Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров

© А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков

МГТУ им. Н.Э. Баумана
arka@inteltec.ru, dmitryb@inteltec.ru, mvvbox@yandex.ru, simakov99@mail.ru

Аннотация

В статье изложен нейросетевой метод кластеризации коллекций текстовых документов на основе карт Кохонена. Также предложен метод синтеза аннотаций для формируемых кластеров, в основе которого лежит идея поиска устойчивых словосочетаний.

1 Введение

Отличительной особенностью методов кластеризации является способность автоматически выделять группы в потоке входных данных. В контексте обработки текстов на естественном языке это свойство является особенно привлекательным, когда возникает необходимость оперативно выделить группы в большом массиве текстовых документов. Эта задача, например, актуальна для информационно-поисковых систем, когда пользователю необходимо дать поверхностное представление обо всем списке найденных документов, не принуждая его просматривать большую часть этих документов.

В такой постановке задачи дополнительно возникает необходимость синтеза аннотаций кластеров, кратко отражающих тематику их документов.

2 Основные методы кластеризации

Объекты, с которыми мы оперируем — это тексты или части документов, т.е. некоторые фрагменты текстов. Каждый из документов представляется набором слов (или словосочетаний, или других единиц языка). В этом наборе могу встречаться конструкции, которые не должны влиять на результаты поиска: некоторые слова общей лексики, предлоги и другие строки. Имеются также слова, которые непосредственно влияют на отнесение документа в какую-либо категорию, – это термины. Каждый термин является элементарным признаком, множество терминов составляет пространство. Множество документов – это множество точек или векторов этого пространства. Координатами точки являются величины значимости каждого термина для данного документа. Величина значимости может оцениваться различными моделями. Часто используются следующие модели оценки значимости:

- Хэммингово расстояние, 0 означает, что термин в данном документе не встречается, 1 – встречается,
- относительная частота встречаемости слова,
- модель TF-IDF.

При кластеризации документов следует стремится к тому, чтобы схожесть документов, попадающих в кластер, была смысловой. При этом численно схожесть задается следующими способами:

1. множество документов, значения близости между любыми двумя элементами которого не меньше определённого порога $d^2(X_i, X_j) \geq D_{min}$, где $X_i, X_j \in X$ – множество всех векторов признаков;

2. значение близости между любым документом множества и центроидом кластера не меньше определённого порога, центроид кластера рассчитывается как среднее арифметическое между всеми документами кластера: $C_S = \frac{1}{|S_l|} \sum_{X_i \in S_l} X_i$, где $S_l \in S$ – множеству кластеров;

3. В качестве критерия качества разбиения для алгоритмов с заданным числом кластеров используются критерии:

$Q_D(S) = \sum_{j=1}^{k} \sum_{X_i \in S_l} d^2(X_i, \overline{X}_j)$ – сумма внутриклассовых дисперсий,

$Q_M(S) = \sum_{j=1}^{k} \sum_{X_i \in S_l} d^2(X_i, X_j)$ – сумма попарных внутриклассовых расстояний между элементами.

Для разбиения с заранее неизвестным числом классов критерий качества задают в виде комбинации двух критериев, один из которых характеризует внутриклассовый разброс наблюдений, второй – возрастающей (или неубывающей) функцией числа классов k. Например, в виде формулы:

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

$$D_U(S) = \left(\sum_{j=1}^{k} \sum_{X_i \in S_l} d^2(X_i, X_j) \right) \cdot (ck(S)),$$

где c – некоторая положительная константа, а k(s) – число классов, при разбиении S.

2.1 Suffix Tree Clustering

Метод Suffix Tree Clustering предложил Орен Замир (Oren Zamir) в диссертационной работе «Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results» [7]. Метод Suffix Tree Clustering кластеризует тексты в виде суффиксного дерева. Суффиксное дерево – дерево, содержащее все суффиксы данной строки. Кластеры образуются в узлах специального вида дерева – суффиксного дерева, которое строится из слов и фраз входных документов [1][6][7].

Достоинства метода:

– высокая скорость работы. По времени и занимаемой памяти дерево строится пропорционально количеству документов O(n) ,
– наглядность представления результатов (извлекаются ключевые слова для обозначения кластеров),
– алгоритм не нуждается в обучении и задании порога срабатывания;

Недостатки метода:

– сравнительно низкая точность (по данным источника в среднем 30%),
– важен порядок слов в документе для определения названий кластеров, состоящих из нескольких слов,
– не выявляется скрытая семантика среди документов, которая может присутствовать не только на текстовом уровне;

проблемы синонимии и омонимии.

Данный метод применяется в поисковой машине Vivisimo (www.vivisimo.com), которая была разработана в университете Карнеги, первоначально как экспериментальный некоммерческий проект.

2.2 Single Link, Complete Link, Group Average

Иерархические методы делят на агломеративные и дивизимные, первые из которых объединяют объекты в множества, а вторые наоборот разделяют единое множество объектов на подмножества [2].

Методы Single Link, Complete Link, Group Average относятся к агломеративным иерархическим методам, которые получили широкое распространение. Пусть задано множество объектов O1, O1,… On., для которых определены признаки в виде матрицы:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(p)} & x_2^{(p)} & \dots & x_n^{(p)} \end{bmatrix}$$

где $x_i^{(j)}$ – значение j-го признака на i-м объекте, соответственно, каждый столбец матрицы $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})'$ характеризует объект O_i, т.е. является результатом статистического обследования объекта по p признакам.

Достоинства методов:

– иерархическая кластеризация.

Недостатки методов

– необходимо задавать порог – максимальное количество документов в кластере,
– высокая вычислительная сложность для алгоритмов Single Link и Group Average – $O(n^2)$, а в Complete Link – $O(n^3)$, где n – количество документов.

2.3 K-means

В основе K-means (К-средних) лежит итеративный процесс стабилизирования центроидов кластеров. Основной характеристикой кластера является его ценроид и вся работа алгоритма направлена на стабилизирование или, в лучшем случае, полное прекращение изменения центроида кластера [2].

Достоинства метода:

– низкая вычислительна сложность – O(knT), где n – число документов, k – число кластеров, T – количество итераций,
– метод не нуждается в обучении и при необходимости может накапливать сведения для дальнейшего увеличения точности работы – использование Байесовских оценок параметров кластеризации.

Недостатки метода:

– требуется задание количества кластеров, как минимум на начальных этапах – до использования априорной информации,
– сравнительно низкая точность,
– в том случае, когда центроиды кластеров выбираются случайным образом, результаты, получаемые над одной и той же выборкой документов, будут отличаться. Это может происходить по причине неудовлетворительной работы генератора случайных чисел и вследствие равномерного распределения документов в пространстве – без явных областей сгущения;

2.4 Concept Indexing

Этот метод используется для уменьшения размерности пространства признаков. В пространстве признаков, размерность которого уменьшена, выполняется стандартное отображение множества документов [8].

Основное отличие метода от других заключается в том, что алгоритм может быть обучаемым или необучаемым.

Необучаемый алгоритм поиска кластеров заключается в непосредственном разбиении на k кластеров (k-way) или методом рекурсивной бисекции. В начале процесса непосредственного разбиения про-

извольно выбираются k документов, которые являются центрами для кластеров, затем к этим центрам относят документы, которые имеют наивысшее значение коэффициента близости и производят пересчёт центроидов возникших кластеров. Процесс продолжается до исчерпания всех документов или ограничиваются определённым числом итераций.

Метод рекурсивной бисекции подразумевает разбиение всего объема документов на 2 части, затем эти части тоже подвергаются разбиению, по необходимости. Процесс продолжается рекурсивно до получения k кластеров. Алгоритм рекурсивной бисекции использует значение «суммарного несходства» (aggregate dissimilarity) между документами в кластере, для того чтобы решить какой кластер разбивать дальше. Это значение равно:

$$AggregateDissimilarity = |S_i|^2 \left(1 - \|C_i\|_2^{\ 2}\right)$$

где $|S_i|$ - объем кластера. $||C_i||_2$ – норма центроида i-го кластера.

Обучаемый вариант метода производит классификацию приходящих документов. Документы распределяются по предопределённым кластерам – классам. Кластеры определяются своими центроидами, присвоенными кластерам на этапе обучения. Практически обучение заключается в фиксировании предопределённого количества центроидов кластеров, а классификация заключается в отнесении документа к кластеру, расстояние до центроида которого, наименьшее.

Достоинства метода:

- алгоритм рекурсивной бисекции обладает высокой скоростью работы – выше линейной (O(N×log k)), где N – число документов, k – число кластеров,
- для определения точности алгоритма используется величина RI (retrieval improvement):

$$RI = \frac{\sum_{d \in D} n_d^r}{\sum_{d \in D} n_d^o}$$

где n_d^r – количество документов, принадлежащих классу d в «уменьшенном» пространстве признаков размерности r, n_d^o – количество документов, принадлежащих классу d в оригинальном пространстве признаков размерности,

- использует значения матрицы tf•idf;

Недостатки метода:

- один из вариантов метода – обучаемый, что для автоматических систем это, безусловно, недостаток, однако обучение заметно улучшает качество работы,
- требуется задание количества кластеров, на которые будет разбиваться множество документов.

2.5 SOM

Самоорганизующиеся карты (Self-Organizing Maps, SOM) Кохонена выполняют обобщение предъявляемой информации. Наиболее полное изложение теории SOM можно найти в фундаментальной работе Kohonen T. Self-Organizing Maps. В результате работы нейросети Кохонена получается образ, представляющий собой карту(как правило, двумерную) распределения векторов из обучающей выборки [3][4].

Данная сеть обучается без учителя на основе самоорганизации. Известно значительное число модификаций данной сети, в самом общем виде алгоритм самоорганизации следующий: По мере обучении вектора весов нейронов стремятся к центрам кластеров – групп векторов обучающей выборки. На этапе решения информационных задач сеть относит новый предъявленный образ к одному из сформированных кластеров, указывая тем самым категорию, к которой он принадлежит.

Сеть Кохонена состоит из одного слоя нейронов. Число входов каждого нейрона равно размерности входного образа. Количество нейронов определяется той степенью подробности, с которой требуется выполнить кластеризацию набора библиотечных образов. При достаточном количестве нейронов и удачных параметрах обучения сеть Кохонена может не только выделить основные группы образов, но и установить структуру полученных кластеров. При этом близким входным образам будет соответствовать близкие карты нейронной активности.

Обучение начинается с инициализации весов нейронов путем задания случайных значений или на основе обучающей выборки, векторы $\{X_i\} \in \Re^n$. В дальнейшем происходит процесс самоорганизации, состоящий в модификации весов при предъявлении на вход векторов обучающей выборки. Для каждого нейрона можно определить его расстояние до вектора входа $\{W_j\} \in \Re^n$:

$$\left\| X_i - W_c \right\| = \min_j \{ X_i - W_j \},$$

где W_c – нейрон-победитель W_c. На текущем шаге обучения t будут модифицироваться веса нейронов из окрестности нейрона W_c:

$$W_j^{t+1} = W_j^t + \alpha(t) \cdot h_{ci}(t) \cdot (X_i - W_j^t).$$

Первоначально в окрестности $h_{ci}(t)$ любого из нейронов находятся все нейроны сети, в последствии эта окрестность сужается. В конце этапа обучения подстраиваются только веса самого ближайшего нейрона. Темп обучения $\alpha(t) < 1$ с течением времени также уменьшается. Образы обучающей выборки предъявляются последовательно, и каждый раз происходит подстройка весов. Нейронная сеть Кохонена может обучаться и на искаженных версиях входных векторов, в процессе обучения искажения, если они не носят систематический характер, сглаживаются.

Для наглядности представления карты нейроны Кохонена могут быть упорядочены в двумерную матрицу, при этом под окрестностью нейрона-победителя принимаются соседние (по строкам и

столбцам) элементы матрицы. Результирующую карту можно представить в виде двумерного изображения, на котором различные степени возбуждения всех нейронов отображаются квадратами различной площади.

Вычислительная сложность сети SOM $O(dN^2)$, где n – число векторов в обучающей выборке, d – размерность входных векторов.

В 1996 предложена модификация алгоритма SOM, названная WEBSOM, нашедшая практическое применение в кластеризации больших массивов текстовой информации.

2.6 ART

Сети АРТ позволяют проверить, соответствует ли «новый» образ «старому», что невозможно выполнить другими типами нейронных сетей [5, 9]. Так, например, многослойный персептрон, обучающийся по методу обратного распространения, запоминает весь пакет обучающей информации, при этом образы обучающей выборки предъявляются в процессе обучения многократно. Попытки затем обучить персептрон новому образу приведут к модификации синаптических связей с разрушением структуры памяти о предыдущих образах. Особенностью нейронных сетей с адаптивным резонансом является то, что они сохраняют «пластичность» при запоминании новых образов, и, в то же время, предотвращают модификацию старой памяти. Нейросеть имеет внутренний детектор новизны – тест на сравнение предъявленного образа с содержимым памяти. При удачном поиске в памяти предъявленный образ классифицируется с одновременной уточняющей модификацией синаптических весов нейрона, выполнившего классификацию. О такой ситуации говорят, как о возникновении адаптивного резонанса в сети в ответ на предъявление образа. Если резонанс не возникает в пределах некоторого заданного порогового уровня, то успешным считается тест новизны, и образ воспринимается сетью, как новый. Модификация весов нейронов, не испытавших резонанса, при этом не производится.

Сеть АРТ-1 является классификатором входных двоичных образов по нескольким сформированным сетью категориям. Решение принимается в виде возбуждения одного из нейронов распознающего слоя, в зависимости от степени похожести образа на шаблон критических черт данной категории. Если эта степень похожести невелика, т.е. образ не соответствует ни одной из имеющихся категорий, то для него формируется новый класс, который в дальнейшем будет модифицироваться и уточняться другими образами, формируя свой шаблон критических признаков. Для описания новой категории отводится новый, ранее не задействованный нейрон в слое распознавания. Эта особенность сетей АРТ позволяет использовать их для автоматической классификации текстов.

Используемые в сети ART компоненты вектора признаков являются дихотомическими, что не позволяется задавать значимость терминов. Кроме того, сеть ART имеет высокую чувствительность к шуму, что делает ее непригодной для задачи кластеризации текстов.

3 Предложенная модель нейросети

Предложенная модель многослойной SOM позволяет сохранить достоинства обычной SOM, такие как простое обучение и быстрая работа для обученной сети. Основной идеей было создание такой сети, которая имеет гибкую и динамическую архитектуру, способную адаптировать свой размер и согласно данному набору данных.

После обучения сети, Таким образом, тексты представляются не единым объектом, а разделяются на несколько отдельных сегментов, что в принципе также позволяет множественное размещение одного и того же документа в различных тематических ветвях, используя алгоритмы размытых множеств.

Многослойная SOM представляет собой структуру из множества уровней, каждый из которых состоит из нескольких независимых растущих SOM. Начиная с самого верхнего уровня, каждая карта растет в размерах, чтобы представить детальное представление на каждом уровне. После улучшений глубины детализации, кластеры анализируются, чтобы определить какие из них представляют данные с минимальным уровнем детализации. Эти кластеры, которые являются слишком разнотипными входными данными, расширяются на новой растущей SOM на нижележащем уровне, который представляет данные более детально.

Такая карта снова растет в размерах, пока определенные улучшения качества представления данных не достигнуты. Кластеры представленные на уже однородном наборе данных не требуют дальнейшего расширения на более низкие уровни.

Т.е. многослойная SOM представляет собой дерево из нескольких уровней, каждый узел которого содержит SOM. SOM на первом уровне имеет размерность 1x1 и приближенно организует входящие данных. Каждому кластеру из карты первого уровня соответствует SOM второго уровня, представляющих собой более детальное представление данных и т.д. Входные данные для одной карты представляет собой набор, которые попал в связанный кластер на верхнем уровне. Кластеры второго уровня аналогично связанны с SOM третьего уровня, которая обеспечивает достаточно разряженное представление для входных данных. Карты имеют различные размеры, согласно структуре данных. Нулевой уровень обеспечивает представление всего набора данных и контроль за процессом роста дерева.

Каждый слой иерархии представляет собой обычную карту Кохонена, состоящую из одного слоя нейронов. Число входов каждого нейрона равно размерности входного образа. Обозначим $X=(x_1 \dots x_N)$ множество входных векторов, каждый из которых представляет семантический образ документа, каждый вектор нормализован $\|x_i\| = 1$, W

– матрица связей размерности mЧn, элементами которой являются весовые векторы.

Обучение начинается с задания случайных значений матрице связей W. В дальнейшем происходит процесс самоорганизации, состоящий в модификации весов при предъявлении на вход векторов обучающей выборки. Для каждого нейрона можно определить его расстояние до вектора входа.

$$d_m = \left\| x_k - w_{ij} \right\| \quad (1)$$

Далее выбирается нейрон w=w*, для которого это расстояние минимально.

$$d_{min} = \min_{\substack{i=1..m \\ j=1..n}} \left\| x_k - w_{ij} \right\| \quad (2)$$

На текущем шаге обучения t будут модифицироваться только веса нейронов из окрестности нейрона w*:

$$w_i^j(t+1) = w_i^j(t) + \alpha(t) \cdot h_{ci} \cdot \left[x(t) - w_i^j(t) \right],$$

где $h_{ci} = exp\left(- \frac{\left\| r_c - r_i \right\|}{2 \cdot \delta(t)^2} \right)$

Первоначально в окрестности любого из нейронов находятся все нейроны сети, в последствии эта окрестность сужается. В конце этапа обучения подстраиваются только веса самого ближайшего нейрона. Скорость обучения $\alpha(t) < 1$ с течением времени также уменьшается. Образы обучающей выборки предъявляются последовательно, и каждый раз происходит подстройка весов.

Для наглядности представления карты нейроны Кохонена могут быть упорядочены в двумерную матрицу, при этом под окрестностью нейрона-победителя принимаются соседние (по строкам и столбцам) элементы матрицы. Результирующую карту можно представить в виде двумерного изображения.

При обучении карты ее точный размер не известен, для определения необходимости увеличения размера карты предложен следующий критерий:

Для каждого кластера вычисляется коэффициент «различности» векторов в кластере m_{ij}.

$$m_{ij} = \frac{1}{n_C} \cdot \sum_{x_i \in C_{ij}} \left\| w_{ij} - x_i \right\| \quad (3)$$

где C_{ij} – множество векторов, отнесенных к нейрону матрицы ij, n_c – мощность множества C_{ij}.

Для 0-го слоя данных критерий можно сформулировать следующим образом:

$$m_0 = \frac{1}{N} \cdot \sum_{x_i \in X} \left\| w_0 - x_i \right\| \quad (4)$$

где N – общее количество входных векторов.

В процессе обучения возможно изменение размера карты, как предложено. Определяется нейрон w', для которого максимальный m_{ij} из всех.

$$m_{max} = max\left(\frac{1}{n_C} \cdot \sum_{x_i \in C_{ij}} \left\| w_{ij} - x_i \right\| \right) \quad (5)$$

Далее, из соседних нейронов определяется тот, для которого расстояние наибольшее, и между этим нейронов и w' вставляется новый ряд или строка в матрицу W в зависимости, от того какая из метрик меньше.

$$d_{max} = \max_{w \in C_N} \left\| w - w' \right\| \quad (6)$$

где C_N – множество соседних нейронов. При этом для веса добавленных нейронов инициализируется как среднее от весов их соседей.

В качестве критерия остановки роста размеров карты предложен следующий критерий:

$$M = \frac{1}{m \cdot n} \sum_{\substack{i=1...n \\ j=1...m}} m_{ij} \text{ , } M < \tau 1 \cdot m_u \quad (7)$$

где $\tau 1$ – числовой коэффициент из $0 < \tau 1 < 1$, m_u – коэффициент «различности» для кластеров вышестоящего уровня. При этом уменьшение $\tau 1$ приводит к увеличению результирующей карты, и, соответственно, к большей детализации представления, и наоборот.

Учитывая, что в данном случае количество кластеров ограничено сверху и снизу значениями M_{min}, M_{max}, то возможно оценить разброс значений m_u и выбрать фиксированный размер карты.

После завершения обучения карты, каждый кластер должен быть проверен – нужно ли создавать карту следующего уровня или нет. Т.е. должен быть критерий оценки, насколько содержащиеся в кластере тексты «различны». Для этого предложен следующий критерий:

$$m_{ij} < \tau 2 \cdot m_0, \text{ card } C_{ij} < \mu \quad (8)$$

где $\tau 2$ – числовой коэффициент, $0 < \tau 2 < 1$, выбираемый эмпирически, C_{ij} – множество входных векторов, отнесенных к данному кластеру, μ – ограничение минимального числа текстов в кластере, т.к. с точки зрения уточнения результатов поиска кластеры с малых числом текстов бесполезны, данный коэффициент лежит в пределах 10..25.

Если критерий не выполняется, т.е. $m_{ij} \geq \tau 2 \cdot m_0$ или card $C_{ij} > \mu$, то создается следующий уровень кластера, для которого повторяется процесс обучения. Коэффициент $\tau 2$ позволяет управлять детальностью кластеризации, уменьшение этого коэффициента приводит к увеличению числа уровней к иерархии и, соответственно, к повышению детализации представления, и наоборот.

Для формирования кластеров в заданных пределах M_{min}, M_{max} можно использовать алгоритм k-means (в т.ч. и модификацию алгоритма для размытых множеств), для определения оптимального количества кластеров k_{opt} с точки зрения минимизации внутриклассовых дисперсий.

4 Экспериментальная проверка модели

Для проведения эксперимента была выбрана коллекция текстов, представляющих собой блоки новостной информации. Общее количество текстов – 542.

Эксперимент проводился на ПЭВМ со следующими основными параметрами: один процессор Pentium 4, 3 Гц, объем ОЗУ 1 Гб.

В проведенных экспериментах искусственно использовалась плоская кластеризация вместо иерархической, что обусловлено целями эксперимента.

Время кластеризации коллекции текстов составило 18 минут.

В результате кластеризации, программа сформировала 19 кластеров. Размерность входных векторов составила 1577. Распределение количества текстов по кластерам с ранжированием показано на рис. 1, среднее количество текстов в кластере 28,52, медиана 22, дисперсия 399.26.

Рис. 1. Размеры кластеров

Для оценки качества кластеризации использовалась экспертная оценка. Для каждого кластера определялась преобладающая тема текстов и, исходя из этого, оценивалось количество правильно кластеризованных документов в каждом кластере. Результата оценки представлены на рис. 2. Процент правильно кластеризованных текстов по всей коллекции составил 59,06%. Среднее количество неправильно кластеризованных текстов в кластере составило 11, 68, медиана 7, дисперсия 139, 67.

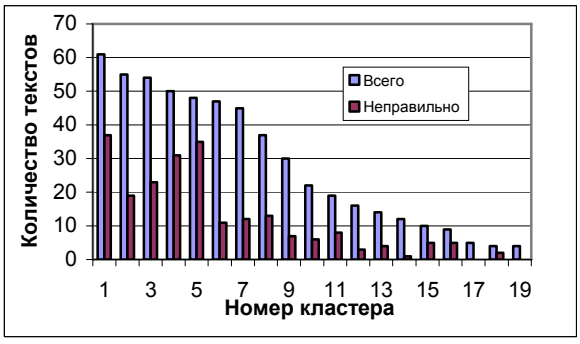

Рис. 2. Оценка качества кластеризации

Проведенный эксперимент показал пригодность предложенной модели для задачи кластеризации текстовых коллекций. Использование SOM с динамической архитектурой позволяет кластеризовать большие коллекции текстов, вместе с тем, для уменьшения времени кластеризации предпочтительно проводить обучений нейросети на выборке из всей коллекции текстов.

5 Метод синтеза аннотаций кластеров

Предложенный метод синтеза аннотаций кластеров реализует подход, в соответствии с которым в основу аннотации закладываются фрагменты текстов, подлежащих кластеризации. Таким образом, реальный синтез связанного текста не выполняется, а задача сводится к поиску текстового фрагмента, наиболее релевантного для всего кластера. В таком случае синтаксическая связанность аннотации гарантируется, поскольку полагается, что обрабатываемые тексты содержат связанные тексты. Исходя из соображений практической применимости, к формируемым аннотациям предъявляется два дополнительных требования.

Лаконичность: фрагменты текста, включаемые в аннотацию, должны быть короткими фразами, содержащими 2-3 слова. В противном, случае при употреблении в аннотациях развернутых предложений, восприятие результата кластеризации будет затруднено. Напомним, что кластеризация в нашем случае необходима для более лаконичного представления результата поиска документов по большой текстовой коллекции (более 1 млн. текстов). Таким образом, пользователь, получая набор кластеров, каждый из которых охарактеризован краткими фразами, имеет возможность быстро сориентироваться в списке документов, возвращенных в результате поиска.

Контекстная независимость: форма, в которой записан фрагмент, включаемый в аннотацию, не должна предполагать наличия контекста. Предположим, что релевантной аннотацией для текстов кластера является фраза «начало массового производства». Если данный фрагмент изъять из его контекста *«Почти сразу же поступил ответ и от Samsung, сообщившей о начале массового производства на втором своем заводе седьмого поколения»*, получим словосочетание «начале массового производства», использование которого в роли аннотации кластера без контекста неприемлемо.

Предложенный метод синтеза аннотаций полагает, что для каждого кластера сформированы списки ключевых многословных терминов, имеющих наибольшую информационную значимость для кластера. Каждое слово термина представлено канонической формой, например, приведенному выше фрагменту будет соответствовать термин «начало, массовый, производство». Процедура синтеза разделена на следующие три этапа:

1. поиск релевантных текстовых фрагментов для каждого термина,
2. выбор нормализованного фрагмента,
3. выбор наиболее предпочтительного фрагмента.

Рассмотрим указанные этапы подробнее.

5.1 Поиск релевантных текстовых фрагментов

Поскольку термин образован каноническими формами слов, для поиска всех вхождений термина в текст кластера необходимо также привести все слова текста к их каноническим формам. Для этих целей использовался морфологический анализатор, функционирование которого описано в [10]. Преобразованный таким образом в сплошную цепочку канонических форм текст кластера подлежит сканированию, в процессе которого по критерию точного совпадения отыскиваются все вхождения термина в обрабатываемый текст. Найденные вхождения термина объявляются кандидатами на роль аннотации кластера.

5.2 Выбор нормализованного фрагмента

На данном этапе из всех найденных кандидатов необходимо отобрать один единственный, соответствующий двум сформулированным выше критериям о лаконичности и контекстной независимости.

Для осуществления такого отбора предложен эвристический критерий, сформулированный согласно следующим соображениям.

1. Поскольку для комфортного восприятия аннотации кластеров должны быть предельно краткими (2-3 слова), то целесообразно включать в аннотации фразы, представленные именными группами (глагольные группы представленные глаголом, согласованным с наречием, не являются информативными, а при включении в группу дополнения теряется свойство лаконичности аннотации).
2. Для обеспечения контекстной независимости формы, в которой записан фрагмент, необходимо отбирать кандидатов, в которых употребляются слова в именительном падеже.

Все фрагменты-вхождения термина фильтруются согласно этим двум критериям следующим образом. Сначала из всех вхождений термина остаются только те, среди слов которых присутствует существительное в именительном падеже. Затем, если среди оставшихся вхождений присутствуют те, в составе которых есть прилагательное в именительном падеже, то все прочие вхождения удаляются. В качестве примера рассмотрим термин «ценный, бумага» (напомним, что каждое слово термина представлено своей канонической формой). В кластере, содержащем тексты с проспектами эмиссий ценных бумаг коммерческих банков, найдены следующие вхождения данного термина (см. рис. 3).

ценных бумаг
ценной бумаги
ценным бумагам
ценных бумагах
ценные бумаги
ценными бумагами
ценной бумагой
ценными бумага
ценной бумаге
ценная бумага

Рис. 3. Пример вхождений термина

После фильтрации данного перечня первым критериями данный список сокращается до следующего (см. рис. 4).

ценной бумаги
ценные бумаги
ценными бумага
ценная бумага

Рис. 4. Первый шаг фильтрация

Как видно из рис. 4 в результат кроме всего прочего попали фрагменты, содержащие грамматические ошибки (выделено курсивом), однако применение второго критерия оставляет в данном списке только два корректных фрагмента (см. рис. 5).

ценные бумаги
ценная бумага

Рис. 5. Второй шаг фильтрации

Тем не менее, после двух шагов такой фильтрации может остаться большое количество вариантов (например, если группа представлена двумя существительными, первое из которых – главное слово – употребляется в различных формах, а второе – подчиненное слово – употребляется в единственной форме, например, в форме родительного падежа, совпадающей с формой именительного падежа омонима). Пример результирующего списка для того же кластера приведен на рис. 6.

проспекта эмиссии
проспект эмиссии
проспекте эмиссии
проспекты эмиссии
проспекту эмиссии
проспектом эмиссии

Рис. 6. Влияние омонимии

Если морфологический анализатор обладает способностью снятия омонимии, то описанная проблема не является актуальной и список фрагментов на рис. 6 был бы сокращен до двух, выделенных жирным шрифтом. Однако в нашем случае морфологический анализатор не был в состоянии оставить у слова *«эмиссии»* единственную форму родительного падежа, поэтому потребовался дополнительный критерий, позволяющий фильтровать подобные списки.

Суть данного критерия заключается в следующем. Все слова, составляющие фрагмент, разделяются на две следующие группы.

- Не изменяющиеся слова. К этой группе относятся слова, употребляющиеся в одних и тех же формах в большинстве анализируемых фрагментов.
- Изменяющиеся слова. К этой группе относятся слова, форма которых изменяется от одного фрагмента употребления термина к другому.

Для примера на рис. 6 к первой группе относится слово *«эмиссии»*, а ко второй – все приведенные формы слова *«проспект»*. Слова первой группы фиксируются, так что полагается, что их нормализованная форма найдена. Для каждого изменяющегося слова отыскивается форма именительного падежа.

Итоговым результирующим нормализованным фрагментом текста для исходного термина объявляется фрагмент с набольшей группой неизменяемых слов и с наибольшим количеством изменяемых слов, употребленных в форме именительного падежа. Для примера на рис. 6 этому критерию удовлетворяют именно выделенные жирным шрифтом фрагменты, из которых произвольным образом выбирается один единственный.

5.3 Выбор предпочтительного фрагмента

К данному этапу для обрабатываемого кластера имеется список терминов, с каждым из которых связан нормализованный фрагмент. Задача данного этапа выбрать из всех терминов наиболее предпочтительный так, что связанный с ним фрагмент образует аннотацию кластера. Для этих целей в данной работе взят за основу критерий TF-IDF при условии, что кластер документов рассматривается как один единый текст. Таким образом, частота термина рассматривается не в контексте одного документа, а в контексте всего кластера.

Поскольку желательно избегать употребления терминов из одних кластеров в аннотациях к другим, необходимо применять более строгие санкции к терминам, которые встречаются более чем в одном кластере. Более того, предпочтение следует отдавать терминам, имеющим равномерное распределение в текстах кластера, т.е. следует избегать включения в аннотацию терминов, встречающихся с одной стороны часто в одном документе кластера и редко – в другом. Также важно поощрять термины, покрывающие наибольшее число документов кластера. В связи с этими замечаниями, предложена следующая модификация функции TF-IDF, оценивающая предпочтительность термина для включения его в аннотацию.

$$p_t^c = \frac{f_t^c}{\sum_i f_i^c} \cdot \frac{C}{C_t} \cdot \frac{D_t^c}{D^c} \cdot \frac{1}{1 + log\left(1 + \sigma_t^c\right)}, \qquad (9)$$

где f_t^c – частота употребления термина *t* в текстах кластера *c*, f_i^c – частота употребления *i*-го термина в текстах кластера *c*, C – количество кластеров, C_t – количество кластеров, в которых встречается термин *t*, D_t^c – количество документов в кластере *c*, в которых встречается термин *t*, D^c – количество документов в кластере *c*, σ_t^c – среднее отклонение частоты употребления термина *t* в текстах кластера *c*.

Таким образом, из всех терминов кластера отбирается единственный, с максимальным значением предпочтительности p_t^c. Аннотацией кластера объявляется нормализованный фрагмент, связанный с данным термином.

6 Экспериментальная проверка метода

Эксперименты с методом синтеза аннотаций были выполнены над искусственной коллекцией, содержащей 4 группы документов из разных предметных областей:

bank: проспекты эмиссии ценных бумаг коммерческих банков (170 текстов),

crime: новости с криминальными сводками УВД некоторых регионов РФ (154 текста),

itnews: новости с анонсами ИТ-продуктов (315 текстов),

law: распоряжения кабинета министров Республики Татарстан (97 текстов).

В рамках каждого кластера были синтезированы списки из двух и трех словных именных групп, составленных из канонических форм слов исходных текстов. Далее в каждом списке было оставлено по 20 наиболее часто встречающихся словосочетаний, которые в итоге были объявлены терминами соответствующего кластера. Над каждым списком терминов выполнялись описанные выше процедуры синтеза аннотаций, отдельно для двух и трех словных терминов.

Поставленный таким образом эксперимент, с изначально явно непересекающимися кластерами, гарантирует корректную трактовку семантики кластеров экспертом и, как следствие, минимизацию ошибки при оценке результата аннотирования. Дополнительно такой эксперимент позволяет управлять ошибкой результата кластеризации, начиная с 0%, когда кластеры полностью соответствуют описанным четырем группам, и, заканчивая 100%, когда документы из всех групп равномерно распределены по четырем группам одинакового размера.

В таблице 1 приведены результаты синтеза аннотаций на основе двух словных терминов. Для каждого кластера отбиралось три термина с наибольшим весом p_t^c.

Каждая строка таблицы содержит результаты синтеза аннотаций для соответствующего кластера. В колонках таблицы отражены результаты для трех серий экспериментов, где приведен процент смешивания исходных эталонных групп. Первой колонке соответствует эксперимент без смешивания, что соответствует идеальному результату кластеризации, последней колонке соответствует результат кластеризации, когда 40% каждой группы равномерно заполнено содержимым из других групп. В каждой ячейке таблицы приведены первые три аннотации с максимальным весом p_t^c, полученные для соответствующего кластера в соответствующем эксперименте. Каждой аннотации поставлено в соответствие значение величины p_t^c.

Таблица 1. Синтез аннотаций на основе двух словных терминов

	0%	30%	40%
bank	уставной фонд: 0.22 собрание акционеров: 0.14 ценные бумаги: 0.12	ценные бумаги: 0.06 уставной фонд: 0.05 собрание акционеров: 0.03	уставной фонд: 0.043 ценные бумаги: 0.04 собрание акционеров: 0.02
crime	мера пресечения: 0.12 розыскные мероприятия: 0.11 обнаружение трупа: 0.09	мера пресечения: 0.027 розыскные мероприятия: 0.027 обнаружение трупа: 0.02	мера пресечения: 0.0209 розыскные мероприятия: 0.0204 ***изъятие наркотиков***: 0.015
itnews	технические характеристики: 0.05 жесткий диск: 0.038 карта памяти: 0.02	жесткий диск: 0.044 ***мобильный телефон***: 0.016 технические характеристики: 0.011	жесткий диск: 0.018 технические характеристики: 0.01 карта памяти: 0.006
law	республика Татарстан: 0.176 министр республики: 0.15 кабинет министров: 0.071	республика Татарстан: 0.076 министр республики: 0.03 кабинет министров: 0.029	республика Татарстан: 0.045 кабинет министров: 0.02 министр республики: 0.019

Для всех серий экспериментов видно, что фактически первые три наилучшие аннотации сохраняются независимо от качества кластеризации, что позволяет судить о применимости предложенного метода (во всем эксперименте появилось только две новые аннотации, они в таблице выделены жирным курсивом). Вместе с тем с ухудшением качества кластеризации значения всех весов уменьшаются, что является закономерным, поскольку с ростом степени смешения кластеров уменьшаются все четыре множителя выражения (9).

Акцентируя внимание на формулировках самих аннотаций, можно заключить о практической применимости предложенных критериев о лаконичности и контекстной независимости, а также о методе, синтезирующем аннотации, удовлетворяющие этим критериям.

В экспериментах с трех словными аннотациями имеют место закономерности, аналогичные приведенным в таблице 1. Отличие заключается лишь в самих формулировках:

для **bank** выделены аннотации: «выпуск ценных бумаг» и «уставной фонд банка»,

для **crime** выделены аннотации «труп неизвестного мужчины» и «тяжкие телесные повреждения»,

для **itnews** выделены аннотации «технология прямой печати» и «время автономной работы»,

для **law** выделены аннотации «министр республики Татарстан» и «постановления кабинета министров».

7 Заключение

Предложенная модель кластеризации текстов в ходе экспериментов продемонстрировала свою работоспособность. Использование семантического образа документа в качестве входного вектора для нейросети позволяет адекватно представлять содержание документа и получать в результате кластеризации результаты, пригодные для практического применения.

Предложенный метод синтеза аннотаций гарантирует синтаксическую связанность формируемых фраз, поскольку берет за основу связанные фразы текстов, подлежащих кластеризации. С точки зрения человека, аннотации, удовлетворяющие предложенным критериям лаконичности и независимости от контекста, воспринимаются комфортно и в полной мере могут дать поверхностное представление о предметной области документов, заключенных в кластере.

Свойства предложенных методов позволяют рекомендовать их совместное применение в виде единого подхода при построении информационно-поисковых систем нового поколения, обеспечивающих представление больших списков текстовых документов в виде лаконичного набора поименованных кластеров, дающих представление о структуре и составе найденного массива.

Литература

[1] Cao G., Song D., Bruza P. Suffix Tree Clustering on Post-retrieval Documents Information. The University of Queensland, 2003.

[2] Jain A. Dubs R., Clustering methods and algorithms,1988 // Prentice-Hall Inc.

[3] Kohonen T., Self-Organizing Maps, Springer, Berlin, 1995, 3rd extended edition, 2001.

[4] Kohonen T., et al. Self-Organization of a Massive Document Collection. Kohonen Maps. Elsevier, 1999.

[5] Mucoz Alberto, “Compound Key Word Generation from Document Databases Using A Hierarchical Clustering ART Model ”, Intelligent Data Analysis, Elsevier Science Inc., 1997.

[6] Zamir O. and Etzioni, O. Web Document Clustering: A Feasibility Demonstration. In Proceedings of ACM/SIGIR’1998.

[7] Zamir O. Clustering Web Documents:A Phrase-Based Method for Grouping Search Engine Results // A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy. University of Washington, 1999.

[8] Прикладная статистика: Классификации и снижение размерности : справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин ; под ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с. : ил.

[9] Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа / А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Пятой всероссийской научной конференции (RCDL'2003). – СПб. : НИИ Химии СПбГУ, 2003. – С. 140–149.

[10] Обучение морфологического анализатора на большой электронной коллекции текстовых документов / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : Труды Седьмой всероссийской научной конференции (RCDL'2005). – Ярославль : Ярославский государственный ун-т, 2005. – С. 173–181.

The method of clustering texts collections and clusters annotating

A. Andreev, D. Berezkin, V. Morozov, K. Simakov

The article contains the method of neural clustering collections of text documents on the basis of Self-Organization Map (SOM). We also propose a method for synthesizing annotations formed clusters, based on a method of finding collocations and stable phrases.