Знания-Онтологии-Теории (ЗОНТ-09)

Классификация математических документов с использованием составных ключевых терминов*

В.Б.Барахнин[1, 2], Д.А.Ткачев[1]

[1]*Институт вычислительных технологий СО РАН, пр. Академика Лаврентьева, д. 6, г. Новосибирск, Россия.*
[2]*Новосибирский государственный университет, ул. Пирогова, д. 2, г. Новосибирск, Россия*

bar@ict.nsc.ru, relk-tda@yandex.ru

Аннотация. *Интенсивный рост объема электронных документов делает актуальной задачу автоматической кластеризации новых документов с целью вовлечения их в процесс научной работы. От качества работы процесса кластеризации зависит корректность формирования целевых групп документов на основе библиографических признаков и полнотекстового содержания и в конечном итоге это выливается в экономию времени научного сотрудника, которое он тратит на поиски необходимого для исследования материала. Рассматривается вопрос, каким образом производить выделение ключевых слов из содержательной части электронного документа, выделять ли отдельные ключевые слова, либо производить выделение ключевых словосочетаний, относящихся к предметной области. Показано, что более оптимальным алгоритмом для использования является FRiS-алгоритм, при его сравнении с жадным алгоритмом.*

Ключевые слова: кластеризация документов, классификация документов, FRiS-функция, жадный алгоритм,

1 Введение

Наблюдаемый в последнее время бурный рост объема научной информации, представленной в электронной форме (прежде всего, в виде интернет-документов), делает актуальным решение задачу разработки методики автоматизированного вовлечения электронных документов в научно-информационный процесс. Одним из важнейших этапов этого процесса является классификация документов, поскольку при отсутствии классификационных признаков поиск документа человеком или его обработка интеллектуальной информационной системой может опираться только на простую проверку вхождения тех или иных терминов в текст документа. К сожалению, даже наиболее структурированные документы – журнальные статьи – далеко не всегда содержат классификационные признаки, к тому же классификатор источника может не совпадать с классификатором, используемым создателями интеллектуальной информационной системы.

В рассматриваемой ситуации требуется провести автоматическую классификацию (или, если говорить точнее, категоризацию) документа, исходя непосредственно из его содержания, приписав документу код(ы) того или иного классификатора предметной области.

* Работа выполнена при частичной поддержке РФФИ: проекты 07-07-00271, 08-07-00229, 09-07-00277, президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.

Еще одним методом работы с электронными научными документами, получающим все более широкое распространение, является создание исследователем картотек библиографических описаний статей, книг и т.д., представляющих для этого исследователя. Электронная форма представления документов позволяет организовывать интегрированные картотеки путем объединения ресурсов совместно работающих исследователей.

С целью освобождению ученых от рутинной работы по просмотру большого числа документов, львиная доля которых в конечном итоге может оказаться не интересной для данного исследователя, целесообразно автоматизировать процесс отбора публикаций, которые могут представлять интерес для конкретного исследователя или группы совместно работающих исследователей. Автоматизация фактически сводится к нахождению по данному документу класса схожих по содержанию документов, иными словами, их кластеризации.

Таким образом, возникает задача разработки методик автоматической классификации, то есть категоризации и кластеризации документов научной тематики.

2 Ключевые слова как основная характеристика содержания документа

В процессе автоматической категоризации документов ключевые слова являются основной характеристикой, отражающей содержание документа (здесь и далее, если не оговорено противное, выражение «ключевые слова» употребляется в значении, принятом в теоретической информатике, а не в смысле «авторских» ключевых слов, являющихся одним из элементов библиографического описания документов). Важную роль ключевые слова играют и при кластеризации документов (хотя в этом случае следует принимать во внимание и другие элементы библиографического описания документа, например, фамилии его авторов).

Однако непосредственное использование ключевых слов для классификации документов осложнено тем обстоятельством, что в некоторых случаях факт употребления того или иного термина в тексте документа отнюдь не означает, что этот термин характеризует его содержание. Достаточно типична ситуация, когда некоторый термин ***T*** употребляется в контекстах типа «*цель данной статьи показать, что теория* ***T*** *неверна*» или «*нами установлено*, что применение методики ***T*** *для исследования данного явления нецелесообразно*». Прямое выявление таких ситуаций означает проведение семантического анализа текста документа, что чрезвычайно трудно с алгоритмической точки зрения и требует больших затрат вычислительных ресурсов». Поэтому важно выделить условия, когда риск употребления ключевых слов в подобных контекстах минимизирован.

Представляется, что возникновение описанной ситуации наиболее редко встречается при работе с документами, относящимися к предметной области «математика». Действительно, естественнонаучные, а тем более общественнонаучные теории постоянно развиваются, причем нередко ранее принятые концепции впоследствии полностью отвергаются, что находит отражение в соответствующих публикациях, порождая вышеупомянутые контексты. В этом смысле математика является счастливым исключением, поскольку она оперирует, согласно терминологии Канта, априорными синтетическими суждениями, что обеспечивает последовательное (практически без отвержения полученных ранее результатов) развитие математической науки. На практике это означает, в частности, что содержание относящихся к математике документов, имеющих биографический характер, т.е. содержащих сведения об основных результатах, полученных тем или иным ученым, вполне адекватно передается посредством извлеченных из этих документов ключевых слов. Отметим, что задача автоматической категоризации биографий математиков весьма актуальна при создании информационно-справочных систем по истории науки (см., например, [5]).

Кроме того, наличие набора аксиом и логическая строгость выводов позволяет в статьях, относящихся к «чистой» математике, обходиться практически без обоснования причин выбора того или иного метода получения результата, что также приводит к минимальной вероятности возникновения упомянутых выше контекстов. В публикациях же, посвященных прикладной математике, нередко встречается подробное обсуждение альтернативных методов исследования, поэтому от возникновения нежелательных контекстов более застрахованы аннотации или краткие рефераты статей, которые обычно содержат описание лишь «положительных» аспектов исследования.

Итак, мы очертили область, априорно наиболее приемлемую для классификации научных документов на основании входящих в них ключевых слов: биографии математиков; статьи по «чистой» математике; аннотации и краткие рефераты статей по прикладной математике. Сказанное, разумеется, не означает, что для иных предметных областей науки

указанный подход даст заведомо неприемлемые результаты, однако степень применимости его к другим предметным областям выходит за рамки данной работы.

Еще одна проблема, возникающая в процессе индексирования документов, состоит в выборе структуры списка ключевых слов: должен ли он состоять исключительно из одиночных слов или он может включать в себя и составные выражения? Конечно, составные ключевые слова более адекватно описывают предметную область, но при их использовании значительно усложняется морфологический анализ текста. Более того, в некоторых работах, например, в статье [11], содержащей подробный обзор современных методов классификации документов с использованием ключевых слов, утверждается, что использование одиночных ключевых слов является «наиболее приемлемым». Такой подход при наличии качественных средств морфологического анализа представляется недостаточно обоснованным, по крайней мере, для коллекций документов, относящихся к какой-либо определенной тематике (собственно говоря, такая оговорка сделана и в [11]), тем более, что серьезные теоретические недостатки использования одиночных ключевых слов: возможность ложной координации, ложных синтагматических связей и др., – указаны еще в классической монографии [10].

Ниже мы изложим разработанные нами методики кластеризации и категоризации математических документов, основанные на использовании составных ключевых слов.

3 Составление списка ключевых слов, их классификация и извлечение из текстов

Важной задачей обработки текстовых документов, без решения которой практически невозможна автоматизация процесса извлечения из них информации и знаний, является координатное индексирование, т.е. извлечение из текстов ключевых слов. Как отмечено в [1, с. 280], координатное индексирование документов может производиться автоматически, поскольку оно дает почти такие же результаты, как и ручное, но имеет перед ним ряд преимуществ:

- обеспечивает единообразие индексирования, почти невозможное для человеческого интеллекта;
- обходится, по меньшей мере, в три раза дешевле.

Ввиду того, что в русском языке имена существительные и прилагательные при склонении изменяют свою форму, разработка эффективного алгоритма автоматизации извлечения ключевых слов является нетривиальной задачей, ибо необходимо учитывать и те случаи, когда слова, образующие термин (т.е. ключевое слово), находятся не только в именительном, но и в косвенных падежах.

В настоящий момент доступно web-приложение, генерирующее по запросу xml-документ с перечнем входящих в данный запрос математических терминов (как источник терминов используется тезаурус [3], построенный на основе «Математической энциклопедии»). Так как в русском языке имена существительные и прилагательные при склонении изменяют свою форму, то подсчет вхождений в текст терминов – словосочетаний из заданного набора является нетривиальной задачей. В основу алгоритма работы web-приложения положено использование двух индексов, содержащих триады: «номер текста» – «позиция в тексте» – «номер слова из лексического словаря» и «номер термина» – «позиция слова в термине» – «номер слова из лексического словаря» [2]. При этом если первый индекс встречается практически во всех информационно-поисковых системах, то введение второго индекса, позволяющее резко повысить эффективность алгоритма, имеет оригинальный характер. Индекс терминов размещается в хранилище данных web-приложения наряду с их списком и пополняется по мере пополнения (изменения) этого списка.

В качестве словаря словоформ web-приложение использует affix-файл свободно распространяемого программного продукта Ispell. В использованной при реализации версии (0.99g2) объем словаря составляет более 137,2 тысяч базовых слов, а полное число образуемых из них словоформ превышает 1,321 миллиона, при этом словарь постоянно пополняется, в том числе и за счет специальных терминов, предлагаемых пользователями словаря.

Следует отметить, что описанный подход не лишен недостатков. Во-первых, анализ текста требует достаточно больших затрат машинного времени (что определяется, в частности, объемом словаря словоформ). Во-вторых, создание полноценного тезауруса, необходимого для категоризации документов, предусматривает классификацию всех его терминов (т.е. понятий, встречающихся в предметном указателе «Математической энциклопедии»), число которых составляет около 27 тысяч, или, по крайней мере, классификацию всех дескрипторов словаря, в качестве которых выступают термины, являющиеся названиями статей энциклопедии

(остальные термины, встречающиеся в той или иной статье, получают код классификатора соответствующего дескриптора). Разумеется, проведение такой классификации требует немалых трудозатрат экспертов, состав которых должен охватывать практически все области фундаментальной математики (в настоящее время проведена классификация терминов, относящихся к следующим разделам «Классификации математических сущностей» (MSC2000): «Дифференциальные уравнения», «Уравнения в частных производных», «Численный анализ», «Механика жидкости» и т.п.). Наконец, в-третьих, предметный указатель «Математической энциклопедии») содержит сравнительно небольшое терминов, относящихся к областям приложения математических методов, даже наиболее традиционным, поэтому он нуждается в дополнении силами экспертов.

В силу указанных причин возникла необходимость в создании «упрощенного» словаря ключевых терминов, состав которого был бы непосредственно привязан к классификатору MSC2000. В качестве такого словаря может выступать список разделов классификатора MSC2000, содержащему, в частности, большое количество разделов, описывающих разнообразные области приложения математики. Разумеется, такой список содержит лишь наиболее общие термины, поэтому его использование наиболее целесообразно для классификации текстов, в которых предметная область описывается преимущественно целых разделов математической науки. К таким текстам относятся, прежде всего, биографии ученых-математиков, а также «визитные карточки» математических организаций (институтов, факультетов, изданий и т.п.).

Извлечение терминов из названий разделов классификатора MSC2000 (здесь и далее речь идет о переводе классификатора на русский язык) является не совсем тривиальной задачей, поскольку достаточно типично название раздела вида

«*Линейная и полилинейная алгебра; теория матриц (материалы, не классифицируемые на более конкретном уровне)*».

Поэтому при составлении словаря были исключены вспомогательные выражения, а также исследованы и нужным образом преобразованы выражения вида

«*слово1 и слово2 слово3*» и «*слово1 слово2 и слово3*».

Например,

«*Математический и гармонический анализ*» → «*Математический анализ*»; «*Гармонический анализ*»,

но

«*Метеорология и физика атмосферы*» → «*Метеорология*»; «*Физика атмосферы*».

Или

«*Голоморфные отображения и соответствия*» → «*Голоморфные отображения*»; «*Голоморфные соответствия*»,

но

«*Вариационные методы и оптимизация*» → «*Вариационные методы*»;

«*Оптимизация*».

Тип преобразования определялся автоматически посредством анализа окончаний слов, образующих преобразуемое выражение.

Извлеченным таким образом терминам естественным образом приписывались коды классификатора того раздела, в название которого эти термины входят.

4 Кластеризация математических документов

В качестве шкал для определения меры сходства между двумя документами целесообразно использовать атрибуты библиографического описания данных документов: авторы; заглавие; название журнала или издательства; год выхода; том, номер, страницы (для публикаций в периодических изданиях); аннотация; коды классификатора; ключевые слова (в узко-библиографическом значении термина) и т.п. Количественная характеристика меры сходства определяется на множестве документов D следующим образом:

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

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

$$m(d_1,d_2) = \sum a_i \, m_i(d_1,d_2), \qquad (2)$$

где i – номер элемента (атрибута) библиографического описания, a_i – весовые коэффициенты, причем $\sum a_i = 1$ (см., например, [7]), $m_i(d_1,d_2)$ – 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-ых 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 (2) может определяться предполагаемой апостериорной достоверностью данных соответствующей шкалы. Например, полное (или даже «почти полное») совпадение значений атрибута «авторы» документа d_1 и документа 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.

В работе [4] было установлено, что среди «классических» алгоритмов кластеризации для решения поставленной задачи лучшие результаты дает так называемый жадный алгоритм [8]. Суть его такова: в матрице подобия, посредством которой каждой паре документов (d_1, d_2) ставится в соответствие коэффициент подобия $S(d_1, d_2)$, находят строку, 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.

Однако использование нового метода кластеризации, основанного на использовании функции конкурентного сходства (FRiS-функции) [6], позволяет существенно повысить качество кластеризации. В этом методе 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.

Кратко изложим методику сравнительного тестирования описанных алгоритмов кластеризации. Тестирование алгоритмов проводилось на электронной базе данных «Сибирского математического журнала», содержащей библиографические описания статей журнала, вышедших в период с 2000-го по 2005-й годы (порядка 700 записей). Статьям в указанной базе данных кроме стандартных атрибутов (название, автор, год издания и т.п.) приписаны соответствующие коды классификатора MSC-2000. Это факт позволил разбить всю работу на два этапа:

1. Нахождение оптимального алгоритма кластеризации. В качестве меры на пространстве документов используется определённая ранее конструкция, однако сравнение ведётся 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 результатов кластеризации с использованием жадного алгоритма и FRiS-алгоритма показало, что FRiS-алгоритм дает лучшую точность кластеризации. Ниже приведены результаты кластеризации базы данных «Сибирского математического журнала» при помощи жадного алгоритма и FRiS-алгоритма.

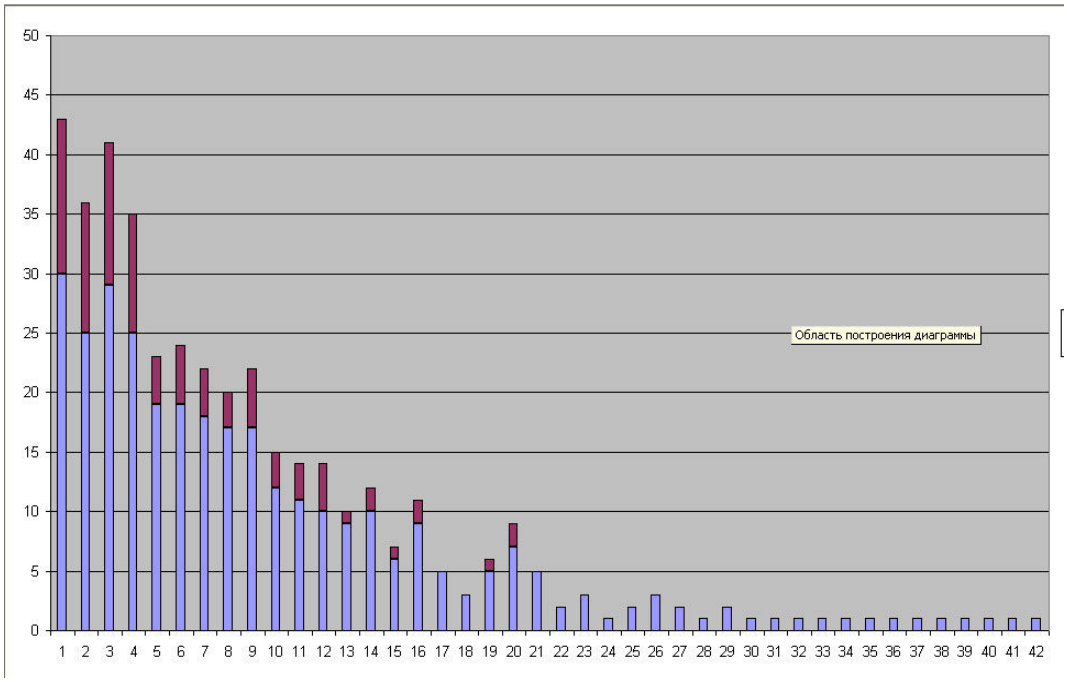

Рис.1. Жадный алгоритм

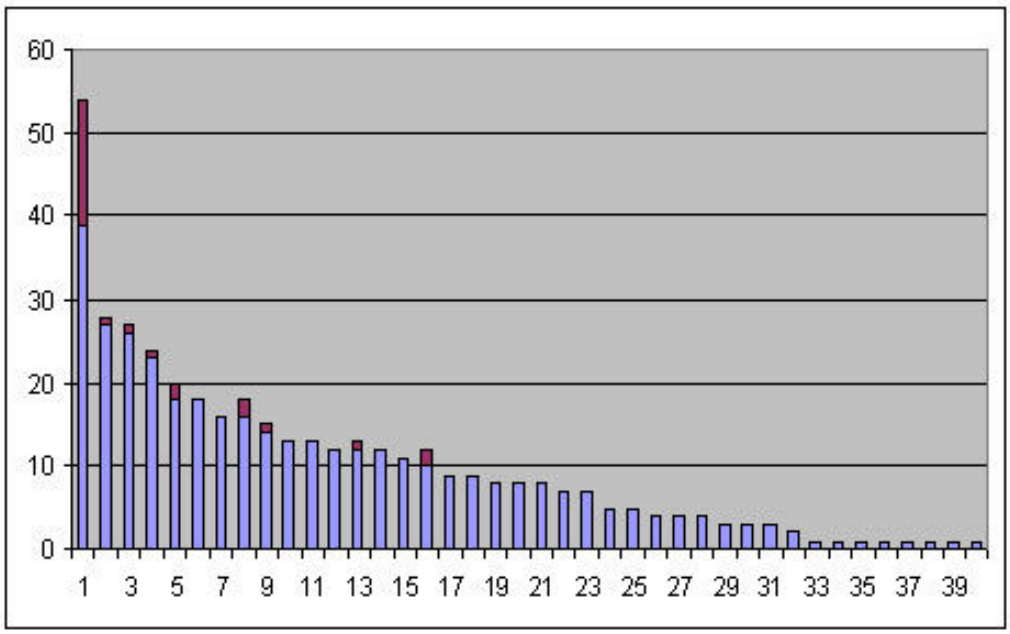

Рис. 2. FriS-алгоритм

На гистограммах отображен состав полученных кластеров. В качестве критерия принадлежности публикации к кластеру использовался его код классификатора из MSC-2000. Если в 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» (верхняя часть столбиков) в кластерах при кластеризации FRiS-алгоритмом существенно ниже, нежели в случае жадного алгоритма, разбиение на кластеры более равномерно, а доля одноэлементных кластеров существенно меньше. Отметим, что большой шум в 1-м кластере на рис.2 объясняется наличием в выборке документов из близких разделов 2-го уровня MSC-2000 (раздел 76Mxx «Основные методы в

механике жидкости» частично поглотил разделы 76Bxx «Несжимаемая невязкая жидкость» и 76Nxx «Сжимаемые жидкости и газовая динамика»).

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.

Далее были проведены вычислительные эксперименты с целью определения оптимального способа задания меры сходства на множестве документов. Использовалась формулу (1), где в качестве шкал использовались следующие атрибуты библиографического описания:

- авторы;
- ключевые слова («авторские», являющиеся элементом библиографического описания);
- аннотация (рассматриваемая как совокупность входящих в нее ключевых слов из предметного указателя «Математической энциклопедии»).

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 определённых 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 путём 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 трёх 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 четырёх терминов тезауруса в аннотации и как минимум 3/5 этих терминов совпадают, то соответствующий весовой коэффициент при атрибуте «аннотация» 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.

5 Категоризация биографий математиков

Тестирование предложенной методики категоризации математических документов было проведено на коллекции биографий выдающихся математиков, в основу которой положен биографический раздел «Математического энциклопедического словаря» [9]. Коллекция содержала 686 биографий. Ключевые слова, извлекаемые из биографий, представляли собой названия разделов русской версии классификатора MSC2000, полученные по описанной выше методике.

В итоге те или иные коды классификатора получили 656 биографий. Классифицированные биографии соответствуют 487 разделам классификатора MSC2000, всего установлено 11587 ссылок на биографии, Ни один код не получило всего лишь 30 биографий (менее 5 % от общего числа), что свидетельствует о высокой эффективности предложенного алгоритма.

Заметим, что для данной задачи был проведен эксперимент по использованию в качестве классификационных признаков одиночных ключевых слов. Так как в задачах категоризации, очевидно, есть смысл использовать только те ключевые слова, которые относятся только к одному разделу классификатора, то нам пришлось удалить из списка ключевых слов,

полученного разбиением составных слов на отдельные термины, те слова, которые относятся сразу к нескольким разделам классификатора. В итоге осталось всего лишь 162 уникальных термина, с помощью которых классификационные коды получили 357 биографий, всего установлено 887 ссылок на биографии, а ни один код не получило почти 48 % биографий. Это показывает нецелесообразность использования в задачах категоризации одиночных ключевых слов.

6 Заключение

Были предложены методики кластеризации и категоризации математических документов, основанные на использовании составных ключевых слов.

Для задачи кластеризации был разработан и протестирован способ задания меры сходства документов, основывающийся на сравнении атрибутов библиографического описания данных документов. Проведено исследование различных алгоритмов кластеризации документов с целью выявления оптимального алгоритма. В ходе тестирования было выявлено, что оптимальным для данной задачи является FRiS-алгоритм, хотя приемлемые результаты дает и жадный алгоритм.

Литература

[1] Арский Ю.М., Гиляревский Р.С., Туров И.С., Черный А.И.: *Инфосфера: Информационные структуры, системы и процессы в науке и обществе.* М., ВИНИТИ, 1996.

[2] Барахнин В.Б., Куперштох А.А.: Алгоритм координатного индексирования электронных научных документов. *Труды международной конференции «Вычислительные и информационные технологии в науке, технике и образовании»*, с. 228-232, Казахстан, Павлодар, 20-22 сентября 2006 г. – Т. I.

[3] Барахнин В.Б., Нехаева В.А.: Технология создания тезауруса предметной области на основе предметного указателя энциклопедии. *Вычислительные технологии,* с. 3-9, 2007, Т. 12, Специальный выпуск 2.

[4] Барахнин В.Б., Нехаева В.А., Федотов А.М.: Методика отбора публикаций из библиографических баз данных на основании меры сходства. *Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–07)*, с. 88-94, Новосибирск, 14-16 сентября 2007 г. – Т. 2.

[5] Барахнин В.Б., Федотов А.М.: Методика построения информационно-справочной системы по истории математической науки. *Электронные библиотеки.* – 2007. – Т. 10. – Вып. 1. – http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2007/part1/BF.

[6] Борисова И.А., Загоруйко Н.Г.: Функции конкурентного сходства в задаче таксономии. *Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–07)*, с. 67-76, Новосибирск, 14-16 сентября 2007 г. – Т. 2.

[7] Воронин Ю.А.: Начала теории сходства. *Новосибирск: Наука.* Сибирское отделение, 1991.

[8] Кормен Т., Лейзерсон Ч., Ривест Р.: *Алгоритмы: построение и анализ*. М., МЦНМО, 2001.

[9] Математический энциклопедический словарь. М., *Советская энциклопедия*, 1988.

[10] Михайлов А.И., Черный А.И., Гиляревский Р.С.: *Основы информатики*. М., Наука, 1968.

[11] Пескова О.В. Автоматическое формирование рубрикатора полнотекстовых документов. *Труды Десятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2008)*, с. 139-148, Дубна, 2008.