УДК 65.012

Шумейко А. А., Сотник С. Л.

ОБ ИСПОЛЬЗОВАНИИ РАСПРЕДЕЛЕНИЯ СЛУЖЕБНЫХ СЛОВ ПРИ ПРОВЕДЕНИИ ЭКСПЕРТИЗЫ ПИСЬМЕННОЙ РЕЧИ

Статья посвящена проблеме определения авторства анонимного документа на основе распределений авторских инвариантов на многомерной сфере. В основе проведенных исследований лежит гипотеза о постоянстве распределения служебных слов автора. Проведенная серия экспериментов подтвердила эффективность приведенного метода

Исследование подчерка является одной из важнейших задач криминалистики [1], однако повсеместное использование компьютерной техники привело к тому, что большую часть документов составляют напечатанные документы, к которым все методы, наработанные экспертами-почерковедами, не годятся. С другой стороны, широкое использование электронных средств связи, приводит к тому, что часто известны тексты (письменная речь) сообщений преступников или потенциальных жертв, и требуется их идентификация по имеющемуся банку материалов. В случае использования неискаженной речи весьма эффективными могут быть методы судебной психодиагностики, однако, использование технологий искажения спектральных характеристик речи существенно усложняет их использование.

Для решения такого рода задач используется атрибуция текста, то есть соотнесение тексту соответствующих атрибутов, к которым определяется автор. Текст, авторство которого подлежит определению, называют анонимным. С первого взгляда, наиболее естественным путём выявления авторских особенностей представляется фиксация внешних деталей авторского стиля, присущих тому или иному человеку и, в частности, любимых слов, терминов, а также фразеологических оборотов и выражений. Однако, выбор таких деталей неизбежно субъективен, и, кроме того, не гарантирует от ошибки в случае подражания, использующего именно внешние детали авторского языка. Часто при анализе анонимного текста все надежды связываются с некоей информацией самого текста, который может дать сведения о политических, идейных, эстетических, религиозных и т.п. взглядах автора текста или какие-то сведения о личной жизни автора. Иногда это действительно служит очень хорошим материалом для атрибуции текста. Однако такие детали могут быть сознательно добавлены анонимом.

Таким образом, единственный путь к решению проблемы атрибуции сводится к выявлению подсознательных особенностей речи конкретного автора. Эти особенности пытаются выявить путём применения формально-количественных методов.

Первые пробные шаги на этом пути были предприняты ещё в начале XX века ([2],[3]). Наиболее чётко необходимость поисков новых путей и отказа от "субъективно-атрибутивной" практики стала ощущаться в 50--60-е гг. прошлого века. Следует отметить работу А.Т. Фоменко [3], в которой приводится грамотная постановка задачи. В этой работе рассматривается характеристика, названная «авторским инвариантом», которая определяется следующим образом:

1.Для одного автора данная характеристика должна иметь минимальную вариацию, т.е. ее колебания могут быть признаны несущественными;

2.Для разных авторов отклонение характеристики от своего среднего должно быть существенным, чтобы было возможным статистически верно разделить двух разных авторов;

3.Она должна быть достаточно "массовой", интегральной, чтобы слабо контролироваться автором на сознательном уровне. Другими словами, она должна быть его "бессознательным параметром", коренящимся настолько глубоко, что автор даже не задумывается о нем. А если бы даже задумался, то не смог бы долго его контролировать и в результате довольно быстро вернулся бы в прежнее устойчивое и типичное для него состояние» [4].

В ходе серьезной и кропотливой работы было установлено, что удельный вес служебных слов (союзов, предлогов и частиц) для каждого автора является величиной с малой вариацией. В работе использовалась достаточно простая характеристика – процент содержания служебных слов в тексте. В нее вошло 55 служебных слов (14 союзов, 38 предлогов и 17 частиц). Характеристика варьируется от 16% до 30% в целом, и в пределах 1-2% для конкретного автора. По утверждению авторов, различие в 2-3% для двух текстов может быть серьезных основанием полагать, что их писали разные авторы.

ПРЕДЛОГИ:

в	на	с	за	к
по	из	у	от	для
во	без	до	о	через
со	при	про	об	ко
над	из-за	из-под	под	

СОЮЗЫ:

и	что	но	а	да
хотя	когда	чтобы	если	тоже
или	то есть	зато	будто	

ЧАСТИЦЫ:

не	как	же	даже	бы
ли	только	вот	то	ни
лишь	ведь	вон	то-есть	нибудь
либо	уже			

Однако нужно заметить, что в данном исследовании используются интегральные характеристики, то есть веса каждого служебного слова в своем множестве (предлоги, союзы, частицы) считаются равными, что абсолютно неприемлемо при анализе письменной речи конкретного анонима или автора.

В работе [5] предлагается для исследования задачи идентификации анонимного текста использовать теоретико-вероятностные методы. Предложенный метод основан на идее представления каждого текста в виде точки (вектора) в многомерном пространстве (размерность пространства совпадает с числом служебных слов). Для нормализации (то есть независимости от размера текста), каждый вектор нормируется единицей. Таким образом, каждая категория (тексты, принадлежащие одному автору) представляет собой пучок единичных векторов. Точки (концы орт) будут собраны в области, которые локализуются в случае, если категория правильно собрана. Каждая категория описывается одним вектором, который равен нормированной сумме всех векторов категории. Проверка соответствия данного документа соответствующей категории сводится к вычислению скалярного произведения орта документа и орта категории. Чем ближе это число к единице, тем более вероятно, то документ соответствует этой категории.

Для построения файла статистики последовательно обрабатываются все тексты, принадлежащие одной категории (одному автору). По множеству слов каждого обрабатываемого файла b^{ν} строится множество уникальных (неповторяющихся) служебных слов и их счетчики - $\left(w_i^{\nu}, n_i^{\nu}\right)$. После этого данные для каждого файла отдельно нормируются

$$\hat{n}_i^{\nu} = \frac{n_i^{\nu}}{\sum\limits_i \left(n_i^{\nu}\right)^2}.$$

После этого, упорядочиваем все слова для каждого документа в одном и том же порядке и находим сумму всех векторов $n_i = \sum_j n_i^v$ и нормируем ее единицей

$$\hat{n}_i = \frac{n_i}{\sum_i \left(n_i\right)^2}.$$

Для полученной центральной точки категории формируем структуру, содержащую соответствующее служебное слово и его координату $\left(w_i, \hat{n}_i\right)$.

Определение подходящей категории. Категории, на принадлежность к которым мы проверяем текст, обозначим через $C_j\,(j = 0,...,M-1)$. Для каждого слова w_i в каждом файле статистики находим это слово и соответствующую координату орта $\hat{m}_j\left(w_i\right)$ (здесь j (j=0,1,..,M-1)-номер категории). Кроме того, пусть

$$\left\langle \hat{b}, \hat{C}_j \right\rangle = \sum_i \hat{n}_i \hat{m}_j\left(w_i\right)$$

скалярное произведение орт $\hat{C}_j = \left\{\hat{m}_j\left(w_i\right)\right\}$ и $\hat{b} = \left\{n_i\right\}$. Заметим, что в обоих случаях номер i соответствует одному и тому же слову w_i. Если слово в одном из этих файлов отсутствует, это значит, что его координата равна нулю.

Проверка категорий на корректность и выяснение пересечений категорий. Проверка проводится последовательно среди выбранных категорий. Пусть проверяется на соответствие категории C_v и C_μ. Для соответствующих орт $\hat{C}_v$, $\hat{C}_\mu$, то есть единичных векторов, строим вектор разности и вектор суммы $\vec{\Delta}_{v,\mu} = \hat{C}_v - \hat{C}_\mu = \left\{\hat{m}_v\left(w_i\right) - \hat{m}_\mu\left(w_i\right)\right\}$ и $\vec{\Xi}_{v,\mu} = \frac{1}{2}\left(\hat{C}_v + \hat{C}_\mu\right) = \frac{1}{2}\left\{\hat{m}_v\left(w_i\right) + \hat{m}_\mu\left(w_i\right)\right\}$. Конец вектора полусуммы численно совпадает с координатами этого вектора. Обозначим его через $\Xi_{v,\mu}$. Проведем через точку $\Sigma_{v,\mu}$ плоскость с нормальным вектором $\vec{\Delta}_{v,\mu}$

$$\Omega_{v,\mu} = \left\langle \vec{\Delta}_{v,\mu} \cdot \left(\mathrm{P} - \Xi_{v,\mu}\right) \right\rangle = 0.$$

Эта плоскость разделяет категории. Для того чтобы векторный метод корректно разделял две этих категории, нужно, чтобы все точки (документы) одной категории находились с одной стороны плоскости, то есть, если $b \in C_v$, то

$$\left\langle \vec{\Delta}_{v,\mu} \cdot \left(\hat{C}_v - \Xi_{v,\mu}\right) \right\rangle \left\langle \vec{\Delta}_{v,\mu} \cdot \left(\hat{b} - \Xi_{v,\mu}\right) \right\rangle \geq 0.$$

В координатной форме это условие будет иметь вид

$$\sum_i \left(\hat{m}_v\left(w_i\right) - \hat{m}_\mu\left(w_i\right)\right)\left(\hat{n}_i - \frac{1}{2}\left(\hat{m}_v\left(w_i\right) + \hat{m}_\mu\left(w_i\right)\right)\right) \geq 0.$$

Точки, в которых это условие не выполняется нужно рассмотреть на вопрос принадлежности к категории C_μ.

Возможна ситуация, когда категории имеют непустое пересечение, то есть присутствуют тексты разных авторов. В этом случае нужно выделить непустое пересечение этих категорий.

Для решения этой проблемы предлагается следующий метод. Рассмотрим категории C_ν и C_μ. Разделим их плоскостью $\Omega_{\nu,\mu}$, и все точки, лежащие с одной стороны соберем в новые две категории C_ν^* и C_μ^*.

Пусть

$$d\left(B,\Omega_{\nu,\mu}\right)=\frac{\left|\left\langle\vec{\Delta}_{\nu,\mu}\cdot\left(B-\Xi_{\nu,\mu}\right)\right\rangle\right|}{\left|\vec{\Delta}_{\nu,\mu}\right|}=\frac{\left|\sum_i\left(\hat{m}_\nu\left(w_i\right)-\hat{m}_\mu\left(w_i\right)\right)\left(b_i-\frac{1}{2}\left(\hat{m}_\nu\left(w_i\right)+\hat{m}_\mu\left(w_i\right)\right)\right)\right|}{\sqrt{\sum_i\left(\hat{m}_\nu\left(w_i\right)-\hat{m}_\mu\left(w_i\right)\right)^2}}$$

расстояние от точки $B=\{b_i\}$ до плоскости $\Omega_{\nu,\mu}$.

Если выполняется условие (то есть, после отсечения данных обе категории отодвигаются друг от друга)

$$\begin{cases} d\left(C_\nu^*,\Omega_{\nu,\mu}\right)-d\left(C_\nu,\Omega_{\nu,\mu}\right)>0, \\ d\left(C_\mu^*,\Omega_{\nu,\mu}\right)-d\left(C_\mu,\Omega_{\nu,\mu}\right)>0, \end{cases}$$

то категории C_ν и C_μ имеют непустое пересечение $\widetilde{C}$, которое можно определить следующим образом, $b\in\widetilde{C}$ если $b\in C_\nu$ и при этом $\left\langle\vec{\Delta}_{\nu,\mu}\cdot\left(\hat{C}_\nu-\Xi_{\nu,\mu}\right)\right\rangle\left\langle\vec{\Delta}_{\nu,\mu}\cdot\left(\hat{b}-\Xi_{\nu,\mu}\right)\right\rangle<0$ или если $b\in C_\mu$ и при этом $\left\langle\vec{\Delta}_{\nu,\mu}\cdot\left(\hat{C}_\mu-\Xi_{\nu,\mu}\right)\right\rangle\left\langle\vec{\Delta}_{\nu,\mu}\cdot\left(\hat{b}-\Xi_{\nu,\mu}\right)\right\rangle<0$.

В соответствии с результатами данного исследования, была проведена серия экспериментов, которая подтвердила эффективность приведенной методологии.

Результаты работы позволяют утверждать, что использование предложенного метода анализа распределения служебных слов может использоваться экспертами-криминалистами для установления авторства анонимных материалов письменной речи.

Литература

1. Криминалистика и научно-судебная экспертиза. Под ред. Д.Х. Панасюк. М. 1950.
2. Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного неизвестного автора. Стилеметрический этюд. // Известия отд. русского языка и словесности Имп.Акад.наук, Т.XX, кн.4, 1915.
3. Марков А.А. Об одном применении статистического метода. // Известия Имп.Акад.наук, серия VI, Т.X, N4, 1916, с.239.
4. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Предисловие А.Т. Фоменко. // Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т. 2, – М.: Изд-во МГУ, 1996. – с.768-820.
5. Шумейко А.А., Усенко С.А. Об использовании вероятностных методов распределения служебных слов при проведении экспертизы письменной речи.// Слідча діяльність: проблеми теорії та практики. Дніпропетровський державний університет внутрішніх справ, 2008.- с.112-117.

Статтю подано 5.11.2009