Поиск неестественных текстов

© Е.А. Гречников, Г.Г. Гусев, А.А. Кустарев, А.М. Райгородский

Яндекс, Лаборатория комбинаторных и вероятностных методов,
{grechnik, gleb57, kustarev, raigorodsky}@yandex-team.ru

Аннотация

В работе описывается метод определения неестественного происхождения документа, основанный на изучении статистики встречаемости пар соседних слов в тексте. Тестирование показывает, что метод может быть использован как отдельно, так и для существенного улучшения результатов уже известных методов определения спама по контенту.

1 Введение

1.1 Постановка задачи

Требуется построить алгоритм, определяющий, написан ли данный документ человеком или же является автоматически сгенерированным либо модифицированным. Под модификацией документа понимается следующее:

- текст является результатом работы синонимайзера – программы, заменяющей отдельные слова на синонимы, или иной системы уникализации контента;
- текст является результатом работы автоматического переводчика с иностранного языка на русский.

Эта задача актуальна, в первую очередь, для проблемы нахождения поискового спама. Многие сайты используют системы уникализации контента и бессмысленную накачку документов ключевыми словами для повышения собственных позиций в поисковой выдаче.

Отличить автоматически сгенерированные или модифицированные тексты от написанных человеком на глаз обычно несложно. Трудность заключается в поиске автоматического алгоритма решения задачи.

Говоря менее формально, нужно на машинном уровне научиться определять степень «бредовости» текста. Идея решения состоит в исследовании корреляций соседних слов в исходном документе.

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

1.2 Работы по схожей тематике

Методы определения поискового спама можно довольно грубо разделить на две части: анализ контента самой страницы и анализ входящих и исходящих ссылок, топологии сети в окрестности документа. Предлагаемый нами метод, очевидно, относится к первой группе.

В ряде статей неоднократно рассматривались методы обнаружения спама, основанные на анализе содержания страницы. В работе [5] был построен классификатор спама, использовавший несколько таких признаков, как сжимаемость документа, средняя вероятность триграмм, доля частых слов в документе и другие – которые затем были собраны в дерево решений при помощи алгоритма машинного обучения C4.5[6].

Довольно близко к нашей работе по теме и методу находится работа [1], в которой анализируются пары слов, находящихся на данном расстоянии друг от друга, после чего по всем таким парам строится общая гистограмма частоты.

В работе [2] анализировались частоты встречаемости в тексте слов, имевших большую коммерческую и рекламную привлекательность.

Наконец, в работе [4] метод определения спама основан на сравнении моделей языка в исходном документе и в документе, его цитирующем – то есть находится на пересечении двух рассматриваемых нами областей.

2 Описание метода

Мы будем работать с множеством 2000 наиболее распространенных слов русского языка. Рассмотрим матрицу (A_{ij}) размера 2000 на 2000, в которой на пересечении i-й строки и j-го столбца стоит частота встречаемости в языке пары слов с номерами i и j. Частота встречаемости пары вычисляется по фиксированной базе текстов, в данном случае использовалась база *ruscorpora,* объем которой – 41298 документов [8].

Пусть A_i и A_j – суммы по строкам и столбцам матрицы (A_{ij}) соответственно. Определим функцию $Cor(i, j)$ по формуле $Cor(i, j) = (A_{ij}/A_i) + (A_{ij}/A_j)$. Функция Cor(i, j) измеряет степень «сочетаемости» слов с порядковыми номерами i и j.

Была выдвинута следующая гипотеза: в неестественном тексте должно быть нарушено распределение пар в тексте по функции Cor. Более

точно, количество редких, нехарактерных для языка пар должно быть завышено по сравнению со стандартом, а количество частых пар – занижено.

Эта гипотеза подтверждается следующей таблицей. В ней четыре столбца с числами соответствуют четырем текстам – один является оригинальной газетной статьей, а три других – его синонимизированными модификациями. Были использованы три различных программы-синонимайзера, найденных в Интернете.

Cor≥0.1	115	92	87	76
0.1>Cor≥0.01	502	350	317	309
0.01>Cor≥0.001	341	291	219	290
0.001>Cor≥0.0001	98	148	73	159
0.0001>Cor≥0.00001	12	18	19	39
0.00001>Cor	2	3	2	6

Основываясь на данных таблицы, объявим теперь пары с $Cor<10^{-4}$ редкими, а пары с $Cor>10^{-2}$ – частыми и заметим, что в данном примере искомая гипотеза тогда подтверждается. Стоит также отметить, что во всех других рассмотренных нами случаях (всего было аналогичным образом проанализировано 45 текстов) число редких пар в синонимизированном дубликате оказалось завышенным по сравнению с исходным текстом.

3 Применение и результаты

Мы приводим два метода использования полученных данных. В первом число редких пар в тестируемом тексте сравнивается с данными, полученными из заведомо хороших текстов, после чего делается вывод о качестве тестируемого документа. Второй метод использует машинное обучение при помощи алгоритма TreeNet [3,7], в котором в качестве факторов используется число пар в тексте, Cor которых лежит в том или ином диапазоне. Мы также сравниваем эффективность наших факторов с классическими факторами, использованными в работе [5] для определения спама по контенту страницы.

3.1 Сравнение с нормальными текстами

Рассмотрим базу заведомо качественных и возможно более разнообразных текстов русского языка (в данном случае использовалась база *ruscorpora* [7]). Если на вход дан тестируемый текст T, то найдем 10 ближайших к нему по длине текстов из базы качественных документов. Затем определим три числа: N(T) – число редких пар в тексте T, M(T) – среднее арифметическое числа редких пар в 10 ближайших по длине к T текстах, D(T) – квадратный корень дисперсии набора чисел редких пар из 10 ближайших к T текстов. В этом случае число (N(T) – M(T))/D(T) измеряет «степень неестественности» текста T.

Тестирование проводилось на выборке из 165 текстов, полученных вручную с помощью трех синонимайзеров из Интернета, а также 41298 оригинальных текстов из базы *ruscorpora (*для контроля ошибки метода). При тестировании текста из базы *ruscorpora* обязательно выбирались 10 текстов, ближайших к нему по длине, но не совпадающих с ним.

Результаты применения этого метода оказались следующие: условие (N(T) – M(T))/D(T)>3 позволяло корректно идентифицировать 41.5% некачественных текстов. При этом ошибка на базе текстов *ruscorpora* составила немногим более 2.3% (978 текстов из 41298 были признаны неестественными).

3.2 Машинное обучение

Гораздо лучшие результаты были получены при применении машинного обучения, позволившего подобрать нужную формулу для проверки качества текста автоматически.

Тестирование проводилось на обучающей и тестовой выборках, не содержащих совпадающих текстов. Обучающая выборка состояла из 2000 заведомо оригинальных документов, а также из 250 некачественных документов, из которых 25 автоматически сгенерированных и 25 синонимизированных были найдены в Интернете, а 200 были изготовлены вручную при помощи трех различных программ-синонимайзеров. Тестовая выборка состояла из 500 заведомо оригинальных и 245 некачественных текстов, из которых 25 автоматически сгенерированных и 25 синонимизированных были найдены в Интернете, а 145 были изготовлены вручную при помощи трех программ-синонимайзеров.

В качестве алгоритма машинного обучения был выбран TreeNet ([3],[7]) с потенциалом «сумма логарифмов вероятностей ошибки классификации». Число итераций алгоритма выбиралось с тем, чтобы минимизировать ошибку на тестовом множестве. В качестве других параметров были выбраны: шаг регуляризации – 0.01, доля обучающей выборки, по которой шло обучение на каждом шаге – 0.5. Результатом работы TreeNet являлась не бинарная классификация, а число, определяющее вероятность нахождения элемента в данном из двух классов.

В качестве факторов использовалось число пар в тексте, Cor которых лежал в данном диапазоне: от 0 до 10^{-7}, от 10^{-7} до 10^{-6} и так далее до диапазона Cor>1. По окончанию работы TreeNet фиксировалось два порога, соответствующих ошибке в 1% и 5% на тестовом подмножестве из 500 оригинальных документов. Затем для этих порогов рассматривалась полнота на соответствующем подмножестве неестественных документов.

Результаты оказались следующие:

- при ошибке в 1% было правильно опознано 77.95% неестественных текстов,

- при ошибке в 5% было правильно опознано 90.61% неестественных текстов.

Для сравнения эффективности метода с методами, уже описанными в литературе, были реализованы 10 контентных факторов, описанных в статье [5]. Для машинного обучения снова использовался алгоритм TreeNet. Результаты классического метода таковы:

- при ошибке в 1% было правильно опознано 90.61% неестественных текстов,
- при ошибке в 5% было правильно опознано 96.73% неестественных текстов.

После этого был проведен тест на «совместную эффективность»: к классическим 10 факторам, показавшим хороший результат, были добавлены еще два новых фактора – число «редких» и число «частых» пар в определениях п.2. Результаты оказались следующие:

- при ошибке в 1% было правильно опознано 93.06% неестественных текстов,
- при ошибке в 5% было правильно опознано 97.95% неестественных текстов.

4 Заключение

Подводя итог проведенным тестам, можно сделать вывод, что описанный нами метод может успешно использоваться для поиска неестественных текстов. Хотя он существенно проигрывает классическому методу в эффективности, добавление новых факторов позволяет улучшить результаты классического метода на несколько процентов (в нашем случае – на 2.45% и 1.22% соответственно). Иными словами, в проведенном нами тестировании около четверти еще не пойманного классическими методами спама подпадает под сферу действия новых факторов.

Литература

[1] J. Attenberg, T. Suel. Cleaning search results using term distance features. In *Proceedings of AIRWeb 2008*, pages 21-24, ACM.

[2] A. Benczur, I. Biro, K. Csalogany, and T. Sarlos. Web spam detection via commercial intent analysis. In *Proceedings of AIRWeb 2007*, pages 89-92, New York, NY, USA, 2007, ACM.

[3] J.H.Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29(5):1189-1232, 2000.

[4] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proc. of the 1st Int. Workshop on Adversarial Information Retrieval on the Web*, pages 1-6, 2005.

[5] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. 15th WWW*, pages 83-92, 2006.

[6] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan-Kauffman, 1993.

[7] G. Ridgeway. Generalized Boosted Models: A guide to the gbm package. *http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf*

[8] Национальный корпус русского языка – *www.ruscorpora.ru*.

Detection of Artificial Texts

Evgeny A. Grechnikov, Gleb G. Gusev,
Andrei A. Kustarev, Andrei M. Raigorodsky.

We present a method of artificial text search based on analysis of frequency of word pairs. This method can be used either for improving results of well-known content spam classifiers or independently.