Влияние представления Веб-сайта на качество классификации

© Михаил Кондратьев

Санкт-Петербургский Государственный Университет
mik_k@inbox.ru

Аннотация

В статье описаны результаты экспериментов по анализу влияния объема информации, используемой для построения описания Веб-сайта, на качество классификации. Эксперименты проводились на основе дорожек по классификации Веб-сайтов от РОМИП'2003 и РОМИП'2004.

1. Введение

Одним из популярных методов доступа к информации является использование структурированных электронных тематических каталогов. Создание и поддержка таких каталогов – непростая задача, все подходы к решению которой можно разделить на две группы:

- Использование труда экспертов в предметной области для классификации информационных ресурсов вручную;
- Использование машинных методов автоматической классификации информационных ресурсов.

Первая группа подходов позволяет получать каталоги содержащие минимальное число ошибок, но высокая стоимость в значительной мере ограничивает масштабируемость этого подхода. Природа Интернет, где огромный объем доступной информации подвержен динамичным изменениям [10], обуславливает значительный интерес к методам автоматической классификации информационных ресурсов Интернет.

Вообще, задача автоматизации тематической классификации текстовой информации изучается на протяжении нескольких

десятков лет, и за последние десять лет ряд работ был посвящен классификации в контексте Веб. В то же время, число работ? посвященных проблеме создания и поддержки каталогов Веб-сайтов, а не Веб страниц, относительно невелико. Нам не известно ни одной крупномасштабной попытки провести сравнительный анализ эффективности решения этой задачи (кроме РОМИП'2003).

Формально процесс классификации можно разделить на несколько шагов:

- Построение описаний категорий, т.е. отображения категории в общее n-мерное пространство
- Построение описаний классифицируемых ресурсов, т.е. отображения ресурса в общее n-мерное пространство
- Вычисление степени соответствия между категориями и ресурсами, как функции от их представлений в n-мерном пространстве.

Математическая статистика предлагает множество метрик для вычисления степени соответствия. Вопросы отображения текстового ресурса в n-мерное пространство много обсуждались как контексте текстового поиска, так и текстовой классификации. Поэтому для целей этой работы мы решили использовать один из широко известных способов классификации многомерных точек и один из стандартных способов отображения текстовой информации в точки многомерного пространства.

Предметом исследования этой работы являлся вопрос, какую текстовую информацию лучше всего использовать для представления сайта. Наивные подходы – использовать только стартовую страницу или текст со всех страниц сайта имеют ряд интуитивно слабых мест. Например, стартовая страница может совсем не содержать текста, а полное представление большого сайта может быть слишком расплывчатым и подходить под описание всех категорий.

Семинар РОМИП [3] предоставляет удобную возможность не только провести крупномасштабную оценку, но и сравнить полученные результаты с результатами других систем. В рамках участия и подготовки к семинару был поставлен ряд экспериментов с использованием различных методик построения профилей, для чего использовались коллекции сайтов РОМИП'2003 и РОМИП'2004.

2. Методы построения профилей

В рамках этой работы мы рассматривали следующие виды профилей;

2.1. Стартовая страница сайта

В качестве представления сайта использовалось содержимое его стартовой страницы. Интуитивно кажется, что вся основная информация должна быть доступна со стартовой страницы.

Очевидным достоинством этого подхода является небольшой объем информации, используемой для описания сайта, что сказывается на быстродействии.

2.2. Окрестность стартовая страницы сайта

Сайт представляется как стартовая страница и ее окрестность, т.е. все страницы, которые достижимы по ссылкам со стартовой страницы.

Мотивацией к использованию этого подхода служило предположение, что страницы, доступные по ссылкам со стартовой страницы раскрывают темы, упомянутые на стартовой странице.

2.3. Страницы в корне сайта

Для построения профиля использовались страницы, расположенные в корне сайта.

В основе такой стратегии лежит предположение, что основные страницы, посвященные тематике сайта, лежат в одном каталоге – в корне, тогда как служебные страницы и страницы посторонней тематики могут быть сгруппированы по каталогам. Аналогичная идея высказывалась в работе [7], где для классификации предлагалось задействовать URL адрес ресурса.

2.4. Текст ссылок и текст в окрестности ссылок

Интересным подходом является стратегия формирования классифицируемого профиля как множества текстов ссылок и некоторой текстовой окрестности вокруг них. При постановке экспериментов с коллекцией РОМИП использовалась окрестность ссылки радиусом в 10 слов. В случае если окрестности перекрывались, текст добавлялся только один раз.

Исходя из предположения, что ссылка и ее окрестность несут достаточно информации для принятия решения о переходе по ней, отобранные данные должны достаточно точно описывать страницы сайта. Дополнительным преимуществом является малый размер полученного профиля.

3. Исследовательский прототип “Золушка”

Разработанный нами прототип системы основан на свободно распространяемом классификаторе rainbow [2]. Rainbow предоставляет довольно эффективную реализацию многих известных алгоритмов классификации, но для целей этой работы мы использовали лишь классификацию методом Байеса.

Для построения профилей сайтов, обеспечения взаимодействия с rainbow и обработки результатов классификации был разработан набор средств на языке Java.

Отметим, что на данный момент система не производит никакого морфологического анализа текста. В частности, вследствие отсутствия стемминга разные словоформы рассматривались как разные термы.

Несмотря на использование простых инструментов, выполнение заданий РОМИП’2004 не поставило перед нами каких-либо вычислительно неразрешимых задач. Для вычислений использовался компьютер с процессором Pentium 4 2,4GHz и объемом оперативной памяти 2 Gb. Построение профилей, обучение классификатора и категоризация ресурсов заняли менее суток.

4. Эксперименты на основе коллекции РОМИП’2003

На этапе подготовки к участию в РОМИП’2004 был поставлен ряд предварительных экспериментов с различными способами построения профилей. Целью подготовки было определить оптимальные стратегии для применения их во время участия в РОМИП’2004.

Предварительные эксперименты проводились по следующим правилам: в качестве обучающего множества использовалось обучающее множество, применявшееся в семинаре РОМИП’2003. Для классификации использовалась часть коллекции РОМИП,

оцененная во время проведения РОМИП'2003. На основе сравнения решения, принятого классификатором, и решения асессоров делалось заключение об эффективности каждой из методик.

В ходе предварительных экспериментов для каждой из стратегий построения профилей были выделены присущие ей недостатки.

4.1. Стартовая страница сайта

Достаточно большой процент главных страниц не содержит текстового наполнения. По оценкам, приведенным в [5], 17 процентов главных страниц сайтов не имеют информативного текстового содержимого. Вместо текстового наполнения страница может содержать только графические изображения, может быть выполнена с использованием технологии flash или может представлять из себя frameset.

Большое количество стартовых страниц имеет минимальное количество текстовых данных, недостаточное для принятия корректного решения классификатором.

Эта стратегия показала наихудший результат из всех рассматривавшихся.

4.2. Окрестность стартовой страницы сайта

Коллекция narod.ru характеризуется большой зашумленностью данных, что отмечалось участниками РОМИП'2003 [4]. Даже ссылки с главной страницы с достаточно высокой вероятностью ведут на документы, не относящиеся к основной тематике или содержащие шумовые данные. Примерами последних являются ссылки на коллекции анекдотов или галерей друзей.

Вследствие указанной особенности коллекции, добавление информации из окрестности к тексту главной страницы практически не улучшило точность классификации по сравнению с профилями, построенными на основе только стартовой страницы сайта. Вместе с тем полученные профили значительно превышали в объеме профили на основе стартовой страницы и требовали большего времени при построении.

4.3. Страницы в корне сайта

В используемой коллекции очень большая часть страниц располагается в корне сайта, что ведет к значительному объему профилей сайта и сильной зашумленности данных.

4.4. Текст ссылок и текст в окрестности ссылок

Получаемые в соответствии с этим способом профили сайта имеют наименьший объем. Вследствие этого, отсутствие морфологического анализа текста или попавшие в профиль шумовые данные могут значительно ухудшить результаты классификации.

Многие сайты используют ссылки для навигации между разделами. При построении профилей необходимо избегать многократного добавления информации о навигационных и подобных им ссылках в профиль. В силу временных ограничений и специфики коллекции при подготовке к семинару не было возможности исключать повторяющиеся фрагменты страниц на основе анализа DOM дерева документа (как, например, в работе [6]). Для уменьшения количества шумовых данных использовался следующий упрощенный подход: информация, связанная со ссылкой, не добавлялась в профиль, если он уже содержит ссылку с таким же текстом и окрестностью. Требование совпадения и текста ссылки, и текста окрестности должно исключить ссылки, встречающиеся в документах сайта в одном и том же контексте

Поставленные эксперименты показали, что наилучшим результатом обладает стратегия построения профилей, использующая информацию о ссылке и ее контексте. Для дальнейшего исследования этой стратегии были построены профили, использующие окрестности глубиной 2 и 3, то есть включающие информацию со страниц, достижимых из главной за два или три перехода. Преимуществом этой стратегии является малый результирующий объем профилей, позволяющий проводить классификацию за минимальное время.

На Рис. 1 приведены диаграммы, характеризующие точность классификации. Видно, что в экспериментах с профилем, построенным на основе текста ссылок документа и текстовой окрестности ссылки, наблюдается небольшой рост точности

классификации при переходе от окрестности глубины 1 (главная страница сайта и те страницы, на которые присутствуют ссылки) к окрестностям большей глубины. Видно так же, что при переходе от окрестности глубины 2 к окрестности глубины 3 прирост точности совсем незначителен. По-видимому, это происходит из-за увеличения влияния шума и поэтому можно предположить, что дальнейшее увеличение глубины также малополезно, хотя мы не проводили такого эксперимента.

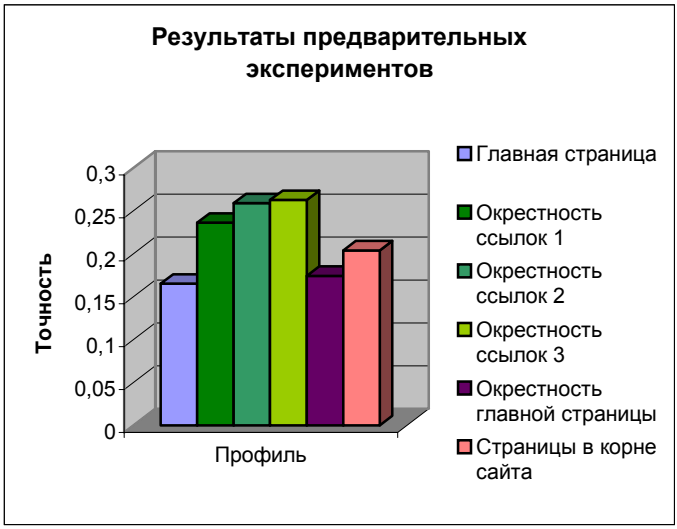

Рис. 1

Сравнение результатов, полученных в ходе экспериментов, позволило принять решение о выборе оптимальной стратегии построения профилей. Исходя из данных, приведенных на диаграмме, наиболее удачной представляется стратегия, использующая текст ссылок и их текстовое окружение. Как уже отмечалось, размер полученных профилей был минимален среди рассматривавшихся. На Рис. 2 приведен размер профилей, построенных для всей коллекции РОМИП'2003.

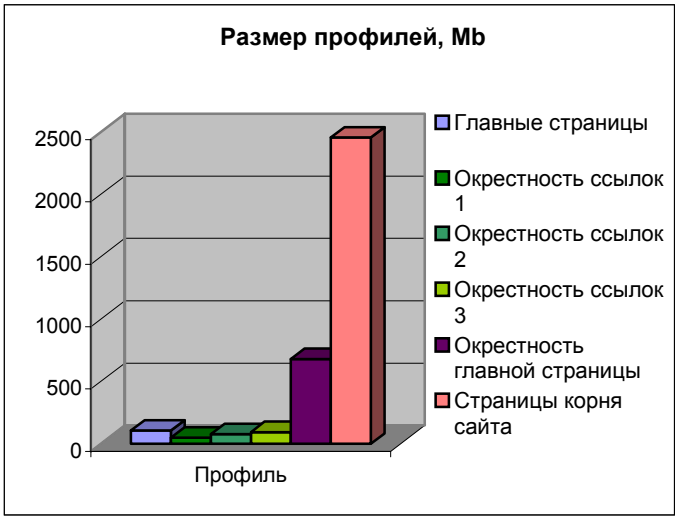

Рис. 2

В тоже время, коллекция Dmoz, часть которой использовалась как тренировочное множество для обучения классификатора в семинаре РОМИП'2004, кажется менее зашумленной, чем коллекция narod.ru. Косвенным подтверждением этого заключения является более жесткий механизм классификации документов в коллекции Dmoz. С учетом этого предположения было принято решение, что подход, основанный на построении профиля как текста главной страницы и ее окрестности, может использоваться для обучения классификатора. При этом классифицировать следует профили, построенные на основе текстовой окрестности ссылок, как показавшие себя наиболее удачными при работе с коллекцией narod.ru

5. Участие в РОМИП'2004

Целью участия в РОМИП'2004 являлась проверка результатов, полученных в ходе подготовительных экспериментов и сравнение результатов с результатами других участников.

5.1. Подготовка профилей

При проведении экспериментов для коллекций narod.ru и Dmoz были построены два вида профилей. Первый профиль формировался на основе текста ссылок и их окружения. Использовались данные главной страницы и документов из ее окрестности радиуса 3, конкатенирующиеся затем в один текстовый файл. Второй вид профилей строился как конкатенация текста главной страницы и страниц ее окрестности глубины 1. Размеры полученных профилей составили 110 и 1261 мегабайт для первого и второго видов профиля соответственно. При подготовке текста HTML теги вырезались из текста профилей.

5.2. Выполненные эксперименты

В рамках РОМИП'2004 были поставлены три эксперимента и их результаты переданы в оргкомитет для оценки. Использовались следующие комбинации профилей:

	Профиль обучающей коллекции	**Профиль рабочей коллекции**
Прогон 1	текст ссылки и ее окрестность	текст ссылки и ее окрестность
Прогон 2	страница и ее окрестность	текст ссылки и ее окрестность
Прогон 3	страница и ее окрестность	страница и ее окрестность

Следует отметить, что подготовка всех видов профилей для обеих коллекций и их классификация заняли менее 8 часов.

5.3. Анализ результатов

Результаты экспериментов оценивались двумя способами: weak и strong. В случае strong ресурс считается релевантным, если он признан релевантным всеми асессорами, в случае weak – если ресурс оценен как релевантный хотя бы одним из асессоров.

	weak	**strong**
Прогон 1: Текст ссылок – текст ссылок		
F1 (macro average)	**0.234**	**0.180**
Precision (macro average)	**0.272**	**0.153**
Recall (macro average)	**0.206**	**0.219**

F1	**0.173**	**0.149**
Precision	**0.319**	**0.201**
Recall	**0.174**	**0.210**
Прогон 2: Окрестность страницы – текст ссылок		
F1 (macro average)	**0.250**	**0.185**
Precision (macro average)	**0.239**	**0.136**
Recall (macro average)	**0.263**	**0.285**
F1	**0.176**	**0.130**
Precision	**0.219**	**0.119**
Recall	**0.218**	**0.238**
Прогон 3: Окрестность страницы – окрестность страницы		
F1 (macro average)	**0.190**	**0.124**
Precision (macro average)	**0.183**	**0.092**
Recall (macro average)	**0.197**	**0.189**
F1	**0.146**	**0.121**
Precision	**0.249**	**0.132**
Recall	**0.177**	**0.229**

В прогоне 1 как для обучения, так и для классификации использовались профили на основе текста ссылок. В прогонах 2 и 3 для обучения использовались профили сайта, построенные как окрестность стартовой страницы глубиной 1, а для классификации в прогоне 2 использовались профили на основе текста ссылок.

Наилучшие результаты были показаны в первом и втором прогоне, причем наблюдается незначительное преимущество прогона 2. Более подробный анализ данных показал, что это превосходство относительно нестабильно (см. таблицу). Так, например, по метрике F1 прогон 2 показывает лучший или равный результат лишь в 24 категориях из 38, оцененных асессорами. Из таблицы так же видно, что прогон 2 показал несколько лучшие значения полноты (28 категорий из 38), но уступил в точности прогону 1 (33 категории из 38). Это может говорить о том, построенные профили содержали достаточно много общей информации, что уменьшило точность классификатора.

Сравнение результатов прогонов 1 и 2			
	прогон 2 лучше	равны	прогон 2 хуже

	прогона 1	(+/- 3%)	прогона 1
Точность	5 кат.	13 кат.	20 кат.
Полнота	16 кат.	12 кат.	10 кат.
F1	9 кат.	15 кат.	14 кат.

Наихудший результат был получен при использовании текста стартовой страницы и ее окрестности как для обучения, так и для классификации (значение метрики F1 хуже или равно в 29 категориях из 37 по сравнению с прогоном 2). Сравнение результатов прогонов 2 и 3 позволяет сделать вывод, что использование текста документов из окрестности главной страницы для построения профилей мало эффективно, если коллекция содержит большой процент шумовых данных.

Итоговый усредненный результат во многом обусловлен тем, что для многих категорий (в среднем 9 категорий на прогон) система не смогла найти ни одного относящегося к ней сайта. Причины такого поведения пока не очевидны и требуют дополнительного исследования.

Полученные результаты уступают результатам, полученным другими участниками РОМИП'2004, что объясняется так же следующими причинами:
- при классификации акцент делался только на один аспект задачи, при этом не учитывались многие другие проблемы категоризации веб сайтов. Так, например, не проводилось никакого морфологического анализа текстов, что может иметь особенно большое значение при работе с профилями малого размера.
- использованные алгоритмы не являются наилучшими, а средства классификации содержат ошибки.

В тоже время, на ряде категорий были получены результаты, превосходящие результаты других участников. По количеству таких категорий лучше всего себя показал прогон 2 (6 категорий из 38 в оценке weak). Такая ситуация косвенно подтверждает превосходство стратегии, выбранной для прогона 2, по сравнению с прогонами 1 и 3 (каждый из них показал лучший результат на 1 категории).

Стратегия, использующая для построения профиля текст ссылок и их контекста, показала наилучшие результаты как на стадии предварительных экспериментов, так и в экспериментах

РОМИП’2004. Исходя из этого можно сделать вывод, что именно эта стратегия построения профилей является наилучшей из рассматриваемых и позволяет наиболее точно извлечь релевантную тематике сайта информацию.

5.4. Дальнейшая работа

Анализ результатов позволяет выделить основные направления дальнейшего развития использованных технологий классификации.

Необходимо совершенствовать алгоритмы выделения релевантной информации из гипертекстовых документов. Интересными, например, представляются различные способы ранжирования по важности фрагментов гипертекстовых документов ([6],[8]). Применение этих способов должно улучшить результаты за счет фильтрации рекламной информации, навигационных ссылок и элементов дизайна.

Для профилей, использующих текстовое окружение ссылки, необходимо ввести зависимость веса термов от расстояния до ссылки ([9]). Существенным недостатком применявшейся стратегии является и то, что профиль, использующий только информацию из текста ссылок и их окрестности, практически не несет информации о стартовой странице сайта. Учитывая, что решение о принадлежности сайта той или иной тематике часто принимается по стартовой странице, необходимо расширить профили за счет добавления данных, характеризующих ее содержимое.

Хотя мы и не надеялись достичь (да в общем и не ставили своей целью) уровня качества классификации, который демонстрируют специализированные системы, но все же мы не ожидали, что отставание будет столь значительным. В ближайшем будущем мы планируем добавить в систему хотя бы упрощенную поддержку русской морфологии, чтобы отставание “Золушки” было не столь заметно.

6. Заключение

В статье описаны результаты экспериментов по анализу влияния объема информации, используемой для построения описания Веб-сайта, на качество классификации. Эксперименты проводились на основе дорожек по классификации Веб-сайтов от РОМИП’2003 и РОМИП”2004.

На стадии предварительной подготовки было отобрано несколько методов построения профилей. Результаты наших экспериментов показали, что наилучшим из рассматривавшихся является способ, использующий для построения профиля текст ссылок и их окрестностей. Этот метод позволяет выделить информацию, являющуюся наиболее релевантной основной тематике сайта.

7. Литература

[1] Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска. *Программирование, 28(4):* 226-242, 2002.

[2] Сайт классификатора rainbow, http://www-2.cs.cmu.edu/~mccallum/bow/

[3] Сайт Российского семинара по Оценке Методов Информационного Поиска (РОМИП) http://romip.narod.ru/

[4] Труды РОМИП'2003, http://romip.narod.ru/romip2003/index.html

[5] John M. Pierre Practical Issues for Automated Categorization of Web Sites, *ECDL 2000 Workshop on the Semantic Web, 2000*

[6] Lan Yi, Bing Liu, Xiaoli Li Eliminating Noisy Information in Web Pages for Data Mining, *SIKGDD '03, 2003*

[7] Lawrence Kai Shih, David R. Karger Using URLs and Table Layout for Web Classification Tasks, *WWW2004, 2004*

[8] Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma Learning Block Importance Models for Web Pages, *WWW2004, 2004*

[9] Soumen Chakrabarti, Kunal Punera, Mallela Subramanyam Accelerated Focused Crawling through Online Relevance Feedback, *WWW2002, 2002*

[10] Steve Lawrence, C. Lee Giles Accessibility of information on the web, *Nature, 400:107-109,* 1999

Influence of representation of the Website on quality of classification

Kondratyev Mikhail

Saint-Petersburg State University
Mikhail.Kondratyev@sun.com

Abstract

In article different approaches for building Website profiles for classification and influence of volume of the information on the classification results are analysed. Experiments were carried out on the basis of Website classification tasks from RIRES'2003 and RIRES'2004.