Модел. и анализ информ. систем. Т. **20**, № **2** (2013) 80–91

УДК 004.738.52

Единая модель для геоклассификации веб-сайтов

Волков А.Н.

ООО «Яндекс», 119021, Россия, г. Москва, ул. Льва Толстого, д. 16

e-mail: ark-kum@yandex-team.ru

получена 12 декабря 2012

Ключевые слова: геоклассификация, машинное обучение

Работа представляет новый подход к задаче определения регионального фокуса веб-сайтов (геоклассификации). В отличие от традиционных подходов к многозначной классификации, когда для каждого класса (региона) обучается по отдельной классификационной модели, предлагаемый подход основан на обучении всего одной модели, которая используется для всех регионов одного типа (например, для городов). Такой подход становится возможным благодаря использованию "относительных" факторов, которые показывают, как некоторый выбранный регион соотносится с другими регионами для заданного веб-сайта. Классификатор задействует большой набор разнородных факторов, которые до этого момента не использовались вместе для геоклассификации веб-сайтов с применением машинного обучения. Оценка качества демонстрирует преимущество нашего подхода "по одной модели на тип региона" перед традиционным подходом "по одной модели на регион". Отдельный эксперимент демонстрирует способность описываемого классификатора успешно детектировать регионы, которые отсутствовали в обучающей выборке (что невозможно при использовании традиционных подходов).

1. Введение

Одна из проблем, с которыми в наши дни сталкиваются поисковые системы – это проблема регионализации поиска. Пользователи задают поисковые запросы, ответы на которые должны быть разными, в зависимости от региона пользователей. Правильный ответ на запрос «официальный сайт президента» зависит от страны пользователя, задавшего запрос, а запрос «доставка пиццы» зависит от положения пользователя ещё сильнее. Многие веб-ресурсы (такие, как веб-сайты местных фирм, ресторанов, театров) релевантны только пользователям из определённых регионов. Чтобы выдавать релевантные ответы на регионально-зависимые запросы, поисковая система должна знать не только регион пользователя, но и региональный фокус веб-сайтов.

В этой работе описывается решение задачи определения регионального фокуса веб-сайта (геоклассификации). Для каждого веб-сайта мы хотим найти множество

регионов, где живут люди, которым интересно содержимое веб-сайта. Решение проблемы нахождения регионального фокуса веб-страницы было впервые предложено Ding et al. [4]. Их подход, по большей части, основывался на снятии омонимии топонимов и устранении неоднозначности регионально-зависимых сущностей (например, телефонных и почтовых кодов) в текстах веб-страниц. Последующие работы Amitay et al. [1] и Zong et al. [11] основывались на распространении показателей доверия найденных топонимов вверх по дереву регионов для нахождения наиболее вероятного общего предка (например, для определения страны на основе нескольких упоминаний городов). В работе Pyaling et al. [7] различные факторы использовались для последовательного улучшения результата. Все эти работы не использовали машинное обучение и не учитывали в текстах страниц слова, про которые заранее не было известно, что связаны с регионами. Тем не менее, несколько работ по геоклассификации других типов веб-контента, таких, как онлайн-фотографии, помеченные пользователями [3, 10], или твиты [2] использовали более традиционный подход к текстовой классификации и, для каждого региона из обучающего множества, обучали классификационную модель, использующую слова в качестве факторов [9]. Несмотря на более прочное теоретическое обоснование, такой подход, очевидно, страдает от нехватки обучающих данных о более редких регионах, жители которых предоставляют недостаточный для обучения хорошего классификатора объём регионально помеченного контента.

В отличие от вышеупомянутых подходов, мы предлагаем классификационную систему, где факторы, основанные на извлечении сущностей, и текстовые факторы используются вместе для осуществления геоклассификации веб-сайтов с высокими показателями качества. Более того, мы используем одну модель на тип региона (т.е. «города», «страны») вместо отдельной модели на каждый регион. Такой подход решает проблему нехватки обучающих данных и позволяет удовлетворительно классифицировать даже веб-сайты, относящиеся к регионам, которые практически (или полностью) отсутствуют в обучающей выборке. Наш метод может быть использован для регионов любого типа (страна, штат, область, город и т.д.), но система, описанная в данной статье, использует два типа регионов – город и страна.

Данная статья организована следующим образом: В разделе 2 приводится описание нашего метода классификации, включая описание источников данных, факторов и алгоритма классификации, использующего эти факторы. В разделе 3 демонстрируется преимущество нашего классификатора, использующего одну модель на тип региона, над классификатором, строящим отдельную модель для каждого региона. Раздел 4 завершает статью и описывает возможности по расширению и улучшению описанного подхода.

2. Описание метода

Цель нашей работы – для каждого веб-сайта $w \in W$ найти множество регионов $R(w) \subset \hat{R}$, жителям которых интересно содержимое этого веб-сайта (здесь W – это множество всех исследуемых веб-сайтов, а $\hat{R}$ – множество всех рассматриваемых регионов). Это стандартная задача многозначной (multi-label) классификации. Каждый объект (веб-сайт) может принадлежать нескольким классам (регионам) или не принадлежать ни одному классу.

Большинство работ по региональной классификации веб-ресурсов, применявших машинное обучение, использовали для многозначной классификации набор бинарных классификаторов, обученных по схеме «один против всех». Такой подход имеет ряд проблем. Во-первых, такая система способна корректно классифицировать только веб-сайты, относящиеся к регионам, которые присутствовали в обучающей выборке. Более того, даже если регион присутствовал в обучающей выборке, но количество примеров было небольшим, качество классификации веб-сайтов, относящихся к такому региону, будет низким. Во-вторых, вычислительная сложность подхода «по модели на каждый регион» растёт линейно в зависимости от количества регионов и, следовательно, от уровня детализации таксономии регионов.

В данной работе мы представляем систему, которая использует машинное обучение для построения одной классификационной модели и затем использует эту модель для выявления разных регионов с таким же или лучшим качеством, чем традиционные системы, строящие для каждого региона отдельную модель. Система обучает ранжирующую модель, которая стремится для каждого веб-сайта и региона вычислить релевантность региона веб-сайту. Это похоже на то, как работают алгоритмы обучения ранжированию (Learning to Rank) [6], вычисляющие релевантность документа запросу. Главным преимуществом нашего подхода к проблеме геоклассификации является использование одной модели на все регионы одного типа (например, одной модели на все города) для проведения многозначной классификации. В следующих разделах мы опишем нашу систему классификации более детально.

2.1. Единая классификационная модель

Наша система решает проблемы с определением редких регионов и вычислительными затратами, меняя объект классификации. Вместо обучения набора моделей, каждая из которых классифицирует *веб-сайт* как релевантный или нерелевантный по отношению к соответствующему региону, наша система обучает одну модель, которая классифицирует *пары <сайт, регион>*. В результате задача многозначной классификации превращается в задачу бинарной классификации, существенно снижая вычислительную сложность. Применяя классификационную функцию, мы вычисляем для каждой пары вероятностную оценку релевантности. Важно отметить, что единые модели получаются лучше обученными, так как им достаётся намного больше положительных примеров, чем моделям «по одной на регион».

2.2. Источники данных и алгоритмы извлечения данных

Наша система классификации использует источники данных разных типов. Это и региональная информация, извлекаемая из документов с использованием алгоритмов извлечения сущностей, и данные, внешние по отношению к содержимому документов веб-сайта, и информация, извлекаемая из полного текста проиндексированных документов. Наша система не использует, явным образом, факторы, основанные на полном тексте документов. Вместо этого используются результаты внешнего, по отношению к системе, текстового классификатора [8], использующего слова документов в качестве признаков.

Идея использования текстового классификатора состоит не только в том, чтобы

узнать региональные свойства веб-сайтов. Мы также хотим получить тематические аспекты, заложенные в словах документов. Тематика документов может помочь нам определить «универсальность» веб-сайта и повлиять на отнесение веб-сайта к локальному (конкретный город или страна) или глобальному (общестрановый или международный сайт) классу. К примеру, сайты о программировании и электронные библиотеки обычно глобальны, в то время как веб-сайты кинотеатров и ресторанов локальны. Все источники данных, перечисленные в этом разделе, полезны для классификации, т.е. их удаление ухудшает качество классификации.

Ниже приводится список используемых источников данных, а также детали их обработки.

- **Количество упоминаний названия региона в заголовках страниц; количество упоминаний почтовых и телефонных кодов региона в адресных блоках.** Мы формируем запросы к поисковой системе Яндекс, мгновенно получая адреса всех страниц, на которых почтовый/телефонный код некоторого города расположен рядом с названием этого города и т.н. маркерами адресных блоков (словами «тел.», «ул.», «телефон» и т.д.). Названия регионов в заголовках страниц ищутся аналогичным образом. Результат для каждого веб-сайта и региона вычисляется как суммарное количество региональных меток соответствующего типа, найденных на страницах веб-сайта. Значения суммируются отдельно для контактных страниц, для главной страницы и для всех страниц.

- **Количество упоминаний телефонных номеров, принадлежащих некоторому региону.** Телефонные номера извлекаются при помощи алгоритма извлечения сущностей, который применяется ко всем проиндексированным веб-страницам.

- **Регион, соответствующий IP-адресу веб-сайта.** IP-адреса сайтов преобразуются к регионам, используя внутреннюю базу данных, собранную из нескольких источников.

- **Регион, соответствующий домену верхнего уровня веб-сайта.** Каждый веб-сайт имеет имя, состоящее, в том числе, из *домена верхнего уровня* (*TLD*). Каждый национальный домен верхнего уровня соответствует определённой стране, но некоторые так называемые омонимичные домены (например, .tv, .cd, .dj, .fm, .me, .ws) считаются *красивыми* и используются для веб-сайтов, расположенных за пределами стран, для которых они были предназначены. Мы не учитываем такие домены для предотвращения ложных срабатываний.

- **Региональный результат контентного классификатора** Используется сторонняя реализация Байесова классификатора [8], строящая классификационную модель для каждого региона, присутствующего в обучающей выборке.

- **Языковая статистика веб-сайта.** Для каждого языка, который мы можем определить, мы считаем количество страниц сайта, использующих данный язык, а также долю этих страниц.

- **Тематический результат текстового классификатора** (упомянутого ранее) [8]. Для каждой темы источник выдаёт одно двоичное значение, показывающее, имеет ли веб-сайт соответствующую тематику.

- **Домен верхнего уровня.** Между доменами верхнего уровня и странами существует, кроме явной связи стран с национальными доменами, и неявная связь. Например, домены .mil и .gov используются, в основном, правительственными сайтами США; украинские сайты нередко используют домен .ru. Поэтому, мы также используем «сырые» данные о домене верхнего уровня в качестве фактора при определении страны.

2.3. Гео-кодирование

Некоторые данные связаны с регионами очевидным образом. Другие же имеют лишь неявную связь с регионами. Такие данные должны быть гео-кодированы перед использованием. Гео-кодирование – это процесс преобразования неявно региональных данных в явные регионально-специфичные признаки. Примером гео-кодирования может быть преобразование фактора «страницы веб-сайта содержат почтовые индексы: 141701, 127560, 141707» к фактору «почтовые индексы: {Долгопрудный: 2, Москва: 1}». Для некоторых источников данных региональное кодирование тривиально, но бывают довольно сложные случаи. В качестве примера можно привести данные о статистике языков страниц сайта. Языки, очевидно, имеют связь с регионами (странами), но нельзя просто преобразовать фактор «на веб-сайте найдено 60 страниц на английском языке и 50 – на испанском» в фактор «языки страниц: {Великобритания: 60, Испания: 50}», так как Великобритания и Испания – далеко не единственные страны, где используются английский или испанский языки.

Гео-кодирование жизненно важно для нашей системы, так как система должна иметь представление о регионах и их связях. Система должна знать, что факторы «найдены почтовые коды 127560, 127274», «найдены телефоны +7(495)1234567, 8(499)1234567» и «найдено 6 страниц со словом Москва в заголовке» относятся к одному региону – городу Москве, в то время как фактор «найден телефон +7 (7172) 74 55 24» относится к совсем другому региону.

2.4. Переход к относительным факторам

Чтобы создать единую модель, которая будет работать для нескольких классов, региональные факторы должны быть преобразованы из «абсолютных» факторов, которые зависят от конкретных регионов, в «относительные». В формулах, описывающих относительные факторы, не упоминаются конкретные регионы. Вместо этого используются формулы, вычисляющие значение фактора относительно некоторого выделенного «текущего» региона. Пример абсолютного фактора: «Количество упоминаний региона Москва». Примеры относительных факторов: «Значение региона, имеющего максимальное значение, кроме текущего региона», «Значение текущего региона, делённое на сумму значений остальных регионов».

2.5. Вектор значений факторов

Каждый источник данных поставляет значения для финального вектора значений факторов. Простые источники данных могут давать одно или несколько численных значений, которые просто помещаются в финальный вектор. Для источников данных, дающих несколько значений (таких, как языковая статистика), в финальный вектор, помимо самих значений, помещаются доли этих значений. Региональные источники (источники, которые для каждого веб-сайта и региона дают некоторое значение) дают по 15 значений для финального вектора значений факторов.

Опишем конкретные формулы для значений факторов, полученных из региональных источников. Пусть задан определённый веб-сайт, и мы хотим получить значения для определённого регионального источника данных. Пусть $v(r)$ – это значение, которое источник данных сопоставляет некоторому региону r (например, количество телефонов из региона r). Пусть $p(r)$ – это априорная вероятность встретить сайт из региона r (частота региона в обучающем множестве). Значения этих функций для множества регионов R просто равняются сумме значений функций для регионов множества: $v(R) = \sum v(r), r \in R$, $p(R) = \sum p(r), r \in R$. Определим несколько множеств регионов: All – это множество всех регионов; Cur – это регион текущей пары <сайт,регион>; $Rest = All \setminus Cur$ - множество всех регионов, кроме текущего; $Rival$ – «регион-конкурент» – не текущий регион с самым большим значением: $Rival = argmax\ value(r), r \in Rest$; $RRest = Rest \setminus Rival$ – это множество всех регионов, кроме текущего и региона-конкурента. Определим функции $ratio(R) = v(R)/v(All)$, $rel(R) = v(R)/v(Cur)$, $nratio(R) = ratio(R) \cdot p(All)/p(R)$. Когда 4 функции ($v$, $ratio$, rel, $nratio$) применяются к 4 региональным множествам (Cur, $Rival$, $Rest$, $RRest$) мы получаем 16 значений, но, так как $rel(Cur) \equiv 1$, мы исключаем эту комбинацию.

2.6. Обучение модели

Для обучения классификационной модели мы создаём таблицу значений факторов, с вектором факторов для каждой *гипотезы* вида <сайт, регион>. Рассматриваются не все комбинации сайтов и регионов, так как это приведёт к нежелательному линейному росту размера таблицы и длительности вычислений. Вместо этого для каждого веб-сайта формируется множество *регионов-кандидатов*. Регион включается в список кандидатов для некоторого веб-сайта, только если есть какая-либо связь между сайтом и регионом (например, один из источников данных даёт для данного веб-сайта и региона ненулевое значение). Благодаря такому подходу, среднее количество регионов-кандидатов для веб-сайта намного меньше полного количества регионов. Более того, добавление в классификатор новых регионов практически не влияет на количество регионов-кандидатов.

Для машинного обучения используется алгоритм MatrixNet [5], основанный на градиентном бустинге полных двоичных деревьев решений. Применяя обученную модель к вектору факторов, соответствующему паре <сайт, регион>, мы получаем вероятностную оценку релевантности региона веб-сайту. Множество регионов-кандидатов сайта фильтруется: регион попадает в результирующее множество, если полученная для него оценка релевантности оказывается выше порога, подобранного

при обучении. Пороги выбираются так, чтобы максимизировать F_1-меру результата на обучающем множестве.

2.7. Многоуровневая система классификации

Страны и города довольно сильно различаются как регионы. Поэтому, хотя мы могли бы использовать одну модель для детектирования и городов, и стран, лучше иметь отдельные модели («страны», «города») для разных типов регионов.

Кроме того, есть две принципиально разные задачи, связанные с геоклассификацией. Первая – это многозначная классификационная задача отнесения веб-сайтов к конкретным регионам. Другая задача – это бинарная классификационная задача определения того, принадлежит ли сайт к какой-нибудь конкретной стране (или городу), или не имеет регионального фокуса (национальный или международный сайт). Для второй задачи используются дополнительные бинарные классификаторы. Задача «общестранового» классификатора – отделить сайты, которые должны быть отнесены к какому-то городу, от сайтов, которые не должны быть отнесены к конкретным городам. Аналогично «международный» классификатор отделяет сайты, которые должны быть отнесены к какой-то конкретной стране, от международных сайтов, которые не должны быть отнесены к конкретным странам. Результаты этих бинарных классификаторов – числа, характеризующие вероятность того, что веб-сайт не должен быть отнесён ни к одному городу (стране). Между всеми классификаторами («города», «страны», «общестрановый» и «международный») есть связи – классификаторы используют результаты друг друга. Схема зависимостей классификаторов (а также используемые ими признаки) изображена на рис. 1.

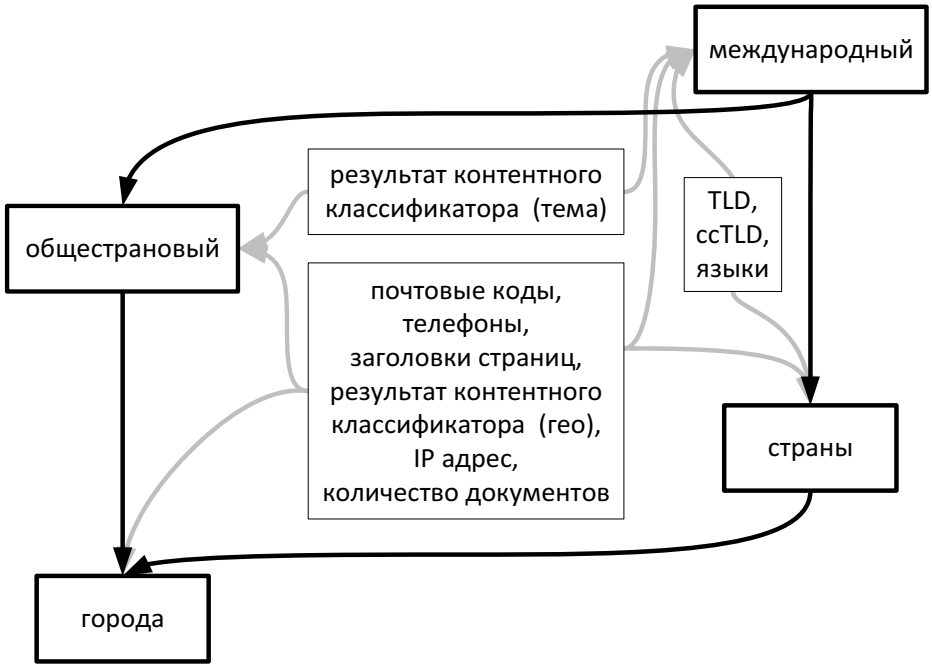

Рис. 1. Зависимости между классификаторами и источниками данных

3. Экспериментальные результаты

3.1. Набор данных

В качестве обучающего множества использовался каталог веб-сайтов, поддерживаемый компанией Яндекс. Каталог поддерживается и пополняется профессиональными редакторами и содержит большую коллекцию релевантных веб-сайтов, разбитую по большому множеству тематических и региональных рубрик. Редакторы каталога вручную приписывают веб-сайтам региональные и тематические категории. Для экспериментов использовались отнесения сайтов к регионам уровня города, страны, а также международного уровня. Экспериментальный набор данных содержал все доступные региональные отнесения и состоял из примерно 115000 веб-сайтов, отнесённых к 344 регионам разного уровня (285 городов, 58 стран и регион «планета Земля»).

3.2. Условия экспериментов

3.2.1. «Одна модель на тип региона» против «модели на каждый регион»

Как было сказано ранее, нашей целью было построение классификационной модели, которая могла бы устранить проблему нехватки обучающих данных для слабо представленных регионов. Поэтому базовая система, которую мы использовали для сравнения, была практически идентична нашей системе (использовала те же самые векторы признаков и алгоритм машинного обучения), но строила для каждого региона отдельную модель. Сложность (количество настраиваемых параметров) такой базовой системы была на порядки выше, чем сложность нашей системы, использующей «одну модель на тип региона», что давало ей (базовой системе) преимущество, позволяя настраивать каждую модель на особенности конкретного региона.

3.2.2. Текстовые факторы против остальных факторов

Обычно наша система использует в качестве фактора результат внешнего текстового классификатора. Мы знаем, что этот классификатор внутри строит по модели на каждый регион. К этому можно придраться, сказав, что система, использующая результат такого классификатора, не следует идее «одна модель на тип региона». Это не так, поскольку, с точки зрения нашей системы, внешний текстовый классификатор – это непрозрачный источник данных, внешний по отношению к системе. Тем не менее, мы всё-таки решили оценить качество нашей системы без текстовых факторов. Мы также отдельно измерили качество текстового классификатора. Нашей целью было показать, как объединение разных факторов в одной системе позволяет достигнуть лучшего качества классификации.

3.3. Оценка качества классификации

Для оценки качества нашей системы и сравнения с другими системами мы использовали F_1-меру – широко используемую меру оценки качества классификации, основанную на мерах точности и полноты. Использование F_1-меры позволило нам

Таблица 1. Микро- и макроусреднённая F_1-мера

	Страны	**Города**		**Междунар.**	**Общестран.**
		микро	**макро**		
наша система	**0.93**	**0.83**	**0.68**	**0.65**	**0.70**
с. без текстовых факторов	0.92	0.79	0.66	0.59	0.66
с. «по модели на регион»	0.93	0.80	0.51	0.65	0.70
текстовые факторы	0.91	0.63	0.34	0.56	0.63

получить одну меру качества, которую можно использовать для сравнения с другими результатами исследований. Таблица 1 показывает микроусреднённую F_1-меру для регионов каждого типа (города, страны, а также категории «общестрановый» и «международный»). Для городов также было вычислено значение макроусреднённой F_1-меры. Мы хотели посмотреть, как размер обучающей выборки для региона влияет на качество классификации. Рисунок 2 сравнивает значения F_1-меры различных систем для городов с разными размерами обучающей выборки. Рисунок 3 показывает для каждого города прирост качества нашей системы в сравнении с базовой системой. Мы выбрали города для демонстрации преимущества нашей системы классификации. Страны, обычно, имеют достаточно веб-сайтов, чтобы построить хороший классификатор, используя только веб-сайты из этих стран. Города намного чаще не имеют достаточного количества обучающих данных и, следовательно, должны сильнее выигрывать от использования нашего подхода.

3.4. Результаты

Экспериментальные результаты показывают, что наша система, использующая «одну модель на тип региона», даёт лучшие результаты, чем система, использующая «одну модель на каждый регион». Оба графика подтверждают наше предположение о том, что наша система работает лучше обычной системы для слабо представленных регионов. Для регионов с большим объёмом обучающей выборки наша система работает так же или незначительно хуже. Качество классификации по странам и категориям «международный»/«общестрановый» одинаково (см. табл. 1), а качество для крупных городов падает не более чем на 0.015 F_1-меры (см. рис. 2 и рис. 3) по сравнению с системой «одна модель на каждый регион». На регионах же, которые имеют менее 500 положительных примеров в обучающей выборке, наша система превосходит систему с «одной моделью на регион» . Преимущество становится сильнее с уменьшением количества положительных примеров для региона. Качество классификации остаётся довольно высоким, даже когда регион почти (или полностью) отсутствует в обучающей выборке. Микроусреднённая F_1-мера нашей системы всего на 0.03 выше, чем у обычной системы, но *макро*-усреднённая F_1-мера, показывающая среднее по регионам качество, увеличивается значительно – на 0.17.

3.5. Классификация без обучения

Интересной особенностью нашей системы является возможность определения регионов, отсутствующих в обучающем множестве. Классификационная модель не получает какой-либо информации о том, к какому веб-сайту и региону относятся

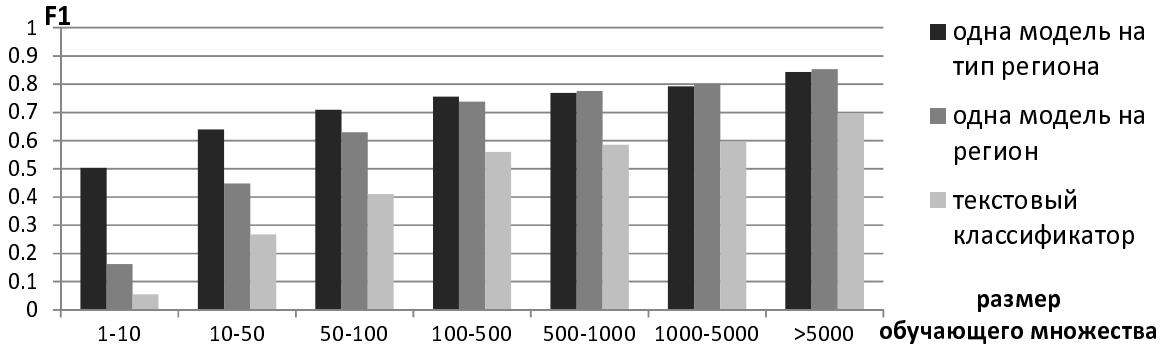

Рис. 2. Зависимость качества классификации от объёма обучающего множества

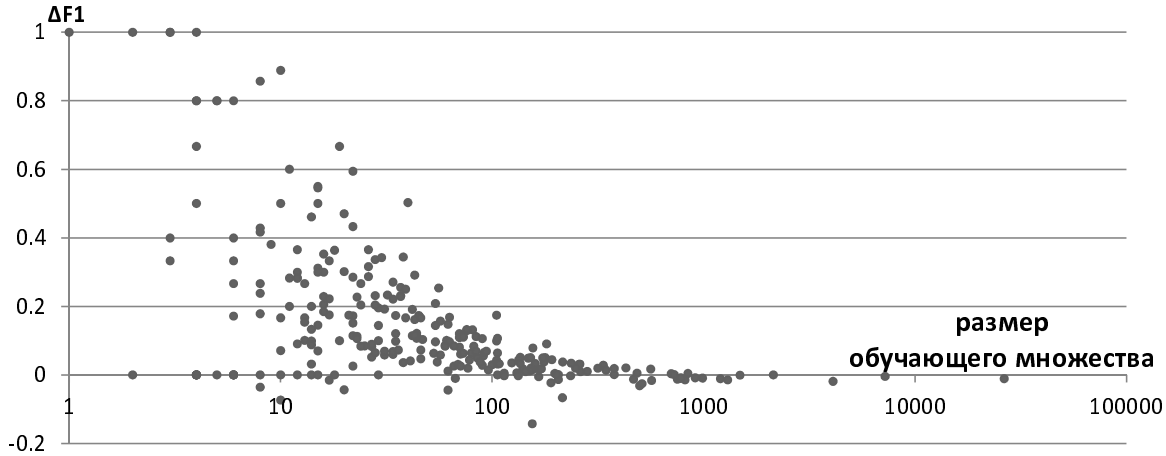

Рис. 3. Прирост F_1-меры при использовании нашей системы

элементы вектора значений факторов. Модель лишь использует факторы, отражающие связь между веб-сайтом и регионом и поэтому никак не зависит от конкретного рассматриваемого региона. Мы можем использовать модель для проверки релевантности веб-сайту любого региона, а не только такого, на котором была обучена модель. Чтобы доказать такую возможность, мы провели следующий эксперимент. Множество городов из обучающей выборки было случайным образом разбито на обучающее и тестовое множество (так, чтобы веб-сайты поделились в отношении 2 : 1). Мы обучили городскую классификационную модель, используя веб-сайты, относящиеся к регионам из обучающего множества, и применили её к тестовому множеству. Качество классификации оказалось таким же, что подтвердило возможность классификации веб-сайтов, относящихся к регионам, отсутствующим в обучающем множестве.

4. Выводы

В данной работе была представлена система региональной классификации, использующая классификационные модели, построенные для разных типов регионов, а

не для отдельных регионов. Это позволило нам объединить разные виды факторов (факторы, основанные на извлечении сущностей и текстовые факторы) в одной классификационной системе и улучшить качество классификации для веб-сайтов, относящихся к регионам, которые имеют недостаточное количество примеров в обучающей выборке. Система смогла успешно определить даже регионы, которые не были представлены в обучающей выборке.

В будущем планируется улучшить качество классификации, используя анализ связей между веб-сайтами.

Список литературы

1. Amitay E., Har'El N., Sivan R., and A. Soffer. Web-a-where: geotagging web content. SIGIR. ACM, 2004. P. 273–280.

2. Cheng Z., Caverlee J., and Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. CIKM, 2010. P. 759–768.

3. Crandall D. J., Backstrom L., Huttenlocher D., and Kleinberg J. Mapping the world's photos. WWW. ACM, 2009. P. 761–770.

4. Ding J., Gravano L., and Shivakumar N. Computing geographical scopes of web resources. VLDB, 2000.

5. Gulin A. and Karpovich P. Greedy function optimization in learning to rank., 2009.

6. Liu T.-Y. Learning to rank for information retrieval // *Foundations and Trends in Information Retrieval.* 2009. 3.

7. Pyalling A., Maslov M., and Braslavski P. Automatic geotagging of russian web sites. WWW, 2006. P. 965–966.

8. Pyalling A., Maslov M., and Trifonov S. Automatic classification of websites. RCDL, 2008.

9. Qi X. and Davison B. D. Web page classification: Features and algorithms // *ACM Comput. Surv.* 2009. 41.

10. Serdyukov P., Murdock V., and van Zwol R. Placing flickr photos on a map. SIGIR, 2009. P. 484–491.

11. Zong W., Wu D., Sun A., Lim E.-P., and Goh D. H.-L. On assigning place names to geography related web pages. JCDL. ACM, 2005. P. 354–362.

Unified Classification Model for Geotagging Websites

Volkov A.N.

Yandex LLC
Leo Tolstoy St., 16, Moscow, 119021, Russia

Keywords: geotagging, classification models, machine learning

The paper presents a novel approach to finding regional scopes (geotagging) of websites. Unlike the traditional approaches, which generally involve training a separate classification model for each class (region), the proposed method is based on training a single model which is used for all regions of the same type (e.g. cities). This approach is made possible by the usage of "relative" features which indicate how a selected region matches up to other regions for a given website. The classification system uses a variety of features of different nature that have not been yet used together for machine-learning based regional classification of websites. The evaluation demonstrates the advantage of our "one model per region type" method versus the traditional "one model per region" approach. A separate experiment demonstrates the ability of the proposed classifier to successfully detect regions which were not present in the training set (which is impossible for traditional approaches).

Сведения об авторе:
Волков Алексей Николаевич,
ООО «Яндекс», разработчик программного обеспечения;
Московский Физико-Технический Институт, аспирант