УДК 340.11(3)

В. Б. Барахнин [1], В. А. Нехаева [2], А. М. Федотов [3]

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия
E-mail: [1] bar@ict.nsc.ru; [2] nekhaeva@ngs.ru; [3] fedotov@nsu.ru

О ЗАДАНИИ МЕРЫ СХОДСТВА ДЛЯ КЛАСТЕРИЗАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ *

В работе решается задача автоматизации процесса отбора текстовых документов научной тематики, которые могут представлять интерес для конкретного ученого-исследователя или группы совместно работающих исследователей. В качестве шкал для определения меры предлагается брать атрибуты библиографического описания документов (авторы, ключевые слова, аннотация). Значения весовых коэффициентов в формуле для вычисления меры сходства определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы.

В качестве потенциально пригодных для решения поставленной задачи были проанализированы три классических метода кластеризации документов: кластеризация путем нахождения клик в полной матрице подобия документов, кластеризация по методу Роккио и метод, базирующийся на так называемом жадном алгоритме, а также новый алгоритм Н. Загоруйко, основанный на использовании функции конкурентного сходства (так называемой FRiS-функции). В ходе тестирования было выявлено, что оптимальным для данной задачи является FRiS-алгоритм, хотя приемлемые результаты дает и жадный алгоритм.

Ключевые слова: мера сходства, кластеризация текстовых документов.

Введение

Ежегодно в мире публикуются миллионы научных статей. Даже в узкоспециализированных отраслях науки просматривать весь объем информации практически невозможно. Вследствие этого широкое распространение получили электронные носители информации о новых научных публикациях, в частности:

- базы данных Реферативных журналов;
- базы данных «Current Contents»;
- специализированные сетевые базы данных типа Zentralblatt MATH.

Помимо этого у каждого исследователя за годы его работы образуется картотека библиографических описаний статей, книг и т. д., представляющих для него интерес. Основной критерий их отбора – личные интересы ученого. В настоящее время такие картотеки хранятся, как правило, на электронных носителях. Это позволяет организовывать интегрированные картотеки путем объединения ресурсов совместно работающих исследователей.

Таким образом, возникает задача автоматизации процесса отбора публикаций из электронных баз данных, которые могут представлять интерес для конкретного исследователя или группы совместно работающих исследователей. Для нахождения нужной статьи исследователь обращается либо к реферативным журналам, либо к их электронным аналогам. Так как существуют достаточно эффективные алгоритмы поиска конкретной публикации на электронных носителях, то наиболее актуальной среди проблем информационного поиска на

* Работа выполнена при частичной финансовой поддержке РФФИ (проекты № 06-07-89060, 06-07-89038, 07-07-00271), президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

данный момент является задача нахождения по данному документу класса схожих по содержанию документов.

В данной работе решается задача автоматизации процесса отбора публикаций из библиографических баз данных, которые могут представлять интерес для конкретного исследователя или группы совместно работающих исследователей. С этой целью на первом этапе работы исследуются алгоритмы измерения меры сходства между двумя документами электронной базы данных, а также алгоритмы кластеризации документов, составляющих эту базу. В качестве шкал для определения меры предлагается брать атрибуты библиографического описания документов.

Под кластеризацией мы, согласно [Солтон, 1979], понимаем процесс разбиения множества документов электронной базы на классы, при котором элементы, объединяемые в один класс, имеют большее сходство, нежели элементы, принадлежащие разным классам (в нашем случае сравнение документов при формировании кластеров ведется по атрибутам их библиографического описания). Каждый кластер при этом обычно описывается с помощью одного или нескольких идентификаторов, называемых профилем или центроидом. Профиль кластера может быть представлен некоторым формальным объектом, расположенным в центре кластера, или любым представительным объектом, способным характеризовать остальные объекты этого кластера (подробнее о различных методах определения центроидов будет рассказано ниже). С использованием понятия центроида можно искать похожие документы, сравнивая поисковые запросы сначала с профилями кластеров, а затем, проверяя записи, входящие в кластеры, имеющие очень близкие профили.

Задание меры сходства на множестве документов

Как уже упоминалось, в качестве шкал для определения меры сходства между двумя документами мы используем атрибуты библиографического описания данных документов. Перечислим основные элементы библиографического описания заносимых в картотеку документов: авторы; заглавие; название журнала или издательства; год выхода; том, номер, страницы (для публикаций в периодических изданиях); аннотация; коды классификатора; ключевые слова.

Изложим алгоритм определения меры сходства нового документа с документами из имеющегося набора (т. е. личной библиографической базы). Количественная характеристика меры сходства определяется на множестве документов D следующим образом:

$$m: D \times D \rightarrow [0, 1],$$

причем функция m в случае полного сходства принимает значение 1, в случае полного различия – 0. Вычисление меры сходства осуществляется по формуле вида

$$m(d_1, d_2) = \sum a_i \, m_i(d_1, d_2), \qquad (1)$$

где i – номер элемента (атрибута) библиографического описания, a_i – весовые коэффициенты, причем $\sum a_i = 1$ (см., например, [Воронин, 1991]), $m_i(d_1, d_2)$ – мера сходства по i-му элементу (иными словами, по i-й шкале). Поскольку в описываемой ситуации практически все шкалы – номинальные, то мера сходства по i-й шкале определяется следующим образом: если значения i-х атрибутов документов совпадают, то мера близости равна 1, иначе 0. При этом необходимо учитывать, что значения атрибутов могут быть составными. В таком случае $m_i = n_{i1}/n_{i0}$, где $n_{i0} = \max\{n_{i0}(d_1), n_{i0}(d_2)\}$, а $n_{i0}(d_j)$ – общее количество элементов, составляющих значение i-го атрибута документа d_j, n_{i1} – количество совпадающих элементов.

Заметим, что изложенный алгоритм измерения меры сходства, может быть положен в основу некоторой экспертной системы, обладающей определенными продукционными правилами. Так, значения весовых коэффициентов a_i в формуле (1) может определяться предполагаемой апостериорной достоверностью данных соответствующей шкалы. Например, полное (или даже «почти полное») совпадение значений атрибута «авторы» документов d_1 и d_2 более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одного автора). В такой ситуации мы можем увеличивать значение соответствующего весового коэффициента в формуле (1) с одновременным пропорциональным уменьшением других коэффициентов.

Методы кластеризации документов

Основная проблема кластеризации документов заключается в таком разнесении документов по группам, при котором элементы каждой группы были бы настолько сходны друг с другом, чтобы в некоторых случаях можно было пренебречь их индивидуальными особенностями. В частности, производить поиск в систематизированном файле гораздо легче, чем в несистематизированном, ибо группы документов, профили которых не имеют сходства с поисковым предписанием, не включаются в углубленный процесс поиска. При кластеризации документов важно прийти к разумному компромиссу относительно размера кластеров, избегая как формирования большого числа очень мелких кластеров (что снижает эффективность кластеризации как выделения множеств схожих документов), так и небольшого количества очень крупных классов (что может вызвать уменьшение точности поиска).

Принято различать ряд задач классификации: формирование кластеров на основе сведений (свойств и характеристик) о классифицируемых объектах; отнесение объектов к сформированным кластерам или кластерам, находящимся в процессе формирования; извлечение информации, необходимой для идентификации и описания классов документов. Собственно формирование классов выполняется обычно на основе сопоставления векторов документов, причем класс определяется как множество всех объектов, имеющих достаточно высокие значения коэффициента подобия. Составление характеристик класса эквивалентно построению профиля; отнесение объектов к классам зависит от степени подобия между идентификаторами объектов и профилями классов.

В настоящей работе мы исследуем методы кластеризации, использующие в качестве критерия для сравнения только заранее заданные элементы библиографического описания документов, не учитывая индивидуальные поисковые возможности данных документов и мнения потребителей об их полезности.

В качестве потенциально пригодных для решения поставленной задачи были проанализированы три классических метода кластеризации документов: кластеризация путем нахождения клик в полной матрице подобия документов [Солтон, 1979], кластеризация по методу Роккио [Там же] и метод, базирующийся на так называемом жадном алгоритме [Кормен и др., 2001], а также новый алгоритм, основанный на использовании функции конкурентного сходства (FRiS-функции) [Борисова, 2007]. Кратко изложим суть перечисленных алгоритмов.

Процесс нахождения клик основан на построении полной матрицы подобия, посредством которой каждой паре документов (d_1, d_2) ставится в соответствие коэффициент подобия S (d_1, d_2). Обычно выбирается пороговое значение T, и матрица подобия приводится к бинарному виду путем замены всех коэффициентов подобия таких, что S (d_1, d_2) $\geq T$, единицей, а всех остальных – нулем. Далее искомые классы определяются как клики, которые могут быть получены из бинарного ряда подобия.

В алгоритме Роккио построение матрицы подобия заменяется проверкой плотности пространства некоторых документов. В качестве возможных центров кластеров выступают только те документы, которые по результатам вычислений оказались расположенными в плотных зонах пространства. Кластеризуемый документ относят к тому классу, подобие с центроидом которого оказалось наиболее высоким.

При использовании жадного алгоритма в матрице подобия находят строку (или столбец – матрицам симметрична), сумма компонент которой будет максимальной. Документ, соответствующий этой строке, объявляют центром первого кластера и включают в кластер все документы, коэффициенты подобия к которым больше либо равно некоторого наперед заданного порогового значения. Далее выбрасывают все попавшие в кластер документы, вычеркивая из матрицы соответствующие строки и столбцы, после чего процесс повторятся несколько раз, пока все документы не будут кластеризованы.

В методе кластеризации с использованием функции конкурентного сходства при определении меры сходства между двумя документами рассматривается конкурентная ситуация: решение о принадлежности документа d к первому кластеру принимается не в том случае, когда расстояние r_1 до этого кластера «мало», а когда оно меньше расстояния r_2 до конкурирующего кластера. Для вычисления меры конкурентного сходства, измеренной в абсолютной шкале, используется нормированная величина $F_{12} = (r_2 - r_1)/(r_2 + r_1)$, называемая функцией

конкурентного сходства или FRiS-функцией (от Function of Rival Similarity). Разумеется, на первоначальном этапе кластеризации, когда конкурирующих кластеров еще нет, приходится работать с некоторой модификацией (редукцией) FRiS-функции, использующей виртуальный кластер-конкурент. Суть алгоритма состоит в том, что с использованием редуцированной FRiS-функции в качестве центроидов выбираются центры локальных «сгустков» распределения документов, после чего формируются линейно разделимые кластеры.

Выбор оптимального алгоритма

В качестве практической цели применения проанализированных алгоритмов стояла задача автоматизации процесса отбора публикаций из электронных баз данных, которые могут представлять интерес для конкретного исследователя или группы совместно работающих исследователей.

Тестирование алгоритмов проводилась на электронной базе данных «Сибирского математического журнала», содержащей библиографические описания статей журнала, вышедших в период с 2000 по 2005 г. Статьям в указанной базе данных кроме стандартных атрибутов (название, автор, год издания и т. п.) приписаны соответствующие коды классификатора из «Классификации математических сущностей» (MSC-2000). Это факт позволил разбить всю работу на два этапа.

1. Нахождение оптимального алгоритма кластеризации. В качестве меры на пространстве документов используется определенная ранее конструкция, однако сравнение ведется по одному-единственному атрибуту – кодам классификатора. Поскольку совпадение данных кодов для группы документов является объективным критерием совпадения тематики данных документов, то такую меру можно считать идеальной.

2. Задание меры на множестве документов, которая после кластеризации базы даст результат, близкий к результату с использованием меры, определенной в п. 1.

Сравнение трех классических алгоритмов показало [Борисова, Загоруйко, 2007], что метод определение кластеров на множестве клик, полученных из матрицы подобия, показал себя малопригодным для решения поставленной задачи, так как имеет тенденцию к образованию большого количества очень мелких групп. Алгоритм Роккио показал несколько лучшие результаты: поскольку в данном методе кластеризация происходит вокруг выборочных документов, то стало возможным появление достаточно больших классов. Однако вычисленная плотность пространства документов оказалась такой, что и большая часть документов не вошла ни в один кластер.

Более качественный результат показал жадный алгоритм. Его использование привело к формированию кластерного массива, в котором каждый кластер содержит в среднем порядка 6–10 записей (для сравнения: общее число статей в базе данных – порядка 700). При этом, несмотря на необходимость построения матрицы подобия, временные затраты несущественно отличались от требуемых в алгоритме Роккио. Таким образом, по сравнению с методом клик и алгоритмом Роккио жадный алгоритм имеет ряд преимуществ.

1. Отсутствует проблема слишком большого количества больших кластеров.
2. Отсутствует проблема слишком большого количества мелких кластеров.
3. Невозможно появление документов, не попавших ни в один кластер.
4. Нет проблемы определения профилей документов, т. е. центров, вокруг которых формируются кластеры.

Далее было проведено сравнение FRiS-алгоритма с жадным алгоритмом. Выяснилось, что FRiS-алгоритм дает лучшую точность кластеризации. На рис. 1. и 2 приведены результаты кластеризации базы данных «Сибирского математического журнала» при помощи жадного алгоритма и FRiS-алгоритма соответственно.

На гистограммах отображен состав полученных кластеров. По горизонтальной оси отложены условные номера кластеров (соответствующие тем или иным разделам классификатора MSC-2000), по вертикальной – количество документов в кластере. В качестве критерия проверки правильности отнесения публикации к кластеру использовался его код классификатора из MSC-2000. Если коды классификатора центроида кластера содержались в числе кодов

классификатора данной записи, то мы полагали, что запись была отнесена к кластеру правильно.

Как нетрудно заметить, величина «шума» (отображаемая в верхней части столбиков) в кластерах при кластеризации FRiS-алгоритмом существенно ниже, нежели в случае жадного алгоритма. Более того, разбиение на кластеры более равномерно, а процент одноэлементных кластеров существенно ниже.

К сравнительным недостаткам FRiS-алгоритма следует отнести необходимость вручную задавать число кластеров в разбиении, а также несколько большую вычислительную сложность – $O(kN^2)$, где k – задаваемое пользователем число кластеров, – по сравнению с $O(N^2)$ у жадного алгоритма. Однако при кластеризации крупных баз такое увеличение сложности становится не столь существенным, к тому же для создания системы, автоматизирующей процесс отбора научных публикаций, кластеризацию базы данных требуется проводить только единожды. Таким образом, в качестве оптимального алгоритма для решения задачи кластеризации баз данных научных публикаций был признан FRiS-алгоритм.

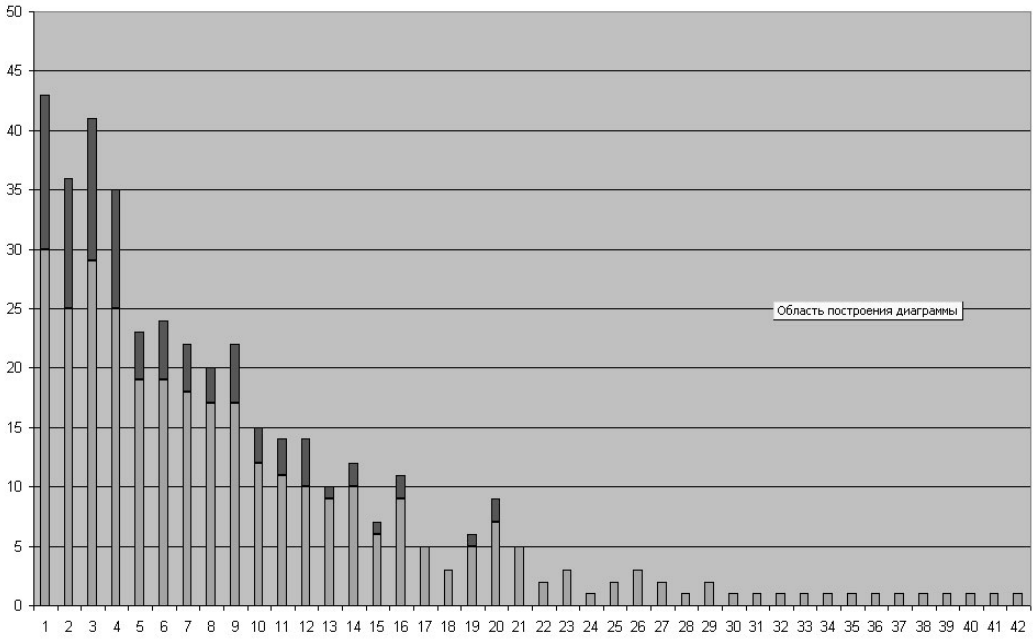

Рис. 1. Жадный алгоритм

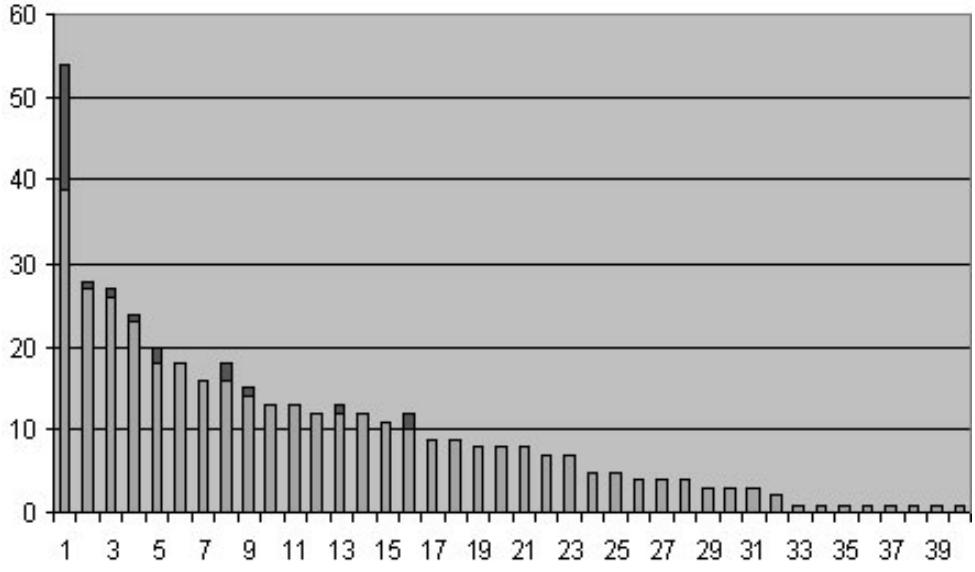

Рис. 2. FriS-алгоритм

Выявление оптимального способа задания меры сходства на множестве документов

Для задания меры на множестве документов мы применили формулу (1), где в качестве шкал использовались следующие атрибуты библиографического описания: авторы; ключевые слова; аннотация.

Так как сравнение аннотаций в явном виде (т. е. как текстовых строк), очевидно, бессмысленно, то в качестве отдельной подзадачи решался вопрос выделения терминов из общего текста аннотаций. В настоящий момент доступно web-приложение [Барахнин, Куперштох, 2006], генерирующее по запросу xml-документ с перечнем входящих в данный запрос математических терминов (как источник терминов используется тезаурус [Барахнин, Нехаева, 2007], построенный на основе «Математической энциклопедии»). Таким образом, после интеграции этого приложения в кластеризующую документы программу, данная подзадача была исчерпана, и аннотации стало возможным сравнивать между собой как прочие составные атрибуты.

Кроме того, при задании меры был принят во внимание тот факт, что значения весовых коэффициентов в формуле (1) определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы и в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных.

Для определения весового коэффициента при каждом из атрибутов была проведена кластеризация выборок из базы данных «Сибирского математического журнала». Нами были рассмотрены выборки различной мощности, а в качестве критерия истинности применялся результат кластеризации, полученный с мерой, основанной на кодах MSC-2000.

Как показал эксперимент, наибольшее сходство с результатом кластеризации при мере, базирующейся на кодах классификатора, было достигнуто путем введения следующих продукционных правил.

1. Если каждый из документов d_1 и d_2 имеет более двух авторов и как минимум $^2/_3$ из числа авторов совпадают, то соответствующий весовой коэффициент при атрибуте «авторы» мы полагаем равным единице.

2. Если каждый из документов d_1 и d_2 содержит более трех ключевых слов и как минимум $^3/_4$ этих слов совпадают, то соответствующий весовой коэффициент при атрибуте «ключевые слова» мы полагаем равным единице.

3. Если каждый из документов d_1 и d_2 содержит более четырех терминов тезауруса в аннотации и как минимум $^3/_5$ этих терминов совпадают, то соответствующий весовой коэффициент при атрибуте «ключевые слова» мы полагаем равным единице.

В противном же случае мы полагаем коэффициент при атрибуте «авторы» равным 0,2, а при атрибутах «ключевые слова» и «аннотация» равным 0,4.

Интересно отметить, что эти правила оказались оптимальными как для жадного алгоритма, так и для FRiS-алгоритма.

Заключение

Был разработан и протестирован способ задания меры сходства документов, основывающийся на сравнении атрибутов библиографического описания данных документов.

Также было проведено исследование различных алгоритмов кластеризации документов с целью выявления оптимального алгоритма для разбиения массива записей электронной базы с информацией о научных публикациях, на кластеры, содержащие в себе статьи по сходной тематике. Тестирование алгоритмов проводилось на электронной базе данных «Сибирского математического журнала», содержащей в себе библиографическое описание статей, выходивших с 2000 по 2005 г. В ходе тестирования было выявлено, что оптимальным для данной задачи является FRiS-алгоритм, хотя приемлемые результаты дает и жадный алгоритм.

Список литературы

Барахнин В. Б., Куперштох А. А. Алгоритм координатного индексирования электронных научных документов // Тр. междунар. конф. «Вычислительные и информационные техноло-

гии в науке, технике и образовании». Казахстан, Павлодар, 20–22 сентября 2006. Павлодар, 2006.Т. 1. С. 228–232.

Барахнин В. Б., Нехаева В. А. Технология создания тезауруса предметной области на основе предметного указателя энциклопедии // Вычислительные технологии. 2007. Т. 12. Спец. вып. 2. С. 3–9.

Борисова И. А., Загоруйко Н. Г. Функции конкурентного сходства в задаче таксономии // Материалы Всероссийской конференции с международным участие «Знания – Онтологии – Теории» (ЗОНТ–07), Новосибирск, 14–16 сентября 2007. Новосибирск, 2007. Т. 2. С. 67–76.

Воронин Ю. А. Начала теории сходства. Новосибирск: Наука. Сиб. отд-ние, 1991. 128 с.

Кормен Т. и др. Алгоритмы: построение и анализ / Т. Кормен, Ч. Лейзерсон, Р. М. Ривест. М.: МЦНМО, 2001. 960 с.

Солтон Дж. Динамические библиотечно-информационные системы. М.: Мир, 1979. 560 с.

Материал поступил в редколлегию 22.03.2008

V. B. Barakhnin, V. A. Nekhayeva, A. M. Fedotov

Similarity Determination for Textual Documents Clusterization

The problem of computerized selection of textual documents on scientific subjects is solved that could be of interest for an individual researcher or a research team. Attributes of bibliographical description (authors, keywords, abstract) are proposed to be used as scales for the measure determination. The values of weight coefficients in the formula for calculating the similarity measure are determined by the assumed a posteriori reliability of the respective scale data.

Three classical document clusterization methods have been analysed in order to find the ones potentially feasible for the solution of the formulated problem: clusterization by finding cliques in the full matrix of documents similarity,clusterization by Rocchio method and the method based on the so-called greed algorithm as well as the new method suggested by N.Zagoruiko based on employing the function of a rival similarity (the so-called FRiS-function). Testing showed that FRiS algorithm proved to be the most efficient one for this problem although the greed algorithm also yields acceptable results.

Keywords: similarity, clusterization of textual documents.