УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 150, кн. 4 Физико-математические науки *2008*

УДК 004.912+004.822+004.855.5

АВТОМАТИЧЕСКАЯ РУБРИКАЦИЯ ТЕКСТОВ: МЕТОДЫ И ПРОБЛЕМЫ

М.С. Агеев, Б.В. Добров, Н.В. Лукашевич

Аннотация

Статья посвящена рассмотрению трех основных технологий рубрикации текстов (ручного рубрицирования, рубрицирования на базе знаний, рубрицирования на базе машинного обучения), описываются их преимущества и возникающие проблемы. Рассматриваются два метода автоматической рубрикации текстов, направленные на преодоление недостатков существующих методов, приводятся данные о результатах их оценки на общедоступных коллекциях. Первым методом является метод, основанный на большом лингвистическом ресурсе – тезаурусе РуТез и комплексе автоматической обработки текстов АЛОТ. Второй метод представляет собой метод машинного обучения, порождающий описания рубрик в виде булевских формул.

Ключевые слова: обработка документов, автоматическая рубрикация, тезаурус, машинное обучение.

Введение

Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией. Распространенность больших информационных коллекций делает необходимым развитие автоматических методов рубрикации.

Известны две основные технологии автоматической рубрикации:

– методы, основанные на знаниях (также именуемые «инженерный подход»), при применении которых правила отнесения текстов к рубрикам строятся инженерами по знаниям в форме булевских выражений, правил продукций и т. п.

– методы на основе машинного обучения, при применении которых используется коллекция документов, предварительно отрубрицированная человеком. Алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества отрубрицированных текстов.

В настоящее время можно наблюдать рост научных работ, посвященных применению методов машинного обучения для автоматической рубрикации текстов. Приводятся высокие оценки результатов работы таких методов [1–4].

Однако, как отмечалось в ряде работ [5–9], для больших рубрикаторов – 500 и более рубрик – из-за трудности формирования качественной непротиворечивой обучающей коллекции единственно работающим подходом в настоящее время является так называемый «инженерный» подход [9–11], подразумевающий ручное описание смысла каждой рубрики. Например, в компании Рейтер, предоставляющей текстовые коллекции, на которых продемонстрированы многие высокие результаты технологий машинного обучения, используется технология, сочетающая работу системы автоматической рубрикации, основанной на знаниях, с последующим просмотром редакторами [8].

В данной работе мы проведем анализ трех основных технологий рубрикации текстов (ручного рубрицирования, рубрицирования на базе знаний, рубрицирования на базе машинного обучения), рассмотрим их преимущества и возникающие проблемы.

Мы опишем также две технологии автоматической рубрикации текстов, направленные на преодоление недостатков автоматических методов рубрикации, и приведем данные о результатах их оценки на общедоступных коллекциях.

Первой технологией является технология, основанная на большом лингвистическом ресурсе, – Тезаурусе РуТез и комплексе автоматической обработки текстов АЛОТ [12]. Эта технология позволяет быстро настраивать систему рубрикации на новый рубрикатор и новую предметную область за счет больших объемов предварительно собранных знаний о языке и мире.

Второй технологией является метод машинного обучения ПФА [13], который автоматически порождает описания рубрик в виде булевских формул, что дает возможность применения этих результатов как в качестве первичного описания для инженерных методов рубрикации, так и при использовании автоматизированных режимов при взаимодействии с экспертами предметной области.

Изложение материала построено следующим образом. В первом разделе опишем существующие методы оценки качества автоматической рубрикации текстов и проанализируем достигнутые результаты на общедоступных текстовых коллекциях. Во втором разделе для каждой технологии рубрикации проведем анализ основных проблем, влияющих на качество рубрикации. В третьем разделе представим разрабатываемые нами технологии автоматической рубрикации, направленные на преодоление проблем существующих подходов рубрикации.

1. Методы рубрикации текстов и их оценка

Оценка качества автоматической классификации производится путем сравнения с эталонной («правильной») классификацией набора документов. Для оценки систем автоматической классификации в качестве эталона используется коллекция документов, отрубрицированных вручную.

Для оценки эффективности работы систем рубрицирования используются такие характеристики, как полнота и точность [14].

Полнота r (recall) – это отношение R/Q, где R – количество текстов, правильно отнесенных к некоторой рубрике, а Q – общее количество текстов, которые должны быть отнесены к этой рубрике.

Точность p (precision) – это отношение R/L, где R – количество текстов, правильно отнесенных системой к некоторой рубрике, а L – общее количество текстов, отнесенных системой к этой рубрике.

Метрика F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера для данного запроса (рубрики) вычисляется по формуле

$$F = 2 \Big/ \left(\frac{1}{p} + \frac{1}{r} \right).$$

Для оценки эффективности методов машинного обучения для задачи автоматической рубрикации текстов используются стандартные корпусы текстов, классифицированных по заданным рубрикаторам.

Рассмотрим результаты рубрикации для наиболее популярных англоязычных и русскоязычных корпусов текстов.

1.1. Исследование методов рубрикации на коллекции Reuters-21578.

Большое число исследований эффективности методов автоматической рубрикации

проводится на популярной коллекции финансовых сообщений информационного агентства Рейтер – Reuters-21578 [15], которая была специально создана для тестирования методов автоматической рубрикации текстов. Для этой коллекции характерны следующие особенности:

– тексты сообщений небольшие по величине и принадлежат узкой предметной области финансовых и биржевых новостей;

– рубрикатор, включающий 135 рубрик, относительно прост, без иерархии, причем обычно [1, 16] для тестирования используются лишь 10 наиболее частотных рубрик;

– присвоение рубрик проводилось с контролем качества работы экспертов. В частности, 40% из имеющихся 21578 документов не рекомендуются к использованию из-за того, что присвоение рубрик к ним признано некачественным. Оставшиеся 12902 документа помечены как «качественно отрубрицированные».

Для 10 наиболее частотных рубрик коллекции Reuters-21578 результаты применения машинного обучения весьма высоки – в среднем около 84% F-меры. Сравнительные исследования эффективности методов машинного обучения на коллекции Reuters-21578 [1, 2, 4] показали, что наиболее эффективным методом является метод опорных векторов SVM по сравнению с методами Байеса, ближайших соседей, Rocchio, деревьев решений C4.5, нейронных сетей, байесовских сетей.

Для выяснения эффективности методов на полном рубрикаторе мы провели эксперименты [17] по классификации коллекции Reuters-21578 методом SVM [2]. На первых 10 рубриках результаты аналогичны опубликованным в [2] (табл. 1). Однако, уже на 11-й рубрике результаты значительно ниже. В среднем по 50 наиболее частотным рубрикам значение F-меры составляет 56%.

В 2004 г. Ф. Деболь и Ф. Себастьяни [16] опубликовали детальное исследование качества классификации коллекции Reuters-21578 в зависимости от используемого алгоритма машинного обучения, подмножества рубрик и способа усреднения оценок. Оказалось, что:

1) выбор способа оценки и множества рубрик влияет на результат сильнее, чем выбор метода машинного обучения;

2) качество классификации частотных рубрик значительно выше, чем низкочастотных;

3) усреднение по парам документ-рубрика (микроусреднение) [14] дает более высокий результат, чем усреднение по рубрикам (макроусреднение) – этот вывод формально следует из предыдущего, так как высокочастотные рубрики дают больший вклад в микроусредненную метрику, чем макроусредненную;

4) лучший результат для 90 рубрик – всего около 50% F-меры в среднем по рубрикам.

Таким образом, при детальном рассмотрении системы рубрикации, основанные на машинном обучении, имеют серьезные проблемы даже на относительно простом рубрикаторе: 50% F-меры означает, что только половина документов получила правильные рубрики.

1.2. Исследование методов рубрикации на коллекции РОМИП. Среди российских исследователей популярным способом оценки эффективности систем автоматической рубрикации текстов является участие в Российском семинаре по методам информационного поиска РОМИП (http://romip.ru).

В течение 2003–2007 гг. в дорожках классификации РОМИП использовались 5 коллекций документов и три рубрикатора объемом 160–240 рубрик:

– Сайты интернет: narod.ru ($\sim$700000 документов), DMOZ ($\sim$300000 документов) и by.web ($\sim$1500000 документов).

Табл. 1

Результаты классификации коллекции Reuters-21578 различными алгоритмами машинного обучения для наиболее частотных рубрик. Для каждой рубрики приведены количество положительных примеров, результаты различных модификаций алгоритма SVM [2, 6, 7] и результат алгоритма построения формул [18]

Номер по числу примеров	Рубрика	Количество примеров	Джоахимс (Joachims) p/r b.p.*	Дюма и др. (Dumais et al.) p/r b.p.	Наш SVM	ПФА
1	earn	3964	98.20	98.00	97.79	90.70
2	acq	2369	92.60	93.60	95.69	82.01
3	...	2108			83.72	56.06
4	money-fx	717	66.0	74.50	72.83	58.54
5	grain	582	91.30	94.60	89.00	88.89
6	crude	578	86.00	88.90	82.82	69.31
7	trade	486	69.20	75.90	77.45	64.52
8	interest	478	69.80	77.70	75.57	56.59
9	ship	286	82.00	85.60	74.55	69.60
10	wheat	283	83.10	91.80	89.59	89.74
11	corn	237	86.00	90.30	86.31	90.32
12	dlr	175			69.81	51.79
13	money-sup	174			74.01	48.54
14	oilseed	171			65.96	78.57
15	sugar	162			88.54	85.37
16	coffee	139			92.72	91.80
17	gnp	136			83.57	75.56
18	veg-oil	124			77.56	70.97
19	gold	124			64.48	61.54
20	soybean	111			61.56	74.70
21	nat-gas	105			61.03	44.44
22	bop	105			69.13	53.52

*p/r b.p. (precision/recall breakeven point) – точка равновесия, в которой полнота примерно равна точности. Используется как мера качества классификации, аналогичная F-мере.

– Нормативно-правовые документы РФ: 2004–2006 гг. – ∼64000 документов, 2007 г. – ∼300000 документов.

Задачи автоматической рубрикации текстов РОМИП имеют следующие особенности:

– коллекции документов и рубрикаторы имеют широкий спектр тематики;

– значительное число рубрик;

– для оценки рубрики присваиваются документам большим количеством экспертов, зачастую с низким контролем качества.

Оценка результатов автоматической классификации документов по коллекциям РОМИП приведена в табл. 2. Для каждой из дорожек, проводившейся в 2003–2006 гг., приведено количество «прогонов», то есть фактически алгоритмов (включая вариации параметров), примененных участниками для выполнения дорожек. Приведены также наибольший и наименьший показатели метрики F-мера, усредненной по рубрикам.

Участники дорожек классификации РОМИП 2003–2007 гг. применяли разные методы машинного обучения: SVM (во множестве вариаций, с оптимизацией различных параметров), нейронные сети, ПФА, некоторые модификации метода Rocchio [19, 20].

Табл. 2

Результаты участников дорожек классификации РОМИП 2003–2006 гг.

Дорожка	Год	Прогоны	$\max F$	$\min F$
Классификация сайтов narod.ru	2003	5	**45%**	11%
Классификация нормативных документов	2004	9	30%	7%
	2005	32	43%	7%
	2006	26	**47%**	7%
Классификация сайтов DMOZ	2005	8	30%	15%
	2006	5	**48%**	9%
Классификация страниц DMOZ	2005	8	35%	5%
	2006	7	**54%**	2%

Анализ публикаций РОМИП, подкрепленный нашим собственным опытом участия, позволяет интерпретировать приведенные данные как оценку сложности задачи классификации следующим образом:

– достигнуть результатов на 20% ниже максимума можно, применяя вполне стандартные методы и алгоритмы классификации;

– для получения более высоких результатов при помощи методов машинного обучения требуются существенные дополнительные усилия по оптимизации параметров алгоритмов, глубокой обработке исходных документов и настройке на особенности конкретной задачи.

Легко видеть, что приведенные результаты – около 50% F-меры в среднем по всем рубрикам – характерны также и для коллекции Reuters-21578.

2. Проблемы методов классификации текстов

Традиционно считается, что несоответствие результатов автоматической классификации ожидаемым, разумным критериям соответствия документов рубрикам вызвано несовершенством самих методов автоматической классификации. Данное предположение является основной мотивацией для разработки более совершенных моделей представления текста и методов автоматической классификации.

Однако определение основной тематики текста и выбор адекватных рубрик является сложной проблемой и для человека. Трудность ручного рубрицирования и неоднозначность выбора адекватных рубрик является проблемой, порождающей многие проблемы автоматического рубрицирования.

Поэтому сначала мы рассмотрим проблемы ручного рубрицирования, а затем перейдем к описанию проблем автоматических методов рубрицирования.

2.1. Проблемы ручного рубрицирования. Характерными особенностями ручного рубрицирования являются:

– высокая точность рубрицирования (как показывает практика, процент документов, в которых проставлена явно неправильная рубрика, мал);

– низкая скорость обработки документов;

– низкая полнота рубрицирования.

Обычно специалисты по рубрикации проставляют рубрики, характеризующие основное содержание документа, хотя документ может быть отнесен и к ряду других рубрик. В результате получается, что при сравнении результатов рубрикации разными экспертами одних и тех же документов процент совпадения проставленных рубрик может оказаться весьма низким – 60%, то есть похожие документы могут получить достаточно разные наборы рубрик. Такая ситуация усугубляется при увеличении величины и иерархической сложности рубрикатора.

Непоследовательность ручного рубрицирования становится серьезной проблемой для настройки разного типа систем автоматического рубрицирования, поскольку затрудняется построение формальных правил отнесения документов к той или иной рубрике.

Представляется, что основными причинами непоследовательной работы экспертов-индексаторов при рубрицировании по большим классификаторам является:

1) сложность ориентации в большом классификаторе (эксперт может не знать или забыть о существовании более близкой по смыслу рубрики);

2) неуверенность эксперта, который обычно является специалистом по ограниченному кругу вопросов, при необходимости принимать точное решение по вопросам, в которых он менее компетентен (например, специалист по строительству будет менее компетентен в вопросах финансов и наоборот). В этом случае эксперт может поставить более широкую рубрику (что не очень плохо), ошибочную рубрику или не ставить на всякий случай никакой рубрики;

3) сложность в принятии решения о важности/неважности побочных тем для содержания документа;

4) наличие неформализованных ограничивающих правил рубрицирования.

Суть последней 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.

Таким образом, на наш взгляд, создание достаточно большой, последовательно отрубрицированной текстовой коллекции является серьезной организационной проблемой.

2.2. Проблемы методов машинного обучения. При разработке системы автоматической рубрикации, основанной на машинном обучении, необходима коллекция документов, размеченная экспертами по рубрикам. Для эффективного обучения рубрицированию по большому рубрикатору требуется большее число размеченных документов. Важной особенностью такой размеченной коллекции является то, что разметка должна быть выполнена последовательно, то есть эксперты должны применять одни и те же принципы отнесения текстов к рубрике, чтобы похожие документы получали похожие рубрики.

Однако для многих возникающих на практике задач, где требуется автоматическая классификация текстов, коллекция классифицированных документов либо отсутствует, либо имеет недостаточный объем. В этом случае методы машинного обучения неприменимы, и затраты на создание обучающей коллекции адекватного объема весьма высоки. Кроме того, при низкой степени согласованности проставления рубрик методы машинного обучения дают весьма низкие результаты.

Проблема создания обучающей коллекции достаточного объема и качества обостряется с увеличением количества рубрик. Распределение количества документов по рубрикам существенно неравномерно, поэтому бóльшая часть рубрик содержит весьма мало документов.

Таким образом, факторами, усложняющими или делающими невозможным применение методов машинного обучения для автоматической рубрикации текстов, являются следующие:

• множество примеров отсутствует и не может быть создано в короткое время;

• множество примеров существует, но при их создании отсутствовали требования к качеству, например, документы отрубрицированы их авторами, то есть людьми, которые не имеют согласованного взгляда на содержание каждой конкретной рубрики;

• множество примеров противоречиво и (или) недостаточно для большинства рубрик (очень большие классификаторы) – такая ситуация может возникнуть и при едином руководстве ручной рубрикацией;

• множество примеров для обучения взято из близкой, но другой коллекции.

Кроме того, попытки использования методов рубрикации, основанных на машинном обучении, в автоматизированных режимах с участием экспертов-индексаторов сталкиваются с проблемой плохой объяснимости результатов машинного обучения, невозможностью продемонстрировать эксперту конкретные слова или словосочетания, которые привели к выбору данной рубрики.

2.3. Проблемы автоматического рубрицирования с использованием экспертного описания рубрик. К достоинствам методов, основанных на знаниях, относится высокая эффективность и «прозрачность» алгоритма – результаты обработки легко интерпретировать, то есть понять, почему документ был отнесен к данной рубрике. Для реализации этих методов фактор непоследовательного рубрицирования коллекции не является существенным. Основным недостатком этого класса методов является высокая трудоёмкость описания рубрик — до 8 человекочасов на одну сложную рубрику [9].

Проблемы автоматического рубрицирования с использованием «инженерного подхода» связаны со следующими обстоятельствами:

• для автоматической рубрикации нужно вручную создать образ рубрики как некоторое выражение на основе слов и (или) терминов реальных текстов, неполный учет вариантов употребления слов в тексте может привести к проблемам автоматической рубрикации;

• при автоматической обработке конкретных текстов могут возникнуть достаточно серьезные проблемы анализа языкового материала, контекста употребления того или иного слова, требующие привлечения обширных знаний о языке и предметной области, которые очень трудно описать в действующих программных системах автоматической рубрикации.

Так, серьезной проблемой, приводящей к появлению ложных рубрик или нехватке правильных рубрик, является *многозначность слов*, то есть употребление слова в тексте не в том значении, на которое рассчитывал эксперт, составляя образ рубрики.

Еще одной неприятной проблемой является так называемая проблема *ложной корреляции*. Ложная корреляция может возникнуть в случаях, когда для отнесения текста к рубрике необходимо присутствие в тексте двух логических элементов. Например, для рубрицирования по рубрике «Экономические реформы» необходимо присутствие в тексте двух тематических элементов: темы экономики и темы реформы. Ложная корреляция и, соответственно, неправильное отнесение текста к данной рубрике возникает в тех случаях, когда такие тематические элементы присутствуют в тексте, но не имеют отношения друг к другу. Например, такая ситуация может произойти, если в тексте речь шла о судебной реформе и были упомянуты некоторые экономические вопросы.

Сложной является и ситуация, которую можно обозначить как *рубрикация по несущественному элементу*. Текст отнесен к рубрике по слову или словосочетанию, которое по сути соответствует содержанию рубрики, но в данном тексте это

опорное слово или словосочетание употреблено случайно или в каком-то специфическом контексте, из-за чего текст становится нерелевантным рубрике. Например, текст может быть ошибочно отнесен к рубрике «Средства массовой информации» на основе следующего фрагмента: «*Около 40 человек умерли во Франции в результате установившейся в стране жары... Правительство и* ***средства массовой информации*** *следят за ситуацией...*».

Таким образом, при инженерном подходе к рубрикации после создания образов рубрик необходимо проводить несколько этапов тестирования сделанных описаний рубрик.

3. Методы преодоления проблем автоматической рубрикации

Описанные проблемы классификации текстов могут быть частично решены за счет сочетания различных подходов и использования взаимодополняющих преимуществ различных методов.

Нами развиваются следующие подходы к построению систем автоматической рубрикации:

• оптимизация «инженерного подхода» – повышение скорости и качества описания рубрик – за счет накопления знаний о языке (употребляемые термины, их синонимы, разные значения слов) и мире (отношения между терминами) в лингвистическом ресурсе, не зависимом от конкретного рубрикатора;

• использование машинного обучения в сочетании с «инженерным подходом»: для автоматизации описания рубрицирующих правил, уточнения правил описания рубрик, определения расхождений между описаниями экспертов.

Опишем эти технологии подробнее.

3.1. Технология рубрицирования, основанная на экспертном описании рубрик. Для решения задачи рубрицирования по большим классификаторам применяется комплекс из нескольких компонентов [11, 12]:

– большой лингвистический ресурс – Тезаурус по общественно-политической тематике, специально предназначенный для автоматической обработки и автоматических выводов о содержании текста;

– программный комплекс АЛОТ (Автоматической Лингвистической Обработки Текста), позволяющий получать тематическое представление содержания текста;

– технология описания смысла рубрики посредством понятий тезауруса.

3.1.1. База знаний. Общественно-политический тезаурус РуТез (далее — Тезаурус РуТез) является основой тематического анализа в рамках АЛОТ.

Общественно-политический тезаурус как ресурс для автоматической обработки текстов обладает следующими основными особенностями.

Во-первых, Тезаурус РуТез – это иерархическая сеть понятий, которая включает значительно больше понятий, терминов, отношений, синонимов, чем традиционные тезаурусы для ручного индексирования. В настоящее время Тезаурус РуТез содержит 90 тыс. терминов, 34 тыс. понятий, связанных более чем 133 тыс. тезаурусных отношений. Во-вторых, важнейшей особенностью является интеграция Тезауруса РуТез в процесс автоматической обработки текстов, что позволяет организовать обратную связь, анализируя результаты обработки.

Понятийная сеть Тезауруса РуТез включает до 10 уровней иерархии. Ценным свойством Тезауруса является возможность использования транзитивности иерархических отношений (с учетом иерархии в Тезаурусе РуТез описано 850 тыс. отношений, то есть в среднем каждое понятие связано с 28 другими).

3.1.2. Описание смысла рубрики понятиями тезауруса. При создании образа рубрики каждая рубрика R описывается дизъюнкцией альтернатив, каждый дизъюнкт D_i представляет собой конъюнкцию:

$$R = \bigcup_i D_i = \bigcup_i \Big[\bigcap_j K_{ij}\Big] = \bigcup_i \Big[\bigcap_j \Big(\bigcup_k d_{ijk}\Big)\Big]. \quad (1)$$

Конъюнкты K_{ij}, в свою очередь, описываются экспертами с помощью так называемых «опорных» понятий Тезауруса РуТез d_{ijk}. Для каждого опорного понятия задается правило его расширения $f(\cdot)$, определяющее, каким образом вместе с опорным понятием учитывать подчиненные ему по иерархии понятия. Выделяется несколько способов расширения: без расширения, полное расширение по дереву иерархии Тезауруса РуТез, расширение только по родовидовым связям, расширение по всем связям по иерархии вниз на один шаг.

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

Отметим, что некоторые отношения в Тезаурусе РуТез снабжены пометкой «аспект», что при автоматическом расширении ведет к приостановке флага «необходимость подтверждения» – рубрика не будет выводиться для текста при наличии только «неподтвержденных» понятий, при наличии же подтверждения – подтвержденные понятия учитываются в полной мере.

Следует подчеркнуть, что в данной методологии достаточно хранить только опорные понятия, полное же описание рубрики может быть каждый раз пересчитано заново при изменении Тезауруса РуТез.

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

3.1.3. Автоматическое рубрицирование на базе построения тематического представления текста. Значимость термина для содержания текста определяется в результате построения так называемого *тематического представления текста*, слабо зависящего от величины и типа текстов [11].

Тематическое представление текста моделирует тематическую структуру текста посредством объединения близких по смыслу терминов текста в так называемые тематические узлы. Каждый тематический узел имеет центр тематического узла (обычно наиболее частотный термин или термин из заглавия документа). Тематические узлы подразделяются на основные тематические узлы, соответствующие основной теме документа, и локальные тематические узлы, соответствующие темам отдельных фрагментов документа (подробнее см. [11, 12]).

В зависимости от того, элементом какой структуры тематического представления оказывается понятие d Тезауруса РуТез, формируется оценка значимости $\omega(d; D)$. Окончательно вес понятия для текста определяется добавлением стабилизирующего фактора, учитывающего частотность понятия в документе. Оценка релевантности содержания текста рубрике (вес рубрики) рассчитывается путём соотнесения документа с булевской формулой описания рубрики [11] с учётом:

– информации о весах понятий в тексте, входящих в описание рубрики (учет тематической структуры текста);

– возможной многозначности терминов текста (понятия, для которых не удалось найти подтверждения многозначного значения, получают низкий вес);

– расстояний между понятиями в тексте (рядом расположенные понятия из одного конъюнкта дают больший вклад в вес рубрики).

Алгоритм рубрицирования работает следующим образом. Для всех понятий Тезауруса РуТез, найденных в тексте, определяется множество рубрик, которые могут быть определены в тексте. Для каждой рубрики происходит расчет ее веса. В результирующем множестве остаются рубрики, вес которых превосходит задаваемый заранее для коллекции порог.

Данная технология была использована для построения систем автоматической рубрикации по большим рубрикаторам – до 3000 рубрик, в том числе:

– рубрикации правового законодательства РФ (3000 рубрик);

– рубрикации нормативных документов по президентскому классификатору нормативно-правовых актов (1100 рубрик);

– системы классификации текстов по выборной тематике (450 рубрик) и др.

3.1.4. Оценка метода на дорожке классификации web-страниц РОМИП'2007. В рамках дорожки классификации web-страниц по коллекции ROMIP.BY РОМИП'2007 мы провели эксперимент по применению технологии экспертного описания рубрик. Одной из мотиваций для данной работы было получить оценки трудоемкости для построения описания рубрикатора.

В задачу дорожки классификации РОМИП'2007 входила рубрикация по 247 рубрикам рубрикатора DMOZ сайтов белорусского интернета. Было созданы описания 234 рубрик, на что потребовалось 8 человеко-часов двух экспертов (2 эксперта по 4 часа).

Рубрики описаны в виде объединения своих основных тем – дизъюнктов: $R = \bigcup\limits_i D_i$. Для 234 рубрик описано 265 дизъюнктов.

Каждый дизъюнкт представляется конъюнкцией $D_i = \bigcap\limits_j K_{ij}$ – всего 334 конъюнкта (см. (1)). Далее, расширением по иерархии тезауруса было получено полное представление рубрики, где уже для всех рубрик задействовано 40161 концептов (естественно, с учетом повторения) и 107897 текстовых входов.

Таким образом, заранее накопленные в Тезаурусе РуТез знания дали возможность за 8 часов работы сопоставить рубрикатору более 100 тысяч слов и выражений.

При описании рубрик классификатора эксперты в основном ориентировались на формулировку рубрики. В единичных случаях эксперты заходили на сайт dmoz.org для уточнения объема рубрики.

Тезаурус РуТез покрывает практически все предметные области, отражаемые в деловой прозе – нормативных актах, СМИ федерального уровня. Поэтому для решения задачи описания рубрик потребовалось ввести в Тезаурус РуТез дополнительно только восемь понятий для описания специфических экстремальных видов спорта.

Рассмотрим пример описания, сделанного экспертами для рубрики № 135 «Спорт – Боевые искусства».

Опорное булевское выражение состоит из одного понятия *БОЕВЫЕ ИСКУССТВА* (Е) с меткой «(Е)» полного расширения по Тезаурусу РуТез.

В состав расширенного булевского выражения входят помимо исходного следующие понятия: *АЙКИДО, ДЖИУ-ДЖИТСУ, ДЗЮДО, КАРАТЭ, САМБО, ДЗЮДОИСТ, КАРАТИСТ, САМБИСТ.*

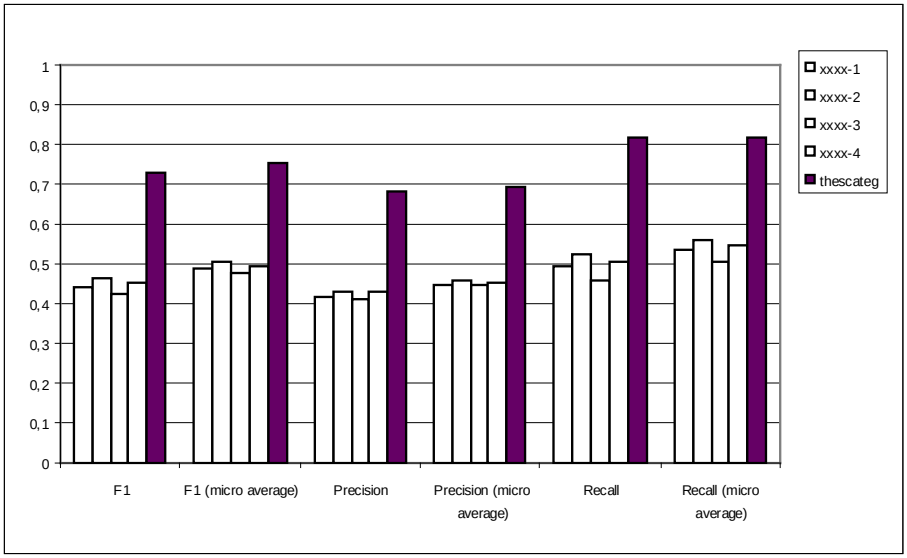

Рис. 1 РОМИП'2007: классификация веб-страниц (таблица релевантности [or])

Понятия Тезауруса РуТез (*ДЗЮДОИСТ, КАРАТИСТ, САМБИСТ*), соответствующие людям, входят в рубрику с пометкой подтверждения, поскольку появление соответствующих слов в тексте еще не означает, что текст посвящен боевым искусствам.

Рассмотрим пример более сложного описания на примере рубрики № 43 «домашний ремонт»:

(*РЕМОНТ* (N)
OR *КАПИТАЛЬНЫЙ РЕМОНТ* (N)
OR *ТЕКУЩИЙ РЕМОНТ* (N)
OR *РЕМОНТНО-СТРОИТЕЛЬНЫЕ РАБОТЫ* (N))
AND
(*ЖИЛОЕ ЗДАНИЕ* (L)
OR *ЖИЛОЕ ПОМЕЩЕНИЕ* (L)
OR *КВАРТИРА* (L)).

Здесь пометка «(L)» означает, что предусматривается только расширение по родовидовым отношениям, пометка «(N)» – отсутствие расширения.

На рис. 1 приведены результаты работы данного алгоритма по сравнению с алгоритмами других участников данной дорожки.

Отметим, что не вполне корректно сравнивать результаты нашего прогона «thescateg» с результатами других систем – из-за разницы в представлении результатов мы представили сортированный список с контролем, чтобы на каждый сайт приходилось не более пяти документов, в то время как другие участники представляли несортированные данные по пять документов на сайт.

Тем не менее достижение показателей качества рубрицирования при нестрогом согласии между экспертами (полнота = 81.7%; точность = 68.2%; F-мера = 72.9%) следует признать весьма успешным для 8 часов трудозатрат экспертов. Эти результаты превышают результаты автоматической рубрикации, полученные за всю историю РОМИП.

Полученные результаты позволяют произвести улучшение путем анализа ошибок и внесения соответствующих изменений в описание рубрик.

По информации организаторов семинара другие участники использовали модификации метода опорных векторов SVM, хорошо зарекомендовавших себя в многочисленных сравнительных экспериментах. Проведенный анализ обучающей коллекции показал, что она не является качественно размеченной коллекцией. Отнесенные к той или иной рубрике сайты могли содержать большое количество страниц, которые по содержанию к этой рубрике не относились и тем самым являлись серьезным шумовым фактором для обучения. При этом достигнутые результаты нашего инженерного подхода показали, что возможно, используя заранее описанные знания, решить задачу рубрикации, невзирая на плохо размеченную коллекцию.

3.2. Метод машинного обучения ПФА, основанный на моделировании логики рубрикатора. Методы построения классификаторов, используемые экспертами при инженерном подходе, подразумевают описание рубрики в виде правил относительно простого вида. Получаемые правила классификации имеют простой смысл и легко поддаются интерпретации.

В то же время широко используемые алгоритмы машинного обучения получают представления рубрики, которые трудно или вообще невозможно понять и интерпретировать.

Мотивацией для разработки нового метода машинного обучения была необходимость создания алгоритма машинного обучения, который строил бы правила описания рубрики, которые можно легко интерпретировать и использовать для автоматизации описания рубрик в «инженерном подходе».

Данная постановка задачи отличается от классической задачи построения автоматической процедуры классификации текстов, максимизирующей метрики качества рубрицирования — полноту и точность. В нашем случае важной метрикой качества алгоритма является также экспертная оценка соответствия полученных правил классификации смыслу рубрики.

Нами разработан алгоритм построения формул ПФА [3, 17, 18], который на основе отрубрицированной коллекции текстов строит описание рубрики в виде формул, аналогичных используемым экспертами при «инженерном подходе».

3.2.1. Описание алгоритма построения формул. Алгоритм машинного обучения ПФА [13, 18] строит формулы описания рубрики в виде булевских формул фиксированной структуры над элементами текста — словами или понятиями Тезауруса РуТез. Различные модификации алгоритма строят формулы вида:

$$1.\ U = \bigcup_{i=1}^{k} \bigcap_{j=1}^{J_i} t_{ij} \quad \text{(основной алгоритм)};$$

$$2.\ U = \bigcup_{i=1}^{k} \Big(\bigcap_{j=1}^{J_i} t_{ij} \Big\backslash \bigcup_{m=1}^{M_i} t'_{im} \Big);$$

$$3.\ U = \Big(\bigcup_{i=1}^{k} \bigcap_{j=1}^{J_i} t_{ij} \Big) \Big\backslash \Big(\bigcup_{i=1}^{k} \bigcap_{j=1}^{J'_i} t'_{ij} \Big),$$

где t_{ij}, t'_{ij} – множества документов, содержащих некоторое понятие Тезауруса РуТез (или в общем случае некоторый *терм* – элемент векторного представления документов). Конъюнкции, составляющие формулу, имеют длину J_i от 1 до 3.

Табл. 3

Примеры описания рубрик коллекции Рейтер булевскими формулами. Формулы построены методом машинного обучения ПФА на основе обучающей коллекции

Рубрика	Формула	Полнота	Точность
Gold	*OUNCE*	67%	83%
Gold	*OUNCE* OR *GOLD*	100%	50%
Coffee	*COFFEE*	100%	85%
Soybean	*SOYBEAN*	94%	62%
Wheat	*WHEAT*	99%	82%
Corn	*CORN*	100%	82%
Alum	(*ALUMINIUM* AND *TONNE*) OR *ALUMINIUM* OR *ALUMINUM* OR (*ALUMINA* AND *TONNE*) OR *ALCOA*	97%	60%

3.2.2. Экспериментальное и аналитическое исследование алгоритма. Исследование [13] основано на общедоступных коллекциях текстов, отрубрицированных экспертами по заданному рубрикатору. Результаты работы алгоритма построения формул сравнивались с результатами работы других методов классификации текстов. Были проведены следующие эксперименты.

1. Эксперименты на коллекции Reuters-21578 показали, что:

– качество классификации (F-мера) сравнимы с SVM;

– создаваемые формулы описания рубрик соответствуют содержанию рубрики.

Исследование рубрик коллекции Рейтер с помощью метода ПФА показало, что многие рубрики описываются коротким запросом – зачастую из одного-двух слов – с высокой полнотой и точностью, сравнимой по качеству классификации с SVM. В табл. 1 приведены результаты классификации данным методом. В табл. 3 приведены примеры запросов – формул описания рубрик, построенных алгоритмом ПФА.

При наличии столь простых правил отделения рубрик, как в приведенных примерах, любой алгоритм машинного обучения будет показывать достаточно высокие результаты. Однако для многих задач классификации текстов правила отделения рубрик гораздо сложнее и не могут быть описаны требованием наличия определенных ключевых слов.

2. Эксперименты на коллекции РОМИП'2004 – дорожке тематической классификации нормативных документов РФ Российского семинара по Оценке Методов Информационного Поиска 2004 года (60015 документов, 170 рубрик). Эксперименты показали, что [19]:

– алгоритм построения формул показал лучший результат по сравнению с 6 другими алгоритмами классификации текстов;

– создаваемые формулы описания рубрик соответствуют содержанию рубрики.

3. В 2007 г. с помощью метода ПФА мы проанализировали качество рубрикации обучающей коллекции задания по рубрикации РОМИП и показали массовое наличие явно нерелевантных страниц в обучающей коллекции, что позволило спрогнозировать низкие результаты технологий машинного обучения, поэтому возникла необходимость выполнения задания с помощью инженерного подхода.

Алгоритм ПФА позволяет косвенно оценить, насколько верна гипотеза о возможности описания рубрики булевской формулой над элементами языка. С одной стороны, очевидно, что если построенные формулы описывают рубрику с высокой

полнотой и точностью, то рубрика представима в виде булевской формулы. В обратную сторону данное утверждение тоже верно, что требует доказательства. В исследовании [18] доказана теорема о том, что при определенных предположениях относительно рубрики и параметрах разработанного алгоритма возможно построение описания рубрики, близкого к оптимальному. Получены оценки параметров алгоритма, при которых достигается заданный уровень полноты/точности и длины формулы.

Существенной проблемой метода ПФА является наличие высоких требований к вычислительным ресурсам на этапе поиска оптимальных конъюнктов. В настоящее время мы ведем работы по оптимизации перебора терминов на этом этапе.

Заключение

Мы рассмотрели проблемы и преимущества основных технологий рубрикации текстов: ручного рубрицирования, автоматического рубрицирования, основанного на знаниях, и автоматического рубрицирования на базе машинного обучения.

Ручное рубрицирование при больших объемах информации требует значительных финансовых и организационных затрат. Серьезной проблемой ручного рубрицирования является обеспечение последовательного рубрицирования, особенно при использовании больших иерархических рубрикаторов.

Существенным условием эффективной работы методов машинного обучения является наличие большой, качественно отрубрицированной коллекции, создание которой очень трудоемко и не всегда возможно. Кроме того, часто методы машинного обучения порождают трудно интерпретируемые результаты, что затрудняет анализ результатов экспертами и использование результатов в автоматизированных режимах.

Методы рубрикации, основанные на знаниях, требуют работы специалистов по созданию описаний рубрик, а также многоэтапной процедуры тестирования этих описаний. Однако наличие отрубрицированной коллекции для этих методов значительно менее существенно. Таким образом, при выполнении практических задач рубрикации необходимо анализировать имеющиеся ресурсы для правильного выбора подходящей технологии.

В статье также описаны реализуемые нами методы автоматического рубрицирования, которые направлены на преодоление указанных недостатков автоматических технологий.

Метод автоматической рубрикации, использующий знания, описанные в Тезаурусе РуТез, и технологию автоматической обработки текстов АЛОТ, может быть достаточно быстро настроен на новую предметную область и новый рубрикатор за счет больших объемов предварительно собранных знаний о языке и мире. Эти возможности технологии убедительно продемонстрированы при тестировании на семинаре РОМИП'2007.

Метод автоматической рубрикации ПФА использует технологию машинного обучения для порождения описаний рубрик в виде булевских формул, что дает возможность как предъявлять эти формулы экспертам предметной области, так и использовать их в качестве предварительного этапа для инженерных технологий автоматического рубрицирования.

В дальнейшем для преодоления проблем каждого отдельного подхода к рубрицированию наиболее перспективным является разработка комплексных подходов, сочетающих разные технологии. Так, ручное рубрицирование может стать более последовательным, если будет базироваться на результатах предварительной автоматической обработки. Применение инженерных технологий с контролем методов

машинного обучения даст возможность сразу учитывать особенности коллекции, позволит быстрее создавать описания рубрик и допускать меньше неточностей в описании.

Summary

M.S. Ageev, B.V. Dobrov, N.V. Loukachevitch. Automatic Text Categorization: Methods and Problems.

The paper is devoted to analysis of three techniques of text categorization (manual text categorization, knowledge-based text categorization and machine learning). Their advantages and problems are described. Two approaches are considered, intended to overcome problems of automatic text categorization. Their evaluation on public collections is presented. The first method is based on a large linguistic resource: RuThes Thesaurus and ALOT document processing technique. Another one is machine learning method of text categorization, generating descriptions of categories in form of Boolean formulas.

Key words: document processing, automatic text categorization, thesaurus, machine-learning.

Литература

1. *Dumais S., Platt J., Heckerman D., Sahami M.* Inductive learning algorithms and representations for text categorization // Proc. Int. Conf. on Inform. and Knowledge Manage. – 1998. – P. 148–155.
2. *Joachims T.* Text Categorization with Support Vector Machines: Learning with Many Relevant Features // Proc. ECML-98, 10th Europ. Conf. on Machine Learning. – 1998. – URL: http://www.cs.cornell.edu/people/tj/publications/joachims_98a.ps.gz.
3. *Lewis D.* Applying Support Vector Machines to the TREC-2001 Batch Filtering and Routing Tasks // Proc. TREC-2001 Conf. – NIST Special Publication, 2001. – P. 286–294.
4. *Yang Y., Liu X.* A re-examination of text categorization methods // Proc. of Int. ACM Conf. on Research and Development in Information Retrieval (SIGIR-99). – 1999. – P. 42–49.
5. *Ageev M., Dobrov B., Loukachevitch N.* Text Categorization Tasks for Large Hierarchial Systems of Categories // SIGIR 2002 Workshop on Operational Text Classification Systems / Eds. F. Sebastiani, S. Dumas, D.D. Lewis, T. Montgomery, I. Moulinier. – Tampere: Univ. of Tampere, 2002 – P. 49–52.
6. *Dumais S., Lewis D., Sebastiani F.* Report on the Workshop on Operational Text Classification Systems (OTC-02) // SIGIR-2002. – Tampere, Finland, 2002. – URL: http://www.sigir.org/forum/F2002/sebastiani.pdf.
7. *Lewis D., Sebastiani F.* Report on the Workshop on Operational Text Classification Systems (OTC-01) // ACM SIGIR Forum. – New Orleans, 2001. – V. 35, No 2. – P. 8–11.
8. *Rose T., Stevenson M., Whitehead M.* The Reuters Corpus Volume 1 – from Yesterday News to tomorrow's Language // Proc. of the Third Int. Conf. on Language Resources and Evaluation, Las Palmas de Gran Canaria. – 2002. – URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.956.
9. *Wasson M.* Classification Technology at LexisNexis. // SIGIR 2001 Workshop on Operational Text Classification. – 2001. – URL: http://www.daviddlewis.com/events/otc2001/presentations/otc01-wasson-paper.txt.
10. *Hayes P.J., Weinstein S.P.* Construe: A System for Content-Based Indexing of a Database of News Stories // Proc. of the Second Annual Conf. on Innovative Applications of Intelligence. – 1990. – URL: http://portal.acm.org/citation.cfm?id=653070.

11. *Добров Б.В., Лукашевич Н.В.* Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая нац. конф. по искусственному интеллекту: Труды конф. – М.: Физматлит, 2002. – Т. 1. – С. 178–186.

12. *Добров Б.В., Лукашевич Н.В.* Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всерос. конф. по электронным библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции»: Труды конф. – Петрозаводск, 2001. – С. 78–82.

13. *Агеев М.С., Добров Б.В., Макаров-Землянский Н.В.* Метод машинного обучения, основанный на моделировании логики рубрикатора // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 5-й Всерос. науч. конф. RCDL'2003 - СПб.: НИИ Химии СПбГУ, 2003. – С. 150–158.

14. *Агеев М.С., Кураленок И.Е.* Официальные метрики РОМИП'2004 // Сб. трудов «Российский семинар по оценке методов информационного поиска» / Под ред. И.С. Некрестьянова. – СПб.: НИИ Химии СПбГУ, 2004. – С. 142–150.

15. *Lewis D.* Reuters-21578 text categorization test collection. Distribution 1.0. – URL: http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt.

16. *Debole F., Sebastiani F.* An Analysis of the Relative Hardness of Reuters-21578 Subsets // Proc. of LREC-04, 4th Int. Conf. on Language Resources and Evaluation. – Lisbon, PT, 2004. – P. 971–974. – URL: http://citeseer.ist.psu.edu/691424.html.

17. *Ageev M., Dobrov B.* Support Vector Machine Parameter Optimization for Text Categorization Problems // Information Systems Technology and its Applications (ISTA'2003): Proc. Int. Conf. – 2003. – V. 30. – P. 165–176.

18. *Агеев М.С.* Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов: Дис. ... канд. физ.-матем. наук. – М., 2005. – URL: http://www.cir.ru/docs/ips/publications/2005_diss_ageev.pdf.

19. *Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В.* Экспериментальные алгоритмы поиска/классификации и сравнение с "basic line" // Сб. трудов «Российский семинар по оценке методов информационного поиска» / Под ред. И.С. Некрестьянова. – СПб.: НИИ Химии СПбГУ, 2004. – С. 62–89.

20. Труды РОМИП'2006 / Под ред. И.С. Некрестьянова. – СПб.: НУ ЦСИ, 2006. – 274 с.

Поступила в редакцию
26.02.08

Агеев Михаил Сергеевич – кандидат физико-математических наук, старший научный сотрудник Научно-исследовательского вычислительного центра Московского государственного университета им. М.В. Ломоносова.

Добров Борис Викторович – кандидат физико-математических наук, заведующий лабораторией Научно-исследовательского вычислительного центра Московского государственного университета им. М.В. Ломоносова.

E-mail: *dobroff@mail.cir.ru*

Лукашевич Наталья Валентиновна – кандидат физико-математических наук, старший научный сотрудник Научно-исследовательского вычислительного центра Московского государственного университета им. М.В. Ломоносова.

E-mail: *louk@mail.cir.ru*