Тематическая классификация текстов

© С.В. Панков, С.П. Шебанин, А.А. Рыбаков

ROOKEE
sergey.pankov@re-actor.ru, sergey.shebanin@re-actor.ru,
alexander.ribakov@re-actor.ru

Аннотация

В статье описан алгоритм тематической классификации веб-страниц и веб-сайтов. Представлен результат эксперимента в рамках дорожек по тематической классификации семинара РОМИП 2010.

1. Введение

Тематическая классификация сайтов – задача, становящаяся все более актуальной в связи с постоянно растущим объемом информации в интернете и потребностью в ней ориентироваться. Как отдельную подзадачу ее часто приходится решать в различных задачах информационного поиска. В этом документе описан процесс решения двух таких подзадач: поиск документов по заданной тематике и определение тематики для заданного сайта.

Для классификации текстов в данном исследовании использовалась модификация линейного метода. Выбор был обусловлен простотой реализации и высокой скоростью работы алгоритма.

2. Постановка задачи классификации

Пусть $D = \{d_1 \dots d_{|D|}\}$ – множество документов, $C = \{c_1 \dots c_{|C|}\}$ – множество тематик, $\Phi: D \times C \to \{0,1\}$ – неизвестная целевая функция, которая для пары $\langle d_i, c_j \rangle$ говорит, принадлежит ли документ d_i тематике c_j. Задача состоит в построении функции Φ', максимально близкой к Φ.

При машинном обучении классификатора предполагается еще наличие коллекции $\Omega = \{d_1 \dots d_{|\Omega|}\}$ заранее классифицированных документов, т.е. с известным значением целевой функции.

В нашем случае постановка задачи несколько отличается от классической, т.к. коллекция классифицированных документов зашумлена. Поэтому постановка задачи будет расширена задачей разделения множества документов $\Omega = \{d_1 \dots d_{|\Omega|}\}$, о которых известно, что большая часть из них имеет одинаковую тематику, на два множества Ω' и Ω", относящихся к этой тематике и не относящихся соответственно.

3. Индексация текстовой коллекции

Документы из обеих коллекций: обучающей и тестовой были предварительно проиндексированы и приведены к единому виду. Из текста каждой страницы были убраны блоки с рекламой и служебной информацией (меню, реклама, облака тегов и т.п.). Слова в тексте были нормализованы при помощи морфологического анализатора АОТ[1], стоп-слова и знаки препинания отбрасывались. Т.к. в дальнейшем информация о взаимном расположении слов не использовалась, для каждого документа далее хранился только вектор слов и информация о частоте их встречаемости в тексте.

В индекс также не брались документы, состоящие менее чем из 100 слов, т.к. точность определения тематики сильно зависит от размера документа.

Дополнительно по объединенной коллекции документов для каждого слова было вычислено его значение IDF.

4. Обучение классификатора

4.1. Выделение тематического ядра сайта

Особенностью обучения классификатора для РОМИПа является то, что в обучающей выборке дана информация только о тематике сайтов в целом, а не страниц в отдельности. Это поставило перед нами дополнительную задачу выбора кластера страниц, которые хорошо описывают тематику всего сайта.

Отсеивание нетематичных страниц базировалось на предположении, что большая часть страниц сайта относится к тематике всего сайта. Таким образом, задача выделения тематического кластера сайта сводилась к построению бинарного

классификатора, делящего все страницы сайта на две категории: относящиеся к тематике всего сайта и не относящиеся.

Мерой тематической близости между документом и сайтом считалось значение косинуса угла межу N-мерными векторами документа и сайта, где N – число слов в словаре коллекции.

Для каждого документа строился вектор весов слов, входящих в него. Вес каждого слова w_i в документе j рассчитывался по формуле:

$$P_{ij} = count_{ij} \cdot IDF_{w_i},$$

где $count_{ij}$ – число вхождений слова в документ, IDF_{w_i} – обратная частота слова в коллекции. После расчета веса каждого слова в документе, вектор нормируется:

$$P_{ij}^{*} = \frac{P_{ij}}{\sqrt{\sum_{k=1}^{m} P_{kj}{}^{2}}}$$

Аналогичным образом строится вектор и для всего сайта, при этом текст сайта получается объединением текстов всех входящих в него документов.

Решение о включении документа в тематическое ядро сайта принималось на основе пороговой оценки тематической близости. Были опробованы три различных варианта:

1. Выбор фиксированного процента наиболее тематичных документов.
2. Выбор всех документов со значением близости выше среднего по сайту.
3. Выбор всех документов со значением близости выше медианы значений близости по сайту.

В итоге по результатам тестирования на предыдущих результатах РОМИП обучающих наборов документов, построенных каждым из трех способов, был выбран последний, как показавший лучшие результаты.

4.2 Обучение

С целью повышения скорости обучение производилось методом регрессии, т.е. сразу по всей коллекции.

По каждой заданной тематике строился набор документов на основе тематического ядра сайтов из обучающего набора. Все документы из одной тематики объединялись в один большой документ, по которому строился вектор слов с весами:

$$O_{ij} = count_{ij} \cdot IDF_{w_i},$$

где $count_{ij}$ – число вхождений слова в тематику. Затем веса слов внутри тематики нормируются:

$$O_{ij}^{*} = \frac{O_{ij}}{\sqrt{\sum_{k=1}^{m} {O_{kj}}^{2}}}$$

5. Классификация

5.1 Классификация веб-страниц

В процессе классификации для каждого документа из тестовой выборки рассчитывалась его близость к каждой тематике, как косинус угла между вектором документа и вектором тематики.

При составлении результатов, в список документов, относящихся к заданной тематике, попали все документы, для которых заданная тематика являлась наиболее близкой по сравнению с остальными, но не менее чем 0,2.

5.2 Классификация веб-сайтов

При классификации веб-сайтов использовалась информация о тематике документов, принадлежащих сайту. Для каждого документа отбрасывались все тематики с близостью менее 0,2, затем тематики всех документов объединялись в единый список для всего сайта, на основе этого списка строился вектор частот встреч каждой тематики на сайте. После нормализации отбрасывались все компоненты вектора со значением менее 0,3. По требованию задания РОМИП вектора дополнительно сокращались до 5 ненулевых значений.

6. Результаты и заключение

На рисунках приведены результаты участников дорожек РОМИП 2010 классификации веб-станиц и веб-сайтов по метрике AND.

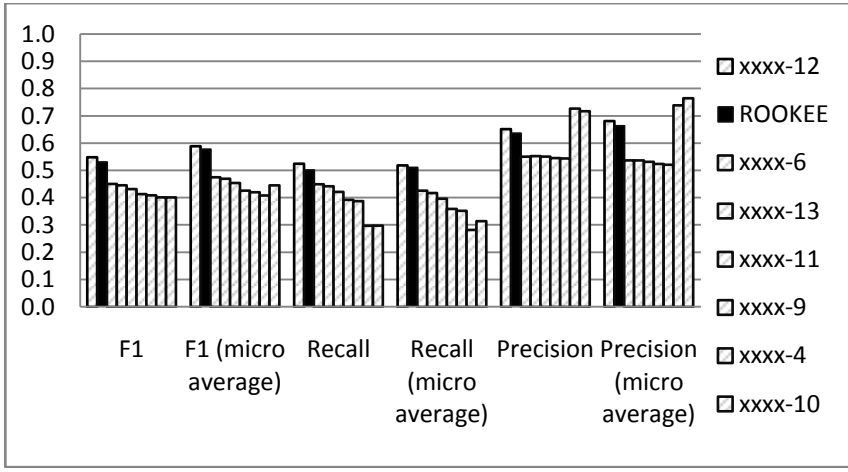

Рис. 1 Сравнительные результаты классификации страниц, оценка AND

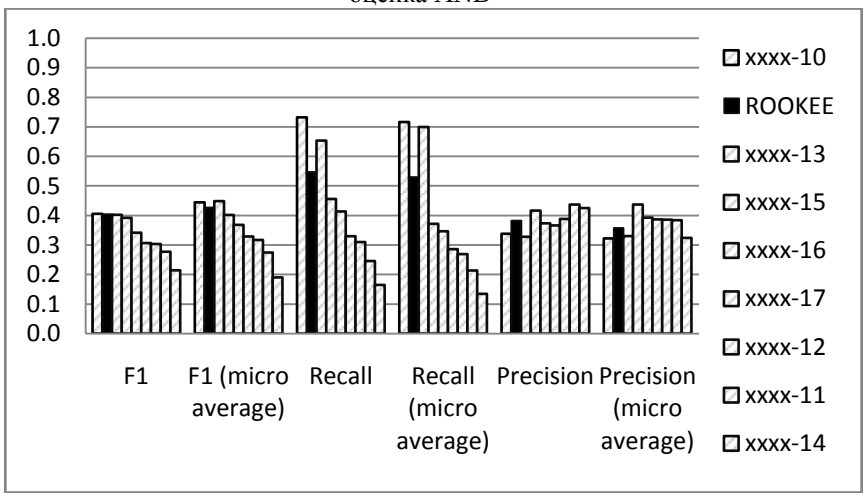

Рис. 2 Сравнительные результаты классификации сайтов, оценка AND

По этой оценке, релевантными считались только те результаты, которые были признаны релевантными всеми асессорами. В обеих дорожках метод показал неплохие результаты, что позволяет сделать вывод о том, что относительно простые методы при соответствующей подготовке обучающей выборки способны показывать хорошие результаты.

7. Выводы

Участие в РОМИПе показало, что потенциал линейного метода классификации еще до конца не исчерпан, и с помощью него можно добиваться неплохого качества классификации. Однако, ему есть еще куда развиваться, в частности, практически не использовалась информация о структуре страниц и взаимном расположении слов. Планируется также использовать наработки в области семантического анализа документа, например, использовать в качестве координат векторов целых многословных понятий вместо отдельных термов.

Также планируется продолжать улучшать метод выделения тематического ядра, т.к. замечены существенные недостатки текущего метода на сайтах с политематичным контетом.

Литература

[1] AOT Web site, 2010. http://aot.ru/

[2] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1? pages 1–47, 2002.

[3] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, 2008. Introduction to Information Retrieval.