Метод тематической кластеризации масштабных коллекций научно-технических документов[1]

Д.А. Девяткин, Р.Е. Суворов, И.В. Соченков

Аннотация. В статье представлены результаты исследования в области методов тематической кластеризации научно-технических документов. Сформулированы требования к реализации методов кластеризации масштабных коллекций документов в поисково-аналитических системах. Предложен метод и разработан алгоритм тематической кластеризации масштабных коллекций научно-технических документов в поисково-аналитической системе. Выполнено экспериментальное сравнение результатов работы предложенного метода с несколькими классическими методами кластеризации текстов.

Ключевые слова: кластеризация, классификация, дескриптор, спектральный индекс, тематическая значимость.

Введение

По прогнозу компании EMC объем информации, опубликованной в Интернете к 2020 г., возрастет относительно объема 2011 г. в 50 раз [1]. Многократно увеличатся размеры массивов данных, с которыми работают информационные системы. Ручная обработка таких массивов невозможна, поэтому используются автоматизированные, либо автоматические системы обработки информации. Среди таких систем выделяется важный класс – поисково-аналитические системы, предназначенные для обработки информации на естественном языке. В настоящее время функции поисково-аналитических систем не ограничиваются простым поиском и организацией доступа к информации. Современная поисково-аналитическая система должна предоставлять пользователям инструменты для всестороннего анализа содержащихся в ней данных, отслеживания динамики целого ряда показателей, связанных с обрабатываемыми документами.

Одним из важных аспектов обработки данных в поисково-аналитической системе является тематический анализ. Система должна разделять коллекции документов по их тематической принадлежности. При этом в качестве коллекций могут выступать как целые массивы данных, загруженные в систему, например, из Интернета, так и некоторые подмножества этих коллекций, в частности, пользовательские подборки документов или их автоматические выборки за некоторый период времени. Тематический анализ в зависимости от потребностей пользователя или особенностей коллекции может выполняться на основе автоматической классификации в соответствии с априорно заданной таксономией, с помощью поиска тематически похожих документов и нечетких дубликатов, а также с применением методов кластеризации информации [2].

Настоящая работа посвящена разработке экспериментального метода распределенной кластеризации больших коллекций текстовых

[1] Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту № 14.514.11.4024 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

документов. В качестве документов выступают научно-технические публикации в электронном виде, доступные в открытых источниках Интернета. Задача кластеризации предполагает разбиение исходной коллекции документов на группы, в каждой из которых содержатся документы близкой тематики. При этом изначально тематическое разбиение не задано: алгоритм самостоятельно должен выделить тематические группы документов (кластеры) и охарактеризовать их некоторым образом, например, с помощью типичных документов-представителей или множеств ключевых слов и словосочетаний [3].

В разработанном методе для снижения вычислительной сложности используются инвертированные списки, обычно применяемые для хранения индексированных данных в системах обработки информации. Реализация метода кластеризации текстовой информации тесно интегрирована с алгоритмами и структурами данных в поисково-аналитической системе обработки научно-технической информации. Это одно из важных требований к разрабатываемому методу кластеризации наряду с высокими характеристиками качества: интерпретируемостью результатов и чувствительностью разбиения к параметрам настройки алгоритма.

1. Постановка задачи исследования

Сформулируем задачу кластеризации. Пусть c – конечное множество объектов, подлежащих кластеризации: $c = \{d_i\}$. $\rho: c \times c \to [0,\infty)$ – функция расстояния, на основе которой определяется сходство и различие между объектами. Необходимо найти множество кластеров K и алгоритм кластеризации $a: c \to K$, чтобы каждый кластер состоял из близких объектов, и объекты разных кластеров существенно различались.

В методах кластеризации текстов, в качестве исходных данных используют либо объектно-признаковую матрицу D, либо матрицу смежности текстов $M=\{m_{ij}|m_{ij}=\rho(d_i,d_j)\}$, рассчитанную на основе некоторой заданной функции различия или сходства ρ. Предметом настоящего исследования является метод кластеризации текстов на основе их тематического сходства.

На практике решение задачи кластеризации усложняется следующими факторами:

– число кластеров $|K|$, как правило, неизвестно заранее;

– разбиение на кластеры неоднозначно, оно зависит от выбранного метода кластеризации, пространства признаков и критериев определения различия и сходства между текстами, а также настроечными параметрами алгоритма.

– оценка качества кластеризации существенно зависит от субъективных представлений о сходстве и различии документов, т.е. от личных предпочтений эксперта;

– не существует единых критериев оценки качества кластеризации.

Существуют графовые (алгоритм определения связных компонент (клик) графа, Роккио (Rocchio), ФОРЭЛ), иерархические (агломеративная кластеризация), статистические методы (k-средних, EM-алгоритм) кластеризации [4]. Однако многие из указанных методов обладают недостатками, препятствующими их применению на практике в поисково-аналитических системах на больших объемах исходных данных:

1. Из-за вычислительной сложности порядка $O(|c|^2)$ от объема коллекции время работы многих алгоритмов на коллекциях в десятки тысяч документов достаточно велико, а коллекции, содержащие миллионы документов, не могут быть обработаны в принципе.

2. В качестве входных данных методы используют матрицы типа M или D, размеры которых для коллекций, состоящих из миллионов документов, могут значительно превосходить объемы доступной оперативной памяти современных компьютеров.

3. Большинство методов не предполагают эффективной реализации для вычислительных машин с параллельной архитектурой.

4. Большинство методов предполагает лишь такой характер разбиения $c \to K$, что каждый документ может относиться только к одному кластеру, однако в реальных массивах исходных данных часто встречаются документы, лежащие на стыке нескольких тематических групп (Рис. 1).

На основе вышесказанного, сформулируем основные требования, предъявляемые к методу тематической кластеризации текстов и его реализации в поисково-аналитической системе для предоставления функций тематического анализа:

Рис. 1. Нечеткое разбиение на кластеры

– линейная вычислительная сложность, пропорциональная объему коллекции;

– возможность эффективной параллельной реализации алгоритма;

– характеризация кластеров, полученных в результате работы метода, с помощью ключевой лексики, содержащейся в документах этого кластера;

– инкрементность: возможность пополнения коллекции и определения подходящих тематических групп для новых документов без повторения процедуры кластеризации пополненной коллекции целиком;

– интерпретируемость результатов кластеризации экспертом: соответствие ключевой лексики кластеров и документов из этого кластера, а также однородность тематики среди всех документов кластера, определяемая на основе экспертных оценок.

Целью настоящего исследования является разработка метода тематической кластеризации, удовлетворяющего вышеприведенным требованиям.

2. Метод тематической кластеризации

В основе разработанного метода тематической кластеризации лежит векторное представление текста документа. В качестве пространства признаков выступает множество лексических дескрипторов (ЛД) – отдельных лексем или словосочетаний в канонической форме (главное слово приведено к словарной форме, а форма зависимых слов подчинена управлению главного слова), характеризующие лексический состав текста [5,6]. В качестве значений признаков рассматривается информационная значимость ЛД в текстах документов, определяемая на основе ряда вероятностных критериев [7,8]. Для определения информационной значимости используется подход, предложенный в работах [9, 10] и примененный авторами для решения близкой задачи тематической классификации.

Рассмотрим следующие величины:

1. Инверсная частота ЛД w в произвольном множестве текстов τ:

$$IDF(w,\tau) = \log_2 \frac{|\tau|}{m(w,\tau)}, \qquad (1)$$

где $m(w,\tau)$ – число документов в τ, содержащих ЛД w.

2. Вес ЛД в тексте документа d [11]:

$$lTF(w,d) = \log_{S(d)}\left(C(w,d)+1\right), \qquad (2)$$

где $C(w,d)$ – количество вхождений ЛД w в текст $T(d)$ документа d, $S(d)$ – общее количество вхождений всех ЛД в текст документа d.

3. Изменение информативности ЛД w документа при отнесении его к некоторому тематическому подмножеству документов σ некоторой (общетематической) коллекции c:

$$\Delta\tilde{I}(w,c,\sigma) = IDF_N(w, c\setminus\sigma) - IDF_N(w,\sigma), \qquad (3)$$

$$\Delta\tilde{I}^{+}(w,c,\sigma) = \Delta\tilde{I}(w,c,\sigma)\cdot X(\Delta\tilde{I}(w,c,\sigma)), \qquad (4)$$

где $X(z)$ – функция Хевисайда.

4. Информационная значимость ЛД w текста документа d в коллекции c определяется формулой:

$$v(w,d,c) = lTF(w,d)IDF(w,c), \qquad (5)$$

представляющей собой модификацию известной формулы TF-IDF.

5. Для ЛД w документа d, относящегося к тематическому подмножеству σ коллекции c, определяется характеристикой тематической значимости (ХТЗ) [9]:

$$\tilde{v}(w,d,c,\sigma) = lTF(w,d)\Delta\tilde{I}^{+}(w,c,\sigma). \qquad (6)$$

Каждому документу d соответствует вектор значений признаков

$V(d,c)=(v(w_1,d,c),v(w_2,d,c),...,v(w_n,d,c))$, определяющий информационную значимость ЛД w_i в тексте этого документа.

На практике рассматриваются не все ЛД, а только наиболее значимые: $\hat{W}(d)=\{w \mid v(w,d,c)>\lambda\}$, где λ – параметр метода. В другом варианте можно рассматривать θ наиболее значимых ЛД, полагая веса остальных ЛД в документе равными 0 и исключая их, таким образом, из рассмотрения. Значение θ выбирается как параметр алгоритма, либо вычисляется для каждого документа как некоторый процент от общего числа ЛД в нем.

Для решения задачи кластеризации используется матрица тематического сходства текстов $M=\{m_{ij} \mid m_{ij}=SIM(d_i,d_j)\}$, рассчитанная на основе заданной функции различия ρ:

$$SIM(d_i,d_j)=1-\rho(d_i,d_j).$$

При построении матрицы M для оценки тематического сходства между документами d_i и d_j используется расстояние Хэмминга:

$$SIM_{Ham}(d_i,d_j)= 1-\frac{\sum_{w\in\hat{W}(d_i)\cap\hat{W}(d_j)}\left|v(w,d_i,c)-v(w,d_j,c)\right|}{\sum_{w\in\hat{W}(d_i)\cap\hat{W}(d_j)}\left(v(w,d_i,c)+v(w,d_j,c)\right)}, \quad (7)$$

где d_i, d_j – документы, тематическое сходство которых оценивается; c – коллекция текстов, к которой относятся документы d_i, d_j; $\hat{W}(d_i)$, $\hat{W}(d_j)$ – множества значимых ЛД документов d_i и d_j, соответственно.

Заметим, что при использовании вышеприведенной формулы матрица M является симметричной.

На основе построенной матрицы M производится формирование множества «ядер» кластеров K – наборов документов, чья мера близости $SIM(d_i,d_j)$ превышает некоторый порог SIM_1. Этот шаг предполагает построчный просмотр матрицы и выборку наиболее тематически сходных документов. При этом если d_i относится к ядру некоторого кластера $k\in K$, и для некоторого d_j выполнено $SIM(d_i,d_j)>SIM_1$, то d_j также будет отнесен к ядру кластера k. В противном случае ($SIM(d_i,d_j)<SIM_1$) d_j образует ядро нового кластера.

В ходе обработки матрицы M указанным способом будет получено множество ядер кластеров K. Ядра, содержащие менее N элементов, расформировываются, а отнесенные к ним документы подвергаются процедуре классификации [3] по оставшимся кластерам $K'=\{k\in K \mid |k|>N\}$. При этом сначала для множества K' выполняется процедура слияния тематически близких ядер кластеров. Для этого на основе ЛД документов, входящих в ядро кластера $k\in K'$, строится множество значимых ЛД этого кластера:

$$W(\sigma,c)=\left\{w \mid \Delta\tilde{I}^{+}(w,c,\sigma)>\Delta\tilde{I}^{+}_{MIN}\right\}$$

и соответствующий ему вектор значений признаков

$$J(\sigma,c)= $$

$$=\left(\Delta\tilde{I}^{+}(w_1,c,\sigma),\Delta\tilde{I}^{+}(w_2,c,\sigma),...,\Delta\tilde{I}^{+}(w_{|W(\sigma,c)|},c,\sigma)\right),$$

представляющий собой дескриптор тематического кластера. Затем на основе расстояния Хэмминга выполняется сопоставление вектора $J(\sigma_k,c)$ каждого ядра кластера с остальными:

$$DIST_{Ham}(\sigma_k,\sigma_l)= \frac{\sum_{w\in W(\sigma_k,c)\cup W(\sigma_l,c)}\left|\Delta\tilde{I}^{+}(w,c,\sigma_k)-\Delta\tilde{I}^{+}(w,c,\sigma_l)\right|}{\sum_{w\in W(\sigma_k,c)\cup W(\sigma_l,c)}\left(\Delta\tilde{I}^{+}(w,c,\sigma_k)+\Delta\tilde{I}^{+}(w,c,\sigma_l)\right)}. \quad (8)$$

В качестве альтернативы для оценки тематического сходства двух кластеров может использоваться косинусообразное расстояние.

Процедура слияния тематически близких ядер выполняется методом сравнения всех кластеров со всеми остальными. В результате формируется множество кластеров K''.

Для классификации документов, не вошедших в ядра кластеров, используются дескрипторы кластеров $J(\sigma,c)$. Процедура классификации состоит в оценке тематической близости «класс» - документ и применении решающего

правила в соответствии с подходом, предложенным в [9]. Для оценки тематической близости «класс» - документ применяется модифицированная ХТЗ, вычисляемая по формуле:

$$RIC(\sigma_k, d, c) = \frac{\sum_{w \in W(\sigma_k, c) \cap W(d)} \tilde{v}(w, d, c, \sigma_k)}{\sum_{w \in W(\sigma_k, c)} \Delta\tilde{I}^{+}(w, c, \sigma_k)}, \quad (9)$$

где $W(\sigma_k,c)$ – множество значимых ЛД тематического подмножества документов σ_k (соответствующего ядру кластера $k \in K''$); $W(d)$ – множество ЛД документа (не обязательно значимых).

Величина, заданная формулой (9), неотрицательна и нормирована на единицу. Она оценивает величину совокупной информационной значимости ЛД в тексте документа d относительно совокупной информационной значимости ЛД тематического подмножества документов σ_k. Заметим, что в отличие от работы [9], в которой для каждого классифицируемого документа выбирается подходящий тематический класс, решается противоположная задача: для заданного тематического подмножества выбираются соответствующие ему документы. В качестве критерия соответствия используется пороговое решающее правило: документ d относится к кластеру k, тогда и только тогда, когда $RIC(\sigma_k,d,c) > RIC_{MIN}$, где RIC_{MIN} – пороговое значение (параметр метода). Такой критерий обеспечивает «размытость» кластеризации: классифицированные на этом этапе документы могут быть отнесены к нескольким группам одновременно. Если подобный эффект является нежелательным, то решающее правило заменяется правилом следующего вида: документ d относится к кластеру k, тогда и только тогда, когда $RIC(\sigma_k, d, c) = max_{i=1}^{|K''|}\{RIC(\sigma_i, d, c)\}$, что обеспечивает единственность соотнесения с тематической группой.

Заметим, что при реализации метода тематической кластеризации на практике нет необходимости в полном построении и хранении матрицы M оценок тематического сходства, поскольку формирование ядер кластеров выполняется за один проход по строкам матрицы. Поэтому документы могут обрабатываться параллельно и независимо друг от друга: строки матрицы (списки тематически похожих документов для заданного эталона) вычисляются по мере необходимости. После обработки строка матрицы может быть удалена из памяти, что снижает требования к ее объемам, необходимой для решения задачи. В разработанном методе кластеризации изначально было решено не использовать матрицу «документы–признаки», поскольку применение этой матрицы потребует больших затрат памяти для ее хранения. Например, считая, что для каждого документа в среднем выбирается 50 значимых ЛД, для коллекции в 1 млн документов потребуется построить и затем обработать матрицу в 50 млн элементов, что является ресурсо-затратным и вычислительно трудоемким, а также создает трудности при параллельной реализации метода

Изложенный подход позволяет эффективно решать задачу кластеризации и «дополнительной классификации» с применением инвертированных индексов, как это будет показано в следующем разделе статьи.

3. Алгоритм тематической кластеризации

Опишем алгоритм тематической кластеризации крупных коллекций научно-технических документов в распределенной вычислительной среде. Приводимый далее алгоритм и структуры данных интегрированы в поисково-аналитическую систему. В этой системе документы помещаются в пополняемые коллекции.

На предварительном этапе из электронного представления документов выделяется текст, который подвергается лингвистическому анализу: морфологическому (нормализация и определение морфологических признаков словоупотреблений) и синтаксическому (для выделения составных ЛД - словосочетаний).

Затем для каждого документа строится спектральный индекс, соответствующий вектору значений признаков $V(d,c)$, элементами которого являются пары *<идентификатор ЛД, вес ЛД в документе>*. Спектральный индекс документа помещается в базу данных (в качестве ключа выступает идентификатор документа). Спектральный индекс также инвертируется и помещается в инвертированный спектральный индекс коллекции, ключом в котором выступают идентифика-

торы ЛД, а значениями – списки пар < *идентификатор документа, вес ЛД в документе* >. Подобный подход нашел свое применение в области информационного поиска и впоследствии был адаптирован для задач поиска нечетких дубликатов документов [11,12].

Необходимая статистическая информация о частотах встречаемости ЛД коллекции накапливается в счетчиках частот встречаемости.

Компонент системы, осуществляющий построение и хранение вышеприведенных структур данных, назовем индексатором. В распределенной поисково-аналитической системе существует, как правило, более одного индексатора, и они функционируют на разных серверах, взаимодействуя удаленно. При этом каждый из них хранит некоторое подмножество коллекции документов.

При накоплении достаточного количества документов в системе или по требованию запускается процедура тематической кластеризации. При этом выполняются следующие действия:

1. Модуль кластеризации запрашивает спектральный индекс документа из индексатора и пересылает его остальным индексаторам с запросом на поиск тематически похожих документов.

2. Индексаторы осуществляют поиск в инвертированном спектральном индексе, ранжируют документы по убыванию тематического сходства и отсекают документы, для которых $SIM(d_i,d_j)<SIM_1$.

3. Модуль кластеризации агрегирует полученные от индексаторов списки в один (этот список соответствует строке матрицы M). Затем обрабатывает этот список для построения ядер кластеров. Затем модуль кластеризации очищает полученный спектральный индекс и список похожих документов.

Шаги 1–3 выполняются для всех документов во всех индексаторах. Обработка может быть распараллелена по документам естественным образом.

В модуле кластеризации для каждого формируемого ядра кластера ведется список идентификаторов документов, отнесенных к этому ядру. Для каждого обработанного документа номер ядра соотнесенного кластера заносится в хэш-таблицу, ключом которой является идентификатор документа.

После построения ядер кластеров модуль кластеризации выполняет их фильтрацию (построение K' из K) и слияние близких ядер в соответствии с критерием, заданным формулой (8), – формирование K''.

Для построения дескрипторов тематических кластеров у индексаторов запрашиваются спектральные индексы документов, вошедших в ядра кластеров. На следующем этапе построенные дескрипторы рассылаются индексаторам в качестве запроса на поиск документов, близких к выделенным кластерам. На заключительном этапе полученные результаты агрегируются и сохраняются.

Предложенный алгоритм для краткости назовем алгоритмом «тематической кластеризации - классификации» (Topic Clustering – Classification (TCC)).

В соответствии с приведенным алгоритмом могут быть кластеризованы произвольные подмножества коллекций документов, например, опубликованные в определенный период, полученные из одного источника и т.п. Алгоритм TCC эффективен на множествах документов размером до 1 млн. Он обеспечивает полноту исходных данных (эффективное сопоставление «всех со всеми» за счет применения распределенных инвертированных спектральных индексов), но вместе с тем достаточно сильно нагружает каналы передачи данных. В силу этого при необходимости кластеризации коллекций, содержащих более 1 млн документов, целесообразно равномерно (случайным образом) распределить документы по индексаторам (1-3 млн на каждый индексатор), а затем выполнить вышеописанный алгоритм локально на данных каждого индексатора без сетевой пересылки. В силу равномерности распределения документов статистические характеристики частот ЛД на каждом индексаторе будут примерно одинаковы (следствие из законов Ципфа-Мандельброта)[7]. Это обеспечит сходное разбиение на кластеры для каждого индексатора. Впоследствии дескрипторы кластеров, полученные на разных индексаторах, могут быть объединены методом, с помощью которого выполняется слияние ядер кластеров.

4. Оценка вычислительной сложности алгоритма

Оценим вычислительную сложность разработанного алгоритма. Пусть c – исходный массив документов, D_{avg} – средний размер спектрального индекса документа, D_W – количество ЛД во множествах значимых ЛД $\hat{W}(d)$ документов в среднем по коллекции.

Вычислительная сложность построения инкрементальных инвертированных индексов составляет $O(|c|D_{avg}\ log(D_{avg}))=O(|c|)$, где $D_{avg}log(D_{avg})$ – константа: сложность сортировки инвертированного индекса одного документа. Сложность помещения инвертированного индекса одного документа в инвертированный индекс коллекции линейна: $O(D_{avg})$ за счет применения хэш-таблиц и методов сбалансированного многопутевого слияния [13].

Вычислительная сложность получения одного спектрального индекса – в среднем константа (за счет использования хэш-таблиц) $O(1)$.

Количество уникальных ЛД, содержащихся во всех текстах коллекции, определяет размер U словаря хэш-таблицы, поиск в котором для каждого документа занимает константное время $O(D_W)$. При поиске производится слияние списков элементов, соответствующих найденным ЛД, и выполняется однопроходный расчет оценки тематического сходства. При этом длины этих списков линейно зависят от размера коллекции $|c|$. Верны следующие утверждения:

1. Размер словаря U много больше D_{avg} (в среднем, на 2-3 порядка) [7]. Это означает, что для каждого ЛД соответствующий список элементов в инвертированном индексе растет на 2-3 порядка медленнее, чем размер коллекции.

2. Во множестве $\hat{W}(d)$ содержатся информационно-значимые ЛД, редко встречающиеся в документах коллекции, что еще больше замедляет темпы роста списков в инвертированном индексе, участвующих в оценке тематического сходства.

3. Количество документов, которое может быть обработано одним индексатором, ограничено объемами доступной памяти. По соображениям ускорения процесса индексации алгоритм реализуется в распределенном варианте, а значит, в каждый индексатор попадает ограниченное количество документов.

В силу этого сложность поиска похожих документов для одного эталонного в инвертированном индексе имеет оценку $O(1)$. Совокупная сложность операции поиска в инвертированном спектральном индексе для всех документов коллекции составляет $O(|c|)$, что соответствует оценке сложности формирования ядер кластеров.

Оценка сложности формирования дескрипторов кластеров:

$$O\left(\sum_{i=1}^{|K'|}|\sigma|D_{avg}\right)\le O\left(|c|\right).$$

Процедура слияния тематически близких ядер кластеров требует сравнения всех кластеров со всеми остальными, что соответствует оценке $O(|K'|^2)$. Однако параметры кластеризации подбираются таким образом, чтобы количество кластеров $|K'|$ не было слишком большим даже в масштабных коллекциях (поскольку получившееся разбиение должно быть обозримо экспертом, так как иначе задача теряет свой смысл). В силу этого на практике процедура слияния тематически близких ядер кластеров не требует больших вычислительных затрат и явно не зависит от размера коллекции.

Таким образом, итоговая вычислительная оценка сложности алгоритма составляет $O(|c|)$, что соответствует оценкам наилучших алгоритмов в этой области [14].

5. Методы оценки качества работы алгоритма кластеризации

На заключительном этапе работы алгоритма кластеризации производится автоматическая проверка полученного разбиения на кластеры на соответствие следующим критериям.

Требование близости объектов одного кластера – есть сумма средних внутрикластерных расстояний, которая характеризует степень связности элементов кластера [14]:

$$\Phi_0=\sum_{k=1}^{|K'|}\frac{1}{|\sigma_k|}\sum_{i=1}^{|\sigma_k|}\sum_{j=i}^{|\sigma_k|}\rho(d_i,d_j)$$

.

Требование тематического различия отдельных кластеров – есть сумма межкластерных

расстояний, которая характеризует меру сцепления отдельных кластеров друг с другом:

$$\Phi_1 = \sum_{k=1}^{|K''|} \frac{1}{|\sigma_k|} \sum_{i=1}^{|\sigma_k|} \sum_{d \in S_k} \rho(d_i, d),$$

где $S_k = c \setminus \sigma_k$.

Отношение пары функционалов $O_q = \frac{\Phi_1}{\Phi_0} \to \max$ – критерий оценки результатов работы алгоритма кластеризации [15]. Этот критерий позволяет оценить качество кластеризации при фиксированном способе расчета ρ.

Следующие критерии инвариантны относительно выбора ρ. Средний размер кластера вычисляется как среднее арифметическое:

$$\omega = \frac{\sum_{k=1}^{|K''|} |\sigma_k|}{|K''|}.$$

Максимальный размер кластера $\xi_{\max} = \max(|\sigma_k|, K'')$. Отношение пары функционалов $O_d = \omega / \xi_{\max} \to \max$ – критерий равномерности разбиения. Количество документов, вошедших в кластеры:

$$C_{xk} = |\{d_i \in \sigma_j \mid k_j \in K''\}|.$$

Отношение $O_c = C_{xk} / |c| \to \max$ – критерий покрытия исходного набора документов алгоритмом кластеризации.

Рассмотренные критерии эффективности методов кластеризации имеют относительный характер. Т.е. они имею смысл лишь при сравнении различных методов кластеризации на одном и том же массиве исходных документов.

Введем интегральную оценку качества кластеризации, как произведение набора оптимальных по Парето значений величин O_q, O_d и O_c:

$$O_i = O_q O_d O_c \to \max.$$

6. Экспериментальное исследование работы алгоритма тематической кластеризации

Для проведения экспериментов были проиндексированы две коллекции документов: «Авторефераты по с/х наукам» (AGRO) и «Новости» (NEWS). В коллекцию «Авторефераты по с/х наукам» вошли авторефераты диссертационных работ на русском языке по сельскохозяйственным наукам за период 1960-2000 гг. В коллекцию «Новости» вошли материалы российских информационных агентств по теме «выборы», опубликованные в период с 06.2012 по 01.2013 гг. Целью кластеризации первой коллекции являлась группировка документов по проблематикам исследований внутри предметной области. Целью кластеризации второй коллекции являлось выделение основных тем.

Для сравнительного исследования эффективности предлагаемого метода были проведены эксперименты по кластеризации разработанным методом ТСС, а также методами выделения клик и Роккио [3, 4]. Результаты сравнения методов кластеризации приведены в Табл. 1.

На основе полученных данных можно сделать вывод о том, что предложенный метод показывает лучшие результаты по критерию O_i на больших (св. 20 тыс. документов) коллекциях, главным образом, благодаря более равномерному разбиению исходных документов на кластеры и лучшему покрытию.

Разработанный метод позволяет характеризовать отдельные кластеры при помощи дескрипторов. Примеры дескрипторов кластеров, полученных для коллекций AGRO и NEWS, приведены в Табл. 2.

Заключение

Разработанный метод распределенной кластеризации масштабных коллекций текстовых документов позволяет эффективно организовать кластеризацию больших, динамически изменяющихся коллекций текстовых документов в распределенных системах. Предложенный алгоритм обладает линейной сложностью пропорционально размеру коллекции. Метод позволяет пополнять коллекции новыми документами без повторения полной процедуры кластеризации.

Кластеры, полученные в результате работы метода, могут быть представлены в удобной для экспертной оценки форме – с помощью ключевой лексики.

Предметом исследований авторов, проводимых в настоящее время, является разработка

Табл.1. Результаты экспериментов

Коллекция	AGRO			NEWS		
Метод	Выделение клик	Роккио	TCC	Выделение клик	Роккио	TCC
Тип исходных данных	Матрица смежности	Матрица смежности	Списки схожих документов	Матрица смежности	Матрица смежности	Списки схожих документов
Минимальный размер ядра кластера	15	10	10	6	6	16
Кол-во кластеров	186	197	169	49	50	10
Документов вошло в кластеры	5439	18399	18117	595	737	849
Всего документов	24624			1555		
Максимальный размер кластера	154	1019	925	37	70	212
Средний размер кластера	30	404	411	13	17	86
Связность	0,15	0,0579	0,0589	0,44	0,382	0,331
Сцепление	0,05	0,017	0,018	0,3	0,274	0.294
Интегральная оценка	0,134	0,994	1,019	0,199	0,163	0,249
Интегральная оценка (%)	13%	97%	100%	80%	65%	100%

Табл.2. Примеры дескрипторов кластеров

AGRO/1 кластер		NEWS/1 кластер	
Текст	**Вес**	**Текст**	**Вес**
Дефицит фосфора	0.56	Литовское государство	0.87
Аспект качества	0.52	Реабилитировать	0.87
Кумулятивный эффект	0.51	Массовый иск	0.87
Фосфатный уровень	0.50	Международная комиссия	0.87
Количественная закономерность	0.50	Оценка преступления	0.77
Фосфат почвы	0.49	Советский режим	0.77
Новая единица	0.48	Нацистский	0.72
Последействие дозы	0.48	Советская оккупация	0.68
Удвоить дозу	0.47	Привлечение	0.65

методики экспертной оценки предложенного метода кластеризации и получение экспертных оценок качества кластеризации.

Программная реализация метода была применена для формирования тематических коллекций научных публикаций с целью решения аналитических задач.

Литература

1. EMC [Электронный ресурс] URL: http://www.emc.com/leadership/programs/digital-universe.htm, общий доступ. [Проверено: 12.02.2013].
2. Tryon R.C. Cluster analysis. — London: Ann Arbor Edwards Bros, 1939. — 139 p.
3. Salton G. Dynamic information and library processing. – N.J.: Prentice Hall, 1975. – 523 p.
4. Jain A., Murty M., Flynn P. Data Clustering: A Review// ACM Computing Surveys. – 1999. – vol. 31, no 3. – p. 564 – 323.
5. Вейзе А.А. О ядерных текстах и их получении путем компрессии // Проблемы текстуальной лингвистики /Под. ред. проф. В.А. Бухбиндера. — Киев, 1983.
6. Лотман Ю.М. Структура художественного текста. — М., 1970; Общение. Текст. Высказывание. — М., 1989. – 285с.
7. Пиотровский Р.Г. Текст, машина, человек. - Л.: Наука, 1975. - 327 с.
8. Севбо И. П. Структура связного текста и автоматизация реферирования. — М., 1969. – 132 с.
9. Э. Мбайкоджи Э., Драль А., Соченков И.. Метод автоматической классификации коротких текстовых сообщений // Информационные технологии и вычислительные системы. – 2012. – №3. – с.93 – 102.
10. Тихомиров И.А., Соченков И.В. Метод динамической контентной фильтрации сетевого трафика на основе

анализа текстов на естественном языке // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – Новосибирск: 2008. – Т. 6. № 2. – с. 94-100.
11. Frakes, William B., Baeza-Yates, Ricardo, Information Retrieval: Data Structures and Algorithms. - Englewood Cliffs, New Jersey: Prentice-Hall, 1992. – 504 p.
12. Агеев М.С., Добров Б. В. Метод эффективного расчёта матрицы ближайших соседей для полнотекстовых документов // Вестник Санкт-Петербургского университета / СПбГУ. – СПб.: 2011. – Сер. 10, Вып. 3. – с. 72-84.
13. Седжвик Р. Фундаментальные алгоритмы на C++. Анализ/Структуры данны/Сортировка/Поиск: Пер. с англ./ Роберт Сэджвик. – К.: ДиаСофт, 2001. – 688 с.
14. Dobrynin V., Patterson D., Rooney N. Contextual document clustering // In Proceedings of the 26th European Conference on Information Retrieval Research, LNCS 2997. – Berlin.: Heidelberg:Springer, 2004. – p. 167–180.
15. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. М.: Финансы и статистика, 2003. – 353 с.

Девяткин Дмитрий Александрович. Инженер-исследователь ООО "Технологии системного анализа", аспирант ИСА РАН. Окончил Рыбинский государственный авиационный технический университет им. П.А. Соловьёва в 2011 г. Автор трех научных работ. Область научных интересов: тематическая кластеризация текстовых документов, анализ эмоциональной окраски текста, дистанционное обучение. E-mail: ddeviatkin@gmail.com

Суворов Роман Евгеньевич. Инженер-исследователь ООО "Технологии системного анализа", аспирант ИСА РАН. Окончил Рыбинский государственный авиационный технический университет им. П.А. Соловьева в 2012 г. Автор двух научных работ. Область научных интересов: тематическая классификация текстовых документов, контентная фильтрация, интеллектуальные динамические системы. E-mail: resuvorov@gmail.com

Соченков Илья Владимирович. Научный сотрудник ООО "Технологии системного анализа", инженер-исследователь ИСА РАН. Окончил Российский университет дружбы народов в 2009 г. Автор 25 научных работ. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, защита сетей, контентная фильтрация, компьютерная лингвистика. E-mail: ivsochenkov@gmail.com