

С. И. Круглова, С. Ю. Мельников, Е. С. Сидоров (Москва, ООО «Линфо»). Способ совместного выбора параметров метода импосторов в открытой задаче идентификации авторства текста (русский язык).

Разработке методов определения авторства текста посвящено значительное количество исследований (см. обзор в [1]). В большинстве работ изучается закрытая задача идентификации (анализируемый текст написан одним из известных авторов). Здесь мы рассматриваем идентификацию авторства текста в условиях открытой задачи (анализируемый текст может быть написан как одним из известных, так и неизвестным автором). Для идентификации используется метод импосторов [2, 3]. Предложен практический способ совместного выбора значений двух параметров этого метода, при которых достигается минимум среднего арифметического ошибок первого и второго родов. Значения параметров вычисляются на базе авторских текстов. Приведены результаты экспериментов для текстов на русском языке.

Метод импосторов. Анализируемому тексту X ставится в соответствие вектор характеристик, состоящий из частот n -грамм символов ($n = 1, 2, 3, 4$), слов ($n = 1, 2$), служебных слов (в подпоследовательности служебных слов [4]) ($n = 1, 2$), шаблонов очертаний слов (последовательность значений регистра букв), доли уникальных слов в тексте. Такие характеристики эффективны для определения авторства.

Степень близости $\text{sim}(X, Y)$ текстов X и Y вычисляется как min-max мера [5] между их векторами характеристик:

$$\text{sim}(X, Y) = \min \max(\mathbf{x}, \mathbf{y}) = \frac{\sum \min(x_i, y_i)}{\sum \max(x_i, y_i)}.$$

Степень $IM(X, Y)$ авторской близости текстов X и Y вычисляется с привлечением множества S посторонних, заведомо «чужих» текстов (так называемых импосторов) следующим образом.

Алгоритм

Вход: X, Y — тексты; S — множество импосторов

Параметры: k, n — целые; $\theta, \Delta, rate > 0$

Выход: $IM(X, Y)$

- 1 $Score = 0$
 - 2 Повторяется k раз
 - a. Случайно выбираются координаты вектора $\mathbf{x}$, которые будут использоваться для подсчета степени близости. Доля множества выбираемых координат равна $rate$
 - b. Случайно выбираются тексты $I_1, \dots, I_n \in S$
 - c.
$$Score = Score + \frac{\text{sim}(X, Y) * \text{sim}(Y, X)}{\text{Max}_{1 \leq j \leq n} \text{sim}(X, I_j) * \text{Max}_{1 \leq j \leq n} \text{sim}(Y, I_j)}$$
 - 3 $IM(X, Y) = \frac{1}{k} Score$
-

Пусть M авторов представлены коллекциями своих текстов

$$A_i = \{T_{ij}, j = 1, 2, \dots, |A_i|\}, \quad i = 1, 2, \dots, M.$$

Для анализируемого текста T вычисляются значения

$$GIM_{\Delta}(T, A_i) = \frac{1}{|A_i|} \sum Ind_{\Delta}(T, T_{ij}),$$

где

$$Ind_{\Delta}(T, T_{ij}) = \begin{cases} 1, & IM(T, T_{ij}) \geq \Delta, \\ 0, & IM(T, T_{ij}) < \Delta, \end{cases} \quad i = 1, 2, \dots, M.$$

В случае $GIM_{\Delta}(T, A_i) \geq \theta$ принимается решение, что текст T написан i -м автором, в противном случае считаем, что текст T написан другим автором.

Выбор параметров Δ и θ . Ошибка первого рода возникает в том случае, когда текст T написан i -м автором, но $GIM_{\Delta}(T, A_i) < \theta$. Ошибка второго рода возникает, если текст T написан другим автором, но $GIM_{\Delta}(T, A_i) \geq \theta$. Диапазоны изменения параметров

$$\Delta \in (0, \text{Max}_{i,j,T} IM(T, T_{ij})) \quad \text{и} \quad \theta \in (0, 1)$$

разбиваются равномерно на несколько отрезков, и в узлах полученной сетки вычисляются ошибки первого и второго родов. В качестве анализируемого текста T последовательно выбираются все авторские тексты $\{T_{ij}\}$, при этом идентифицируемый текст удаляется из авторской коллекции. Лучшей парой значений Δ и θ считается та, для которой сумма ошибок 1 и 2 родов по всем $\sum_{i=1}^M |A_i|$ анализируемым текстам минимальна.

Описание экспериментов. Исследования проводились на корпусе авторских текстов на русском языке, содержащем тексты 29 авторов (от 7 до 25 текстов одного автора), средняя длина текста составила 3543 символа. Корпус выбирался из публикаций в Живом Журнале за 2018–2019 гг. по общественно-политической тематике. Множество импосторов состояло из набора русскоязычных твитов общим объемом 500 МБ. При значениях параметров $k = 10$, $N = 20$, $rate = 0,4$ минимальное значение среднего арифметического ошибок достигнуто при значениях $\Delta = 0,9$, $\theta = 0,3$, и составило 0,26 (см. рис.).

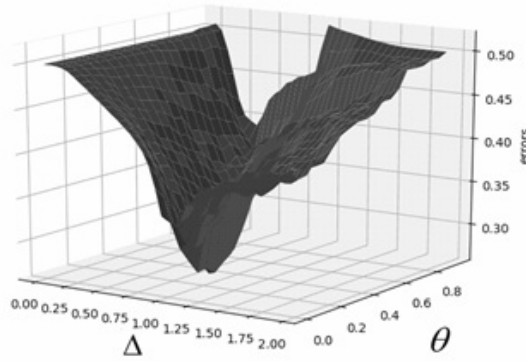


Рис. Среднее арифметическое ошибок первого и второго родов в зависимости от параметров Δ и θ

Заключение. Представлен способ расчета значений параметров метода impostors, при которых на базе авторских текстов достигается минимум среднего арифметического ошибок первого и второго родов.

СПИСОК ЛИТЕРАТУРЫ

1. Романов А. С., Шелупанов А. А., Мещеряков Р. В. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста. Томск: В-Спектр, 2011, 188 с.
2. Seidman S. Authorship verification using the impostors method. – Notebook for PAN at CLEF 2013. In: CLEF 2013 Evaluation Labs and Workshop. (Valencia, September 23–26, 2013.) Working Notes Papers. /Ed. by P. Forner, R. Navigli, D. Tufis. Aachen: CEUR-WS.org, 2013. <http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-Seidman2013.pdf>
3. Khonji M., Iraqi Y. A slightly-modified GI-based author-verifier with lots of features (ASGALF). – Notebook for PAN at CLEF 2014. — In: CLEF 2014 Evaluation Labs and Workshop. (Sheffield, September 15–18, 2014.) Working Notes Papers. /Ed. by L. Cappellato, N. Ferro, M. Halvey, W. Kraaij. Aachen: CEUR-WS.org, 2014. https://pan.webis.de/downloads/publications/papers/khonji_2014.pdf
4. Германович А. В., Мельников С. Ю., Хвостенко В. М. О выборе множества слов, характеризующих авторский стиль арабского текста. — Обозрение прикл. и промышл. матем., 2017, т 24, в. 4, с. 324–325.
5. Koppel M., Winter Y. Determining if two documents are written by the same author. — J. Assoc. Inform. Sci. Technol., 2014, v. 65, is. 1, p. 178–187.