Определение заимствований в тексте без указания источника*

Сафин К. Ф., Кузнецов М. П., Стрижов В. В.
kamil.95@yandex.ru, mikhail.kuznecov@phystech.edu
Московский физико-технический институт

Работа посвящена поиску внутренних заимствований в тексте. Предполагается, что большая часть текста написана одним автором. Необходимо выделить участки текста, написанные другим автором, если таковые имеются. При этом документ анализируется изолированно, то есть источники заимствований неизвестны. В работе предлагается алгоритм, в основе которого лежит использование стилевой функции, определяющей по данному участку текста вероятность его заимствования. Эксперимент проводится на коллекции конкурса PAN-2011.

Введение

Текстовый плагиат является большой проблемой в сфере образования и научных исследований. Развитие сети интернет сделало большой объем информации свободным для плагиата.

В задаче обнаружения плагиата существует два глобальных подхода: выявление «внешнего» и «внутреннего» плагиата. При поиске внешних заимствований есть некоторый корпус, из которого возможны заимствования. То есть задача состоит в попарном сравнении участков подозрительного текста и текстов из корпуса заимствований и нахождении сходств. Задача поиска «внутреннего» плагиата состоит в анализе исключительно подозрительного текста. Решающие алгоритмы должны анализировать стиль письма или же выделять характерные признаки, свойственные данному автору.

Решающие алгоритмы разбивают исходный текст на сегменты и сравнивают текст сегмента со всем текстом. Разбиение проводится по предложениям [1, 2], или же определяется окно заданной ширины, согласно которому производится сегментирование текста [3, 4, 5, 6]. Выбор меры схожести сегмента со всем текстом или, наоборот, меры различия является ядром алгоритма. Работы [1, 4, 2] используют стилистические, синтаксические, лексические характеристики: частотности частей речи, порядок следования частей речи в предложении, пунктуацию, среднюю длину предложения и подобные признаки. Возможно использование символьных n-грамм(чаще других применяют 3-граммы) в качестве признака, а точнее частот их использования [3, 6]. Метод [7] использует кластеризацию абзацев по частоте встречаемости существительных. В 2011 году был проведен конкурс PAN-2011, посвященный поиску плагиата в текстах. Метод Oberreuter [5], являющийся победителем в конкурсе PAN-2011, использует функцию, характеризующую письменный стиль автора. Функция строит вектор частот встречаемости слов во всем документе и в выделенном фрагменте. Эти вектора используются для определения величины отклонения фрагмента от всего текста. Данный алгоритм показал результат 0,32 по F1 мере. Это демонстрирует то, что алгоритма, решающего данную задачу в большинстве случаев, до сих пор нет.

Предлагаемый алгоритм при помощи стилевой функции строит статистическое описание текста, которое используется для классификации фрагментов текста. Стилевая функция должна удовлетворять условиям: на оригинальных сегментах текста иметь небольшой

Научный руководитель: Стрижов В. В. Задачу поставил: Стрижов В. В. Консультант: Кузнецов М. П.

разброс значений, по сравнению со значением на всем тексте, а на заимствованных сегментах иметь значительные отличия.

В работе используются данные конкурса PAN-2011 [9]. Для оценки работы алгоритма определяются микро- и макро- меры качества precision и recall (точность и полнота), и затем вычисляется F1 мера как среднее гармоническое для precision и recall. Данная мера представляет качество классификации сегментов текста.

Постановка задачи

Пусть D — коллекция текстовых документов, d — текстовый документ, s — сегмент текста ($d = \bigcup s_i, d \in D$). Задача состоит в том, чтобы каждому элементарному сегменту сопоставить один из двух классов: $s_i \mapsto \{0, 1\}$. Класс '0' если сегмент считается оригинальным, и класс '1' — если сегмент заимствован.

Описание выборки. В работе используется блок текстовых документов конкурса PAN-2011 [9]. В текстах присутствуют сегменты настоящего плагиата, имитированного плагиата и искуственного плагиата. Где соответственно сегмент текста полностью взят из другого источника, либо заимствованный текст переписан человеком другими словами, либо специально обученный алгоритм строит текст, стараясь повторить стиль автора.

Выборка состоит из 4753 текстов, к каждому из которых прилагается файл с экспертной разметкой заимствованных сегментов.

Критерии качества. В экспериментах используются критерии качества, применявшиеся в PAN-2011 [9]. Обозначим за пару (s, d) последовательность символов, помеченную экспертом как плагиат. $S = \bigcup s_i$ — совокупность всех заимствованных сегментов. За пару (r, d) обозначим последовательность, помеченную алгоритмом как заимствованную. Аналогично, $R = \bigcup r_i$ — совокупность всех сегментов, которые алгоритм классифицировал как плагиат. Рассмотрим меры качества *Precision* и *Recall*:

$$\mathrm{Prec}(S, R) = \frac{1}{|R|} \sum_{r_j \in R} \frac{|\bigcup\limits_{s_i \in S} (s_i \cap r_j)|}{|r_j|}, \quad \mathrm{Rec}(S, R) = \frac{1}{|S|} \sum_{s_i \in S} \frac{|\bigcup\limits_{r_j \in R} (s_i \cap r_j)|}{|s_i|},$$

Данные величины отражают точность (доля правильной классификации по отношению ко всей классификации) и полноту (доля правильной классификации по отношению ко всему плагиату в тексте) работы алгоритма.

Вычиляется *F1-мера* как среднее гармоническое между *Precision* и *Recall*:

$$F1(S, R) = \frac{\mathrm{Prec}(S, R) \cdot \mathrm{Rec}(S, R)}{\mathrm{Prec}(S, R) + \mathrm{Rec}(S, R)}.$$

Вычисляется величина гранулярности

$$\mathrm{gran}(S, R) = \frac{1}{|S_R|} \sum_{s_i \in S_R} |R_{s_i}|,$$

где $S_R = \{s | s \in S \wedge \exists r \in R : \text{r detects s}\}$, $R_S = \{r | r \in R \wedge \text{r detects s}\}$, r detects s: if $r \cap s \neq \varnothing$.

и итоговая мера качества *pladget*:

$$\mathrm{pladget}(S, R) = \frac{F1(S, R)}{\log_2(1 + \mathrm{gran}(S, R))}.$$

Постановка задачи. Для обнаружения плагиата, исходный текст разбивается на элементарные сегменты, каждый из которых будет впоследствии классифицирован как заимствованный или нет. Для каждого фрагмента строится вектор признаков, и проводится нормировка статистики. Затем происходит классификация сегментов текста на основании отклонения их признакового вектора от среднего значения - если отклонение превышает некоторый заданный порог, то сегмент считается заимствованным. Таким образом ставится задача поиска аномалий, т.е. сильных отклонений, в ряде статистики, где статистика строится для каждого сегмента текста.

При обучении на размеченных данных, параметры классификатора настраиваются таким образом, чтобы улучшить меры качества работы алгоритма. При фиксированном способе разбиения текста мера *Granularity* не изменяется, т.к. она зависит от мелкости разбиения. Тогда, для увеличения итоговой меры качества *Pladget*, достаточно улучшить *F1-меру*:

$$F1(S,\ R) = \frac{\mathrm{Prec}(S,R)\cdot\mathrm{Rec}(S,R)}{\mathrm{Prec}(S,R)+\mathrm{Rec}(S,R)} \longrightarrow max.$$

Базовый эксперимент

Целью базового эксперимента являлась проверка гипотезы о том, что заимствованные сегменты текста имеют отличные от среднего вектора значения признаков.

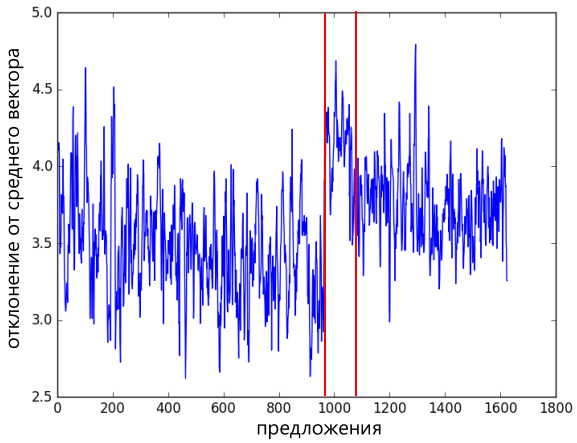
В качестве такого признака была выбрана частота встречаемости слов. Каждому слову ставится в соответствие число:

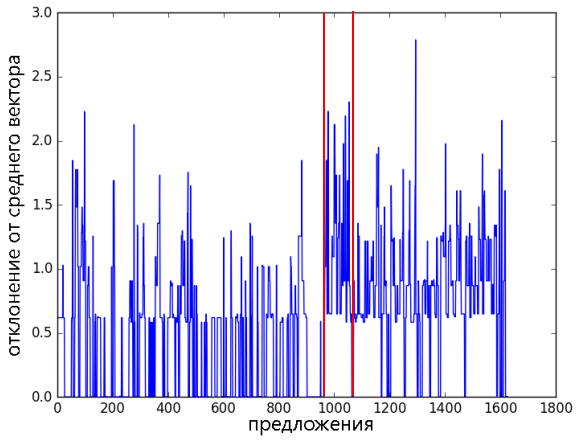
$$\mathrm{fr\_class}_w = \log_2 \frac{n_{\max}}{n_w}$$

где $n_{\max}$ — число вхождений наиболее часто употребимого слова в тексте, n_w — частота вхождений слова w в этом предложении.

В качестве основного признака использовались квантили распределения данной величины внутри окна фиксированной ширины.

Обозначим за $m_j^d = \overline{x_j^d}$ среднее значение j-го признака для документа d, за σ_j^d — среднеквадратичное отклонение. В качестве нормализованных признаков рассматриваются значения $\frac{x_j^d - m_j^d}{\sigma_j^d}$. За элементарные сегменты были выбраны предложения текста. Каждое предложение текста векторизовалось, и затем подсчитывалось отклонение от усредненного по всему тексту вектора. Эксперимент проводился на одном из текстов конкурсной коллекции PAN-2011.





На графике показано отклонение признакового вектора каждого предложения от усредненного вектора. Красными линиями отделены предложения, помеченные экспертом как заимствованные. Видно, что заимствованные фрагменты имеют характерные выбросы из области средних значений отклонения. Однако некоторые предложения, не являющиеся плагиатом, также сильно отличаются от усредненного признакового вектора. На основании этого можно сделать вывод, что использование только данного признака недостаточно для решения поставленной задачи.

Описание алгоритма

Модель Предлагаемый алгоритм работает с частотными признаками, предоставляющими стилевое описание текста. В качестве такого признака выбран признак частоты встречаемости слов, описанный выше.

Исходный текст подвергается предобработке: удаляются незначащие и служебные символы,все буквы переводятся в нижний регистр. Также удаляются стоп-слова, полученные с помощью python-модуля работы с естественными языками nltk.corpus.stopwords.

Сегментирование текста Текст разделяется на базовые сегменты c, которые затем будут классифицированы. С помощью модуля nltk.tokenize текст разбивается на предложения. Одним из настраиваемых параметров алгоритма является минимальная длина сегмента. Если длина очередного предложения меньше минимальной длины сегмента, к этому предложению добавляется следующее за ним - процесс повторяется, пока длина сегмента c не превысит заданную минимальную длину.

Построение статистики и детектирование аномалий Для всех сегментов текста строится ненормализованный количественный вектор v_c рассматриваемого частотного признака. Строится статистика v_c-series на основе отклонения вектора признаков от усредненного по всему тексту вектора (в L1-метрике).

Полученная статистика сглаживается методом скользящего среднего, ширина сглаживания которого также является настриваемым параметром. На основе полученной статистики проводится классификация сегментов. Если в ряде статистики присутствует аномалия, превышающая заданный порог δ, то сегмент, отвечающий этому выбросу классифицируется как заимствованный.

Минимальная длина сегмента, ширина окна сглаживания и порог выброса настраиваются на обучающей выборке путем максимизации *F1-меры*

Алгоритм 1 Intrinsic plagiarism detection

for $c \in C$ **do**
  build v_c using frequency feature
end
$style \leftarrow \frac{1}{|C|} \sum v_c$
scale-and-smooth(v_c-series)
for $c \in C$ **do**
  if $v_c - style < \delta$ **then**
    mark segment c as plagiarism
  end
end

Вычислительный эксперимент

Эксперимент был проведен на выборке текстов из коллекции PAN-2011.

Параметры алгоритма настраивались на 10 текстах, путем максимизации *F1-меры*. Красные участки обозначают заимствованные фрагменты, зеленым обозначен порог выброса значений стилевой функции.

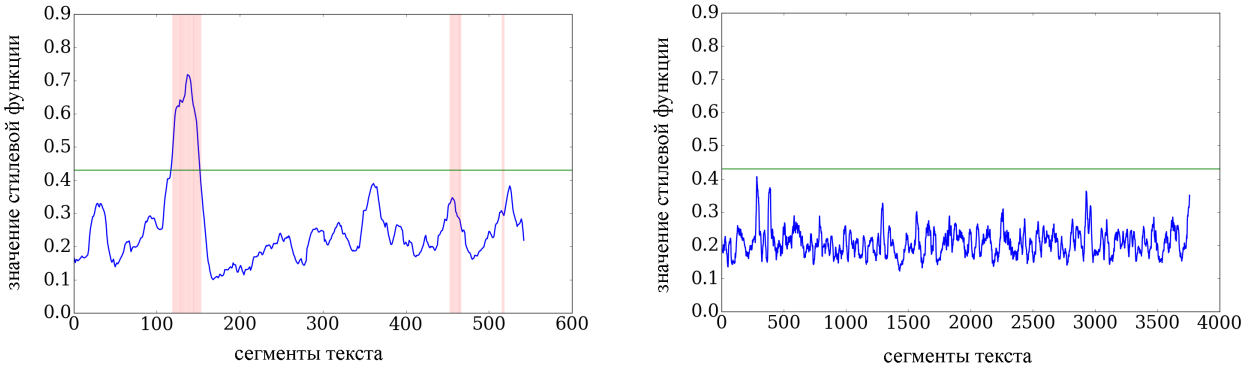


Рис. 1. Результаты на обучающей выборке

Полученные оптимальные параметры: минимальная длина сегмента (в символах) - 300, ширина сглаживающего окна - 20, порог выброса значений - 0,43.

С данными параметрами алгоритм тестировался на выборке из 20 текстов.

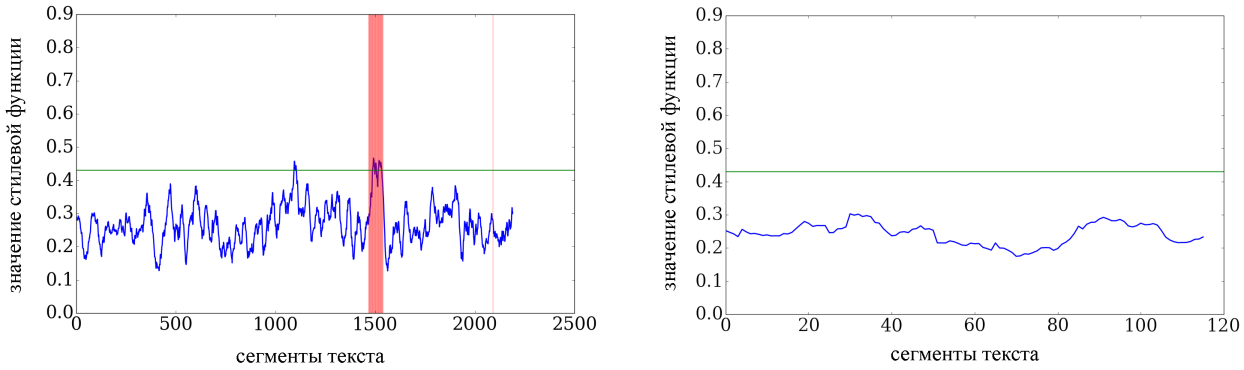


Рис. 2. Результаты на тестовой выборке

Результаты работы алгоритма на обучающей и тестовой выборках приведены в таблице.

	Rec	Prec	F1	Granularity	Pladget
learn	0,44	0,59	0,5	1,0	0,5
test	0,42	0,42	0,42	1,0	0,4

Работа алгоритма сильно зависит от длины документа. При анализе небольших по объему тесктов сглаживание приводит к потере информации об аномальных значениях стилевой функции.

Заключение

Был проведен вычислительный эксперимент по нахождению внутреннего плагиата. На небольшой коллекции документов предлагаемый алгоритм показал более высокие результаты, чем победитель конкурса PAN-2011.

Дальнейшие исследования могут быть направлены на более точную настройку параметров алгоритма и подбор параметров в зависимости от длины рассматриваемого текста.

Литература

[1] Mario Zechner, Markus Muhr, Roman Kern, and Michael Granitzer. External and intrinsic plagiarism detection using vector space models. 2009.

[2] Michael Tschuggnall and Gunther Specht. Countering plagiarism by exposing irregularities in authors grammars. 2013.

[3] Efstathios Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. 2009.

[4] Sven Meyer zu Eissen and Benno Stein. Intrinsic plagiarism detection. 2006.

[5] Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. Approaches for intrinsic and external plagiarism detection. 2011.

[6] Imene Bensalem, Paolo Rosso, and Salim Chikhi. Intrinsic plagiarism detection using n-gram classes. 2014.

[7] Anna Vartapetiance and Lee Gillam. Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. 2012.

[8] Douglas Bagnall. Author identification using multi-headed recurrent neural networks. 2015.

[9] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. An evaluation framework for plagiarism detection. 2011.