Поддержка системы автоматического рубрицирования для сложных задач классификации текстов♣

© М.С. Агеев
Механико-математический факультет МГУ;
Научно-исследовательский вычислительный центр МГУ;
АНО Центр информационных исследований
ageev@mail.cir.ru

Б.В. Добров
Научно-исследовательский вычислительный центр МГУ;
АНО Центр информационных исследований
dobroff@mail.cir.ru

Н.В. Лукашевич
Научно-исследовательский вычислительный центр МГУ;
АНО Центр информационных исследований
louk@mail.cir.ru

Аннотация

В сложных задачах рубрикации единственным способом решения задачи является итерационное уточнение правил рубрицирования. В настоящей работе мы излагаем способы ускорения процедур уточнения рубрикации, базирующиеся на интерактивном использовании локального статистического анализа, выявляющего основные понятия документов, возвращаемых поисковой машиной.

1 Введение

Классификация/рубрикация информации (отнесение документов к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией. В больших информационных коллекциях имеет смысл говорить только об автоматической рубрикации.

Существует два основных подхода для автоматической классификации:

- так называемый, "инженерный" подход, относящийся к группе методов, основанных на знаниях, когда каждая рубрика описывается правилами (с использованием логических выражений, весов) над языковыми единицами, обнаружение которых в тексте требует вывода данной рубрики;

- методика машинного обучения, когда система классификации сначала настраивает параметры рубрикации, обучаясь на примерах, затем "раскладывает" новые документы в соответствии с настроенными параметрами.

Для рубрикаторов простой структуры большинство методов автоматической рубрикации показывают приемлемые результаты.

В настоящее время можно наблюдать всплеск научных работ, посвященных описанию, применению и оценкам методов машинного обучения для автоматической рубрикации текстов. Приводятся высокие оценки результатов работы таких методов [5].

При ближайшем рассмотрении оказывается, что практически все такие методы тестируются на одной и той же текстовой коллекции - это коллекция финансовых сообщений информационного агентства Рейтер [7], которая была специально создана для тестирования методов автоматической рубрикации текстов. Коллекция характеризуется следующими основными чертами: 1) рубрикатор, включающий 135 рубрик, относительно прост, без иерархии; 2) небольшие по величине тексты принадлежат достаточно узкой области финансовых новостей; 3) для обучения представляется более 15 тысяч отрубрицированных вручную документов. Эти особенности коллекции значительно упрощают решение задачи машинного обучения автоматической рубрикации текстов. Кроме того, обычно приводятся результаты только для 10-20 самых частотных рубрик, в то время как аккуратный анализ на всем множестве рубрик показывает ухудшение результатов [1].

Реальные задачи рубрикации текстов в значительной мере отличаются от задачи классификации сообщений на тестовой коллекции агентства Рейтер. Эта проблема – недостижимость хороших результатов классификации при использовании простых методов, без длительного и

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2004, Пущино, Россия, 2004.

трудоемкого этапа настройки – выдвинулась на передний план интересов специальной группы по информационному поиску SIGIR и специальной группы по извлечению знаний SIGKDD. В рамках годовых конференций указанных групп стали проводиться специальные семинары Operational Text Categorization [2, 3, 6].

Во втором разделе данной работе мы опишем проблемы реальной классификации, с которыми мы встретились, решая задачи в рамках проекта Университетская информационная система РОССИЯ (далее – УИС РОССИЯ, http://www.cir.ru) [13], которая поддерживается Научно-исследовательским вычислительным центром МГУ им. М.В.Ломоносова и АНО Центр информационных исследований. В третьем разделе кратко описывается применяемый нами метод рубрикации документов. В четвертом разделе мы изложим методологию решения сложных задач рубрикации во взаимодействии с экспертами предметной области.

2 Реальные задачи классификации текстов

Приведем несколько реальных задач классификации текстов, с которыми нам пришлось столкнуться на практике. Они интересны тем, что их постановки отличаются от «классической», что делает проблематичным применение «классических» методов машинного обучения.

2.1 Массив заранее отрубрицированных текстов отсутствует и не может быть создан вручную за короткое время

Для создающегося Архива социологических данных, предназначенного объединить результаты опросов различных социологических служб, Независимый институт социальной политики (www.socialpolicy.ru) разработал рубрикатор, включающий более 300 рубрик и 4 уровня иерархии. Отсутствует набор отрубрицированных по этому рубрикатору документов и нет ни соответствующих специалистов, ни финансовых возможностей создать набор социологических опросов, достаточный для автоматического обучения по более чем 300 рубрикам.

2.2 Массив отрубрицированных документов есть, но документы отрубрицированы пользователями

Международное научное сообщество RePEc/СоцИоНет поддерживает архив научных публикаций по экономической и социологической тематике (www.socionet.ru). Автор публикации сам помещает свою публикацию в архив и может снабдить ее рубриками классификатора JEL [4], включающего более 700 рубрик по экономической тематике. Для улучшения качества сервиса важно автоматизировать работу по рубрикации публикаций, например, автоматически предлагая автору набор возможных рубрик. Однако по уже отрубрицированным публикациям обучаться невозможно из-за значительной непоследовательности приписанных рубрик. Авторы не обязаны помнить весь рубрикатор, иметь согласованный взгляд на содержание его рубрик. Отсутствие конкретной рубрики в списке рубрик текста не означает, что ее там не должно быть, автор мог про нее просто забыть.

2.3 Массив отрубрицированных документов есть, но достаточен ли он…

При разработке рубрикаторов, предназначенных для стандартизации обмена информацией, часто следуют иным соображениям, стараются учесть как можно больше конкретных направлений, важных для той или иной группы составителей рубрикатора [15]. В результате имеется много реально использующихся рубрикаторов [4, 16], насчитывающих сотни и тысячи рубрик.

В частности, чтобы упорядочить хранение нормативной базы российского законодательства был разработан Классификатор правовых актов, насчитывающий более 1100 рубрик и 3-4 уровня иерархии. Все документы федерального уровня рубрицируются по этому рубрикатору специалистами Главного государственно-правового управления Президента РФ и хранятся в эталонной коллекции нормативных документов НТЦ «Система» (www.systema.ru).

Однако, ко многим рубрикам сложного классификатора даже в большой коллекции документов попадает сравнительно мало документов (менее 1% от всей коллекции), что затрудняет применение методов, использующих машинное обучение. В нашей практике мы имели случай, когда квалифицированные эксперты в коллекции в 10 тысяч правовых актов по рубрикатору из тысячи рубрик только для 50 рубрик отнесли более 100 документов, только для 200 рубрик – более 20 документов.

2.4 Обучиться нужно на одной коллекции текстов, а рубрицировать другую коллекцию

В России имеется 89 субъектов федерации, каждый из которых имеет свою законодательную базу, что в совокупности составляет сотни тысяч документов. Вручную качественно отрубрицировать такое количество документов невозможно, учитывая сравнительно небольшое количество подготовленных соответствующим образом специалистов.

Прямое применение нами лучших методов автоматической рубрикации, основанных на машинном обучении, по результатам ручного рубрицирования документов федерального уровня (10 тысяч документов) не решает задачу, так как много региональных документов не получает ни одной рубрики, в среднем значительно падает количество рубрик на один документ. При

сравнительно небольшой вариативности лексики федеральных и региональных документов к тем рубрикам, где в федеральной коллекции было приписано менее 0.5% документов, программа [5] не относит ни одного документа.

2.5 Общие проблемы получения качественной обучающей коллекции для больших рубрикаторов

В процессе анализа результатов ручного рубрицирования по большим рубрикаторам, даже проводимого высококвалифицированными экспертами, было выявлено три типа проблем ручного рубрицирования.

2.5.1 Проблема определения и соблюдения ограничивающих правил рубрицирования

Суть проблемы заключается в том, что ограничивающие правила рубрицирования, не связанные непосредственно с формулировкой конкретной рубрики, являются серьезной базой для субъективизма:

- об этих правилах забывает часть экспертов,
- для разных рубрик эти правила соблюдаются с разной степенью последовательности,
- эти правила неизвестны пользователю, в большой степени он опирается на буквальную формулировку рубрики.

Выбор такого рода правил напрямую зависит от четкого определения ролей рубрикатора в информационно-правовой системе, взаимодействия этих ролей с ролями других типов информации (например, указателей – действующий или не действующий документ), моделью пользователя системы, сценариями работы различных типов пользователей с рубрикатором.

2.5.2 Проблема документов, отнесенных экспертами к рубрике ошибочно

Процент таких документов в общем количестве документов обычно невелик. Важность нахождения такого рода документов состоит в том, что с большой вероятностью ошибочная рубрика проставлена вместо правильной рубрики, и данный документ не будет найден по правильной рубрике.

2.5.3 Проблема пропущенных экспертами документов

Нахождение пропущенных экспертами документов является непростой задачей, и может усугубляться проблемой существования ограничивающих правил, по которым не приняты окончательные решения, и большого количества «промежуточных документов», для которых неясно, должны ли они принадлежать рубрике или нет.

Важным шагом является определение набора документов-кандидатов в рубрику для дополнительного просмотра. Наиболее оптимальным здесь, видимо, является опора на результаты анализа результатов автоматического и ручного рубрицирования. Представляются полезными следующие шаги:

- первичный анализ результатов автоматического и ручного рубрицирования, выявление рубрик, в которых явление пропуска релевантных документов носит массовый характер;
- консультация с экспертами, что документы, которые показались релевантными, действительно такими являются;
- по результатам классификации исправление результатов автоматического рубрицирования, чтобы достичь максимального значения полноты без снижения содержательной точности (т.е. без дополнительного появления в результатах автоматического рубрицирования явно нерелевантных документов);
- представление экспертам новых результатов автоматического рубрицирования;
- на множестве документов, помещенных в рубрику автоматической системой, но не взятых в рубрику экспертами, эксперты должны просмотреть практически все документы в этом множестве один за другим и решить, каким документам добавить анализируемую рубрику;
- возможно необходимо использовать систему решений не из двух значений (принадлежит рубрике или не принадлежит), а из трех: добавить еще – условно принадлежит – в случаях расхождения между экспертами или неясности решения.

3 Рубрикация с использованием тезауруса

Авторы ранее разработали методы рубрицирования, основанные на знаниях [12] - для построения правил отнесения документа к рубрике используется Общественно-политический тезаурус для автоматического концептуального индексирования [14]. Тезаурус разрабатывается с 1994 года АНО Центр информационных исследований и включает в настоящее время более 30,000 понятий, 75,000 терминов (дескрипторы + синонимы), 115,000 прямых и 850,000 наследуемых отношений между понятиями. Тезаурус покрывает 95-99% терминологии любого русскоязычного текста из коллекции нормативных документов РФ и материалов СМИ с 1991 года.

Правила отнесения документа к рубрике задаются в виде краткого логического выражения, в котором используется лишь малое количество так называемых «опорных» понятий. В процессе рубрицирования это логическое выражение расширяется при помощи связей тезауруса. Отметим, что использование техники опорных понятий на порядок сокращает время описание экспертом рубрики по сравнению с известными по литературе данными [9], однако остается весьма высоким (в среднем 200 рубрик на одного эксперта в месяц), что требует увеличения автоматизации.

Каждая рубрика R описывается дизъюнкцией альтернатив, каждый дизъюнкт представляет собой конъюнкцию. Конъюнкты в свою очередь описываются экспертами с помощью так называемых «опорных» понятий тезауруса:

$$R = \bigcup_i D_i = \bigcup_i \left[\bigcap_j K_{ij} \right] = \bigcup_i \left[\bigcap_j \left(\bigcup_k d_{ijk} \right) \right] \quad (3.1)$$

Для каждого опорного понятия задается правило его расширения $f(\cdot)$, определяющее каким образом вместе с опорным понятием учитывать подчиненные ему по иерархии понятия. Выделяются несколько случаев – без расширения, полное расширение по дереву иерархии тезауруса и расширении только по родо-видовым связям и т.п.

Опорный концепт может быть как «положительным», который добавляет нижерасположенные понятия в описание конъюнкта, так и «отрицательным», который вырезает свои подчиненные понятия. Последовательность учета положительных и отрицательных опорных понятий регулируется заданием специального атрибута. Результатом применения расширения опорных понятий является совокупность понятий тезауруса, полностью описывающая конъюнкт:

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn}) = \bigcup_k d_{ijk}$$

Важным атрибутом является пометка о необходимости «подтверждения». Понятия, требующие подтверждения, не могут самостоятельно выводить рубрику, но могут усиливать эту рубрику, если в тексте встречаются понятия, не требующие подтверждения.

Следует подчеркнуть, что в данной методологии достаточно хранить только опорные понятия, а также понятия, у которых изменен атрибут подтверждения, полное же описание рубрики может быть каждый раз пересчитано заново при изменении тезауруса.

Типичные цифры о параметрах описания: на одну рубрику рубрикатора в среднем приходится 1-2 дизъюнкта, 2-3 конъюнкта, 10-20 опорных понятия («положительных» и «отрицательных»), 200-400 понятий полного описания, то есть 400-800 текстовых входов.

Оценка релевантности содержания текста рубрике (вес рубрики) может быть рассчитана на основе информации о весах понятий в тексте, входящих в ее описание, а также с учетом текстовых связей между понятиями одного конъюнкта и понятиями другого.

В нашем подходе значимость термина для содержания текста определяется в результате построения так называемого тематического представления текста, независимом от величины и типа текстов. Основные этапы построения тематического представления текста таковы (подробнее см. [11, 14]):

- сопоставление текста с Тезаурусом создает для текста терминологический индекс, в котором указываются термины и ассоциированные с ними понятия Тезауруса и в каком месте текста обнаружены;
- нахождение для каждого понятия текста тематически близких понятия и отражение этой информации в, так называемой, тезаурусной проекции текста;
- использование связей между понятиями в тезаурусной проекции для разрешения многозначности терминов;
- построение текстовых связей для каждого понятия текста, то есть фиксация для каждого вхождения понятия трех соседних понятий вправо и трех влево. Выбор таких цифр величина экспериментальная, однако согласуется с экспериментами в области исследования кратковременной памяти;
- построение тематических узлов - групп понятий, близких по иерархии Тезауруса. Тематические узлы строятся вокруг «центров узлов» - тех понятий, которые отличаются своей частотностью или местоположением в начале текста;
- выбор среди построенных тематических узлов - основных тематических узлов, то есть тех, которые моделируют элементы основной темы текста. Выбор производится на основе анализа суммированных текстовых связей тематических узлов.

В зависимости от того элементом какой структуры тематического представления оказывается понятие d тезауруса формируется оценка значимости $\omega(d;D)$. Типичные значения – 0.9 для центра основного тематического узла, 0.7 для элемента основного тематического узла, 0.75 для центра локального тематического узла и т.д. Окончательно вес понятия для текста определяется добавлением стабилизирующего фактора, учитывающего частотность понятия в документе:

$$\theta(d) = \alpha \cdot \omega(d;D) + (1-\alpha) \cdot \frac{freq(d;D)}{\max\limits_c freq(c;D)}, \quad (3.2)$$

$\alpha = 0.7$

Таким образом, оценка релевантности содержания текста рубрике (вес рубрики) может быть рассчитана на основе информации о весах понятий в тексте, входящих в ее описание.

В УИС РОССИЯ вес конъюнкта рассчитывается по формуле:

$$\theta(K_{ij}) = \min\left\{1.0;\ \max\left(\theta(d_{ijk}),\ \chi \cdot \theta(p_{ijm})\right)\right\}, \quad (3.3)$$

где d_{ijk} понятия, не требующие подтверждения, p_{ijm} – понятия, требующие подтверждения, χ -

множитель равный единице если имеются понятия, не требующие подтверждения, и нулю иначе.

Вес дизъюнкта предназначен учитывать не только сумму весов составляющих его конъюнктов, но и меру близости конъюнктов в тексте:

$$\theta(D_i) = \frac{\sum_{j=1}^{m} \theta(K_{ij}) \quad + \quad \sum_{j<k} S(K_{ij}, K_{ik})}{m \quad + \quad C_m^2}, \quad (3.4)$$

здесь

$$S(K_{ij}, K_{ik}) = \min\{1.0; \frac{\sum s(c_{ijq} \in K_{ij}, d_{ikw} \in K_{ij})}{\max s(c \in D, d \in D)}\}$$

- сумма всех текстовых связей между понятиями одного конъюнкта и понятиями другого, деленная на значение максимальной текстовой связи между любыми двумя понятиями текста. Этот член равен обычно единице для сильно связанных конъюнктов и принимает малое значение, если понятия различных конъюнктов обсуждались в разных местах текста.

Вес рубрики представляется максимумов весов входящих в описание рубрики альтернатив. В случае имеющихся иерархических связей между рубриками оценка релевантности нижестоящих рубрик переносится на вышестоящие. Так что при запросе по вышестоящей рубрике будут выходить и документы, к которым были приписаны нижестоящие рубрики.

Алгоритм рубрицирования работает следующим образом. Для всех понятий тезауруса, найденных в тексте, определяется множество рубрик, которые могут быть определены в тексте. Для каждой рубрики происходит расчет ее веса по формулам (3.2), (3.3) и (3.4). В результирующем множестве остаются рубрики, вес которых превосходит задаваемый заранее для коллекции порог.

4 Поддержка системы рубрицирования с использованием тезауруса

Мы видим, что описание рубрики (3.1) может быть смоделировано запросом к информационно-поисковой системе. Естественно использовать это обстоятельство для улучшения качества описания рубрики. Работа по уточнению описания рубрики происходит на основе предварительно загруженных в базу данных документов, отрубрицированных экспертами и программой автоматической рубрицирования.

Интерфейс пользователя УИС РОССИЯ позволяет уточнить запрос, добавив или удалив заинтересовавшее понятие в/из строки запроса (для этого достаточно одного нажатия клавиши «мыши»). Можно также войти в тезаурус, воспользоваться навигацией по иерархии понятий тезауруса для расширения/сужения запроса.

Документы могут быть просмотрены:

- по рубрикам, приписанным экспертами;
- по рубрикам, приписанным программой автоматического рубрицирования;
- по входящим в документ словам/леммам;
- по концептам Общественно-политического Тезауруса (в том числе с расширением по иерархии);
- по булевскому запросу, образованному из всех вышеперечисленных компонентов и операторов И, ИЛИ, НЕ.

Для анализа ошибок используется также правая панель экранного интерфейса, содержащая наиболее характерные термины для множества документов, полученного по запросу (информеры УИС РОССИЯ).

4.1 Интерактивное изменение запроса

Вспомогательные средства поиска, основанные на статистическом анализе запроса и содержания документов, постепенно получают широкое распространение и внедряются в различных поисковых системах. Стоит отметить развитые средства анализа коллекции документов и агрегирования данных развиваемые в течение многих лет компаниями TextWise (www.textwise.com) и Inxight (www.inxight.com). В поисковых машинах Teoma (www.teoma.com) и Vivisimo (www.vivisimo.com) используется кластеризация найденных документов по классификатору тем, описываемых словосочетаниями. Из российских систем можно отметить Галактика-Зум (zoom.galaktika.ru) [10], основанную на выделении наиболее значимых слов и словосочетаний типа прилагательное/местоимение + существительное; различные подходы к визуализации результатов компании «Гарант-Парк-Интернет» (research.metric.ru).

Тематический анализ результатов запроса в УИС РОССИЯ производится при помощи выделения понятий Тезауруса, наиболее характерных (контрастных) для документов, полученных в результате исполнения запроса. Список дескрипторов понятий упорядочивается по убыванию значимости и показывается рядом с результатами запроса в специальном визуальном «информере» (см. Рис.1). Степень важности термина обозначается цветом — более значимые термины имеют более теплые цвета.

4.2 Использование информеров при решении задач классификации

На основе своего опыта мы можем утверждать, что при формировании или модифицировании логической формулы, описывающей рубрику, необходимо производить различные оценки полноты и точности рубрикации. Информеры УИС РОССИЯ позволяют экспертам производить данные оценки интерактивно, что повышает эффективность труда – работа ускоряется и результаты имеют

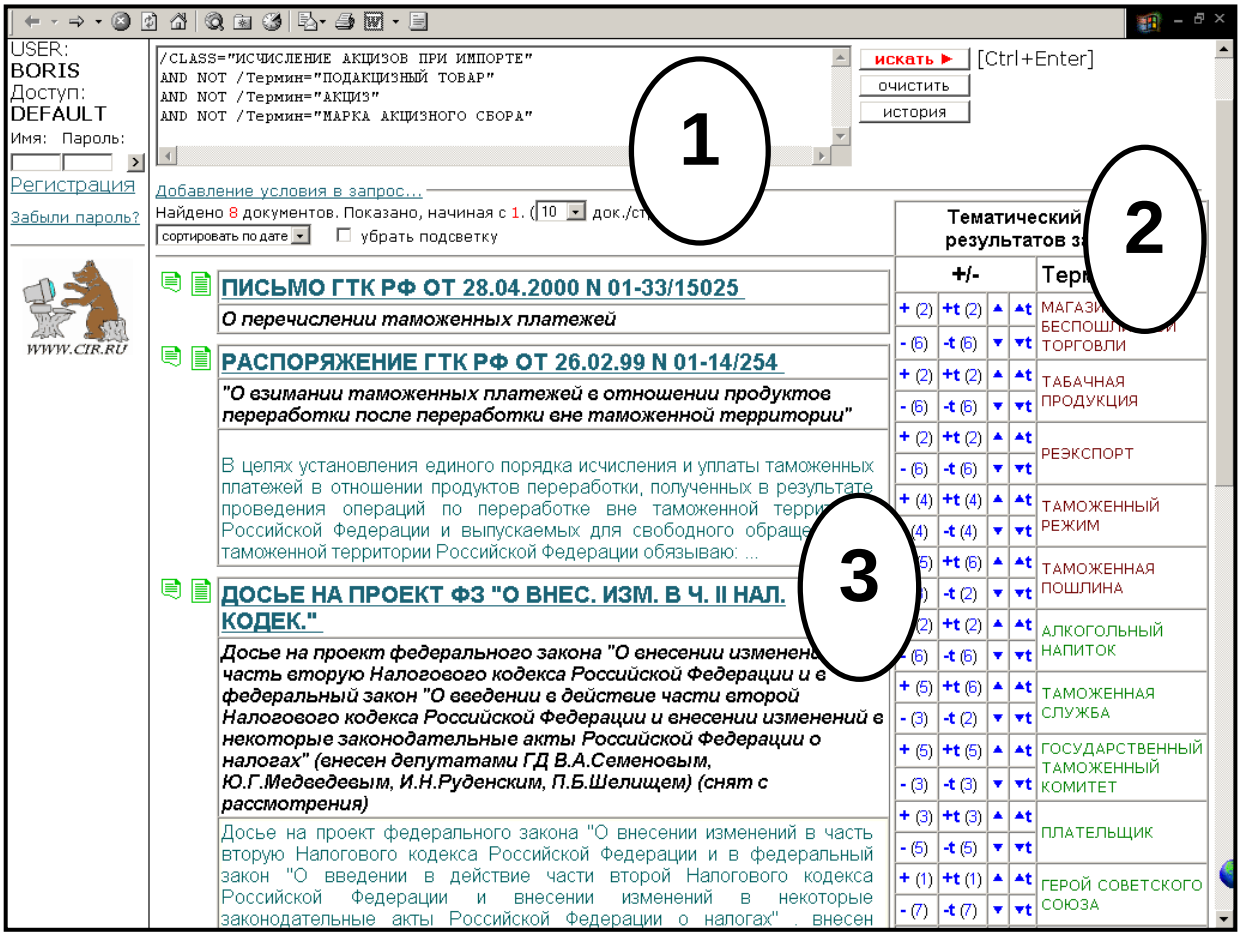

Рис.1. Использование информеров УИС РОССИЯ для интерактивного уточнения описания рубрики.
(1) окно условий запроса; (2) тематический информер;
(3) «ссылки-кнопки» для оперативного добавления условия в запрос

Лучшие показатели по критерию полноты и точности.

Опишем алгоритм работы специалиста по рубрикации для решения различных задач поддержки рубрицирования по сложному рубрикатору.

4.2.1 Создание терминологического описания для рубрики

По содержанию рубрика обычно разделяется на несколько более элементарных единиц, упоминания которых прямо или косвенно должны быть найдены в тексте документов рубрики.

Для того, что составить для рубрики терминологическое описание, необходимо выявить элементарные смыслы рубрики, найти, какими терминами эти смыслы могут выражаться. Далее необходимо записать булевское выражение, в котором термины, выражающие разные составляющие смыслы рубрики, будут соединяться конъюнкцией, а термины, выражающие один и тот же смысл дизъюнкцией.

Для нахождения соответствующих понятий удобно использовать информеры УИС РОССИЯ.

Рассмотрим «модельную» рубрику «ИСЧИСЛЕНИЕ АКЦИЗОВ ПРИ ИМПОРТЕ».

Каждый текст, относящийся к этой рубрике, должен содержать термины, относящиеся к сфере импорта, и термины, относящиеся к сфере акцизов.

Выполняем поиск по рубрике – получаем набор документов, отнесенных к рубрике экспертами.

Выбираем из информера понятия, относящиеся к акцизам: *ПОДАКЦИЗНЫЙ ТОВАР*, *АКЦИЗ*, *МАРКА АКЦИЗНОГО СБОРА*. Удаляем из выдачи документы, содержащие эти понятия, чтобы определить, какие еще термины могут относиться к сфере акцизов.

Собираем теперь понятия, относящиеся к *импорту*. Возвращаемся к запросу по рубрике. Изучаем информер – имеется понятия *ИМПОРТ*. Удаляем документы, включающие этот термин, из выдачи.

Информер больше понятий не дает. Начинаем изучать оставшиеся тексты. В текстах содержатся слова *ввоз, ввезти, ввозить, ввозной*. Убираем эти документы.

В информере появились понятия *ТАМОЖЕННАЯ ПОШЛИНА, ТАМОЖЕННОЕ*

ОФОРМЛЕНИЕ ТОВАРОВ, ГОСУДАРСТВЕННЫЙ ТАМОЖЕННЫЙ КОМИТЕТ. В сочетании с вопросами акцизами эти понятия должны указывать на импорт.

Таким образом, мы получаем формулу:

```
(ПОДАКЦИЗНЫЙ ТОВАР
 или АКЦИЗ
 или МАРКА АКЦИЗНОГО СБОРА)
и
(ИМПОРТ
 или ВВОЗ
 или ТАМОЖЕННАЯ ПОШЛИНА
 или ТАМОЖЕННОЕ ОФОРМЛЕНИЕ ТОВАРОВ
 или ТАМОЖЕННЫЙ КОМИТЕТ)
```

На каждом шаге происходит контроль оставшегося количества документов, процесс уточнения формулы прекращается, если достигнут требуемый уровень ошибки.

Если название рубрики выглядит как состоящее из одного термина, то это часто не означает, что достаточно упоминания этого термина в тексте, чтобы присвоить тексту рубрику. Часто такой текст должен обсуждать какие-то значимые для данного понятия части, свойства и ситуации.

Так, тексты в рубрике «ОБЩЕСТВА С ОГРАНИЧЕННОЙ И С ДОПОЛНИТЕЛЬНОЙ ОТВЕТСТВЕННОСТЬЮ» должны содержать не только термины *общество с ограниченной ответственностью* или *общество с дополнительной ответственностью*, но и обсуждать такие важнейшие аспекты для этих организаций, как создание, регистрация, учредители, уставный капитал, собственность и т.п.

Таким образом, реально рубрика также разлагается на два элементарных смысла, тот что назван в формулировке и что-то вроде «общие вопросы», и описывать рубрику нужно в виде конъюнкции двух частей. Понятия, которые нужно включить во вторую часть конъюнкции, т.е. те которые важны для функционирования первой части, могут быть набраны из правой панели экранного интерфейса. Для упомянутой рубрики на правой панели мы увидим: *УСТАВНЫЙ КАПИТАЛ*, *УЧРЕДИТЕЛЬ*, *РЕГИСТРАЦИЯ ЮРИДИЧЕСКИХ ЛИЦ*, *СОВЕТ ДИРЕКТОРОВ*.

4.2.2 Использование программы автоматической рубрикации для нахождения ошибок ручного рубрицирования

Для нахождения ошибочных документов в совокупности документов, приписанных рубрике экспертами, необходимо убедиться, что каждый из документов рубрики, упоминает явно или косвенно каждый их элементарных смыслов.

Выполним запрос на поиск документов, приписанных экспертами рубрике «СТРАХОВЫЕ ВЗНОСЫ В ПЕНСИОННЫЙ ФОНД».

Выберем и мысленно зафиксируем один из элементарных смыслов. Например, выберем понятие *ПЕНСИОННЫЙ ФОНД*. Нам нужно проследить его наличие в каждом документе рубрики. Для этого будем выбирать на правой панели понятия, которые могут выражать в тексте этот смысл, например, *ГОСУДАРСТВЕННЫЙ ПЕНСИОННЫЙ ФОНД*, *ПЕНСИОННОЕ СТРАХОВАНИЕ*, *ПЕНСИОННЫЙ ФОНД*.

Используя кнопку “-“, удаляем из выдачи документы, содержащие эти понятия. т.е. на множестве документов рубрики выполняем запрос:

```
/CLASS=    "СТРАХОВЫЕ ВЗНОСЫ В
           ПЕНСИОННЫЙ ФОНД"
AND NOT    /Термин="ГОСУДАРСТВЕННЫЙ
           ПЕНСИОННЫЙ ФОНД"
AND NOT    /Термин="ПЕНСИОННОЕ
           СТРАХОВАНИЕ"
AND NOT    /Термин="ПЕНСИОННЫЙ ФОНД"
```

Смотрим еще раз на правую колонку, и если находим еще понятия, соответствующие выбранному элементарному смыслу, то удаляем содержащие их документы и т.д. Так продолжаем, пока правая колонка уже не содержит такого рода понятий.

Если документы уже закончились, то это означает, что выбранный смысл найден в каждом из документов и можно переходить к следующему смыслу.

В противном случае, необходимо вызывать на экран оставшиеся документы и, читая их, понять, какие слова или термины в них указывают на искомый элементарный смысл.

В нашем случае выяснилось, что многие из оставшихся текстов содержат аббревиатуру ПФР. Удаляем из выдачи документы, содержащие найденное слово или термин.

Повторяем процедуру, ища понятия, соответствующие элементарному смыслу, на правой панели, или слова внутри текста.

Находим, что многие оставшиеся тексты содержат формулу «страховые взносы во внебюджетные фонды» и понятие ВНЕБЮДЖЕТНЫЙ ФОНД в правой колонке, удаляем документы с эти понятием.

В результате повторения процедуры остаются документы, отнесение которых к рубрике регулируется не содержимым, но внешними параметрами («Внесение изменений», «Досье на проект») и ошибочные документы.

Для нахождения пропущенных экспертами релевантных документов необходимо сначала сформировать множество документов, в которых весьма вероятно могут находиться такие документы. В качестве такого множества могут служить документы из выдачи процедуры

автоматической рубрикации для данной рубрики и (или) документы, выданные по запросу – булевскому выражению из слов и (или) понятий, сформированному на основе формулировки рубрики, например, (СТРАХОВОЙ ВЗНОС and ПЕНСИОННЫЙ ФОНД).

Из полученной таким образом выдачи документов необходимо удалить документы, приписанные рубрике экспертами.

Результирующее множество документов необходимо изучить. Здесь выполняем следующую процедуру.

По содержанию документы в результирующем множестве могут подразделяться на несколько классов:

- документ явно нерелевантен;
- документ явно релевантен – пропущенный документ найден и должен быть добавлен к множеству документов рубрики,
- документ касается темы рубрики, но акцент документа несколько смещен – таких документов в рассматриваемом множестве может быть достаточно много.

Для рассмотрения последнего типа документов необходимо выполнить следующие шаги:

1) необходимо выяснить, сколько документов, похожих на найденный, включено экспертами в рубрику. Для этого на множестве документов, полученных в результате ручной рубрикации, выполняется булевский запрос из слов и терминов, наиболее полно отражающий суть документа;

2) по всему этому множеству документов должно быть принято решение о включении (не включении) в рубрику;

2а) или все эти документы должны быть включены в рубрику и тогда к рубрике нужно приписать соответствующее правило о включении и добавить найденный документ;

2б) если принято решение не включать такой тип документов, тогда правило не включения также должно быть зафиксировано, а подобные документы, прежде включенные в рубрику, должны быть удалены из нее как ошибочные.

После анализа документа необходимо по возможности как можно точнее описать его основное содержание в виде булевского запроса и удалить всю совокупность аналогичных документов из рассматриваемого множества, после чего начинать рассмотрение следующего документа.

4.2.3 Итерационное повышение полноты автоматического рубрицирования

Для повышения полноты автоматического рубрицирования необходимо найти термины, которые выражают элементарные смыслы рубрики, но не были учтены в текущем терминологическом описании рубрики.

Для этого из множества документов, приписанных рубрике экспертами, вычитается множество документов, помещенное в ту же рубрику при автоматическом рубрицировании, т.е. формируется набор документов рубрики, на котором программа автоматического рубрицирования проработала неудачно.

Пропущенные элементарные смыслы пытаемся найти на правой панели экрана. Удаляем из набора документы, содержащие эти понятия (используем кнопку «-» на правой панели).

Продолжаем поиск дополнительных понятий для включения в терминологическое описание на правой панели.

В некоторый момент мы не можем найти добавления в терминологическое описание ни на правой панели, ни в текстах документов.

Если документы остались, то обычно это документы трех видов:

- документы, отнесенные к рубрике экспертами ошибочно,
- документы вида «внесение изменений, не содержащие в явном виде смысловых элементов рубрики,
- документы, в которых присутствуют все элементарные смыслы рубрики, но рубрика получает при автоматическом рубрицировании слишком небольшой вес (например, потому, что текст большой, а релевантная фраза одна).

4.2.5 Итерационное повышение точности автоматического рубрицирования

Для определения способов повышения точности автоматического рубрицирования необходимо получить набор документов, которые были включены в рубрику в процессе автоматического рубрицирования, но не были включены в рубрику экспертами. Для этого в оболочке УИС РОССИЯ необходимо выполнить запрос по рубрике для документов, отнесенных к этой рубрике в процессе автоматического рубрицирования, а затем удалить из выдачи, те документы, которые были включены в рубрику экспертами.

Полученные документы и необходимо изучить, просматривая их один за другим.

Могут встретиться следующие случаи:

1) очередной документ релевантен – это означает, что программа отработала правильно, а эксперты пропустили документ и не включили его в рубрику

2) для очередного документа непонятно, должен ли он включаться в рубрику – необходимо задать дополнительные вопросы по поводу правил экспертного рубрицирования

3) очередной документ явно нерелевантен.

Для выяснения причин нерелевантности документа, нужно сравнить содержание документа с терминологическим описанием рубрики и выяснить, какие именно термины или совокупности терминов привели к проставлению этой рубрики программой.

Причинами появления нерелевантной рубрики у документа могут быть следующие:

3.1) В терминологическом описании содержится понятие без дополнительных условий, и именно по нему текст был отнесен к рубрике. Если появление

таких нерелевантных текстов по данному понятию – массовое, то в терминологическое описание рубрики необходимо добавить к этому понятию дополнительные условия, в виде тех понятий, которые также должны встретиться в тексте;

3.2) Текст приписан к рубрике на основе двух различных понятий, встретившихся в этом тексте – в терминологическом представлении рубрики была записана конъюнкция этих двух понятий. Совместная встречаемость этих понятий в тексте иногда дает анализируемую рубрику, но достаточно часто приводит к ложной рубрикации. Например, если при описании терминологической формулы для рубрики «Страховые взносы в Пенсионный Фонд» в формулу была бы включена (или случайно образовалась) конъюнкция *ПЛАТЕЖ* и *ПЕНСИЯ*, то часто эта пара понятий давала бы тексты о выплате пенсий, а не о платежах в Пенсионный фонд.

Для исправления возникшей ситуации могут быть сделаны следующие шаги:

3.3) Возможно, можно обойтись без данной пары понятий в конъюнкции терминологического описания. Конъюнкции для каждого понятия из пары нужно сделать уже, не включая неудачную пару.

3.4) Возможно данную пару понятий нужно уточнить дополнительными условиями, т.е. превратить конъюнкцию из пары в тройку

3.5) Возможно, из этих двух понятий нужно образовать более длинный термин. Так, мы пытались сделать терминологическое описание для рубрики «НАЛОГ НА ПРИОБРЕТЕНИЕ АВТОТРАНСПОРТНЫХ СРЕДСТВ», как конъюнкцию *налог* + *приобретение* + *автотранспортное средство*, но затем пришли к выводу, что наилучший результат автоматическое рубрицирование даст, если ввести в тезаурус такой длинный термин и построить терминологическое описание данной рубрики на базе этого термина.

3.6) Ложную рубрику дает неправильно разрешенная многозначность термина, как это было с термином *журнал* для рубрики «ГАЗЕТЫ, ЖУРНАЛЫ» или термином *единый налог* для рубрики «УЧЕТ И ОТЧЕТНОСТЬ ПО ЕСН». Если явление массовое, то может помочь внесение в тезаурус дополнительных однозначных терминов, содержащих обнаруженный многозначный термин, в качестве составной части, например, *журнал учета, кассовый журнал* и т.п.

3.7) Возможно, что текст нерелевантен, потому что существует правило, о том, что такого рода тексты должны относиться к другой рубрике. Данное правило может быть записано в списке правил нормативно-правового рубрицирования.

3.8) Несмотря на все предпринятые усилия, может сохраняться явление так называемой ложной корреляции, когда одна и та же пара терминов в тексте иногда дает правильную рубрику, а иногда нет.

Так, например, при анализе результатов автоматического рубрицирования для рубрики «СОЗДАНИЕ, РЕОРГАНИЗАЦИЯ И ЛИКВИДАЦИЯ ТАМОЖЕН И ТАМОЖЕННЫХ ПОСТОВ» была выявлена группа явно нерелевантных документов, полученных при автоматическом рубрицировании, когда создаются или ликвидируются склады, комиссии, зоны при таможнях.

С этой проблемой очень трудно бороться, однако в наших экспериментах она встречается только в примерно 3% рубрик.

Заключение

В сложных задачах рубрикации существенным становится взаимодействие с экспертами заказчика, так как единственным способом решения задачи рубрикации является итерационное уточнение правил рубрицирования, удовлетворяющих заказчика, в том числе помощь заказчику в прояснении данных правил.

Для этих целей можно применять методы, основанные на знаниях, которые позволяют легко интерпретировать, почему такой-то документ был отнесен к рубрике. Основным недостатком этих методов является высокая трудоемкость, обусловленная необходимостью привлечения экспертов для составления таких правил. Однако, представляется, что это неизбежно, поскольку в реальных задачах рубрикации отмечена значительная непоследовательность исходных данных ручной рубрикации.

Рекомендуется использовать автоматическое рубрицирование в следующих ситуациях:

1) при первичном приписывании рубрик экспертами для повышения полноты рубрицирования:

- предполагается, что эксперт не забудет поставить ту рубрику, которую считает правильной;
- результаты автоматического рубрицирования позволят значительно поднять показатели полноты рубрицирования, поскольку доказательно продемонстрируют наличие в документе и других рубрик;
- при этом целесообразно уделить особое внимание ориентации автоматической обработки на работу в человеко-машинном комплексе – обеспечить высокие показатели точности для рубрик, получающих больший вес;
- эксперт, оценивая результаты машинной обработки, должен иметь возможность руководствоваться четырехзначной логикой – согласен, не согласен, не уверен, не рассматривал.

2) при массовой обработке поступающих документов (смене рубрикатора) с минимальным участием экспертов:

- эксперты обрабатывают только наиболее важные документы;
- остальные документы проходят только машинную обработку;
- правка машинной рубрикации проходит уже на массиве обработанных документов.

Литература

[1] Debole F. & Sebastiani F., An Analysis of the Relative Hardness of Reuters-21578 Subsets // Journal of the American Society for Information Science and Technology, 2004 (http://faure.isti.cnr.it/~fabrizio/Publications/JASIST04.pdf)

[2] Lewis D.D. & Sebastiani F., Report on the Workshop on Operational Text Classification Systems (OTC-01) – SIGIG-2001, New Orleans (http://www.acm.org/sigir/forum/F2001/textClassification.pdf)

[3] Dumais S.T., Lewis D.D. & Sebastiani F., Report on the Workshop on Operational Text Classification Systems (OTC-02) – SIGIR-02, Tampere, Finland (http://www.sigir.org/forum/F2002/sebastiani.pdf)

[4] Journal of Economic Literature Classification System (www.aeaweb.org/journal.html).

[5] Joachims T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features // Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.

[6] Workshop on Operational Text Classification Systems (OTC-03) – The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD03), (http://www.daviddlewis.com/events/otc2003/)

[7] (www.daviddlewis.com/resources/testcollections/reuters21578/)

[8] Adding a New Dimension to Search: The Teoma Difference is Authority (http://sp.teoma.com/docs/teoma/about/searchwithauthority.html)

[9] Wasson M., Classification Technology at LexisNexis // SIGIR 2001 Workshop on Operational Text Classification.

[10] Антонов А.В., Пример задачи поиска «жизненных историй» - НТИ, Серия 1. – 2003. - № 7 – С.12-17. (http://www.viniti.ru/cgi-bin/nti/nti.pl?action=show&year=1_2003&issue=7&page=12)

[11] Добров Б.В., Лукашевич Н.В., Построение и использование тематического представления содержания документов // V национальная конференция с международным участием "Искуственный интеллект-96", Казань, 1996, Том I, С.130-134.

[12] Добров Б.В., Лукашевич Н.В., Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту. КИИ-2002. 7-12 октября 2002, Коломна – М.: Физматлит – Т.1 – С.178-186

[13] Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н., Технологические аспекты организации доступа к разнородным информационным ресурсам в университетсткой информационной системе РОССИЯ // Электронные библиотеки – 2002 – Том.5 – Выпуск 2

[14] Лукашевич Н. В., Салий А. Д., Представление знаний в системе автоматической обработки текстов // НТИ. Сер.2. 1997 No 3.

[15] Маковский А.Л., Новиков Д.Б., Силкина А.В., Симбирцев А.Н., Принципы построения системы классификации правовых актов // Правовой классификатор и правовой тезаурус с законотворчестве и юридической практике / Сост. В.Б.Исаков и др. – М., ГД РФ; Изд-во Гуманитарного университета – 1998. - С.5-28.

[16] «О классификаторе правовых актов» - Указ Президента РФ №511 от 15 марта 2000г.

Improvement of Operational Text Categorization Results with Visual Concept Query Refinement

Mikhail S. Ageev, Boris V. Dobrov, Natalia V. Loukachevitch

For complex operational text categorization tasks only interactive tehniques are possible. In the paper we present the method for improvement of automatic text categorization results using a visual tool that estimates the concept representation of difference between manual and automatic procedures.

♣ Работа частично выполняется при финансовой поддержке РФФИ, грант № 03-01-00472.