Определение географического местоположения интернет ресурсов

©Дмитрий Соловьев, Андрей Калинин

Поиск@Mail.Ru

d.soloviev@corp.mail.ru, kalinin@corp.mail.ru

Аннотация

В статье рассматривается решение задачи определения географического местоположения веб-ресурса. Документ включает описание двух методов привязки сайта к географии: анализ статистики посещаемости и анализ контента страниц ресурса, основанного на скрытой Марковской модели.

1. Введение

С первых дней становления интернета много усилий было потрачено на совершенствование организации, навигации и поиска документов. Возможно, что поиск по инвертированному индексу ключевых слов является, сегодня, одной из наиболее полезных техник, позволяющей пользователям находить информацию по заданной теме. В тоже время, глобальное расширение интернета приводит к тому, что количество найденной информации, получаемой пользователем при поиске с использованием только ключевых слов, слишком велико. Используя одни и те же слова, разные люди, в зависимости от условий, хотят получить различные результаты. Например, задавая запрос [установка окон ПВХ], человек, проживающий в Хабаровске ожидает увидеть в результатах те страницы, которые относятся к локальным компаниям, занимающихся установкой окон, а не получать страницы компаний расположенных в Екатеринбурге. Такая задача не очень хорошо решается при помощи использования поиска по ключевым словам. В то же время он может стать отправной точкой формирования высокоуровневых семантических запросов, которые могут использоваться для нахождения такой информации. Таким образом, можно сформировать дополнительные метаданные страницы, используя которые, можно повысить качество ответа поисковой машины. Метаданными, описывающими локализацию страниц в поиске, является информация о географическом положении ресурса.

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

Поиск информации, основанный на географических критериях достаточно общая задача. Примеры могут включать: путешественников, которые хотят получить информацию о цели путешествия; аналитиков, подготавливающих отчет о данной местности, планирование расширения бизнеса в других регионах. За более подробным объяснением областей, в которых используется поиск информации с учетом географических критериев можно обратиться к [4]. В любом случае эту задачу можно разбить на две составляющие: определение географического положения ресурса и поиск информации с учетом геоданных. Нужно отметить, что понятие географического положения ресурса не обязательно подразумевает его физическое нахождение в конкретной местности, а больше связанно с отображением его в реальные объекты, например: компании, региональные СМИ, или объединение пользователей по региональным тематикам, например, региональный сайт бесплатных объявлений. Для решения первой части поставленной задачи нужно определиться с источниками и методиками экстракции информации. К таким техникам можно отнести: извлечение информации из каталогов или анализ базы WHOIS. За подробным описанием таких методик экстракции геоинформации можно обратиться к [5]

Интересным источником географической информации может являться непосредственно сам контент. Например, веб-страницы сайта компании могут содержать адреса и телефоны, как головного офиса, так и региональных ее представительств. Страницы, сообщающие о новых событиях, также могут содержать информацию о месте, где данное событие проводится.

Многие веб-ресурсы рассчитаны на посещение пользователями из определенного региона, например, к таким ресурсам можно отнести городские порталы, местные издания газет. Кроме того, такие ресурсы могут не содержать прямых указаний на свое географическое месторасположение непосредственно в контексте страниц. Поэтому, в этом случае можно использовать информацию, получаемую от пользователей посещающих данный ресурс. В данном случае можно оперировать статистикой посещений пользователей с учетом их региональной привязки.

Нужно отметить, что уровень детализации геопривязки ресурса в зависимости от условий может меняться. В некоторых случаях однозначно определить геопринадлежность веб-ресурса нельзя из-за его нахождения в различных географических контекстах. Одни веб-ресурсы могут получать жесткую привязку к местности, вплоть до номера дома; другие могут содержать только укрупненную информацию: город, район или область.

Целью публикации является разработка и исследование методов привязки ресурса к географии для получения качественных данных, которые могут быть впоследствии использованы для поиска информации с учетом географических критериев. В статье предлагаются к рассмотрению две методики, позволяющие решить задачу определения географической дислокации ресурса, а так же приводятся оценки точности и полноты этих методов. Дополнительно рассматривается возможные направления использования полученных данных. Для решения поставленной задачи мы использовали непосредственно контент веб-сайтов, а также анализ посещаемости ресурсов пользователями.

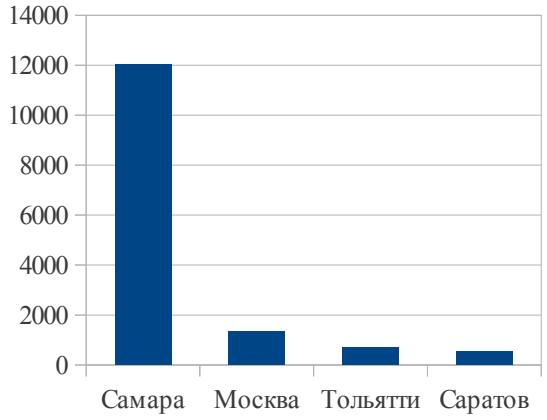
2. Геопривязка ресурса на основе анализа посещаемости пользователей

Одним из источников информации о географическом местоположении ресурса могут служить данные о пользователях и распределение их посещаемости по регионам. Можно предположить, что если сайт интересен и посещаем пользователями одного региона, то он имеет тематическую привязку, направленную на данный регион, и как следствие можно осуществить локализацию данного ресурса. В качестве наиболее характерного примера можно взять электронное издание «Из рук в руки» (irr.ru). Ресурс имеет деление на поддомены третьего уровня, каждый из которых принадлежит определенному городу или региону. В таблице 1 приведены примеры поддоменов третьего уровня, используемых в качестве региональных. Можно предположить, что, например, поддомен samara.irr.ru будет более интересен пользователям, проживающим в Самаре, чем пользователям, проживающим в Нижнем Новгороде. Таким образом, если построить гистограмму распределения количества пользователей посетивших ресурс по регионам, то у рассматриваемого ресурса будет наблюдаться максимум посещений в регионе Самара. На гистограмме 1 приведен пример распределения посещаемости для сайта samara.irr.ru за сутки. Можно аннотировать частоту посещений ресурса для заданного региона как f_r .

Для корректного решения поставленной задачи так же необходимо принять во внимание тот факт, что распределение количества пользователей по регионам не равнозначно, т. е. пользователей, использующих Интернет в Москве больше, чем в

Поддомен	**Регион**
http://saint-petersburg.irr.ru	Санкт-Петербург
http://nizhniynovgorod.irr.ru	Нижний Новгород
http://samara.irr.ru	Самара
…	...

Таблица 1. Региональные поддомены сайта http://irr.ru



Гистограмма 1: Распределение посещаемости сайта samara.irr.ru за сутки

Самаре. Это обстоятельство нужно учитывать при построении гистограммы, и с этой целью были введены нормирующие коэффициенты, отражающие распределение пользователей Интернета по регионам. Таким образом, частотная характеристика региона в гистограмме учитывает еще и коэффициент неравномерности распределения k . Тогда нормализованная частота посещений в регионе F_r будет равна:

$$F_r = k * f_k$$

Следующим шагом определения принадлежности ресурса региону является выбор периода, за который будет рассчитана гистограмма распределения. Путем проведения ряда экспериментов был определен оптимальный период равный одному месяцу. Как показали эксперименты, усреднение гистограммы за полный период может быть не корректно, поскольку в некоторых случаях дает ошибочный результат. Например, при сильных всплесках посещаемости ресурса пользователями других регионов, в коротком промежутке времени. Такое поведение характерно для региональных издательств, когда на ресурсе размещаются публикации описывающие положение дел в других географических регионах. Чтобы исключить влияние всплесков посещаемости, весь диапазон делится на сегменты, и для каждого сегмента F_r^s

рассчитывается нормализованная частота посещений пользователей по регионам F_r^s :

$$F_r^s = k_r * f_r^s$$

- где f_r^s - частота посещений для региона, рассчитанная в рамках одного сегмента.

Далее для каждого сегмента определяется регион с максимальной нормализованной частотой:

$$F_{rmax}^s = maxF^s$$

- где F^s множество значений частот для данного сегмента.

Для каждого региона в рамках одного сегмента вычислим величину R_r^s такую что:

$$R_r^s = \begin{cases} 1, & F_r^s = F_{rmax}^s \\ 0, & F_r^s < F_{rmax}^s \end{cases}$$

По всем сегментам региона рассчитывается агрегированная величина R_{ragr}

$$R_{ragr} = \frac{\sum_{R_{r,i}^s \in N, i=1}^{N} R_{r,i}^s}{N}$$

- где N - общее количество сегментов данных. По результатам определяется принадлежность ресурса региону R , на множестве N_{agr} :

$$site \in R, R_{ragr} = maxN_{agr} \wedge R_{ragr} \geq K_{tr}$$

Значение порогового коэффициента K_{tr} задается таким, чтобы исключить попадание ресурса в регион с низким рейтингом. В наших работах этот коэффициент принимался равным 0,6.

3. Геопривязка ресурса на основе анализа контента страниц ресурса

Как уже было сказано ранее, одним из источников информации о геопривязке сайта может служить сам контент. Как правило, такими источниками являются сайты организаций, на которых публикуется информация о местах их расположения, включающая в себя адреса и телефоны. Извлекая эту информацию из страниц сайтов, можно осуществить геопривязку сайта более точно, чем при помощи метода описанного в предыдущей секции.

Решение задачи извлечения информации, в нашем случае, было разбито на несколько частей:

- определение типовых шаблонов сайтов, на которых может размещаться информация о месте расположения организации;
- извлечение кандидатов для последующей привязки сайта к географической информации;
- фильтрация кандидатов.

3.1 Определение типовых шаблонов

На этом этапе была проанализирована структура сайтов организаций, и на основе полученной информации были отобраны наиболее часто встречающиеся типовые шаблоны сайтов. По результатам анализа можно выделить следующие три этапа:

- Поиск адресов на корневой странице сайта.
- Поиск ссылок на страницу «Контакты».
- Поиск адресов на странице «Контакты».

Как показывают результаты экспериментов, одним из наиболее часто встречающихся мест расположения контактной информации является корневая страница. В тоже время, эта страница может не являться достоверным источником информации, поскольку существуют сайты, например, размещающие объявления, включающие в себя хорошо читаемые адреса, часть из которых может быть включена в главную страницу. В данной работе для адресов, извлеченных с корневой страницы сайта, применяется дополнительная фильтрация.

Еще одним наиболее часто встречающимся местом расположения контактной информации является страница «Контакты». Как правило, на нее существуют ссылки с главной страницы, и они в большинстве случаев подчиняются ряду правил. Например, текст ссылки может содержать слово «контакты», «О нас» и т. д.

3.2 Извлечение кандидатов

Как уже говорилось ранее, при анализе сайта производятся попытки извлечь информацию об адресах из корневой страницы сайта и (или) со страницы «Контакты». Существуют множество подходов к извлечению информации из неструктурированных текстов [6,7]. Мы в своей работе использовали комбинированную методику, основанную на словарном поиске города, вероятно входящего в адрес, и скрытой Марковской модели, которая позволяет оценить последовательность слов окружающих найденный город. Левый и правый контекст оценивался отдельно. Поскольку в данной задаче требуется вычислить только вероятность появления последовательности адреса в окрестности города, то для решения использовался алгоритм «forward-backward».

Если рассматривать элементы почтового адреса как состояния модели, то в рамках локально решаемой задачи количество состояний можно значительно уменьшить, что приведет к упрощению самой модели. Например, можно объединить все модификаторы улиц в одно состояние. В этом случае — улица, шоссе, переулок... образуют состояние ms_{state} . Таким же образом, можно транслировать множество известных географических названий в одно состояние последовательности. Сформированное таким образом множество

tw_{state}	Описывает город, найденный в словаре
cn_{state}	Описывает страну, найденную в словаре
mt_{state}	Описывает один из известных модификаторов города (г., сел...)
ms_{state}	Описывает один из известных модификаторов улицы (ул, ...)
mh_{state}	Описывает один из известных модификаторов дома.(д., ...)
mf_{state}	Описывает модификатор квартиры (кв., офис...)

Таблица 2. Пример состояний скрытой Марковской модели для адреса

состояний модели

$$S = \{s_{1,} s_{2,} .., s_n\}, n = 19$$

Уменьшение количеств состояний модели приводит к необходимости вводить матрицы проекций элемента адреса на состояние модели и в то же время приводит к значительному уменьшению размера обучающего множества, на основе которого определяются последовательности смены состояний $s_{1,}..., s_n$. Затем строится матрица вероятностей переходов между состояниями $P_{s'}(s \mid v)$, где $s' \in S$ - предшествующее состояние системы; $s \in S$ - текущее состояние системы; v - рассматриваемый элемент последовательности, принадлежащий множеству $V = \{v_{1,}..., v_m\}$. Можно обозначить состояние, которое принимает система во время t как q_t, а наблюдаемую величину в момент t как y_t. Элементы матрицы вероятностей перехода из состояния i в состояние j обозначим как $a_{ij} = p(q_{t+1} = s_j \mid q_t = s_i)$, а вероятность получить данные v_k в состоянии j обозначим как $b_j(k) = p(v_k \mid s_j)$. Обозначим данные через $D = d_{1....} d_T$ (последовательность наблюдаемых, d_i принимает значение из V). Также для построения модели нужно учесть начальное распределение $p = \{p_j\}, p_j = p(q_1 = s_j)$. В нашем случае мы по полученной модели $Л = (A, B, p)$ и последовательности D найдем $p(D \mid Л)$. Формально можно записать:

$$p(D \mid Л) = \sum_Q p(D \mid Q, Л) p(D \mid Л)$$

Используя построенную модель и зная возможную точку расположения адреса на странице, найденную при помощи словаря городов, производим оценку контекста, в котором находится найденный город, используя процедуру «forward-backward». За более подробным описанием алгоритма можно обратиться к [1].

3.3 Фильтрация кандидатов

Извлеченные адреса проходят фильтрацию. Первый этап фильтрации заключается в том, что из страницы также извлекается дополнительная информация, как например, телефон, который ставится в соответствие одному или нескольким адресам, извлеченным из страницы.

Одно из сопоставлений это проверка кода региона, указанного в номере телефона на соответствие городу, указанному в адресе.

Второй этап фильтрации включает набор эмпирических правил, которые накладываются на выделенный адрес. К таким правилам, например, относится ограничение на возможное количество цифр, содержащихся в номере дома. После применения ряда правил извлеченный адрес либо принимается, как один из адресов, описывающий местоположение организации, либо отклоняется.

4. Использование извлеченной географической информации

Извлеченная географическая информация методами, описанными ранее, используется для решения задач связанных с ранжированием документов в поиске, а так же привязки найденных сайтов к картографическому сервису.

Для осуществления ранжирования с учетом географической информации, необходимо знать регион пользователя, который, прежде всего, определяется по его IP адресу. Если в регионе пользователя существуют локальные сайты, которые, в том числе, релевантны запросу, то они попадают в региональное ранжирование.

При использовании в сервисе картографии данные проходят дополнительную нормализацию и проверку на соответствие реальным адресам. Данные для проверки берутся из картографической базы.

5. Оценка других источников извлечения географической информации

В нашей работе мы опираемся на два основных источника получения информации о географии сайта: анализ статистики посещений сайта и извлечение информации непосредственно из страниц сайта. Есть так же и другие источники,

	Количество сайтов взятых для анализа	Сайтов получивших географическую привязку	Точность геопривязки веб ресурса	Полнота охвата исходных данных
Анализ контента страниц	20 миллионов.	330604	97%	1,6%
Анализ статистики посещаемости	1 миллион	121609	76%	12%, или от всего множества 0,6%
Суммарно по сайтам.	20 миллионов	440213	80%	2,2%
Суммарно по страницам	3, 9 миллиарда страниц	1,3 миллиарда страниц, получили географический признак	-	33%

Таблица 4: Полнота охвата базы сайтов

Регион	Сайты, попавшие в регион
Санкт-Петербург	spbgu.ru, flot.com, 5-tv.ru, saint-petersburg.ru, newspb.ru
Екатеринбург	oblgazeta.ru, doskaurala.ru, medgorodok.ru, urbc.ru, uralweb.ru
Киев	ati.com.ua, pregnancy.org.ua, football.ua, realt.ua, ukranews.com

Таблица 3: Пример сайтов, приписанных региону по сумме двух методов

которые можно было бы использовать для определения географической принадлежности сайта.

Например, в [5] описаны методы получения информации о месте положения сервера путем анализа маршрута IP пакета. К сожалению, многие сайты располагаются на площадках нелокальных провайдеров, и в полученной таким способом информации, будут содержаться ошибочные результаты.

Так же в качестве источника информации можно рассматривать и некоторые каталоги, в которых организации могут размещать информацию о себе, в том числе и о своем месте расположения. В силу некоторых особенностей устройство таких каталогов таково, что практически отсутствует механизм контроля введенной информации. По этой причине информация, содержащаяся в таких каталогах, может содержать как ошибки, так и подлоги.

6. Эксперименты

Для экспериментов была взята база страниц скачанных из интернета, содержащая порядка 20 миллионов сайтов и 3,9 миллиарда страниц. Из этих данных на основе анализа контента страниц проводилась географическая привязка сайта алгоритмом, описанным в п.3. Так же была проанализирована статистика посещаемости миллиона сайтов (данные Top@Mail.Ru), для определения привязки сайта к его географическому расположению на основе алгоритма описанного в п.2. При анализе статистики из рассмотрения исключались статистически не значимые данные, например, сайты с суточной посещаемостью менее 100 посетителей. После обработки, полученные результаты объединялись и загружались в единую базу для последующего решения задач связанных с ранжированием и отображением найденной информации. Результаты анализа обоих этапов приведены в таблице 3. Также в таблице представлена суммарная характеристика по обоим этапам. Нужно учесть, что при слиянии результатов частично произошло пересечение по сайтам. Так же было рассчитано, какое количество страниц получили географическую привязку для сайтов, принадлежащих регионам. Как видно из таблицы 3, суммарное количество сайтов, получивших геопривязку, не превышает трех процентов. В тоже время этот набор сайтов дает порядка одной трети всех станиц находящихся в базе. На диаграмме 1 приведено распределение сайтов по регионам, как видно из диаграммы самым большим регионом, в который попадают сайты, является Москва.

В качестве примера были взяты два самых крупных региональных города России и один из Украины, для них случайным образом отобрали пять сайтов, получившие в качестве географического признака идентификатор этого города. Результаты представлены в таблице 4

Для оценки точности мы отобрали случайным образом порядка 100 сайтов и проверили точность попадания географической привязки, проставленной суммарно по двум методам, и по каждому методу отдельно, сопоставив ее с реальной информацией доступной на сайте.

7. Выводы

В статье рассматриваются два метода получения информации для геопривязки ресурса: на основе анализа посещаемости пользователей и на основе анализа контента страниц ресурса. Как показали эксперименты, наиболее точным методом является метод, построенный на основе анализа контента. В его случае точность достигает 97%. Это обусловливается рядом ограничений, а именно:

Диаграмма 1: Распределение сайтов по регионам

использованием предопределенных шаблонов для нахождения страницы с адресом; использованием словаря городов; сопоставление номера телефона и города, а так же существующие формальные правила для записи адреса. Все эти ограничения позволяют достичь достаточно высокой точности, при определении географии веб-ресурса. С другой стороны, эти ограничения приводят к снижению полноты, в случаях, если адрес записан без прямого указания города или с неизвестным городом, если страница с контактами расположена по адресу, который не описан в известных шаблонах поиска.

Метод, реализованный на основе анализа статистики посещаемости, обладает большей полнотой относительно анализируемого множества сайтов (12%). В то же время, он обладает рядом недостатков: множество анализируемых сайтов ограничивается только данными, доступными из статистики посещаемости, а это всего 5% от общего множества сайтов; много статистически не значимых сайтов, порядка 87%, которые выпадают из рассмотрения; большая вероятность ошибки, чем в случае использования метода анализа контента, из-за неверного сопоставления IP адреса пользователя его реальному местоположению.

Таким образом, используя эти два метода, в эксперименте получили привязку к географии только 2,2. процента сайтов. В то же время, в количестве страниц это отношение составляет порядка 33% от всех страниц, взятых для анализа. Достаточно сложно оценить количество страниц, которые должны реально получить геопривязку, поэтому мы проводили оценку качества фильтрации региональных сайтов по географическим запросам. Оценки проводились независимо для трех различных регионов. В результате этого эксперимента, мы получили удовлетворительное качество ответов поисковой машины, по всем трем регионам.

Литература

[1] Cappe O., E. Moulines, T. Ruden. Inference in hidden Markov Models. Springer. 2005. 652 p.

[2] Han J. and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, 2006 .

[3] Jones C.B., R. Purves, A. Ruas, M. Sanderson, M. Sester, M.J. van Kreveld, R. Weibel. Spatial Information Retrieval and Geographical Ontologies An Overview of the SPIRIT Project. SIGIR 2002: In SIGIR'02, Tampere, Finland, 387-388. 2002.

[4] Larson R.R. Geographic Information Retrieval and Spatial Browsing. https://sherlock.ischool.berkeley.edu/geo_ir/PART1.html

[5] McCurley K.S. Geospatial Mapping and Navigation of the Web. 10th International World Wide Web Conference (WWW-2001), Hong Kong, ACM Press, p. 221-229 . 2001.

[6] Zheyuan Y. High accuracy postal address extraction from Web pages. 2007

[7] Прокофьев П. А. Использование методов извлечения информации при географической привязке текстов на русском языке. Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции. Труды RCDL. 2009. с. 254-258

Determining the Geographic Location of Internet Resources

© D.V. Soloviev, A.L. Kalinin.

This paper describes extraction of geospatial information for Web resources. The document includes a description of two methods for binding site to geography based on: analysis of visit statistics and analysis of the web page's content by hidden Markov model (HMM).