Автоматическая классификация веб-сайтов

© М.Ю. Маслов, А.А. Пяллинг, С.И. Трифонов

Компания «Яндекс»
{maslov, pyal, trifon}@yandex-team.ru

Аннотация

Предлагается алгоритм автоматической классификации веб-сайтов на основе анализа их текстового содержания. Алгоритм сравнивается с алгоритмом SVM-light на данных веб-каталога Яндекса.

1 Введение

Целью настоящей работы было построение алгоритма автоматической классификации, способного эффективно классифицировать Рунет (сегмент интернета в доменных зонах .ru и .su, и страницы на русском языке вне этих доменов). Результат классификации предполагается использовать при ранжировании ссылок в ответе поисковой системы. Поэтому алгоритм должен быть:

а) Достаточно незатратным по вычислительным ресурсам, чтобы была возможность классифицировать все сайты Рунета по 500-1000 тематическим рубрикам веб-каталога.

б) В то же время, достаточно качественным, сопоставимым наиболее эффективными известными алгоритмами: чем точнее результат классификации, тем это должно лучше сказаться на ранжировании результатов поиска.

Основной особенностью задачи классификации Веба является большой объем информации. Оценить объем Рунета можно с помощью данных поисковой системы Яндекс. К настоящему моменту в поисковом индексе Яндекса содержится порядка $3Ч10^9$ уникальных веб-страниц, с $4Ч10^6$ «живых» сайтов Рунета (под «живыми» понимаются сайты, у которых есть хотя бы одна уникальная веб-страница из индексной базы Яндекса).

Эталонным набором данных, или учителем, при классификации веба обычно служит один из веб-каталогов, таких как Yahoo!, ODP, Каталог@Mail.ru или Яндекс.Каталог [1-2, 12]. Характерный объем «всемирного» веб-каталога – порядка 10^6–10^7 записей (Yahoo!, ODP), а веб-каталога Рунета – порядка 10^5 записей (Mail.ru, Яндекс).

Сопоставление объемов учителя и «генеральной совокупности» веб-страниц в задаче классификации Веба может привести к мысли, что учитель очень мал по сравнению с «генеральной совокупностью». Так, в [3] задача классификации Веба характеризуется как semi-supervised learning, т.е. как промежуточный случай между задачей классификации с учителем и задачей классификации без учителя (кластеризации). Логика такова: в веб-каталоге [2] описано порядка 10^6 веб-страниц. Во всем Вебе их гораздо больше – порядка 10^{10}. Т.е. задача классификации Веба – классификация с очень маленьким учителем.

Однако, веб-каталог как учитель можно трактовать иначе. Можно считать, что если запись веб-каталога ссылается на входную страницу сайта, то описание в веб-каталоге относится не только к этой входной странице, но и к сайту в целом. При таком подходе объем и представительность учителя радикально возрастает. Например, в [1] на настоящий момент содержится порядка 10^5 описаний сайтов (т.е. столько записей, ссылающихся на корневую страницу сайта), что составляет порядка 2% от общего количества сайтов Рунета. При чем эти 10^5 сайтов содержат суммарно порядка 10^9 веб-страниц, что составляет порядка 30% от общего количества уникальных веб-страниц в индексе Яндекса. В этом контексте уже нет речи о малости учителя. И проблему репрезентативности учителя по отношению к «генеральной совокупности» веб-страниц решить гораздо легче.

Таким образом, особенности постановки задачи, решаемой в данной работе, следующие:

а) Рассматривается задача классификации **веб-сайтов**. В отличие от классификации **веб-страниц** – задачи, подразумеваемой в большинстве работ на эту тему.

б) Сайт в обучающей выборке представлен не только входной страницей, но и репрезентативной выборкой своих страниц.

Отметим, что у такого «посайтового» подхода есть следующие проблемы:

а) **Зашумленность**. Сайты нередко содержат заметное число страниц, не относящихся к теме. Характерным примером являются интернет-форумы. Типичный тематический форум состоит из нескольких разделов по теме, и раздела, предназначенного для общения без ограничения темы. Такой раздел часто бывает довольно объемным. Кроме того, форумы часто подвергаются спамовым атакам, что приводит к большой зашумленности плохо модерируемых и немодерируемых форумов.

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

б) **Политематичность сайтов**. Нередко сайт посвящен широкой предметной области, или сразу нескольким предметным областям. Редакторы каталогов в подобных случаях стараются выделить и описать отдельно тематические подразделы, сайтов. Но часто эти подразделы выделить не удается из-за особенностей структуры сайтов. В таких случаях к нескольким тематическим рубрикам относится весь сайт, и тогда каждая его страница получает несколько тем. Но каждая конкретная страница сайта обычно посвящена одной конкретной теме. И в случае сайта с N темами точность учителя оказывается равной 1/N.

2 Related work

В работе [3] констатируется, что методы автоматической классификации, основанные на анализе текстов, в случае веб-страниц работают плохо. И показано, что при использовании текстов входящих и исходящих ссылок страницы, и учета ссылочной структуры веба в целом, совместно с лексикой текстов веб-страниц, качество классификации можно повысить.

Идея об использовании ссылочной информации совместно с текстом страниц при классификации веба используется также в ряде других работ, в частности [5–8].

В [9] показано, что использование аннотаций сайтов, сделанных редакторами веб-каталога [10], позволяет существенно улучшить качество классификации. Далее, предложен алгоритм автоматического построения аннотаций веб-страниц. И показано, что с помощью таких автоматических аннотаций качество классификации тоже можно повысить.

В семинаре РОМИП [11] задачи классификации веб-сайтов и веб-страниц явно разделены (возможно, впервые). Участникам «дорожки» классификации веб-сайтов предоставляется учитель, построенный на основе русскоязычной части веб-каталога ODP [12]. В нем 2100 сайтов, 300 тыс. страниц, с сайта бралось не более 500 страниц, в порядке обхода «в ширину». Проверка качества алгоритмов проводилась на другом множестве: на сайтах с хостинга Narod.ru (в дорожках 2003–2006 гг.) или сайтах из белорусского интернета (в дорожке 2007 г). Таким образом, обучающее и проверочное множество сайтов в РОМИП могут отличаться по своим характеристикам.

В [13] была предложена постановка задачи, аналогичная РОМИП. Набор данных, предоставляемый стипендиатам, был построен на основе каталога Яндекса [1]. В выборке 40 тыс. сайтов, 1.5 млн страниц, с сайта берется тем больше страниц, чем больше сайт, от десятков до тысяч страниц. В выборку страниц с сайта попадают а) страницы, близкие к корню сайта (1/3 страниц) б) случайная выборка страниц из дополнительной части сайта (2/3 страниц) [14] . Этот набор данных был использован, в частности, в работе [15], в которой анализировалась эффективность ряда известных методов при решении задачи классификации сайтов по текстовому содержанию их страниц.

3 Построение проверочной и обучающей выборки. Тестирование классификаторов

Был создан набор данных из 4 млн веб-страниц в соответствии с процедурой [14]. Полученный набор данных делился на две выборки – обучающую выборку (2/3 сайтов) и проверочную (оставшаяся 1/3 сайтов).

Алгоритмы классификации обучались на всех страницах из обучающей выборки. Далее, проводилась классификация страниц на проверочной выборке. Возможным результатом классификации страницы было три варианта:

1. Страница относится к заданной теме
2. Страница не относится к заданной теме
3. Страницу не удалось определить (мало данных)

Если для сайта число страниц по теме было больше, чем число страниц, не относящихся к теме, то считалось, что сайт относится к теме. В противном случае считалось, что сайт к теме не относится. Качество классификаторов определялось F1-мерой полученного результата классификации сайтов.

«Среднее» качество классификаторов оценивалась на 12 случайно выбранных рубриках разных уровней иерархии (см. табл. 1). Кроме того, был проведен расширенный набор экспериментов и проанализированы ошибки классификатора для темы Порно (см. ниже разделы 4.3, 5 и 6). Качественная классификация сайтов по теме Порно востребована в связи с задачей семейной фильтрации при поиске [19]. Тема Порно была синтезирована посредством объединения пяти подтем рубрики Яндекс.Каталог/ http://yacatalog.narod.ru/admin/sprav/themes.xhtml?level=1&parent=0Развлечения/ Личная жизнь. Три из них публичные: это Эротика и секс, Эротические галереи, Эротические игры. И две не публикуются в Яндекс.Каталоге: это Интим-услуги и Порнография.

4 Классификатор сайтов по ключевым словам (КС-классификатор)

Одним из наиболее мощных методов решения задачи классификации текстовых документов считается метод опорных векторов, или SVM (см. напр. обзор [16]). Для сравнения с описываемым алгоритмом мы выбрали реализацию SVM-Light [18], поскольку она является одной из наиболее широко используемых реализаций метода SVM, и одной из наиболее эффективных по вычислительным ресурсам во время обучения, с возможностью применения к очень большим наборам данных [17].

К сожалению, использование данного алгоритма для классификации всех сайтов, известных поисковой машине Яндекс, по 500 темам, нам представляется затруднительным.

Во-первых, алгоритм хранит все входные данные в памяти, в результате возникают ограничения

по числу страниц, на которых строятся правила (в данной работе 3GB оперативной памяти хватило для хранения 600 тыс. страниц, при средней длине страницы 320 слов).

Во-вторых, при малых объемах обучающей выборки алгоритм в наших экспериментах оказался неустойчивым. При использовании 200 тыс. страниц в качестве обучающего множества, F1-мера для темы Порно была ~30%, а F1-мера на выборке из 600 тыс. страниц ~83%.

В связи с этим возникла потребность в создании достаточно простого и устойчивого алгоритма, способного классифицировать большие объемы данных по большому числу тем. В результате был построен классификатор байесовского типа, который и описан в этой статье.

При построении классификатора использовались следующие идеи:

1. Классификация строится по каждой теме независимо друг от друга.
2. Тему документа можно определить по списку встреченных слов (bag of words).
3. Для каждой темы есть слова, характерные для темы, и слова, которые в теме не встречаются.
4. Множество характерных для темы слов можно разбить на 2 части:
   a. Характеристические слова – множество слов, без которых невозможно раскрыть тему. Если в документе нет ни одного слова из этого множества, то документ к теме не относится.
   b. Остальные слова, характерные для темы. Мы не можем определить только по наличию или отсутствию этих слов в документе, в теме документ или нет.
5. Кроме характерных слов, классификатор существенно использует слова, нехарактерные для темы.

Ниже описаны процедуры обучения и собственно классификации сайтов.

4.1 Процедура обучения

1. Из исходного множества документов (4 млн документов) собирается множество словоформ F которые встречаются в исходном корпусе слов больше чем N раз (бралось N = 500, получено 135 тыс. словоформ).
2. Для каждой темы собирается статистика: сколько раз слово встретилось в теме N_w и не в теме $\overline{N_w}$, а также сколько всего слов во всех документах в теме N_{tot} и не в теме $\overline{N_{tot}}$. Далее: $P_w = N_w / N_{tot}$, где P_w – вероятность встретить данное слово. $P^L{}_w = 1 - (1 - P_w)^L$, где $P^L{}_w$ – вероятность того, что в тесте длиной L есть данное слово для документов, которые относятся к заданной теме. Точно так же считается $\overline{P^L{}_w}$ – вероятность того, что в тесте длиной L есть данное слово для документов, которые не относятся к заданной теме. $W^L{}_w = \ln(P^L{}_w / \overline{P^L{}_w})$, где $W^L{}_w$ – вес слова, логарифм вероятности того, что документ относится к заданной теме, при условии, что в документе длиной L встретилось данное слово.
3. Слова сортируются в порядке убывания веса. Учитываются положительные веса для N_{top} первых слов (характеристические слова). N_{top} – подбирается из условия максимальности F1-меры на обучающем множестве. Для разных тем N_{top} разное: чем шире тема, тем N_{top} больше. Для остальных слов с положительным весом вес обнуляется. Для остальных слов с отрицательным весом вес учитывается.

В формуле выше существенным образом участвует длина документа L, различная для различных документов. Однако, для организации эффективного счета полезно заменить изменяющееся значение L на постоянную усредненную величину. Эксперименты показали, что F1-мера результата от такой замены изменяется незначительно.

Средняя длина документов в обучающей выборке была 320 слов на множестве словоформ F. Было протестировано несколько вариантов с разной длиной среднего документа L при вычислении веса $W^L{}_w$. Оптимальный вариант выбран равным 50.

4.2 Процедура классификации

1. Классификация документа. Выделяются все слова из множества F. Вычисляется среднее значение логарифма вероятности того, что документ соответствует теме для слова в документе:

$$P = \frac{\sum_w W^L{}_w * N_w}{\sum_w N_w}$$

Если $P > 0$ – документ в теме, если $P < P_{min}$ – документ не в теме. Максимальная F1-мера результатов классификации была получена при $P_{min} = -0.5$.

2. Классификация сайта. Подсчитывается число документов в теме и вне темы. Если документов в теме больше, чем вне темы – сайт в теме. Если нет – вне темы.

Преимущество построенного классификатора – высокая скорость работы, хорошее качество классификации. При работе создается множество характеристических слов, которые могут в дальнейшем использоваться для классификации запросов, построения тезауруса, иерархического наследования ключевых слов при классификации тем-родителей. Можно использовать ключевые слова, полученные из других источников, в частности, взять список внешних тематических слов и объединить его со списком автоматически определенных слов с максимальным весом.

Таблица 1. Результаты тестирования КС-Классификатора на 12-и случайно выбранных темах

Уровень	Тема	Точность	Полнота	F1-мера
1	Дом	65.3	71.9	68.5
1	Учеба	51.4	79.0	62.3
2	Дом/Все_для_праздника	48.5	76.6	59.4
2	Общество/Религия	82.3	72.7	77.2
2	Hi-Tech/Мобильная_связь	47.2	69.6	56.2
2	Бизнес/Строительство	37.8	62.2	47.1
3	Общество/Власть/Силовые_структуры	41.4	59.0	48.6
3	Hi-Tech/Мобильная_связь/Сотовые_телефоны	46.0	72.7	56.3
3	Спорт/Экстремальный_спорт/Пейнтбол	71.4	88.2	78.9
3	Дом/Все_для_праздника/Фейерверки	57.9	100	73.3
3	Дом/Кулинария/Рецепты	49.4	89.4	63.6
3	Порно	72.7	87.3	79.3

Таблица 2. Анализ ошибок КС-классификатора для темы Порно

Класс ошибки	Описание ошибки	Процент ошибок	Ошибочные сайты
Сайт ошибочно отнесен к теме	Сайт имеет сходную тематику – женское белье, медицина	22%	www.echo-h.ru, marmeladova.com, shop.flirt.ru, www.night-dress.ru, www.chocolatelily.ru
	На сайте порно есть, но с точки зрения Яндекс.Каталога это «развлекательный портал» (типичные разделы: эротика, юмор, обои и т.п.), который к теме Порно редакторы не относят	59%	Anekdots.smeha.net, www.strasti.ru, makecool.ru, www.s-info.ru, www.adamast.ru, www.klassno.com, www.loveorgazm.ru, www.prikol.net, zateynik.ru, bigpenis.com1.ru, www.megaportal.ru, www.erotofun.ru, death.com.ru
	Сайт имеет другую тему, автомат его отнес в Порно, возможно, из-за заспамленных форумов	9%	www.khabensky.net, www.raskolbass.ru
	Неактуальность записи в Яндекс.Каталоге	5%	www.gayvideo.ru
	Сайт недоступен	5%	www.forum.prikola.net
Сайт ошибочно не отнесен к теме	На сайте используется специфический язык (научный, медицинский, торговый)	27%	www.ograna.com, oganezov.boom.ru, www.bondage.ru, www.shorox.ru, www.sexrecord.ru, www.donnasummer.ru, vein.ru, www.fetishshop.ru, www.lesbi.ru
	Много страниц не по теме (форумы с обсуждением других тематик, ошибки типа soft-404 и т.п.)	27%	www.erochat.ru, www.x-time.ru, www.kiss-chat.ru, www.pornovids.ru, ecret-video.ru, photoart.prm.ru, vipxxx-party.ru, www.pickupforum.ru, www.xoxma.ru
	Политематические сайты, большинство страниц относится к другим темам	25%	www.rulz.ru, 1001.vdv.ru, kobelissimo.narod.ru, www.comics.ru, www.babes.ru, www.komi.com, excluz.com, www.krasgay.ru
	Явная ошибка автомата	6%	www.nudecelebs.ru, www.blevota.com
	Ошибочная запись в Яндекс.Каталоге	12%	letstest.info, sexxx6.sbn.bz, www.sexxx.sbn.bz, www.mpeg74.ru
	Сайт недоступен	3%	britva.gay.ru

Работа алгоритма была протестирована на 12 случайно выбранных темах. Результаты представлены в табл. 1.

Сравнение результатов при использовании порно-словника, составленного вручную, представлено в табл. 3.

Таблица 3. Влияние составленного вручную порно-словника

	Точность	Полнота	F1
Авто-список (500 слов)	74.4	88.5	80.8
Авто-список (200 слов) порно-словник (300 слов)	71.4	91.6	80.3

Добавление порно-словника происходило следующим образом. Для каждого слова из порно-словника вычислялся вес слова. Далее, выбиралось 300 слов из порно-словника с максимальными весами. К полученному списку слов добавлялось 200 слов с максимальными весами из списка, полученного автоматически.

4.3 Анализ полученных результатов

Список ключевых слов достаточно сильно изменялся при изменении обучающего множества сайтов, при использовании списка ключевых слов, собранных вручную. Также делались попытки очистить полученные списки слов, удалить случайные слова. Однако алгоритм оказался достаточно устойчивым по отношению к изменению списка ключевых слов. Качество работы изменялось незначительно.

Качество классификатора зависит от количества используемых в классификаторе словоформ. При использовании 200 тыс. словоформ F1-мера результатов классификации упала на 2%. При использовании 50 тыс. словоформ F1-мера результата оказалось также на 1–2% хуже.

Анализ ошибок КС-классификатора для темы Порно приведен в табл. 3.

5 Двухстадийный классификатор

При использовании классификатора по ключевым словам вариация параметров N_{top}, P_{min} позволяет гибко менять соотношение полноты и точности, т.е. получать полноту, близкую к 100%, за счет малой точности.

Сделаем две стадии обучения и классификации сайтов. На первой стадии построим классификатор с максимальной полнотой. На этом этапе отделим все тематические сайты от далеких по теме. Для темы Порно это приведет к тому, что в тему подмешаются сайты с женской одеждой, развлечениями.

Множество сайтов, отнесенных на первой стадии в тему Порно, используем для обучения классификатора на второй стадии. Так как множество не-порносайтов на второй стадии существенно отличается от множества не-порносайтов на первой стадии, то статистика для слов существенно изменится. На второй стадии порносайты будут эффективно отделяться от сайтов с женской одеждой, развлечениями, а не от всего множества не-порносайтов. Это позволяет поднять полноту и точность конечного классификатора.

Результат работы двухстадийного КС классификатора представлены в табл. 4.

Таблица 4 Результаты тестирования двухстадийного классификатора

	Точность	Полнота	F1
Первая стадия	29.5	96.2	45.2
Вторая стадия	77.8	90.9	84.1

Для проверки качества построенного классификатора, была проведена классификация сайтов с помощью классификатора SVM-light. Получена точность 84.4%, полнота 82.4%, F1-мера 83.4%. Таким образом F1-мера двухстадийного классификатора оказалась немного выше, чем при использовании классификатора SVM-light [18].

6 Классификация темы Порно по полным данным

С помощью КС-классификатора был произведен обход всех сайтов из индекса Яндекса.

При обучении классификатора использовалась треть сайтов веб-каталога Яндекса. С помощью построенного классификатора был проведен обход всех страниц всех сайтов находящихся в поисковом индексе Яндекса, тогда как при обучении и предварительном тестировании классификатора использовалась только малая репрезентативная выборка страниц сайта.

Алгоритм классифицировал 4 млн сайтов. Из них 34500 были отнесены к теме Порно. При сравнении результатов классификации с данными веб-каталога Яндекса была получена точность 75.3%, полнота 87%, F1-мера 80.7%.

7 Выводы

В данной работе предложен классификатор сайтов по ключевым словам. Основными преимуществами его является большая скорость работы, простота построения правил, устойчивость, что позволяет использовать его для эффективной классификации больших объемов данных и для большого количества тем. Качество работы классификатора сопоставимо с качеством работы алгоритма SVM-light. Классификатор при своей работе строит множество ключевых слов для темы, которые могут быть использованы отдельно от классификатора.

Литература

[1] Яндекс.Каталог. http://yaca.yandex.ru

[2] Yahoo! Directory, http://dir.yahoo.com

[3] Каталог@MAIL.RU http://list.mail.ru/index.html

[4] S. Chakrabarti. Mining the Web : Discovering Knowledge from Hypertext Data, Morgan-Kaufmann Publishers, 2003.

[5] А.В. Сычев, М.М. Баженов. «Автоматическое пополнение веб-каталога на основе идентификации веб-сообществ с последующей фильтрацией документов по контенту» // Интернет-математика 2007 : Сборник работ участников конкурса. – Екатеринбург : Изд-во Урал. ун-та, 2007. http://download.yandex.ru/IMAT2007/sychev.pdf
[6] G. Attardi, A. Gulli, F. Sebastiani. Automatic Web Page Categorization by Link and Context Analysis // In Chris Hutchison and Gaetano Lanzarone (eds.), Proc. of THAI'99, 1999, 105–119
[7] E. Glover et. al. Using web structure for classifying and describing web pages, Proc. of the 11th international conference on World Wide Web, May 07–11, 2002, Honolulu, Hawaii, USA
[8] P. Calado, M. Cristo, E. Moura et al. Combining link-based and content-based methods for web document classification. – CIKM'03, Nov. 3–8. – [Electronic resource] http://homepages.dcc.ufmg.br/~nivio/papers/cikm03.ps
[9] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, Wei-Ying Ma, Web-page Classification through Summarization
[10] LookSmart Web directory. http://search.looksmart.com (ссылка устарела)
[11] Сайт РОМИП www.romip.ru
[12] Open Directory Project http://www.dmoz.org/
[13] Интернет-математика 2005: Автоматическая обработка веб-данных Москва, 2005. http://company.yandex.ru/grant/list.xml
[14] Описание наборов данных для участников программы научных стипендий Яндекса 2005 г. http://company.yandex.ru/grant/datasets_description.xml
[15] М.В. Киселев. Оптимизация процедуры автоматического пополнения веб-каталога. // Интернет-математика 2005: Автоматическая обработка веб-данных. – М., 2005. http://company.yandex.ru/grant/2005/08_Kiselev_102710.pdf
[16] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.
[17] O. Ivanciuc. Applications of Support Vector Machines in Chemistry, Rev. Comput. Chem. 2007, 23, 291–400.
[18] T. Joachims, Training Linear SVMs in Linear Time, KDD, 2006. http://svmlight.joachims.org/
[19] Семейный поиск Яндекса. http://help.yandex.ru/search/?id=481933

Web sites automatic categorization

Mikhail Maslov, Aleksei Pyalling, Sergey Trifonov

An automatic Web page categorization method is proposed. It's efficiency is evaluated and compared with SVM-light method using Yandex web directory data.