RCO на РОМИП 2008

© Поляков П.Ю., Плешко В.В.
info@rco.ru

Аннотация

Настоящая работа является отчетом об экспериментах, проведенных в рамках семинара РОМИП 2007-2008 годов. Проведены исследования влияния способа отбора терминов в задаче классификации web-страниц и сайтов. Также были проведены эксперименты по кластеризации новостей.

1. Введение

В 2006 году система тематической классификации RCO показала хорошие результаты, улучшив показатели F-меры в среднем на 20% в относительных величинах по сравнению с наилучшими результатами, показанными в 2004-2005 годах [3]. Мы предположили, что это улучшение произошло благодаря более качественному отбору терминов, задействованных в классификации. В частности, было показано, что использование различных вариантов ядра в методе опорных векторов (SVM) [5] не оказывают заметного влияния на качество классификации. Но эти результаты были получены исключительно на коллекции нормативно-правовых документов, поэтому в данной работе будет проверено, удастся ли улучшить результат и в классификации Веб-сайтов и Веб-страниц. Другое положение, которое мы хотим проверить – как влияет на точность и полноту классификации добавление словосочетаний к списку классификационных признаков.

В дорожке кластеризации новостного потока RCO участвует впервые. Предварительные исследования показали, что в этой задаче новостной документ нельзя представлять в виде простого набора слов, необходимо учитывать их положение в тексте, взаимосвязи, семантическую нагрузку. Только в этом случае возможно качественно разбить документы на события и сюжеты. Именно этому моменту уделено внимание в данной работе.

2. Тематическая классификация веб-страниц и веб-сайтов

2.1 Постановка задачи

Участникам было предложено подмножество интернет-каталога dmoz.org (300000 страниц), используя которое в качестве обучающей выборки, требовалось соотнести с категориями каталога dmoz.org (247 категорий) страницы из домена .by (600714 страниц).

Среди особенностей задачи следует отметить «зашумленность» обучающей выборки. Если сайт из обучающей выборки принадлежал категории, то все его страницы относились к положительным примерам этой категории.

2.2 Метод классификации

Исследования проводились в рамках векторной модели представления документов. Во всех прогонах использовался только метод опорных векторов. Приняв во внимание выводы предыдущей работы [3], мы остановились на линейном ядре SVM, так как в этом случае облегчается интерпретация и настройка профилей рубрик. Линейному ядру соответствует обычный линейный классификатор вида **d*****c** > h, где **d** – вектор документа, **c** – вектор профиля рубрики, h – пороговое значение, которое должно превысить скалярное произведение упомянутых векторов для отнесения документа к рубрики. Размерность векторов равна числу терминов, задействованных в классификации. При классификации веб-сайтов вектор **d** представлял собой суперпозицию веб-страниц сайта.

Использовалась реализация SVM-Light [6] с параметрами:
b = 1, j = (число положительных примеров в обучающей выборке) / (число отрицательных примеров в обучающей выборке).

2.3 Отбор терминов

Отбор терминов проводился из набора положительных примеров для каждой из категорий. Было исследовано два способа выделения терминов:

1. Однословные термины;
2. Теоретико-множественное объединение однословных и многословных терминов.

В качестве однословных терминов выделялись все слова документа за исключением служебных частей речи, числительных и

дат. Многословные термины выделялись при помощи алгоритма синтактико-семантического анализа [1], и представляли собой простые именные группы (напр. «подоходный налог, база налогообложения»). Именные группы были усложнены включением в их структуру конструкций с предлогами в соответствии с моделями управления [2] (напр. «налог на добавленную стоимость»).

Выделенные термины подвергались фильтрации для каждой категории отдельно. Фильтрация производилась по информационной значимости термина. За основу информационной значимости термина был выбран коэффициент IG (information gain, см. например [4]):

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_i, \bar{t}_i\}} P(t,c) \cdot \log\left(\frac{P(t,c)}{P(t)P(c)}\right). \quad (1)$$

Нами были взяты только первое и четвертое слагаемые, характеризующие описательную способность термином рубрики. Фильтрация по признаку информационной значимости проводилась следующим образом:

1. Термины упорядочивались по убыванию информационной значимости;

2. Далее были отобраны первые N терминов, сумма информационной значимости которых составила 50% от общей суммы по всем терминам. Влияние числа терминов на качество результата следующее: при увеличении N качество слегка падает, а при уменьшении появляется неустойчивость (существенно усиливается разброс индивидуальных показателей рубрик).

2.4 Взвешивание терминов

В данной работе мы использовали только частотный способ взвешивания термина, так как именно он давал наилучшие результаты в предыдущих работах. При расчете весов терминов в документе создавался вектор, ненулевыми элементами которого служили частоты терминов в документе, отобранных для данной категории. Затем вектор документа приводился к единичной длине.

2.5 Описание прогонов

На оценку представлено 4 прогона обозначенные L, Lpos, LT, LTpos, где

L – обозначает однословные термины,

LT - объединение однословных с многословными,

pos - классификация документов производится только по терминам, дающим положительный вклад в показатель принадлежности документа к рубрике.

Особенность построения профиля рубрик методом SVM такова, что нередко термины, часто встречающиеся именно в документах рубрики, получают отрицательные веса в профиле этой рубрики. Такое поведение противоречит интуитивным представлениям (и Байесовой модели), согласно которым эти термины должны иметь большой положительный вес. Происходит это в SVM вследствии контрастирования границы между документами принадлежащими и непринадлежащими рубрике. Прогоны с параметром pos должны ответить на вопрос, как изменится результат классификации, если согласовать профиль с интуитивными представлениями, то есть исключить из профиля все термины с отрицательными весами.

2.6 Результаты оценки классификации веб-страниц

Традиционно, для более полного анализа, мы предоставляем не только оценки этого года, но также оцениваем результаты работы нового алгоритма по матрицам релевантности предыдущих годов. В таблицах 1-2 показаны оценки F(micro) и F(macro) для всех четырёх прогонов на матрицах релевантности 2005, 2006, 2007, 2008 годов. В последней колонке каждой из таблиц приведено наибольшее значение F-меры среди прогонов, полученных различными методами и оценённых экспертами в соответствующем году. Курсивом помечены результаты прогонов, у которых профили рубрик создавались экспертами вручную.

micro	L	Lpos	LT	LTpos	Best
2008	0.430	0.424	0.413	0.406	0.430
2007	0.363	0.375	0.361	0.374	*0.467*
2006	0.472	0.436	0.448	0.470	0.592
2005	0.409	0.390	0.411	0.431	0.514

macro	L	Lpos	LT	LTpos	Best
2008	0.404	0.401	0.375	0.381	0.404
2007	0.294	0.311	0.291	0.309	*0.443*
2006	0.300	0.300	0.272	0.329	0.296
2005	0.312	0.315	0.308	0.347	0.235

Таблица 1. F1(micro) и F1(macro) веб-страниц по матрицам релевантности 2005-2008 с сильными требованиями к релевантности

micro	L	Lpos	LT	LTpos	Best
2008	0.630	0.634	0.583	0.606	0.634
2007	0.489	0.504	0.474	0.497	*0.752*
2006	0.637	0.669	0.528	0.675	0.512
2005	0.508	0.544	0.455	0.548	0.385

macro	L	Lpos	LT	LTpos	Best
2008	0.608	0.612	0.571	0.592	0.612
2007	0.417	0.432	0.414	0.430	*0.729*
2006	0.483	0.532	0.377	0.519	0.448
2005	0.439	0.480	0.385	0.485	0.350

Таблица 2. F1(micro) и F1(macro) веб-страниц по матрицам релевантности 2005-2008 со слабыми требованиями к релевантности

Эффект перехода к более качественному (по сравнению с [3]) способу отбора терминов, используемых при составлении профилей рубрик, можно оценить, сопоставив наилучшие результаты 2005-2006 годов (колонка Best), с результатами новых прогонов (колонки L, Lpos, LT, Ltpos). Новый метод действительно даёт существенное улучшение качества классификации: порядка 30-40% в относительных величинах.

Также следует отметить, что отбрасывание терминов с отрицательным весом (переход от L к Lpos, и от LT к LTpos) заметно улучшает результат. И это настораживает, так как означает, что SVM-Light сформировал опорные вектора не оптимальным образом с точки зрения тестовой коллекции, и такая простая операция, как отбрасывание отрицательных терминов, улучшает результат на 20%. Возможно, параметры алгоритма были установлены не самым лучшим образом, или сам алгоритм работает неоптимально. В любом случае данный результат даёт основания полагать, что ещё есть возможности по усовершенствованию метода классификации.

Несмотря на то, что словосочетания улучшают качество классификации в среднем всего на несколько процентов, на отдельные рубрики они оказывают очень существенное влияние. В качестве примера возьмём рубрику «Наука->Науки_о_Земле». Прогон L (без словосочетаний) дает полноту r = 0.25 и точность p = 0.081. Прогон LT дает ту же полноту r = 0.25 и гораздо более высокую точность p = 0.214. В первом случае наиболее значимыми терминами профиля являются: БЕРЕГОВЕДЕНИЕ, РЫБИНСКИЙ, СПУТНИКОВЫЙ, КРСУ, ГЕОФИЗИЧЕСКИЙ, ГЕОЛОГИЧЕСКИЙ. Во втором случае – это СПУТНИКОВЫЕ ДАННЫЕ, БЕРЕГОВЕДЕНИЕ, РЫБИНСКОЕ ВОДОХРАНИЛИЩЕ, ГЕОФИЗИЧЕСКОЕ ПОЛЕ, ГЕОЛОГИЧЕСКИЙ,

ИССЛЕДОВАТЕЛЬСКАЯ ПОДВОДНАЯ ЛОДКА, КРСУ, ОЧАГ ЗЕМЛЕТРЯСЕНИЙ. Очевидно, словосочетания – это более узкие термины, с помощью них можно более точно описать область знаний. Однако, если в обучающей выборке интересующая область знаний освещена однобоко, с какой-то одной стороны, в этом случае словосочетания только ухудшат качество классификации, так как система будет больше внимания уделять специальным терминам, которые не задействованы в остальных подмножествах данной области знаний.

Результаты наших прогонов за 2007 год можно сопоставить с результатами классификации, в основе которой лежат профили рубрик, составленных экспертами вручную (курсив в последней колонке таблицы). За счёт чего профили экспертов выигрываеют в 1.5 раза? Чтобы ответить на этот вопрос, мы построили график зависимости полноты от точности классификации отдельных рубрик для прогона LTpos и прогона с экспертными рубриками (Manual) на рис 1. Соотношение полноты-точности в обоих случаях одинаковое, но разброс значений для автоматической классификации сильнее.

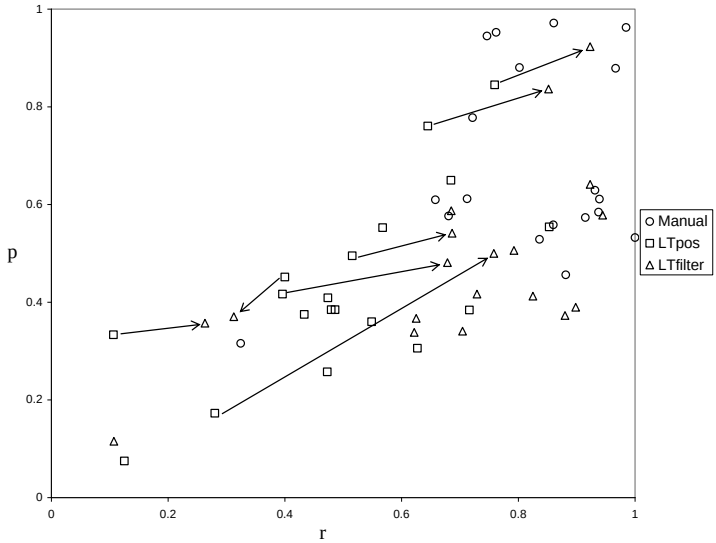

Рис. 1. Зависимость точности классификации рубрик от полноты со слабыми требованиями к релевантности. Manual – профили составлены экспертами вручную, LTpos – автоматически, LTfilter – откорректированные автоматические профили.

Часть точек прогона LTpos лежат в том же диапазоне, что и для прогона с экспертными рубриками, но в то же время имеется много провальных рубрик. Низкое качество классификации этих рубрик, возможно, объясняется неадекватностью обучающей выборки. Посмотрим, какие рубрики оказались провальными. В таблицах 3-4 представлены название и F-мера каждой из тестируемых рубрик для прогонов LTpos и Manual. Верхние рубрики таблицы удалось хорошо классифицировать автоматическими методами. Среди них мы видим прежде всего темы , связанные со спортом. В нижней части таблицы расположены рубрики с низким качеством классификации. Здесь преобладают темы, посвящённые различным наукам, характеризующиеся более широким терминологическим словарём. По-видимому, для этих рубрик обучающая выборка оказалась недостаточной, чтобы выявить все характерные для них термины, поэтому автоматически построенные профили в этом случае оказались неадекватными. В пользу такого вывода также говорит тот факт, что наибольшее превосходство экспертов над машиной мы видим именно для рубрик из нижней части таблицы. Другие интересные наблюдения можно сделать, сопоставив между собой таблицы 3 и 4. Рубрики «Домашнее->Сад_и_огород» и «Игры->Азартные_игры» при сильной оценке находятся внизу таблицы, а при слабой расположены в её верхней части. Этот факт свидетельствуют о том, что ассессоры, оценивающие прогоны, сильно разошлись во мнении при оценке этих рубрик.

Можно ли как-то улучшить автоматически созданные профили, чтобы качество классификации по ним было сравнимо с классификацией по профилям, созданными экспертами? Один из недостатков профилей, созданных автоматом, состоит в том, что значимыми для классификации становятся термины, отражающие специфику текстов из обучающей выборки. И если обучающая и тестовая выборки получены из различных источников, то эти специфичные термины ухудшат качество классификации документов из тестовой выборки. Мы предоставили эксперту автоматически созданные профили из прогона LTpos, он просмотрел первые 50 слов из каждой рубрики и исключил те, которые по его мнению не соответствуют теме. Отчищенные таким образом профили использовались для прогона LTfilter, результаты которого

название рубрики	Manual	LTpos
Спорт->Силовые_виды	0.901	0.735
Спорт->Футбол	0.763	0.709
Спорт->Боевые_искусства	0.873	0.558
Игры->Компьютерные_игры	0.542	0.527
Домашнее->Кулинария	0.437	0.484
Спорт->Зимние_виды	0.700	0.341
Домашнее->Домашний_ремонт	0.367	0.313
Спорт->Пейнтбол	0.676	0.290
Игры->Настольные_игры	0.463	0.282
Домашнее->Личная_жизнь	0.452	0.281
Игры->Ролевые_игры	0.206	0.265
Наука->Науки_о_Земле	0.230	0.222
Наука->Биология	0.273	0.201
Наука->Химия	0.477	0.195
Наука->Агрономия	0.322	0.130
Наука->Математика	0.308	0.119
Домашнее->Сад_и_огород	0.023	0.093
Игры->Азартные_игры	0.122	0.068
Наука->Технологии	0.286	0.064

Таблица 3. F1 рубрик с сильными требованиями к релевантности. Рубрики упорядочены по убыванию качества классификации LTpos.

название рубрики	Manual	LTpos
Спорт->Силовые_виды	0.913	0.800
Спорт->Футбол	0.834	0.698
Игры->Компьютерные_игры	0.751	0.672
Спорт->Боевые_искусства	0.973	0.667
Домашнее->Сад_и_огород	0.320	0.560
Домашнее->Кулинария	0.749	0.505
Игры->Азартные_игры	0.695	0.500
Домашнее->Личная_жизнь	0.633	0.439
Игры->Настольные_игры	0.740	0.435
Игры->Ролевые_игры	0.624	0.429
Домашнее->Домашний_ремонт	0.658	0.427
Спорт->Пейнтбол	0.847	0.424
Наука->Биология	0.601	0.411
Спорт->Зимние_виды	0.839	0.406
Наука->Агрономия	0.921	0.402
Наука->Математика	0.648	0.333
Наука->Химия	0.677	0.214
Наука->Науки_о_Земле	0.705	0.161
Наука->Технологии	0.720	0.094

Таблица 4. F1 рубрик со слабыми требованиями к релевантности. Рубрики упорядочены по убыванию качества классификации LTpos.

micro	and			or		
	r	p	F1	r	p	F1
LTpos	0.63	0.27	0.37	0.55	0.46	0.50
LTfilter	0.84	0.31	0.45	0.77	0.52	0.62
Manual	0.86	0.37	0.51	0.82	0.69	0.75

macro	and			or		
	r	p	F1	r	p	F1
LTpos	0.55	0.24	0.31	0.50	0.43	0.45
LTfilter	0.73	0.28	0.37	0.70	0.48	0.55
Manual	0.82	0.34	0.44	0.82	0.68	0.73

Таблица 5. Полнота, точность и F1 рубрик с сильными (and) и слабыми (or) требованиями к релевантности для микро- и макро-усреднения.

изображены на рис. 1. Стрелочками для некторых рубрик показан переход от LTpos к LTfilter. Как видим, качество классификации заметно улучшилось и приблизилось к тому, что демонстрирует прогон Manual с профилями составлены экспертами вручную. Оценки качества классификации по тестовым выборкам 2007 года для прогонов LTpos, LTfilter, Manual сведены в таблицу 5.

2.7 Результаты оценки классификации веб-сайтов

Оценки F(micro) и F(macro) для прогонов L, LT, LTpos за 2007 год и их сравнение с прогоном Manual, для которого профили рубрик создавались экспертами вручную, представлены в таблице 6. В отличие от веб-страниц, использование профилей экспертов в классификации веб-сайтов не даёт столь однозначного преимущества: прогоны Manual выигрывают в случае слабых, но проигрывают в случае сильных требований к релевантности. Оценки F(micro) и F(macro), полученные в 2008 году, показаны на рис. 2. В отличие от веб-страниц, качество классификации веб-сайтов не улучшается при добавлении словосочетаний в словарь классификационных терминов. При этом, как и следовало ожидать, устранение отрицательных терминов в профилях рубрик улучшает результат, так как в этом случае увеличивается полнота, точность же при классификации веб-сайтов и так высокая.

micro	L	LT	LTpos	Manual
and	0.260	0.246	0.271	*0.233*
or	0.319	0.270	0.377	*0.463*

macro	L	LT	LTpos	Manual
and	0.297	0.296	0.330	*0.266*
or	0.335	0.294	0.406	*0.498*

Таблица 6. F1(micro) и F1(macro) веб-сайтов за 2007 год с сильными (and) и слабыми (or) требованиями к релевантности.

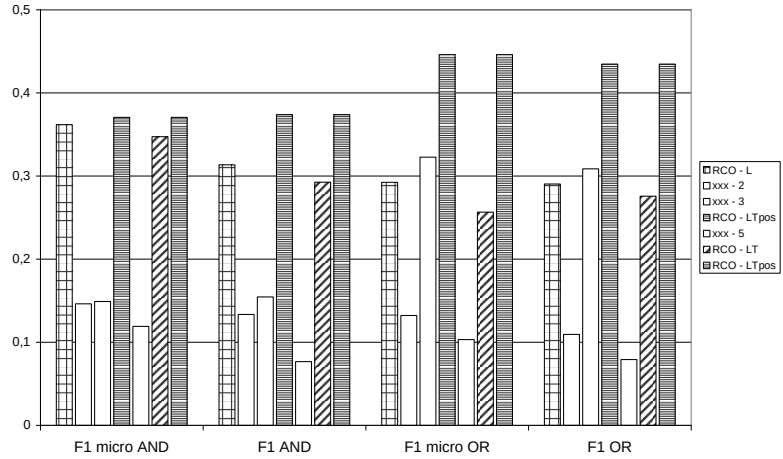

Рисунок 2. Значения F1(micro) и F1(macro) с сильными и слабыми требованиями к релевантности для участников дорожки классификации web-сайтов.

3. Кластеризация новостного потока

3.1 Постановка задачи

Система-участник должна структурировать поток сообщений коллекции в набор сюжетов, связанных ассоциативными связями в "надсюжеты".

3.2 Описание метода

1. Из трёх первых предложений каждого новостного сообщения извлекались идентификационные признаки: слова и словосочетания, исключая стоп-слова. Сообщение представлялось в виде вектора признаков, с учётом частот. При этом именам и географическим названиям придавался больший вес, чем обычным терминам.
2. Между всеми сообщениями вычислялась попарная близость как скалярное произведение двух векторов. Полученная матрица близостей использовалась затем для кластеризации.
3. Для разбиения новостных сообщений на события (topic) использовался агломеративный иерархический кластерный анализ. В кластеризации участвовали сообщения, полученные в течении

определённого промежутка времени T. Близость между документом и кластером вычислялась как средняя близость данного документа со всеми документами кластера. Размер кластеров ограничивался введением специального порога. Число сообщений в кластере увеличивается, пока $\min(c_i) > C$, где C = 0.15 - наперёд заданный порог, c_i – средняя близость между i-м сообщением кластера и всеми остальными документами за выбранный промежуток времени, минимум берётся по всем документам кластера. Порог выбирается исходя из тербований к полноте/точности ведения сюжетной линии.
4. Для выделение дубликатов сообщений (event) применялась описанная в пункте 3 процедура, только вместо порога C использовался порог S = 0.4.
5. Найденные в пункте 3 кластеры объединялись в сюжеты (broadtopic). Для этого сопоставлялись результаты кластеризации двух периодов со сдвигом в один час. Кластеры, близость между которыми превышала порог U = 0.1, объединялись в сюжет.

3.3 Результаты оценки

Мы подали на оценку 2 прогона. В первом прогоне RCO-1 применялся описанный выше алгоритм с временным интервалом T = 24 часа, во втором прогоне RCO-2 мы ограничились кластеризацией новостей на события с временным интервалом T = 1 неделя, без связывания событий в сюжеты. Оценки результатов кластеризации представлены в таблице 7.

	event		topic		broadtopic	
	r	p	r	p	r	p
RCO-1	0.611	0.809	0.724	0.333	0.192	0.784
RCO-2	0.595	0.796	0.659	0.386	0.152	0.800

Таблица 7. Усреднённые по кластерам результаты прогонов RCO-1 и RCO-2.

3. Заключение

В работе показано, что использование многословных терминов в качестве классификационных признаков дает значительный прирост точности при классификации web-страниц, но практически не улучшает результат для web-сайтов. Также было обнаружено, если отбросить из созданного методом SVM профиля термины, имеющие отрицательные веса, то в большинстве случаев качество

классифкации улучшается. Кроме того, значительно повышает качесвтво классификации фильтрация «случайных» терминов, связанных с особенностями обучающей выборки, а не предметной областью.

Литература

[1] *Ермаков А.Е.* Значимость элементов текста в свете теории синтаксической парадигмы // Русский язык: исторические судьбы и современность. II Международный конгресс исследователей русского языка. Труды и материалы. - Москва: МГУ - 2004.

[2] *Ермаков А.Е*. Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003. – Москва, Наука, 2003

[3] *Поляков П.Ю., Плешко В.В.,* RCO на РОМИП 2006 // Труды четвертого российского семинара РОМИП'2006. (Суздаль, 19 октября 2006г.). - Санкт-Петербург: НУ ЦСИ – 2006 - с. 72-79.

[4] *H. Avancini, A. Lavelli, F. Sebastiani, R. Zanoli*. Automatic expansion of domain-specific lexicons by term categorization // ACM Transactions on Speech and Language Processing (TSLP) Discovery – 2006, V.3, No.1 – pp.1-30.

[5] *Burges C.J.C.* A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery – 1998, V.2, No.2 – pp.121-167.

[6] *Joachims T.* Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998.

RCO at RIRES 2007

Polyakov P.Yu., Pleshko V.V.

This article presents report on experiments in IR that were driven as a part of RIRES seminar. The research was taken on different term selection methods that affect quality of web-site and web-page classification task. Also the news clustering task was performed.