Метод машинного обучения, основанный на моделировании логики рубрикатора

М.С. Агеев

ageev@mail.cir.ru
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова.
Механико-математический факультет МГУ им. М.В.Ломоносова.
АНО Центр информационных исследований.

Б.В. Добров

dobroff@mail.cir.ru
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова.
АНО Центр информационных исследований.

Н.В. Макаров-Землянский

nvmz@mail.cir.ru
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова.
Механико-математический факультет МГУ им. М.В.Ломоносова.

Аннотация

В статье описывается алгоритм классификации текстов, основанный на построении описания рубрики при помощи машинного обучения. Алгоритм строит описание рубрики в виде булевской формулы - запроса к полнотекстовой информационной системе. Приводится анализ алгоритма и сравнение его с другими методами машинного обучения. Описываемый метод может использоваться для классификации текстов, экспертной оценки содержания рубрики, оценки сложности описания рубрики.

1. Введение

В настоящее время можно наблюдать всплеск научных работ, посвященных описанию и применению методов машинного обучения для автоматической рубрикации текстов. Предложено много методов для решения данной задачи. Применение методов машинного обучения для классификации текстов очень эффективно при наличии качественно размеченной обучающей коллекции.

Однако, как отмечалось в нескольких докладах на семинаре по практической классификации текстов в рамках конференции SIGIR 2002 [5] 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, подразумевающий тщательной ручное описание смысла каждой рубрики. Что даже при известных процедурах автоматизации (редуцированного описания с использованием тезаурусов) требует высокой трудоемкости. В связи с этим, задача анализа проблем существующих методов автоматической классификации и разработка более эффективных методов является актуальной.

Методы построения классификаторов, используемые экспертами при инженерном подходе, подразумевают описание рубрики в виде правил относительно простого вида [11]. Например, в УИС РОССИЯ [1] применяются булевские формулы фиксированной структуры, в качестве элементов запроса используются термины Тезауруса РуТез [2]. 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.

В основе идеологии инженерного подхода лежит убежденность в том, что рубрикатор создается осмысленно. То есть за каждой рубрикой лежит обычно некий раздел области деятельности, который может быть представлен небольшим вербальным описанием. Мотивацией для данной работы была необходимость создать алгоритм машинного обучения, который бы моделировал смысл рубрики, составленной человеком, по результатам рубрицирования. Необходимым требованием для данного алгоритма было построение правил описания рубрики, которые можно легко интерпретировать.

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.

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

Построение алгоритма, который строит легко интерпретируемые правила описания рубрики важно с теоретической и практической точки зрения.

Наглядное описание содержания рубрики, построенное по коллекции документов, можно рассматривать в качестве краткой аннотации коллекции документов, что позволяет анализировать структуру рубрикатора и выявлять особенности множества отрубрицированных документов. Можно анализировать логику работы экспертов, которые рубрицировали документы вручную и сравнивать логику работы разных экспертов.

Алгоритм позволяет использовать наглядные описания коллекции документов в сочетании с другими методами машинного обучения. Например, можно построить описание набора документов, ошибочно классифицированных некоторым методом машинного обучения. Это позволяет анализировать причины неустойчивой работы методов машинного обучения.

Для больших рубрикаторов сложной, иерархической структуры актуальной проблемой является документирование принципов отнесения документов к той или иной рубрике. Автоматически построенные наглядные описания рубрик можно использовать в полуавтоматической процедуре составления «комментария» к рубрикатору, свободного от субъективности отдельных экспертов.

Автоматически построенное описание рубрики может также использоваться для помощи экспертам, которые составляют описания рубрик при инженерном подходе. Использование таких автоматических помощников позволяет повысить скорость и качество работы экспертов.

2. *Описание алгоритма*

Алгоритм строит описание рубрики в виде булевской формулы фиксированной структуры. Получаемая формула имеет вид:

$$D_{rubr} = \bigcup_{i=1}^{k} \bigcap_{j=1}^{j_i} l_{ij} \qquad (2.1)$$

где l_{ij} - слово в нормальной форме (лемма). Конъюнкции, составляющие формулу, имеют длину j_i от 1 до 3.

То есть подыскиваются альтернативы, в сумме описывающие смысл рубрики, каждая альтернатива может быть задана как пересечение нескольких независимых факторов.

Полученная формула используется в качестве запроса к полнотекстовой информационной системе. При этом каждому элементу формулы — лемме — сопоставляется множество документов, содержащих данную лемму (слово, равное l_{ij} после приведения к нормальной форме).

В данной работе мы использовали коллекцию документов Reuters-21578 [9] для тестирования нашего алгоритма. Результаты тестирования приводятся в разделе 3.

Далее мы опишем алгоритм построения формулы описания рубрики, который строит описание последовательно, без использования значительного по объему перебора вариантов.

Описываемый алгоритм имеет ряд параметров, которые могут влиять на скорость вычисления и качество (полноту, точность, размер) получаемых формул. При описании алгоритма мы указываем конкретные значения параметров, которые выбирались эмпирически, из физического смысла параметров и путем анализа работы алгоритма. В разделе 4 мы приводим краткий анализ влияния некоторых параметров алгоритма на результаты классификации. Подробный анализ влияния параметров алгоритма на результаты работы и тестирование алгоритма на различных коллекциях и рубрикаторах — тема для будущих наших исследований.

2.1. Преобразование текста в вектор

Первым шагом алгоритма является вычисление векторного представления текстов. Все слова, встречающиеся в документах, были приведены к нормальной форме. Слова, встречающиеся менее чем в 5 документах, были усечены. Вес слов вычислялся по формуле TF*IDF [4]. Для каждого слова вычисляются полнота, точность и F-мера описания рубрики запросом, состоящим из одного только этого слова.

Полнота и точность вычисляются на основе матрицы классификации для данной рубрики:

		документы приписаны экспертами	
		да	нет
документы приписаны автоматически	да	a	b
	нет	c	d

Полнота (recall) классификации документов по рубрике вычисляется как отношение количества документов, правильно приписанных (автоматически) к рубрике к общему количеству документов, относящихся к данной рубрике:

$$\text{recl} = \frac{a}{a+c}$$

Точность (precision) классификации документов по рубрике вычисляется как отношение количества документов, правильно приписанных (автоматически) к рубрике к общему количеству документов, приписанных к данной рубрике:

$$\text{prec} = \frac{a}{a+c}$$

F-мера (F-measure) — сводная оценка качества рубрицирования, зависящая от полноты (recl) и точности (prec) следующим образом [12]:

$$F = \frac{1+\beta^2}{\frac{\beta^2}{\text{prec}} + \frac{1}{\text{recl}}} \qquad (2.2)$$

Здесь β — параметр, устанавливающий отношение важности параметров полноты и точности. Для вычисления F-меры слов мы использовали значение $\beta = 1$.

2.2. Вычисление конъюнктов

Следующим шагом работы алгоритма является вычисление конъюнкций, состоящих из двух или трех различных слов. Целью данного шага является вычисление конъюнкций с высокими показателями полноты и точности. Алгоритм перебирает различные пары и тройки слов, и для каждого конъюнкта-кандидата вычисляется полнота и точность описания рубрики этим конъюнктом. Для того чтобы сократить перебор, вычисляются только конъюнкты, состоящие из слов с высокими показателями полноты и точности. Конкретно, берутся первые 100 слов с наибольшим показателем F-меры.

Из списка полученных конъюнктов выкидываются те, для которых количество документов, принадлежащих рубрике, ниже некоторого порога. Для коллекции Reuters-21578 мы использовали порог равный 3 документам. Кроме того, выкидываются все конъюнкты, точность которых ниже 20%.

2.3. Построение дизъюнкции

Следующим шагом алгоритма является построение формулы в виде дизъюнкции элементарных конъюнкций. В качестве элементарных конъюнкций выступают слова и конъюнкты из списка вычисленных на шаге 2.2.

Сначала выбирается первый элементарный конъюнкт - "начало" формулы. Первый элементарный конъюнкт выбирается с максимальным значением функции

$$F_{1stDisj} = \frac{1}{\frac{5}{\text{prec}} + \frac{1}{\text{recl}}} \qquad (2.3)$$

где prec и recl - точность и полнота конъюнкта соответственно. Данная формула - частный случай функции F-мера с весами (2.2). Для первого дизъюнкта точность более важна, чем полнота, поэтому мы добавили коэффициент 5 для точности.

Далее, формула наращивается постепенно новыми элементарными конъюнктами по шагам. На каждом шаге имеется текущая вычисленная формула и рассматривается возможность улучшить эту формулу при помощи добавления нового члена в дизъюнкцию. На каждом шаге вычисляется:

1.1. Список всех элементарных конъюнктов, для которых полнота и точность не ниже 5%

1.2. cntfr: количество документов, принадлежащих рубрике, которые покрыты текущей формулой

1.3. Для каждого элементарного конъюнкта из п.1 вычисляются величины:

1.3.1. addf: количество документов, не описанных текущей формулой, которые содержат данный конъюнкт

1.3.2. addfr: количество документов, не описанных текущей формулой, которые содержат данный конъюнкт и принадлежат рубрике

1.3.3. addprec=100*addfr/addf - точность конъюнкта на непокрытой части рубрики

1.3.4. addrecl=100*addfr/(cntr-cntfr) - полнота конъюнкта на непокрытой части рубрики (здесь cntr - количество документов, принадлежащих рубрике)

1.4. Наилучший конъюнкт-кандидат на добавление к текущей формуле. Лучшим считается конъюнкт с наибольшим значением функции

$$F_{BestConj} = \frac{1}{\frac{wrecl}{\text{recl}} + \frac{waddprec}{\text{addprec}} + \frac{waddrecl}{\text{addrecl}}} \qquad (2.4)$$

$$wrecl = 2,\ waddprec = 10,\ waddrecl = 5$$

Эта формула задает критерий выбора оптимального конъюнкта для добавления в формулу. В формуле содержатся три коэффициента: *вес полноты* (wrecl), *вес дополняющей точности* (waddprec) и *вес дополняющей полноты* (waddrecl).

Вес полноты влияет на "качество" добавляемых элементов, то есть их соответствие общей тематике рубрики. Если положить этот вес равным нулю, то в момент, когда формулой описано уже более 90% документов рубрики в качестве кандидатов на добавление будут попадать слова, слишком специфичные для оставшихся 10% документов. Эти слова могут не иметь отношения к тематике рубрики.

Вес дополняющей полноты влияет на скорость "сходимости" алгоритма, то есть на количество конъюнктов в результирующей формуле. Чем выше этот вес, тем короче получается формула. Если положить этот вес равным нулю, то будут добавляться новые конъюнкты, которые добавляют мало (1-10) новых документов к описанию рубрики.

Вес дополняющей точности влияет на качество классификации коллекции обучения при помощи полученной формулы. Чем больше этот параметр, тем большую точность можно будет получить при фиксированном значении полноты. С другой стороны, слишком высокое значение веса дополняющей точности может привести к переобучению, то есть к излишней подгонке формулы под заданную обучающую коллекцию.

Получаемые формулы зависят от отношения величин данных параметров, но не от абсолютных значений параметров. В разделе 4 мы приведем примеры и покажем, как описанные параметры влияют на результаты работы алгоритма.

К текущей формуле добавляется наилучший конъюнкт, вычисленный на 4 шаге.

Процесс повторяется до тех пор, пока не будет выполнено хотя бы одно из следующих условий остановки:

1. Величина addrecl для наилучшего конъюнкта равна нулю (нет улучшения полноты);
2. Количество дизъюнктов больше 20 (формула слишком сложна);
3. prec<10% и recl>90% (слишком маленькая точность);
4. recl>99% (достигнут хороший результат).

2.4. Усечение формулы

Размер получающейся формулы, полнота и точность получающегося описания и степень "подгонки" формулы под конкретную выборку документов зависят от соотношения параметров алгоритма. Наиболее важными являются параметры, встречающиеся в формулах (2.3) и (2.4).

При наращивании формулы дополнительными конъюнктами полнота растет, а точность - в целом убывает. Алгоритм останавливается тогда, когда либо рубрика покрыта практически полностью (99%), либо когда невозможно найти подходящий конъюнкт. Для получения формулы, реализующей оптимальное соотношение полноты и точности, мы усекаем полученную формулу до того места, где достигается максимум F-меры.

2.5. Модификация алгоритма: построение отрицаний

Исследовалась также работа модифицированного алгоритма, включающего построение описаний рубрики в виде

$$D_{rubr} = \bigcup_{i=1}^{k}\left(\left(\bigcap_{j=1}^{j_i} l_{ij}\right) \setminus \bigcup_{m=1}^{m_i} l'_{im}\right) \qquad (2.5)$$

Для этого после построения конъюнктов вычислялись "уточнения" полученных конъюнктов при помощи отрицания.

Модификация (2.4) является в некотором смысле альтернативой «набиранию/перебору» множества мелких частных случаев проявления рубрики.

Другой рассматриваемой модификацией алгоритма было построение описания рубрики в виде формулы следующего вида:

$$D_{rubr} = \left(\bigcup_{i=1}^{k}\bigcap_{j=1}^{j_i} l_{ij}\right) \setminus \left(\bigcup_{i=1}^{k}\bigcap_{j=1}^{j_i} l'_{ij}\right) \qquad (2.6)$$

В данном случае мы выделяем множество "ошибочных" документов, которые покрываются нашей формулой, но не принадлежат рубрике

$$(rubr') = \left(\bigcup_{i=1}^{k}\bigcap_{j=1}^{j_i} l_{ij}\right) \setminus (rubr) \qquad (2.7)$$

И рассматриваем множество $(rubr')$ как новую рубрику. К рубрике $(rubr')$ применяется алгоритм построения формул и полученная формула

$$D_{rubr'} = \left(\bigcup_{i=1}^{k}\bigcap_{j=1}^{j_i} l'_{ij}\right)$$

вычитается из описания рубрики $(rubr)$ в виде (2.1). В результате получается формула вида (2.6).

NAME	DOC_CNT	Joachims P/R b.p.	Dumais et.al. P/R b.p.	Our SVM	disj formulae
earn	3964	98,20	98,00	97,79	90,70
acq	2369	92,60	93,60	95,69	82,01
...	2108			83,72	56,06
money-fx	717	66,90	74,50	72,83	58,54
grain	582	91,30	94,60	89,00	88,89
crude	578	86,00	88,90	82,82	69,31
trade	486	69,20	75,90	77,45	64,52
interest	478	69,80	77,70	75,57	56,59
ship	286	82,00	85,60	74,55	69,60
wheat	283	83,10	91,80	89,59	89,74
corn	237	86,00	90,30	86,31	90,32
dlr	175			69,81	51,79
money-sup	174			74,01	48,54
oilseed	171			65,96	78,57
sugar	162			88,54	85,37
coffee	139			92,72	91,80
gnp	136			83,57	75,56
veg-oil	124			77,56	70,97
gold	124			64,48	61,54
soybean	111			61,56	74,70
nat-gas	105			61,03	44,44
bop	105			69,13	53,52

Таблица 1. Результаты сравнения метода построения формул с SVM на рубриках Reuters-21578.

3. Результаты

Мы провели тестирование нашего метода построения формул на коллекции Reuters-21578 [9]. Для оценки качества рубрицирования использовался стандартный метод тестирования на ModApte split коллекции Reuters-21578. Этот метод предполагает обучение на фиксированном подмножестве документов Reuters-21578 и тестирование на другом фиксированном подмножестве документов. Для работы используется подмножество, состоящее из 12902 документов, из них 9603=74% используется для обучения и 3299=26% для тестирования.

В таблице 1 приводятся сравнение результатов нашего метода с SVM (Support Vector Machines) [7, 10]. Результаты приведены для рубрик с количеством документов более 100 (столбец doc_cnt). В столбцах "Joachims P/R b.p." и "Dumais et.al. P/R b.p." приводятся результаты, опубликованные в [7] и [6] соответственно.

К сожалению, в указанных работах опубликованы результаты только для наиболее частотных 10 рубрик. В столбце "Our SVM" мы приводим результаты нашей реализации тестов SVM [3]. В столбце "disj formulae" приводятся результаты работы нашего алгоритма построения формул. Использовались формулы вида (2.1).

4. Анализ работы алгоритма

Для многих рубрик получились короткие и понятные формулы.

Например:

Рубрика	Формула	Полнота	Точность
Coffee	/Лемма="COFFEE"	100	84,85
Soybean	/Лемма="SOYBEAN"	93,94	62
Wheat	/Лемма="WHEAT"	98,59	82,35
Corn	/Лемма="CORN"	100	82,35
Alum	(/Лемма="ALUMINIUM" AND /Лемма="TONNE") OR /Лемма="ALUMINIUM" OR /Лемма="ALUMINUM" OR (/Лемма="ALUMINA" AND /Лемма="TONNE") OR /Лемма="ALCOA"	97,14	59,65

Для некоторых рубрик формулы получились весьма длинными. Например:

Рубрика	Формула	Полнота	Точность
Interest	/Лемма="02-333" OR (/Лемма="BANK" AND /Лемма="CUT" AND /Лемма="RATE") OR (/Лемма="PCT" AND /Лемма="MARKET" AND /Лемма="MONEY") OR (/Лемма="REPURCHASE" AND /Лемма="CUSTOMER") OR (/Лемма="DISCOUNT" AND /Лемма="RATE")	61,38	50,84

Результаты показывают, что в целом имеет место следующая зависимость: чем длиннее формула, тем хуже результаты. Это явление можно объяснить тем, что в случае, когда для рубрики существует простое вербальное описание, то, скорее всего, будет существовать и короткая формула, описывающая рубрику. Рубрицирование при помощи такой формулы будет давать результаты, близкие к ручному рубрицированию.

В то же время рубрики, которые сложно описать одним-двумя словами имеют менее четкие, менее формальные границы определения и описать принадлежность рубрике в виде булевской формулы сложно.

4.1. Анализ влияния различных параметров

На примере рубрики "gold" рассмотрим, как некоторые параметры алгоритма влияют на качество описания рубрики. Мы применяли алгоритм построения формулы на рубрике "gold" с различными параметрами и вычисляли метрики качества рубрицирования — полноту и точность — на коллекции обучения и на коллекции тестирования. Мы использовали стандартное разбиение на множество для обучения и множество для тестирования "ModApte split", заданное на коллекции Reuters-21578 [9].

В таблицах результатов для каждого запуска алгоритма мы приводим полученную формулу и вычисленную полноту и точность на разбиении.

Варьируя параметр "вес дополняющей полноты" в формуле (2.4) можно получать формулы различной длины:

Короткую формулу можно получить, задав следующие параметры формулы (2.4): (wrecl= 2, waddprec=10, waddrecl=15). Вес дополняющей полноты addrecl большой, поэтому получается короткая формула:

Формула	***Полнота TRAIN***	***Точность TRAIN***	***Полнота TEST***	***Точность TEST***
/Лемма="OUNCE"	67,02	82,89	53,33	72,73

Попытки повышения полноты путем простого добавления конъюнктов ведут обычно к существенному уменьшению точности:

Формула	***Полнота TRAIN***	***Точность TRAIN***	***Полнота TEST***	***Точность TEST***
/Лемма="OUNCE" OR /Лемма="GOLD"	100	50,27	100	51,72

В данном случае лемма "GOLD" дает сильное улучшение полноты (до 100%) предыдущей формулы, но при этом точность резко падает.

Можно добиться хороших результатов «набирая» дизъюнкциями частные случаи. Слово "gold" встречается в конъюнкции с другими словами. Такую формулу можно получить, установив параметр "вес дополняющей полноты" малым (wrecl=2, waddprec=10, waddrecl=0.1) и установив более жесткие условия на точность первого конъюнкта в формуле (2.3):

Формула

```
/Лемма="MINEWORKER"
OR (/Лемма="OUNCE" AND /Лемма="GOLD")
OR (/Лемма="TON"   AND /Лемма="GOLD")
OR (/Лемма="GOLD"  AND /Лемма="CONTAIN")
OR (/Лемма="GOLD" AND /Лемма="DEPOSIT")
OR (/Лемма="GOLD" AND /Лемма="UNDERGROUND")
OR (/Лемма="GRADE" AND /Лемма="GOLD")
OR (/Лемма="SILVER" AND /Лемма="SHORT")
OR (/Лемма="COIN" AND /Лемма="GOLD")
OR (/Лемма="BULLION" AND /Лемма="GOLD")
```

Полнота TRAIN	*Точность TRAIN*	*Полнота TEST*	*Точность TEST*
97,87	85,98	63,33	76,00

Здесь важно заметить, что для полученной формулы полнота и точность на тестовой коллекции документов сильно ниже полноты и точности на обучающей коллекции. То есть алгоритм слишком сильно "подгоняет" формулу под обучающую выборку.

В качестве альтернативы набору формулы из малочастотных конъюнктов мы испытали возможность построения формул с отрицанием вида (2.5) и (2.6).

Построение формулы с отрицанием вида (2.5) позволяет получить более короткую (по количеству слов) формулу с высокими результатами:

Формула

```
(
/Лемма="GOLD"
 AND NOT /Лемма="NET"
 AND NOT /Лемма="AGREEMENT"
 AND NOT /Лемма="COMMON"
 AND NOT /Лемма="ACCOUNT"
 AND NOT /Лемма="FRANCE"
 AND NOT /Лемма="BLOCK"
 AND NOT /Лемма="UNCHANGED"
 AND NOT /Лемма="BOARD"
 AND NOT /Лемма="DE"
 AND NOT /Лемма="CUT"
)
OR (
       /Лемма="OUNCE"
   AND /Лемма="ORE"
   AND /Лемма="RESOURCE" )
```

Полнота TRAIN	*Точность TRAIN*	*Полнота TEST*	*Точность TEST*
100,00	87,04	76,67	69,70

При этом имеет место большое различие результатов на коллекции обучения и коллекции тестирования.

Модификация алгоритма (2.6) дает формулу

Формула

```
(    /Лемма="OUNCE"
  OR /Лемма="GOLD"
)
AND NOT (
  ( /Лемма="CURRENCY" AND /Лемма="RESERVES")
  OR ( /Лемма="CONVERT" )
  OR ( /Лемма="ROSE" AND /Лемма="RESERVES" )
  OR ( /Лемма="STRENGTH" )
  OR ( /Лемма="RESULTED" )
  OR ( /Лемма="SPECIAL"
       AND /Лемма="RESERVE"
     )
  OR ( /Лемма="95" AND /Лемма="ROSE" )
  OR ( /Лемма="REPAYMENT" )
  OR ( /Лемма="WEEKLY" )
  OR ( /Лемма="END-FEBRUARY"
       AND /Лемма="RESERVE"
     )
)
```

Полнота TRAIN	*Точность TRAIN*	*Полнота TEST*	*Точность TEST*
100,00	61,84	96,67	60,42

Важно отметить, что в данном случае параметры качества рубрицирования не сильно различаются на коллекции обучения и коллекции тестирования.

Полученная формула отражает известный факт [8] что эксперты, рубрицирующие документы Reuters не относят тематику "золотые резервы" к рубрике gold. Соответственно, некоторые конъюнкты, содержащие слово "reserves" вычитаются из описания рубрики.

5. Будущие работы

Для решения задачи автоматической классификации текстов существует два принципиально различных подхода: методы машинного обучения и методы, основанные на знаниях (также иногда именуемые "инженерный подход").

При использовании машинного обучения для построения классификатора используется коллекция документов, предварительно отрубрицированная человеком. Задача алгоритма машинного обучения состоит в построении процедуры классификации документов на основе автоматического анализа заданного множества отрубрицированных текстов.

При использовании методов, основанных на знаниях, правила отнесения документа к той или иной рубрике задаются экспертами на основе анализа рубрикатора и, возможно, части текстов, подлежащих рубрицированию.

Отметим здесь некоторую условность названия "методы, основанные на знаниях". Любые методы автоматической классификации текстов в той или иной форме используют знания о свойствах текста на естественном языке и знания об особенностях текстов, принадлежащих той или иной рубрике. Принципиальная разница между двумя группами методов состоит в том, что методы машинного обучения используют математические методы для извлечения знаний из обучающей коллекции текстов, в то время как "инженерный подход" использует

знания эксперта о свойствах текстов, принадлежащих рубрикам. Знания эксперта основываются, в первую очередь, на предыдущем опыте, в частности, на большой коллекции прочитанных ранее текстов, и во вторую очередь, на части текстов, подлежащих рубрицированию.

В настоящее время можно наблюдать существенный разрыв в исследованиях и в практических методах между двумя указанными подходами к автоматической классификации текстов — методами машинного обучения и методами, основанными на знаниях.

В исследованиях, посвященных применению методов машинного обучения для классификации текстов, применяются универсальные алгоритмы, которые применимы для широкого круга задач анализа и обработки информации. Например, метод SVM успешно используется для задач распознавания образов и оценки плотности сред. Для задачи классификации текстов эти методы работают с абстрактной векторной моделью документа и не учитывают особенностей задачи тематической классификации текстов и структуры рубрикатора. Тем не менее, во многих случаях методы машинного обучения дают весьма высокие результаты.

Во многих случаях, даже при наличии заранее отрубрицированной коллекции документов, методы машинного обучения неприменимы и используется более трудоемкий инженерный подход. Инженерный подход обычно обеспечивает высокое качество рубрицирования и "прозрачность" алгоритма — результаты обработки легко интерпретировать (почему такой-то документ был отнесен к рубрике). К сожалению, при использовании инженерного подхода зачастую совсем не используется ресурс, состоящий в наличии коллекции отрубрицированных текстов.

Наше исследование посвящено сравнению различных методов классификации текстов, выделению положительных сторон и проблем каждого из методов. Целью данных исследований является:

- Построение модели классификации текстов, объясняющей проблемы существующих методов и позволяющей развивать методы классификации.
- Создание методов автоматической классификации текстов, сочетающих в себе преимущества методов машинного обучения и методов, основанных на знаниях.
- Улучшение существующих процедур классификации текстов, использующих инженерный подход. Создание различных помощников для автоматической проверки и коррекции описания рубрик и результатов рубрицирования.

Одним из шагов в нашем исследовании является данная статья. Мы планируем продолжить исследования в области автоматической классификации текстов. Перечислим конкретные планы по развитию и исследованию алгоритма, описанного в данной статье.

1. *Проверка алгоритма на различных коллекциях документов и различных классификаторах*.

В данной статье мы использовали свободно доступную для исследовательских целей коллекцию Reuters-21578. Эта коллекция известна тем, сто обладает довольно простой однородной структурой документов и относительно малым количеством рубрик с простыми правилами отнесения рубрик к документу. Весьма важным является тот фактор, что существует множество опубликованных работ по применению методов машинного обучения, в которых применялась коллекция Reuters-21578, так что есть возможность сравнить наши результаты с результатами других исследователей.

Проведение и публикация исследований с использованием других коллекций затруднено в связи с вопросом об авторских правах и необходимостью получения разрешений на публикацию результатов.

2. *Использование в векторном представлении различных атрибутов документа*:
   2.1. терминов Тезауруса [2]
   2.2. терминов Тезауруса с тематическим расширением по дереву тезауруса
   2.3. словоформ
   2.4. формальных атрибутов документа, таких как источник, автор.

Предполагается, что при использовании указанных атрибутов документа можно повысить качество классификации и автоматически выделять существенные для классификации признаки.

3. *Сравнение получаемых формул с логикой экспертов*.

Для этого предполагается использовать отрубрицированную экспертами коллекцию документов и описания рубрик, полученные при построении автоматического классификатора методами, основанными на знаниях.

Сравнение с логикой экспертов позволит выяснить, как нужно выбирать параметры алгоритма для того, чтобы получить близкие по смыслу формулы.

Интересно также будет проанализировать расхождения: почему признаки, которые подходят по смыслу рубрики, не обладают высокими показателями полноты/точности и наоборот.

4. *Анализ устойчивости алгоритма*.

Планируется выяснить, насколько сильно меняются формулы описания рубрик при изменении разбиения множества документов на множество для обучения и множество для тестирования. Как будут меняться формулы при уменьшении количества документов для обучения? Как будет при этом меняться полнота и точность описания рубрики?

Аналогичная работа по исследованию устойчивости метода SVM была проведена в работе [3]. Ре-

зультаты показали, что устойчивость классификации методом SVM резко падает при уменьшении количества документов, приписанных к рубрике.

5. *Влияние других рубрик на качество классификации*.

Для больших классификаторов сложной структуры характерно наличие большого количества близких по смыслу рубрик. Это отрицательно сказывается на степени согласованности результатов ручного рубрицирования для разных экспертов. Для близких по смыслу рубрик характерно большое отклонение результатов автоматической классификации методами, основанными на знаниях, от результатов ручной классификации документов.

Предполагается, что использование машинного обучения для таких рубрик позволит более четко очертить границы между рубриками и выявлять ошибки ручного рубрицирования.

6. *Сравнение представлений содержания рубрик, полученных различными методами машинного обучения*.

Несмотря на трудность интерпретации правил описания рубрик, генерируемых методами машинного обучения, некоторую информацию о структуре описания все-таки можно выделить.

Например, для линейной SVM и метода Roccio можно выделить относительные веса различных признаков документа.

Планируется сравнить описания рубрик различными методами и выявить причины, по которым тот или иной метод работает лучше.

7. *Автоматическая генерация формул с весами признаков*.

Предполагается включить в генерируемую формулу веса признаков и учитывать веса при рубрицировании, сравнивая полученный вес найденного документа с некоторым порогом.

Полученные формулы будут обобщать линейные SVM, и таким образом можно будет использовать преимущества обоих методов.

6. Заключение

В данной статье описывается алгоритм машинного обучения, который строит правила, разделяющие рубрики, в виде булевских формул — запроса к полнотекстовой информационной системе. Формулы имеют фиксированную структуру.

Для оценки качества работы алгоритма мы провели сравнение с методом Support Vector Machines (SVM). Сравнение производилось на коллекциях документов Reuters-21578. Результаты работы нашего алгоритма вполне сопоставимы с результатами работы SVM. Различие в качестве классификации зависит от свойств рубрики, конкретно, от размера получаемых формул. В статье приводится анализ причин, из-за которых возникает разница в качестве рубрикации.

Результатом работы описываемого алгоритма является формула, позволяющая относить документы к той или иной рубрике. Кроме того, данный алгоритм полезен для решения следующих задач:

- Экспертной оценки содержания рубрики, что может применяться для документирования рубрикатора – составления «комментария», свободного от субъективности отдельных экспертов.
- Оценка ошибки приближения рубрики другими методами в виде явного указания того, что не было учтено.

Список литературы

[1] Добров Б.В., Лукашевич Н.В., Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту – Коломна, 2002.

[2] Добров Б.В., Лукашевич Н.В., Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ. // Третья Всероссийская конференция по Электронным Библиотекам "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - Петрозаводск, 2001 - С.78-82.

[3] Ageev M.S., Dobrov B.V. Support Vector Machine Parameter Optimization for Text Categorization Problems. Proceedings of 2nd International Conference ISTA'2003 "Information Systems Technology and its Applications", LNI 2 GI 2003, pp. 165--176.

[4] Callan J.P., Croft W.B., Harding S.M., The INQUERY Retrieval System // Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications., pp. 78--83, 1992.

[5] Dumais S.T., Lewis D.D., Sebastiani F., Report on the Workshop on Operational Text Classification Systems (OTC-02), 2002. (http://www.sigir.org/forum/F2002/sebastiani.pdf)

[6] Dumais S., Platt J, Heckerman D., Sahami M. Inductive learning algorithms and representations for text categorization. In Proc. Int. Conf. on Inform. and Knowledge Manage., pp. 148--155, 1998.

[7] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137--142, 1998.

[8] Hayes P. Intelligent High-Volume Text Processing Using Shallow, Domain-Specific Techniques. In P. Jacobs (Ed.) Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. Lawrence Erlbaum, Hillsdale, NJ, 1992, pp 227--241.

[9] Reuters-21578 Text Categorization Test Collection

http://www.daviddlewis.com/resources/testcollections/reuters21578/
[10] Vapnik V. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.
[11] Wasson M. Classification Technology at LexisNexis. SIGIR 2001 Workshop on Operational Text Classification. (http://www.daviddlewis.com/events/otc2001/presentations/otc01-wasson-paper.txt)
[12] Yang Y. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, 1999, pp. 69--90.

Machine Learning Method for Text Categorization, Based on Modelling of Classifier's Logic

M.S. Ageev, B.V. Dobrov,
N.V. Makarov-Zemlyanskii

This article describes a machine learning algorithm for text categorization. The algorithm builds a description of each category in the form of a boolean formulae with document terms as atoms. This formulae is used as a query to a full-text information retrieval system. We analyze the algorithm and compare it with another machine learning methods. This method can be used for automatic text categorization, expert's estimation of content of category documents, and for estimation of hardness of category description.