ \min\limits_{P \in \mathcal{P}_{x,y}} \max\limits_{e \in P} w(e), & x \neq y \end{cases} \tag{3.1}$$

является субдоминантной псевдоультраметрикой для веса W.

Утверждение 5. *При построении субдоминантной псевдоультраметрики по формуле (3.1) достаточно оставить в графе G только рёбра, содержащиеся в минимальном остовном дереве графа G.*

Данные два утверждения позволяют эффективно строить субдоминантную псевдоультраметрику для заданной метрики — достаточно построить минимальное остовное дерево (это можно сделать, к примеру, с помощью алгоритма Прима за $O(n^2)$) и найти максимальное ребро на пути между каждой парой вершин.

3.3.2 Задача «Сэндвич»

На построении субдоминантной псевдоультраметрики основано решение следующей задачи. Пусть даны два конечных связных неориентированных графа с общим множеством вершин $G_h = (V, E_h, W_h)$ и $G_l = (V, E_l, W_l)$ и неотрицательными весами всех рёбер. Необходимо найти любую псевдоультраметрику ρ, удовлетворяющую требованиям

$$\rho(u, v) \leqslant W_h(\{u, v\}) \quad \forall \{u, v\} \in E_h \tag{3.2}$$

$$\rho(u,v) \geqslant W_l(\{u,v\}) \quad \forall \{u,v\} \in E_l \tag{3.3}$$

Решение состоит в том, чтобы построить субдоминантную псевдоультраметрику для W_h, после чего удостовериться, что она удовлетворяет нижним ограничениям. В случае, если хотя бы одно ограничение оказывается нарушено, такой псевдоультраметрики не существует ([10]).

3.3.3 Задача поиска ближайшей псевдоультраметрики

Следующая интересующая нас задача формулируется в виде задачи оптимизации: для конечного связного неориентированного графа $G = (V, E, W)$ с неотрицательными весами рёбер необходимо построить псевдоультраметрику ρ, такую, что

$$L_\infty(W, \rho) = \max_{\{u,v\} \in E} |W(\{u,v\}) - \rho(u,v)| \longrightarrow \min_{\rho} \tag{3.4}$$

Искомым минимумом данного функционала будет минимальное число ε такое, что для пары графов $G_l^\varepsilon = (V, E, W - \varepsilon)$, $G_h^\varepsilon = (V, E, W + \varepsilon)$ найдётся решение предыдущей задачи (добавление и вычитание ε к весовой функции графа следует понимать как соответствующее изменение каждого из весов). Нетрудно показать, что в качестве ε достаточно построить субдоминантную псевдоультраметрику ρ_W для M и положить

$$\varepsilon = \frac{1}{2} \max_{\{u,v\} \in E} \{W(\{u,v\}) - \rho_W(u,v)\}.$$

Такой выбор гарантирует, что решение задачи «Сэндвич» ρ^* для G_l^ε и G_h^ε существует, именно оно доставляет минимум функционалу ([10]).

3.4 Связь различных методов построения ультраметрик

Подытожим проведённый выше анализ. Алгоритмы агломеративной кластеризации Ланса-Уильямса с монотонной функцией расстояния по исходной матрице попарных расстояний между объектами порождают ультраметрику ρ в пространстве кластеризуемых объектов. В ультраметризованном пространстве можно, различным образом задавая порог r, получить различные кластеризации, определяемые отношением эквивалентности $[\rho(x,y) < r]$.

Расстояния между объектами в любом ультраметрическом пространстве удобно отображать на дендрограмме, которая отражает его иерархическую природу. При этом выбор порога r эквивалентен выбору вертикальной линии на дендрограмме.

Рассмотрим построение иерархической кластеризации путём ультраметризации пространства с помощью методов, описанных в предыдущем разделе. Для этой цели рассмотрим взвешенный неориентированный полный граф G, заданный следующим образом:

$$G = (X, E, W), \; W(x_i, x_j) = d(x_i, x_j) \tag{3.5}$$

Его вершины отождествлены с объектами выборки, а веса рёбер равны соответствующим расстояниям. К нему можно применить теорию ультраметризации взвешенных графов, изложенную в предыдущем разделе.

Самым простым и естественным вариантом ультраметризации является поиск ближайшей по метрике Чебышёва псевдоультраметрики ρ. При этом в итоговой псевдоультраметрике каждое расстояние изменяет своё значение по сравнению с исходным не более, чем на ε.

Оказывается, данный подход эквивалентен использованию расстояния ближнего соседа (2.6) в схеме Ланса-Уильямса. Сформулируем и докажем соответствующую теорему.

Теорема 3. *Субдоминантная псевдоультраметрика для графа* (3.5) *совпадает с псевдоультраметрикой, порождённой агломеративным алгоритмом кластеризации Ланса-Уильямса с расстоянием ближнего соседа.*

Доказательство. Убедимся, что алгоритм Ланса-Уильямса с пересчётом межкластерных расстояний по формуле (2.6) даёт расстояние между любыми двумя объектами x_i и x_j, равное весу максимального ребра на пути между этими вершинами в остовном дереве для графа (3.5), тогда с учётом утверждений 4 и 5 эквивалентность будет доказана. Для этого покажем, что работа агломеративного алгоритма эквивалентна работе алгоритма Крускала поиска минимального остовного дерева. В самом деле, на каждом шаге работы агломеративного алгоритма выбирается наименьшее расстояние между имеющимися кластерами. Это соответствует тому, что соответствующее ребро в графе G, соединяющее наиболее близкие объекты в данных двух кластерах добавляется в остовное дерево. При этом, как и в алгоритме Крускала, добавляется минимальное ребро, не соединяющее объекты одного кластера. Формула пересчёта (2.6) гарантирует, что на следующих шагах будут учитываться только кратчайшие расстояния между объектами нового кластера и объектами всех остальных кластеров.

Далее, в агломеративном алгоритме расстоянием между объектами x_i и x_j является длина ребра, объединившего кластеры, содержащие данные объекты. Заметим, что среди всех рёбер, содержащихся в остовном дереве графа G на пути между x_i и x_j, это ребро было добавлено последним, и, следовательно, имеет наибольший вес. Теорема доказана. ■

Доказанная теорема говорит об общей природе описанных в предыдущих разделах методов ультраметризации. В частности, L_∞-оптимальная псевдоультраметрика отличается от псевдоультраметрики, построенной алгоритмом из семейства Ланса-Уильямса, на константу $\dfrac{\varepsilon}{2}$.

В следующей главе изучим возможности применения данного алгоритма и других методов для решения задачи кластеризации коротких текстов.

Глава 4

Эмпирический анализ методов кластеризации

В этой главе будут описаны проведённые эксперименты по кластеризации коллекций коротких текстов. Использовались как модельные, так и реальные данные (раздел 4.3). Сопоставлены результаты работы ряда иерархических алгоритмов (раздел 4.4) и алгоритмов плоской кластеризации (раздел 4.5). Особое внимание в исследовании уделено методике измерения расстояний между объектами выборки (раздел 4.2) и методам оценки качества работы алгоритмов (раздел 4.1).

4.1 Методы оценки качества кластеризации

Задача оценки качества кластеризации непроста, поскольку решение задачи кластеризации существенно неоднозначно. Результат кластеризации, оптимальный с точки зрения запрограммированной модели, может оказаться совершенно неприемлемым с точки зрения экспертов предметной области.

Для автоматизированной оценки качества кластеризации используются два класса оценок: *внутренние* и *внешние* оценки. Внутренние оценки качества позволяют сравнивать результаты кластеризации в отсутствие дополнительной информации об истинных классах объектов, т.е. исходя только из имеющихся данных об объектах выборки и построенной кластеризации. Внешние оценки качества опираются на информацию об истинных классах объектов.

В приложениях наиболее адекватную оценку работы алгоритма может дать человек, а лучше эксперт предметной области. Внешние оценки реализуют эту идею, поскольку опираются на разметку истинных классов объектов, построенную вручную. Именно поэтому в исследованиях чаще используются внешние оценки.

Далее будут приведены и проанализированы наиболее часто используемые оценки качества как для плоских, так и для иерархических алгоритмов, порождающих кластеризацию с непересекающимися кластерами. Более полный обзор методов оценки содержится в работе [13].

При построении внешних оценок качества предполагается, что для объектов известны их истинные классы i, и найденные алгоритмом метки кластеров j. Ниже перечислены используемые внешние оценки качества кластеризации:

- F-мера

- Чистота
- Энтропия и перплексия
- Взаимная информация

Отметим, что F-мера может быть использована как для оценки плоской, так и для оценки иерархической кластеризации, тогда как остальные оценки рассчитаны на оценку плоской кластеризации. Рассмотрим подробнее приведённые выше оценки.

4.1.1 F-мера

Введём следующие общепринятые обозначения:

- n — общее число объектов
- n_i — число объектов в истинном классе i
- n_j — число объектов в кластере j
- n_{ij} — число объектов в кластере j, принадлежащих классу i
- $R(i,j) = \dfrac{n_{ij}}{n_i}$ — отклик кластера j на класс i, число объектов в кластере j, принадлежащих классу i, по отношению к числу объектов в классе i
- $P(i,j) = \dfrac{n_{ij}}{n_j}$ — точность кластера j для класса i, число объектов в кластере j, принадлежащих классу i, по отношению к числу объектов в кластере j
- $F(i,j) = \dfrac{2R(i,j)P(i,j)}{R(i,j)+P(i,j)}$ — мера релевантности кластера j классу i

F-мера использует введённую функцию $F(i,j)$ для нахождения наилучшего соответствия между набором кластеров и набором классов:

$$F = \sum_i \frac{n_i}{n} \max_j F(i,j) \tag{4.1}$$

Для оценки иерархической кластеризации максимум берётся по всевозможным кластерам в дереве:

$$F = \sum_i \frac{n_i}{n} \max_{j \in T} F(i,j) \tag{4.2}$$

Именно эта формула будет использоваться при анализе качества иерархических алгоритмов.

4.1.2 Чистота, энтропия, взаимная информация

В [2] указывается следующая проблема F-меры при оценке плоской кластеризации: она наиболее адекватна при правильном выборе числа кластеров, когда можно построить взаимно-однозначное соответствие между классами и кластерами.

Следующие метрики менее зависимы от соответствия числа классов числу кластеров. В соответствии с введёнными выше обозначениями, оценка чистоты кластеров вычисляется по формуле:

$$Purity = \sum_j \frac{n_j}{n} \max_i P(i,j) \tag{4.3}$$

Таким образом, чистота ставит в соответствие кластеру самый многочисленный в этом кластере класс. Чистота находится в интервале $[0,1]$, причём значение $Purity = 1$ отвечает оптимальной кластеризации.

Следующими распространёнными метриками качества кластеризации являются энтропия и перплексия. Вычислим энтропию j-ого кластера:

$$E_j = -\sum_i \frac{n_{ij}}{n_j} \log \frac{n_{ij}}{n_j} \tag{4.4}$$

Тогда общая энтропия кластеризации вычисляется следующим образом:

$$Entropy = \sum_j \frac{n_j}{n} E_j \tag{4.5}$$

Перплексия получается из энтропии, взятой по основанию два, следующим потенцированием:

$$Perplexity = 2^{Entropy} \tag{4.6}$$

Задача алгоритма кластеризации состоит в минимизации энтропии и перплексии.

Недостатком приведённых трёх индексов является то, что они достигают оптимальных значений при числе кластеров, равном числу объектов. Поэтому на практике (например, при подборе оптимального числа кластеров) исследуют скорость изменения значений этих оценок, а не их абсолютную величину.

Другим решением данной проблемы является использование взаимной информации для оценки качества. Формула для часто используемой нормализованной взаимной информации выглядит следующим образом:

$$NMI = \frac{\sum_i \sum_j n_{ij} \log \frac{n n_{ij}}{n_i n_j}}{\sqrt{(\sum_i n_i \log \frac{n_i}{n})(\sum_j n_j \log \frac{n_j}{n})}} \tag{4.7}$$

Значение нормализованной взаимной информации лежит в промежутке $[0,1]$, причём только значение $NMI = 1$ соответствует точному соответствию классов и кластеров. При проведении экспериментов с алгоритмами плоской кластеризации будем пользоваться именно ей.

4.2 Вычисление расстояний между текстами

Для представления данных во всех экспериментах использована векторная модель 2.3. Для решения проблемы разреженности при вычислении величины близости между текстами применялся метод семантического сглаживания (см. 2.5.1). Для построения матрицы семантической близости использовался тезаурус русского языка РуТез-lite [40]. Онтология рассматривалась как неориентированный граф смысловых связей между терминами, в котором вершинами являются слова, а рёбрами — отношения между ними. В этом графе вычислялись кратчайшие расстояния между словами, при этом всем типам отношений присваивались единичные веса.

Для получения из матрицы попарных расстояний между терминами матрицы P необходимо перейти от значений расстояний к значениям близости с помощью некоторой функции $s = s(d)$. В работе [38] анализировались различные варианты преобразования расстояния-близость, и был сделан вывод, что удачным выбором является функция $s(d) = \dfrac{1}{1+d}$. При проведении экспериментов использовалось аналогичное преобразование: каждый элемент матрицы попарных расстояний между терминами заменялся на обратный, при этом нули на диагонали заменялись на единицы, а близость между терминами, между которыми не существует пути в графе, полагалась равной нулю.

Далее для вычисления близости между объектами использовалась формула, аналогичная формуле (2.15), но с учётом нормировки. Для того, чтобы произвести нормировку, для матрицы P применялось разложение Холецкого $P = L^T L$, после чего близости вычислялись по формуле:

$$s(d_i, d_j) = \frac{(Ld_i, Ld_j)}{||Ld_i||||Ld_j||} = \frac{(d_i, Pd_j)}{||Ld_i||||Ld_j||} \tag{4.8}$$

Обратим внимание, что разложение Холецкого матрицы P, вообще говоря, может не существовать (случай, если она не является положительно определённой). Однако по построению матрица P симметрична, все её собственные значения являются действительными числами, поэтому добавление числа, превосходящего модуль наименьшего отрицательного собственного числа, ко всем элементам диагонали сделает её положительно определённой. Данное преобразование можно интерпретировать как увеличение поощрения за совпадения слов в различных ответах респондентов.

Наконец, для получения исходной метрики в пространстве кластеризуемых объектов производится стандартный для косинусной меры переход от близостей к расстояниям по формуле:

$$\rho(d_i, d_j) = \frac{\arccos s(d_i, d_j)}{\pi}. \tag{4.9}$$

4.3 Данные

Опишем использованный в исследовании процесс подготовки текстов на естественном языке к автоматизированному анализу, а также наборы модельных и реальных данных, используемые при эмпирическом анализе.

Для представления данных используется векторная модель с бинарными признаками (см. 2.3). Ответы респондентов очень коротки, в них редко встречаются повторяющиеся слова, поэтому можно считать, что каждый термин в ответе либо присутствует, либо отсутствует.

При предварительной обработке использовалась нормализация слов текстов с использованием открытого корпуса русского языка OpenCorpora. При решении задачи снятия морфологической омонимии был использован метод нормализации, опирающийся на скрытую марковскую модель (см. 2.2).

4.3.1 Генерация модельных данных

Опишем метод порождения исходных данных, учитывающий особенности, присущие реальным текстовым данным. Для проведения экспериментов необходимо породить матрицу термины-документы, а также матрицу семантической близости P.

При генерации матрицы термины-документы фиксируется число документов в коллекции n, а также число кластеров k. Наблюдения над реальными данными — ответами респондентов — показали, что кластеры, как правило, связаны с небольшим (менее 10) ключевых терминов. Помимо ключевых слов в ответах может также встречаться общая лексика, составляющая на практике большинство слов, встречающихся в ответах. При генерации задаётся общая численность слов общей лексики m, а также вероятность α принадлежности каждого слова общей лексики ответу.

Возможна также ситуация, когда ключевые слова одного кластера встречаются в ответах, относящихся к другому, тем самым затрудняя кластеризацию. Поэтому при генерации для каждого кластера задаётся количество принадлежащих ему ключевых слов, а также уровень шума β, имеющий смысл вероятности, с которой ключевые слова могут встречаться в других кластерах.

Каждый ответ порождается как случайное подмножество ключевых слов соответствующего кластера и слов общей лексики, причём хотя бы одно ключевое слово должно встретиться в ответе. Кроме того, каждое ключевое слово, относящееся к другим кластерам, попадает в ответ с вероятностью β. Наконец, учтём, что кластеры могут иметь несбалансированные размеры, что также может усложнить задачу кластеризации.

При порождении матрицы P предполагается, что близки по смыслу ключевые слова, относящиеся к одному кластеру. Кроме того, каждое ключевое слово, относящееся к другим кластерам или общей лексике, с вероятностью γ оказывается близко к данному слову.

Таким образом, модель учитывает:

1. разреженность данных
2. кластерную структуру данных
3. возможный шум в данных и несбалансированность кластеров

Для экспериментов было порождено 5 модельных наборов данных различной структуры. Их характеристики приведены в таблице 4.1.

Набор данных	n	k	m	α	β	γ	**Баланс**
M1	30	3	0	0	0	0	да
M2	50	5	50	0.05	0	0	да
M3	100	5	100	0.05	0.05	0	да
M4	100	5	100	0.05	0.05	0.1	да
M5	100	5	100	0.05	0.05	0.1	нет

Таблица 4.1: Параметры модельных наборов данных

Кластер 1	**Кластер 2**
«наша молодёжь будет жить лучше», «о школьниках, студентах», «Медведев болеет за молодёжь, даёт им работу», «уделял внимание молодёжи»	«про техосмотр», «упрощение системы прохождения осмотров автомобилей», «он и сказал, что техосмотр теперь будут оформлять не в ГАИ, а при ОСАГО»
Кластер 3	**Кластер 4**
«усовершенствование производства, инновации», «модернизация», «надо продолжать процессы модернизации в экономике и политике», «развитие науки»	«о коррупции в рядах чиновников», «о борьбе с коррупцией», «реформы надо продолжать и жестче бороться с коррупцией», «коррупция»

Таблица 4.2: Структура реального набора данных

Первый набор данных является идеальным с точки зрения кластеризации: отсутствует общая лексика, шум в данных и несбалансированность кластеров. Он предназначен для проверки работоспособности алгоритмов кластеризации. Последующие наборы данных получаются путём добавления факторов, затрудняющих кластеризацию и делающих набор данных более правдоподобным: во втором наборе появляется общая лексика, в третьем — пересечение кластеров по ключевым словам, в четвёртом — шум в матрице семантической близости P, в пятом — несбалансированность кластеров.

4.3.2 Описание реальных данных

В качестве реальных данных было рассмотрено множество из 30 ответов на другой вопрос, заданный ФОМ в 2010 году после пресс-конференции Дмитрия Медведева: «Что из того, о чем говорил Д. Медведев на пресс-конференции, Вам больше всего запомнилось и понравилось?» Эти ответы были проанализированы вручную и разбиты на 4 смысловые группы. Некоторые ответы из груп приведены в таблице 4.2.

Кластеры имеют разные размеры: в первом 10 ответов, во втором 6 ответов, в третьем 9 ответов, в четвёртом 5 ответов.

В первом, втором и четвёртом кластере по одному ключевому слову — «моло-

Алгоритм	**UPGMA**		**Single Linkage**		**DIANA**	
Сглаживание	нет	да	нет	да	нет	да
M1	1	1	1	1	0.86	0.99
M2	0.92	0.97	0.80	0.83	0.89	0.98
M3	0.83	0.90	0.61	0.76	0.81	0.88
M4	0.83	0.85	0.61	0.70	0.81	0.84
M5	0.74	0.75	0.53	0.55	0.76	0.77
Реальные данные	0.89	0.91	0.79	0.87	0.90	1.0

Таблица 4.3: Результаты иерархических алгоритмов кластеризации

дёжь», «техосмотр» и «коррупция» соответственно. Третий кластер формируется вокруг ключевых слов «инновация», «модернизация» и «наука».

4.4 Эмпирический анализ иерархических алгоритмов

Для анализа были выбраны 3 различных иерархических алгоритма:

- агломеративный алгоритм Ланса-Уильямса с функцией среднего расстояния (2.9) (англ. *UPGMA*, Unweighted Pair Group Method with Arithmetic Mean)
- алгоритм построения ближайшей по Чебышёвской норме ультраметрики, эквивалентный, согласно теореме 3 агломеративному алгоритму Ланса-Уильямса с расстоянием ближнего соседа (2.6) (англ. *single linkage*)
- дивизивный алгоритм кластеризации DIANA (англ. *DIvisive ANAlysis clustering*)

При этом каждый из алгоритмов запускался как с применением семантического сглаживания ($P \neq I$), так и без использования сглаживания $P = I$, т.е. с использованием стандартной метрики, порождённой косинусной мерой близости (2.4).

Значения F-меры при запусках на каждом из наборов данных приведены в таблице 4.3. Отметим, что в силу случайной природы модельных данных для получения адекватных оценок значения F-меры были усреднены по 10 вариантам генерации данных.

Ожидаемо, с увеличением сложности кластеризуемых модельных данных качество работы всех алгоритмов понижается — если практически все алгоритмы уверенно распознают три кластера в простейшем примере M1, то на зашумлённом наборе M5 менее 80% объектов были правильным образом распределены по кластерам в иерархическом дереве каждым из алгоритмов.

Естественно, для всех алгоритмов, не использующих семантическое сглаживание, результаты на наборах M3 и M4 совпадают —дополнительный шум, внесённый при порождении матрицы P в наборе M4, никак не влияет на ход кластеризации.

Из таблицы вытекает, что применение семантического сглаживания для всех алгоритмов на всех наборах данных позволило улучшить результат в среднем на 5%. Алгоритму DIANA использование семантического ядра P позволило верно распознать все кластеры на реальных данных.

4.5 Эмпирический анализ плоских алгоритмов

4.5.1 Сферический метод K средних

Для анализа был выбран сферический алгоритм K средних ([5]), успешно применяемый для кластерного анализа текстовых коллекций. Известно, что при решении задачи кластеризации в многомерном пространстве гораздо важнее направление вектора, чем его длина, поэтому все векторы, соответствующие ответам, следует нормировать, т.е. поместить на единичную сферу.

Приведём оптимизационные задачи и их решения для стандартного и сферического алгоритма K средних. Общая постановка задачи такова: требуется разбить выборку из n объектов $x_1, x_2, \ldots, x_n$ на K кластеров. Обозначим через $r_1, r_2, \ldots, r_n$ метки принадлежности объектов кластерам ($r_j \in \{1, 2, \ldots, K\}$). Положим также $C_i = \{x_j | r_j = i\},\ i = 1, 2, \ldots, K$ — множества объектов, принадлежащих кластерам. Будем исходить из того, что кластеры не пересекаются, и каждый объект принадлежит ровно одному кластеру:

$$C_i \cap C_j = \varnothing,\ i \neq j, \quad \bigcup_i C_i = \{x_1, x_2, \ldots, x_n\}. \tag{4.10}$$

Будем искать центры кластеров и минимизировать внутрикластерные расстояния между объектами, максимизируя близость объектов кластеров к центрам кластеров, определяя значения меток r_j. Для решения оптимизационной задачи применяется блочно-покоординатная оптимизация по группам дискретных переменных r_j и непрерывных переменных, соответствующих центрам кластеров.

Стандартный алгоритм K средних минимизирует функционал потерь для евклидовой нормы:

$$\sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - \mu_i||^2 \longrightarrow \min_{\boldsymbol{\mu}} \tag{4.11}$$

Итоговые формулы пересчёта имеют следующий вид:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \qquad r_j = \operatorname*{argmin}_i ||x_j - \mu_i|| \tag{4.12}$$

Сферический метод предполагает, что все объекты принадлежат единичной сфере: $x_i \in \mathbb{S}_+,\ \forall i = 1, 2, \ldots, n$ и максимизирует аналогичный функционал для косинусной меры с ограничениями:

$$\sum_{i=1}^{K} \sum_{x_j \in C_i} (x_j, c_i) \longrightarrow \max_{\boldsymbol{c}},\ ||c_i|| = 1 \tag{4.13}$$

Формулы для шагов оптимизации имеют следующий вид:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j, \qquad c_i = \frac{\mu_i}{||\mu_i||}, \qquad r_j = \operatorname*{argmax}_i (x_j, c_i) \tag{4.14}$$

где c_i — единичные векторы, соответствующие «центральным понятиям» коллекции ответов, близость к ним по косинусной мере определяет новую конфигурацию кластеров на очередном шаге алгоритма K средних.

Известна также аналогия стандартного и сферического метода K средних в терминах генеративных моделей коллекций документов. Пусть каждый документ описывается вектором-столбцом $d \in \mathbb{R}^m$, $d = [d_1, d_2, \ldots, d_m]^T$. Рассмотрим гауссову модель: вероятность принадлежности документа d кластеру с средним μ и матрицей ковариации Σ описывается выражением

$$P(d|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(d-\mu)^T\Sigma^{-1}(d-\mu)\right), \tag{4.15}$$

а также модель фон Мизеса-Фишера, в которой вероятность принадлежности документа d кластеру с средним μ и «матрицей ковариации» κ описывается выражением

$$P(d|\mu, \kappa) = \frac{1}{Z(\kappa)} \exp\left(\kappa \frac{d^T\mu}{||\mu||}\right), \tag{4.16}$$

где $Z(\kappa)$ — нормировочная константа распределения.

Известно, что максимизации правдоподобия для модели Гаусса в случае равных для всех кластеров матриц ковариации порождает метод K средних, а модель фон Мизеса-Фишера является аналогом модели Гаусса в случае, когда важно лишь направление векторов, и при равных «матрицах ковариации» κ порождает сферический алгоритм K средних ([37]).

4.5.2 Применение семантического сглаживания

Для применения семантического сглаживания в сферическом алгоритме K средних предлагается следующая модификация. Перейдём в оптимизационной задаче (4.13) к применению ядра сглаживания, заменив x_j на y_j, где y_j определяется формулой:

$$y_j = \frac{Px_j}{||Px_j||}. \tag{4.17}$$

Поскольку формула (4.17) затрагивает только входные данные и гарантирует, что $||y_j|| = 1$, то данная модификация позволяет без изменений применить сферический алгоритм K средних. Такую замену можно интерпретировать как расширение текстов исходных ответов за счёт добавления близких по смыслу терминов к терминам, присутствующим в ответе. Действительно, по построению матрицы P в векторе y_j по сравнению с вектором x_j добавится ряд ненулевых компонент, соответствующих терминам, близким к терминам исходного ответа.

4.5.3 Результаты экспериментов

Будем рассматривать результаты для сферического метода K средних с использованием семантического сглаживания и без него. В таблице 4.4 приведены значения нормализованной взаимной информации, полученные путём запуска сферического алгоритма K средних на модельных и реальных данных. При запусках на модельных данных, как и в предыдущих экспериментах, производится

Сглаживание	нет	да
M1	0.98	1
M2	0.88	1
M3	0.87	0.92
M4	0.87	0.90
M5	0.69	0.70
Реальные данные	0.90	0.97

Таблица 4.4: Результаты сферического алгоритма K средних

усреднение по 10 вариантам генерации исходных данных. Кроме того, учитывается, что результат кластеризации может существенно зависеть от начального приближения, поэтому в реализации алгоритма K средних предусматривается запуск из нескольких начальных приближений с выбором наилучшего по значению оптимизируемого функционала. Отметим, что во всех случаях алгоритму кластеризации передавалось истинное число кластеров в выборке.

Результаты говорят о том, что семантическое сглаживание вновь позволило улучшить результат кластеризации для всех наборов данных. Отметим, что особенно высокий прирост качества наблюдается на реальных данных — кластеры удаётся выделить практически точно.

4.6 Выводы

Эксперименты показали, что применённая методика сглаживания является эффективным способом борьбы с проблемой разреженности. Отметим гибкость этого подхода: матрицу семантической близости удалось успешно использовать для модификации как иерархических алгоритмов кластеризации, так и сферического алгоритма K средних.

Обратим внимание, что агломеративный алгоритм ближайшего соседа показывает наихудшие результаты на всех наборах данных, кроме простейшего набора M1. Методы DIANA и UPGMA демонстрируют близкие результаты, что согласуется с утверждениями работ [32] и [36].

Как и ожидалось, сферический алгоритм K средних с использованием семантического сглаживания демонстрирует высокое качество работы на реальных данных. Именно этот метод будет взят за основу при построении системы для интерактивной кластеризации открытых вопросов, подробно описываемой в следующей главе.

Глава 5

Интерактивная кластеризация

Эта глава посвящена разработке системы, предназначенной для решения задачи автоматизированного кодирования открытых вопросов. Система основывается на построении необходимого разбиения в ходе взаимодействия с экспертом.

Изложение построено следующим образом. В разделе 5.1 обсуждается сущность интерактивного подхода. В разделе 5.2 рассматриваются примеры типовых ситуаций, возникающих при анализе и предлагается набор инструментов аналитика — предметно-ориентированных высказываний, которые он может делать при взаимодействии с системой. В разделе 5.3 производится формализация высказываний аналитика в терминах метода кластеризации. Наконец, в разделе 5.4 приводится небольшой пример работы аналитика с разработанной системой.

5.1 Задачи интерактивной системы кластеризации

Как уже упоминалось выше, результаты кластеризации существенно неоднозначны: по одному и тому же набору данных разные алгоритмы кластеризации могут строить разные разбиения, оптимальные с точки зрения различных критериев. При этом большинство алгоритмов кластеризации не позволяет влиять на ход своей работы, получая на вход выборку и выдавая окончательный результат. В случае, если результат не удовлетворяет требованиям, эксперт может изменить параметры используемого алгоритма или применить другой метод.

Другой подход состоит в том, чтобы построить итеративный процесс взаимодействия эксперта с системой, при котором система будет помогать пользователю

Рис. 5.1: Интерактивный процесс построения кластеризации

достичь желаемого результата. Принцип работы изображён на рисунке 5.1. После запуска системы пользователь видит первичный, полностью автоматически построенный результат кластеризации. Рассматривая эти результаты, он может сделать высказывания, вносящие поправки в результат работы программы, и получить новую кластеризацию, учитывающую экспертное мнение. Этот процесс продолжается до получения оптимального с точки зрения аналитика разбиения на кластеры.

С точки зрения прикладной задачи анализа открытых вопросов, интерактивная система должна предлагать эксперту инструменты для эффективного анализа и построения качественного разбиения ответов респондентов на группы. Для достижения этой цели требуется разработать систему высказываний эксперта, которые будут учитываться системой интеллектуальной обработки данных в интерактивном режиме. Такая организация работы позволит, с одной стороны, ускорить процесс кодирования вопросов за счёт применения вычислительной техники, и, с другой стороны, получить высокое качество результата за счёт использования знаний эксперта.

С точки зрения задачи кластеризации коротких текстов, использование экспертного мнения можно рассматривать как ещё один метод борьбы с проблемой разреженности, при котором недостающую для построения качественной кластеризации информацию система получает от эксперта.

5.2 Разработка системы высказываний эксперта

В следующих разделах предложим ряд инструментов — высказываний эксперта, которые могут быть полезны аналитику при его работе в системе. По ходу рассмотрения будем приводить примеры использования, опираясь на реальные данные — некоторые ответы респондентов на вопрос «Какие события прошедшей недели, о которых сообщалось в средствах массовой информации, больше всего заинтересовали Вас, привлекли Ваше внимание?», который был задан в одном из опросов, проведённых социологической службой ФОМ в 2010 году.

5.2.1 Фиксация и освобождение ответов

Базовым высказыванием будем считать желание пользователя зафиксировать для некоторого подмножества ответов респондентов кластеры, к которым они отнесены в настоящий момент. Фиксация означает, что кластер для данного ответа больше не будет изменяться при применении дальнейших шагов кластеризации. Ещё не зафиксированные ответы, то есть не утратившие способность изменять свой кластер, будем называть *свободными*.

Пусть, например, кластеры были сформированы системой, как показано в таблице 5.1. Видно, что первый кластер обладает смысловой целостностью. Второй кластер условно можно назвать «Катастрофы, пожары, наводнения», и он содержит несколько различных мыслей. Ответы в третьем, по-видимому, были объединены системой из-за общего ключевого слова «убийство». Ответы в первом кластере имеет смысл зафиксировать.

Фиксированные объекты с точки зрения аналитика лучше всего отражают смысл кластера, к которому относятся. Если ответ точно передаёт общую для

Кластер 1	Кластер 2	Кластер 3
«Медведев и журналисты», «пресс-конференция Медведева», «выступление Медведева в Сколково», «Сколково», «Медведев в Сколково»	«катастрофы пожары», «одни катастрофы», «пожары в Свердловской области», «наводнение и пожар в США», «пожар в Канаде», «пожары бушуют по России», «пожар в Гусевке, лес горит», «пожары, наводнения», «наводнение в Якутии»	«убийство террориста бен Ладена», «убийство бен Ладена», «убийство нашего строительного предпринимателя», «нацистскую группировку в Питере арестовали, 20 разбоев и убийство за ними».

Таблица 5.1: Фиксация ответов: пример

многих респондентов мысль, аналитик может захотеть объявить этот ответ представителем данного кластера, зафиксировав его. Автоматизированная процедура кластеризации, учитывая данное высказывание пользователя, более точно сформирует кластерную структуру на следующем шаге.

Обратной операцией является «освобождение» — отмена фиксации для одного или нескольких фиксированных ответов респондентов. После применения этой операции к фиксированному элементу он вновь становится свободным. Эта операция необходима в случае, когда аналитик изменил свою точку зрения на какую-то группу ответов или зафиксировал что-то по ошибке.

Отметим, что в идеале работа аналитика в системе оканчивается тогда, когда каждый ответ фиксирован, т.е приписан к своему кластеру.

5.2.2 Перемещение ответов между кластерами

Пусть алгоритм кластеризации относит ответ к некоторому кластеру, а аналитик считает, что этот ответ этому кластеру не принадлежит. Тогда возможны два случая:

1. кластер, к которому пользователь хотел бы отнести этот ответ, существует. Например, помимо трёх кластеров из приведённого выше примера, система выделила отдельный кластер со следующими ответами:

Кластер 4
«природные катаклизмы»
«землетрясения, катаклизмы»
«снова землетрясение в Японии»
«последствия землетрясения в Японии»
«авария в Японии»

   Хотя ответы данного кластера не содержат ключевых слов «катастрофы», «пожары», «наводнения», их имеет смысл отнести к уже существующему кластеру 2.

2. пользователь хочет сделать этот объект первым в новом кластере. Так, в кластере 3 из предыдущего примера ответ «Убийство бен Ладена» и похожие на него разумно не смешивать с сообщениями о других убийствах, а выделить в отдельный кластер.

Важным частным случаем применения данной процедуры, часто встречающимся на практике, является отнесение ответа-выброса к уже существующему кластеру «Другое». Также пользователь может перенести выделенные ответы в новый кластер. Это может потребоваться пользователю, например, для создания самого кластера «Другое» в начале работы.

Во всех случаях предполагается, что перенесённые объекты автоматически фиксируются в новом кластере.

5.2.3 Удаление ответов

На практике исходные данные опросов часто содержат малоинформативные ответы, которые мешают проведению анализа. Например, в опросе, используемом для примера, среди прочих встречается ответ «сейчас помню, через час не помню». Другим примером неинформативных ответов являются стандартные метки, используемые при проведении опросов, такие как «затрудняюсь ответить», «нет» и «ничего» или пустые ответы. Заметим, что метод кластеризации скорее всего выделит такие ответы в отдельный кластер. Наконец возможен случай, когда аналитика не интересуют повторения одинаковых ответов и ему требуется удалить повторы для удобства работы.

Во всех этих случаях требуется исключить некоторые ответы из рассмотрения, поэтому система высказываний должна содержать команду удаления.

5.2.4 Работа с именами кластеров

Задачей аналитика является осмысленное выделение групп ответов, данных респондентами. Поэтому необходимо предусмотреть возможность задания имён кластеров.

Отметим, что эта функция имеет смысл даже для методов, способных генерировать имена кластеров по их содержимому. Например, рассмотрим следующий кластер, взятый из того же опроса:

Кластер 5
«проиграли в хоккей»
«в хоккей проиграли финнам»
«Хоккей Россия-Чехия 4:7»
«хоккей: финны стали чемпионами, а наши на 4 месте»

Его естественно назвать «Поражение сборной России на Чемпионате мира по хоккею в Словакии», хотя сами ответы многих из этих слов не содержат, поэтому это имя не может быть порождено алгоритмом кластеризации без вмешательства пользователя.

Отметим, что присвоение кластерам имён является скорее косметической процедурой, необходимой для комфортной работы пользователя в системе, нежели высказыванием, которое существенно влияет на работу алгоритма кластеризации.

5.3 Описание интерактивного алгоритма

5.3.1 Формализация системы высказываний

Формализуем систему высказываний, которую должен предлагать интерактивный интерфейс.

Будем считать, что аналитик рассматривает результат кластеризации, полученный на очередном шаге: объекты $x_1, \ldots, x_p$ принадлежат k непересекающимся кластерам $C_1, \ldots, C_k$, которым присвоены имена $s_1, \ldots, s_k$. Позволим эксперту делать следующие элементарные высказывания:

1. присоединить объект x_j к кластеру C_i, $i \in \{1, \ldots, k+1\}$;
2. отменить присоединение объекта x_j к кластеру, к которому он присоединён;
3. удалить объект x_j;
4. удалить кластер C_i, $i \in \{1, \ldots, k\}$;
5. завершить формирование кластера C_i;
6. возобновить формирование кластера C_i;
7. присвоить кластеру C_i имя s_i, $i \in \{1, \ldots, k\}$.

Высказывание 1 реализует только операцию фиксации в случае, когда объект x_j уже отнесён автоматической процедурой к кластеру C_i. В случае же, когда объект x_j лежит в другом кластере, это высказывание реализует сперва операцию перемещения, а затем операцию фиксации. Значение $i = k + 1$ означает отнесение объекта к новому кластеру и увеличение счётчика числа кластеров k, при этом новому кластеру присваивается произвольное имя (в реализованной системе в этом случае по умолчанию используется имя «Cluster $k + 1$»).

Высказывание 2 реализует операцию освобождения.

Высказывание 3 позволяет удалять объекты. Отметим, что в ходе работы аналитика это высказывание о каждом объекте можно сделать не более одного раза, поскольку оно имеет необратимый характер.

Высказывание 4 предполагает удаление кластера вместе со всеми ответами, которые в нём содержатся, и уменьшение счётчика числа кластеров k.

Высказывания 5 и 6 являются операциями фиксации и освобождения для кластеров. Основное отличие от операций для отдельных объектов состоит в том, что автоматическая процедура кластеризации не модифицирует фиксированный кластер, относить к нему ответы может только пользователь вручную. Если же в кластере фиксированы только некоторые объекты, а сам он не фиксирован, то автоматическая процедура не перемещает эти объекты, однако может изменить набор свободных объектов в этом кластере.

Высказывание 7, как уже отмечалось, необходимо для комфортной работы аналитика в системе.

Предполагается, что на одной итерации взаимодействия с системой аналитик может делать высказывание только одного типа среди высказываний 1-6, а также произвольное количество высказываний типа 7. Из этого ограничения следует, например, что на каждой итерации можно создать не более одного нового кластера.

При этом для облегчения работы эксперта в реализованной системе предусмотрена возможность сделать сразу несколько высказываний типов 1-3, т.е., к примеру, можно на одной итерации отнести сразу несколько объектов к кластеру C_i. Таким образом, на основе базовых высказываний пользователь имеет возможность строить более сложные.

Отметим, что все высказывания эксперта, сделанные на очередной итерации, применяются к результату, полученному в результате предыдущих итераций. Эксперту предлагается изменить часть требований к системе кластеров, при этом все остальные требования остаются в силе.

Сформулируем и докажем утверждение о полноте предложенной системы высказываний.

Утверждение 6. *Высказывания 1, 3 и 4 описанной системы высказываний позволяют достичь любой желаемой кластеризации ответов.*

Доказательство. Приведём конструктивный алгоритм получения любой наперёд заданной желаемой кластеризации с кластерами $C_1, \ldots, C_k$. Сперва, пользуясь высказыванием 4, удалим все объекты, которые не вошли ни в один из итоговых кластеров. Далее, за k высказываний типа 1, порождающих новые кластеры, распределим все объекты по кластерам $C_1, \ldots, C_k$. Наконец, удалим ставшие пустыми исходные кластеры с помощью высказывания 3. ■

Таким образом, опираясь на описанную систему высказываний, пользователь может получить любую желаемую кластеризацию объектов. При этом для её достижения потребовались только 3 типа высказываний системы, остальные же необходимы для удобства работы пользователя в системе.

5.3.2 Учёт высказываний эксперта при кластеризации

Опишем теперь, каким образом автоматизированная процедура кластеризации учитывает сделанные пользователем высказывания. Будем придерживаться парадигмы частичного обучения (англ. *semi-supervised learning*), для кластеризации будем пользоваться исследованным в разделе 4.5 сферическим алгоритмом K средних с применением семантического сглаживания.

Изначально пользователю предлагается ввести стартовое количество кластеров, после чего строится первичная кластеризация. При последующих запусках алгоритма кластеризации с учётом пользовательских высказываний имеется три типа объектов (ответов респондентов), учитываемые различным образом при построении кластеризации:

1. Фиксированные пользователем объекты, не принадлежащие фиксированным кластерам — эти объекты, рассматриваются как представители своих кластеров, они не могут менять кластер в ходе работы алгоритма кластеризации. Обозначим их через $x_1, x_2, \ldots, x_l$.

2. Свободные объекты — именно для этих объектов кластер определяется автоматически при запуске алгоритма кластеризации. Обозначим их через x_{l+1}, $x_{l+2}, \ldots, x_{l+q}$.

3. Удалённые пользователем объекты и объекты, относящиеся к фиксированным кластерам — они не принимают участия в работе алгоритма.

Пусть имеется k незафиксированных кластеров. Тогда формулы пересчёта для сферического алгоритма 4.14 принимают вид:

$$\mu_i = \frac{1}{|C_i|}\sum_{x_j \in C_i} x_j,\ c_i = \frac{\mu_i}{||\mu_i||},\ i \in \{1,\ldots,k\} \tag{5.1}$$

$$r_j = \underset{i}{\operatorname{argmax}}(x_j, c_i),\ j \in \{l+1,\ldots,l+q\} \tag{5.2}$$

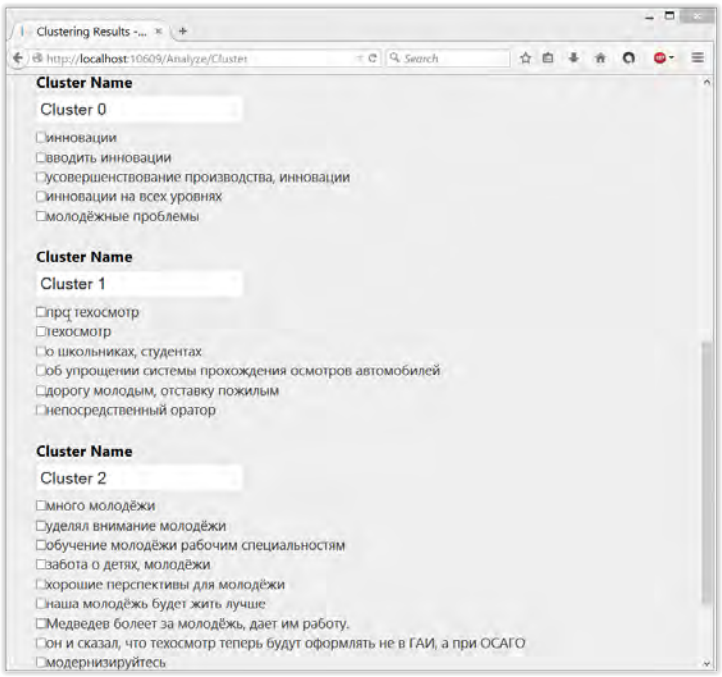
То есть необходимо пересчитывать принадлежность свободных объектов незафиксированным кластерам.

Таким образом, полностью описана интерактивная система, помогающая эксперту в решении прикладной задачи анализа открытых вопросов.

5.4 Эмпирический анализ интерактивной кластеризации

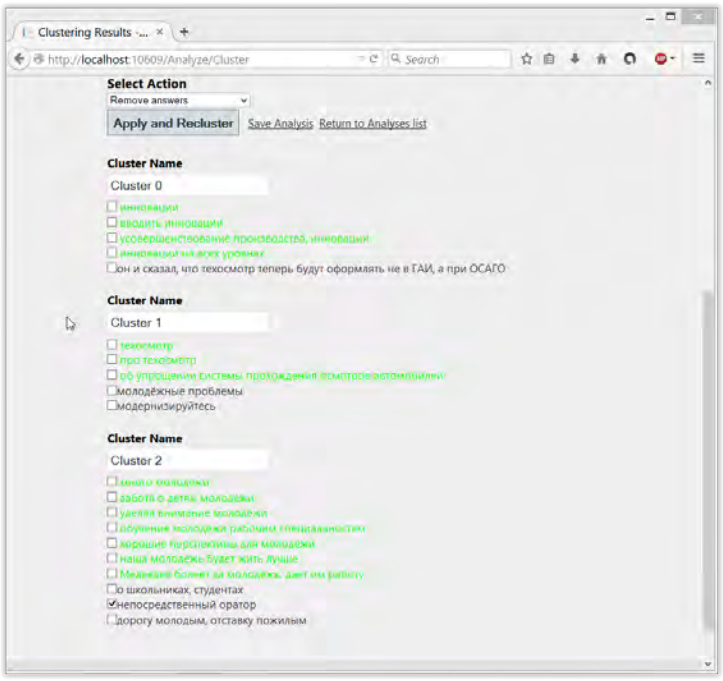
Приведём типовой сценарий использования разработанного интерфейса для анализа другого подмножества ответов на вопрос, заданный ФОМ в 2010 году после пресс-конференции Дмитрия Медведева: «Что из того, о чем говорил Д. Медведев на пресс-конференции, Вам больше всего запомнилось и понравилось?».

Будем считать, что пользователь уже зарегистрирован в системе, а данные открытого вопроса внесены в систему. Начнём анализ, рассмотрим результаты первичной кластеризации:

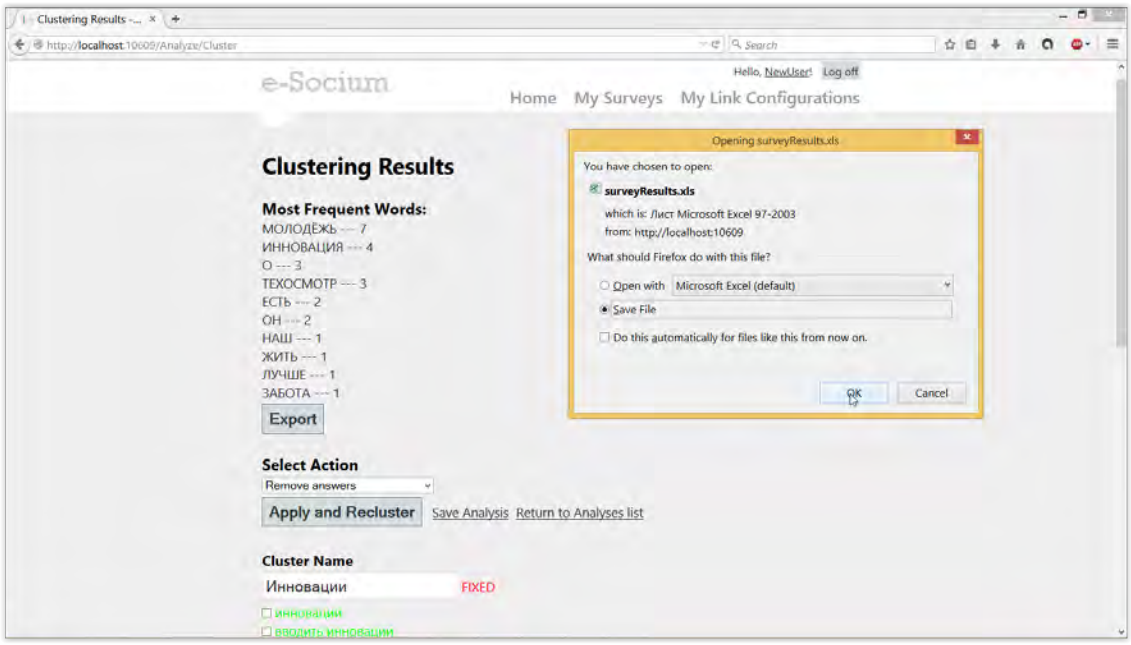


Кластеризация прошла успешно. Результат, безусловно, нуждается в дополнительной обработке, хотя кластеры, посвящённые молодёжи, инновациям и техосмотру уже были успешно обнаружены алгоритмом.

Теперь зафиксируем объекты, правильно отнесённые к кластерам, а затем удалим малоинформативный ответ «непосредственный оратор»:



Как видим, после этого шага была получена правильная кластеризация. Пользователю остаётся дать имена кластерам и экспортировать результат кластеризации в формате Excel:



Таким образом, на данном примере за несколько простых действий с помощью реализованной системы была получена необходимая кластеризация ответов, что демонстрирует эффективность разработанного решения.

Глава 6

Заключение

В ходе исследования были достигнуты следующие результаты:

1. проанализирована задача прикладной области, проведена формализация задачи, выделены основные этапы решения;
2. проведён обзор существующих моделей и алгоритмов для решения задачи кластеризации текстов, а также кластеризации коротких текстов с учётом проблемы разреженности;
3. проведено исследование теории ультраметрических пространств в приложении к задаче кластеризации;
4. доказана теорема об эквивалентности задачи поиска субдоминантной псевдоультраметрики и задачи построения агломеративной кластеризации Ланса-Уильямса с расстоянием ближнего соседа;
5. предложен метод для генерации модельных коллекций коротких текстов;
6. разработана методика применения семантического сглаживания в рамках агломеративных алгоритмов иерархической кластеризации и сферического алгоритма K средних;
7. путём эмпирического анализа доказана эффективность семантического сглаживания как метода борьбы с проблемой разреженности;
8. разработана система высказываний эксперта, позволяющая в ходе интерактивного взаимодействия с процедурой интеллектуальной обработки данных произвести разбиение ответов на группы близких по содержанию;
9. проведена формализация высказываний эксперта на основе алгоритма частичного обучения и показана эффективность предложенного решения;
10. на основе построенной математической модели реализован интерактивный веб-интерфейс для работы пользователя с исходными данными открытых вопросов и проведения их анализа.

Литература

[1] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[2] Nicholas O Andrews and Edward A Fox. Recent developments in document clustering. *Computer Science, Virginia Tech, Tech Rep*, 2007.

[3] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2007.

[4] Steve Branson and Ari Greenberg. Clustering web search results using suffix tree methods. Technical report, Stanford University, Tech. Rep. CS276A Final Project, 2002.

[5] Christian Buchta, Martin Kober, Ingo Feinerer, and Kurt Hornik. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012.

[6] James W Carey, Mark Morgan, and Margaret J Oxtoby. Intercoder agreement in analysis of responses to open-ended interview questions: Examples from tuberculosis research. *Cultural anthropology methods*, 8(3):1–5, 1996.

[7] Marie Chavent, Yves Lechevallier, and Olivier Briant. Divclus-t: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52(2):687–701, 2007.

[8] David Cheng, Ravi Kannan, Santosh Vempala, and Grant Wang. A divide-and-merge methodology for clustering. *ACM Transactions on Database Systems (TODS)*, 31(4):1499–1525, 2006.

[9] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 318–329. ACM, 1992.

[10] Martin Farach, Sampath Kannan, and Tandy Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13(1-2):155–179, 1995.

[11] Benjamin CM Fung, Ke Wang, and Martin Ester. Hierarchical document clustering using frequent itemsets. In *SDM*, volume 3, pages 59–70. SIAM, 2003.

[12] Evgeniy Gabrilovich and Shaul Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI*, volume 5, pages 1048–1053, 2005.

[13] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[14] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[15] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 919–928. ACM, 2009.

[16] Kristin M Jackson and William MK Trochim. Concept mapping as an alternative approach for the analysis of open-ended survey responses. *Organizational Research Methods*, 5(4):307–336, 2002.

[17] Nick Jardine and Cornelis Joost van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240, 1971.

[18] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 775–784. ACM, 2011.

[19] Anjali Ganesh Jivani et al. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938, 2011.

[20] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[21] Jaz Kandola, Nello Cristianini, and John S Shawe-taylor. Learning semantic similarity. In *Advances in neural information processing systems*, pages 657–664, 2002.

[22] Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[23] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22. ACM, 1999.

[24] Julie B Lovins. *Development of a stemming algorithm.* MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[25] Prasenjit Majumder, Mandar Mitra, Swapan K Parui, Gobinda Kole, Pabitra Mitra, and Kalyankumar Datta. Yass: Yet another suffix stripper. *ACM transactions on information systems (TOIS)*, 25(4):18, 2007.

[26] James Mayfield and Paul McNamee. Single n-gram stemming. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 415–416. ACM, 2003.

[27] Massimo Melucci and Nicola Orio. A novel method for stemmer generation based on hidden markov models. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 131–138. ACM, 2003.

[28] Glenn W Milligan. Ultrametric hierarchical clustering algorithms. *Psychometrika*, 44(3):343–346, 1979.

[29] Martin F Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[30] Mehran Sahami and Timothy D Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386. AcM, 2006.

[31] George Siolas and Florence d'Alché Buc. Support vector machines based on a semantic kernel for text categorization. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 5, pages 205–209. IEEE, 2000.

[32] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[33] WT Williams and GN Lance. A general theory of classification sorting strategies: 1. hierarchical systems, 2. clustering systems. *Computer Journal*, 9:10.

[34] Jinxi Xu and W Bruce Croft. Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems (TOIS)*, 16(1):61–81, 1998.

[35] Wen-Tau Yih and Christopher Meek. Improving similarity measures for short segments of text. In *AAAI*, volume 7, pages 1489–1494, 2007.

[36] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524. ACM, 2002.

[37] Shi Zhong and Joydeep Ghosh. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3):374–384, 2005.

[38] Антон Викторович Варламов, Максим Игоревич и Коршунов. Расчет семантической близости концепций с использованием связей в графе ссылок Википедии. 2014.

[39] Евгений Александрович Довгошей, Алексей Альфредович и Петров. Субдоминантная псевдоультраметрика на графах. *Математический сборник*, 204(8):51–72, 2013.

[40] Наталья Валентиновна Лукашевич. Тезаурусы в задачах информационного поиска. *М.: Издательство Московского университета*, 2011.