Оценка влияния извлечения значимой информации на качество классификации web-страниц

© Р.Ф. Кузнецов

Балтийский Государственный
Технический Университет
ruslkuznetsov@gmail.com

Н.В. Мурашов

Санкт-Петербургский Государственный
Университет Информационных Технологий
Механники и Оптики
nikitus_apos@inbox.ru

Аннотация

Целью предстоящей работы является экспериментальное исследование влияния различных подходов извлечения значимой информации на качество классификации web-страниц и проверка гипотезы о том, что выделение этой информации дает положительный результат при построении автоматического рубрикатора интернет-каталога. Кроме того, в работе будет проведено дальнейшее исследование метода извлечения значимой информации предложенного автором в работе [1].

1. Введение

Особенностью представления документов в сети Интернет является наличие на странице, помимо самого текста (содержательной части web-документа) определяющего предмет страницы, большого количества вспомогательных элементов (навигационной части web-документа) призванных обеспечить навигацию по страницам сайта. В отличии от страниц классических печатных книг, где навигационная часть обычно состоит только из номера страницы[1], при том, что остальной текст принадлежит содержательной части, навигационная часть web-страницы состоит из множества элементов, таких как навигационные ссылки, версия для печати, «дорожные знаки»[2], блоки текста с рекламой других разделов сайта, контактные данные компании и т.п. (так называемая служебная информация). Часто эти элементы не имеют прямого отношения к теме страницы и поэтому могут отрицательно влиять на качество информационного поиска [7, 5, 6].

Основываясь на этих предположениях можно сделать вывод, что удаление или уменьшение веса навигационной части может оказать положительное влияние на решение задач информационного поиска, таких как web-поиск, классификация, извлечение текстовой информации и т.п.

В данной работе будет проведено исследование различных методов выделения значимой информации, оценка их влияния на качество классификации, а также дальнейшее изучение метода извлечения значимой информации, основанного на выделении предложений. Особенностью последнего метода является возможность выделения значимой части без использования информации с других страниц сайта. Это может быть полезно в тех случаях, когда информация о других страницах не доступна.

2. Обзор существующих методов

Наметившийся в последнее время значительный интерес исследователей к проблеме выделения значимой части web-страниц, привел к появлению множества работ на эту тему.

Методы, применяемые для выделения шума на веб-страницах можно разделить на два основных типа:

1. методы, основанные на выделении повторяющихся для всех (или части) страниц сайта фрагментов информации [2, 3, 4].
2. методы, основанные на анализе dom-деревьев страниц сайта [7, 5].

Первый метод основан на том предположении, что навигационную информацию от значимой отличает то, что она повторяется на других страницах сайта. То есть, если некий фрагмент информации присутствует на

[1] Помимо номера, на странице также часто размещают имя автора, название книги, название главы и т.п.

[2] «Дорожные знаки» - ссылки, показывающие путь от главной страницы сайта к текущей.

всех (или многих) страницах сайта, с определенной долей уверенности, можно утверждать, что данная информация является навигационной. Это предположение возникло из наблюдений за особенностями размещения данного типа информации на сайтах. Например, навигационные ссылки, контактные данные компании или ссылка «версия для печати» часто присутствуют на всех страницах сайта в неизменном виде.
В работе [2] автор предлагает разбивать страницы на неделимые последовательности символов – токены, после чего искать файлы с одинаковым набором цепочек токенов и удалять найденные последовательности из файлов.
Второй метод исходит из особенностей построения dom-деревьев. Каждой странице в формате HTML соответствует dom-дерево, в котором узлам соответствуют тэги, а листьям - текст или картинки, заключенные в эти тэги. Методы, основанные на анализе dom-деревьев, опираются на некоторые эвристические предположения, о том, какие листья принадлежат содержательной части, а какие используются для служебных целей. Например, обычно, основной текст размещен в центре страницы, а окружающие его текстовые блоки выполняют служебные функции.
Существуют также методы, совмещающие оба этих подхода. Например, в работе [6], автор предлагает на основе dom-деревьев создавать *сжатое дерево структуры (ompressed structure tree, CST)*. Это дерево создается для всего сайта на основе dom-деревьев его страниц. Каждый узел нового дерева образуется при объединении одинаковых узлов dom-деревьев страниц сайта в один узел CST. Объединение происходит последовательно «сверху» для каждого уровня. Объединению подвергаются узлы, образованные одинаковыми тэгами с одинаковыми атрибутами. Далее, для узлов, у которых нет вложенных тэгов (т.е. отсутствуют узлы-«потомки»), происходит объединение листьев с одинаковым содержанием.
Описанные здесь методы, опираются в основном на выделение повторяющихся частей страниц с одного сайта (то есть, помимо рассматриваемой страницы им требуется информация о других страницах сайта). Методы же использующие данные только одной страницы (например, основанные на анализе dom-дерева) опираются на эвристики, описывающие наши представления о «стандартной» структуре HTML-документов.
Методы, основанные на выделении повторяющихся фрагментов страниц одного сайта, показывают более высокие результаты и являются более универсальными, чем методы, основанные на анализе dom-деревьев. Однако для их работы необходима информация обо всех страницах сайта (или хотя бы части из них). Это не всегда возможно и при реализации алгоритма увеличивает время его работы, в связи с необходимостью анализа большого количества страниц. В связи с этим, автор предложил новый алгоритм извлечения значимой информации, описанный в работе [1]. В отличии от используемых выше подходов, данный алгоритм позволяет выделять значимую часть web-страницы без построения dom-дерева и использовать для анализа текста лишь рассматриваемую страницу.

3. Основные методы исследования

В данной работе планируется сравнение методов выделения значимой информации и оценка их влияние на качество классификации. Для сравнения были выбраны следующие три метода:

- метод, основанный на выделении предложений [1]
- метод, использующий токены [2]
- метод, использующий сжатое дерево структуры [6]

Программа исследования разделена на следующие этапы:

- Анализ и улучшение метода, основанного на выделении предложений.
- Разбиение набора данных "Классификация сайтов", предоставленного компанией Яндекс, на обучающую и тестовую выборку.
- Обучение классификатора на обучающей выборке с использованием одного из трех, описанных выше, методов выделения значимой части и применением различных настроек этих методов.
- Обучение классификатора на обучающей выборке без использования метода выделения значимой части.
- Классификация тестовой выборки с использованием получившихся моделей.
- Оценка качества классификации путем анализа полноты, точности и F-меры.
- Выработка рекомендаций по использованию методов выделения значимой части web-страниц при построении автоматического рубрикатора интернет-каталога.

Перед обучением классификатора слова на странице будут приводиться к нормальной форме с помощью стеммера Snowball [8]. Для сокращения размерности пространства признаков планируется использовать метод χ^2 [9]. В качестве метода машинного обучения будет применен PrTFIDF [10].

4. Данные, которые планируется использовать

В качестве исходных данных для обучающей и тестовой выборки планируется использовать набор данных "Классификация сайтов", предоставленных компанией Яндекс.

5. Ожидаемые результаты и методы оценки

В результате этого исследования мы надеемся оценить влияния различных подходов извлечения значимой информации из web-страниц на качество классификации и проверить гипотезу о том, что выделение этой информации дает положительный результат при построении автоматического рубрикатора интернет-каталога. Кроме того, для метода извлечения значимой информации, основанного на выделении предложений, планируется провести исследование эффективности новых эвристик, которые, предположительно, должны повысить качество работы алгоритма.

Литература

[1] Р.Ф. Кузнецов. Извлечение значимой информации из web-страниц с использованием предложений. Сборник тезисов постерных докладов восьмой всероссийской конференции RCDL'2006

[2] М.С. Агеев, И.В. Вершинников, Б.В. Добров. Извлечение значимой информации из web-страниц для задач информационного поиска. Интернет-математика 2005. Сборник работ по программам научных стипендий Яндекса. Москва, 2005.

[3] И. Некрестьянов, Е. Павлова. Обнаружение структурного подобия HTML-документов. Труды четвертой всероссийской конференции RCDL'2002, 38-54, Дубна, Россия, 2002.

[4] S. Gupta, G.E. Kaiser, P. Grimm, M. F. Chiang, J. Starren, Automating Content Extraction of HTML Documents. World Wide Web Journal, January 2004.

[5] C. H. Lee, M.Y. Kan, S. Lai. Stylistic and Lexical Co-training for Web Block Classification. In Proceedings of Workshop on Web Information and Data Management (WIDM '04), Washington, D.C., USA.

[6] Yi, L., Liu, B., Web Page Cleaning for Web Mining through Feature Weighting, in the proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, August, 2003.

[7] L. K. Shih and D. Karger. Using URLs and table layout for web classification tasks. In Proceedings of the 13th International Conference on the World Wide Web, pages 193--202, New York, NY, 2004.

[8] Snowball http://snowball.tartarus.org

[9] Yang Y., Pedersen J. A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference On machine Learning — Nashville, USA, 1997.

[10] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Proceedings of International Conference on Machine Learning (ICML), 1997