

ОБ ОДНОМ ПОДХОДЕ К РАЗРАБОТКЕ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ АВТОРИЗАЦИИ ТЕКСТОВ

В.В. Ретивина

Нижегородский государственный лингвистический университет им. Н.А. Добролюбова
г. Нижний Новгород

В современном литературоведении и историко-литературоведении накоплено огромное количество методов и приемов атрибуции анонимных произведений. Они имеют различные достоинства, но всех их объединяет один недостаток – это зависимость результатов исследования от субъективных представлений и характерных особенностей исследуемого текста. Основной целью внедрения формально-количественных методов в дело атрибуции анонимных текстов является попытка придать процессу исследования более объективный характер. Естественно, что с развитием вычислительной техники появилась возможность использовать более эффективные методы для решения этой проблемы, учитывая различные лингвостатистические параметры.

Цель данной работы – разработка нового подхода к алгоритмическому и программно-техническому обеспечению процесса обработки литературных текстов, позволяющего автоматизировать процесс стилистического разбора текстов, выделив наиболее информативные параметры, и на основе полученных данных проводить информационно-статистический анализ текстов и, в частности, решать задачу по атрибуции текстов (установлению авторства). Характерной особенностью предлагаемого подхода является рассмотрение задачи установления авторства как задачи статистической классификации дискретных объектов, в которой в качестве решающего правила используется принцип минимума информационного рассогласования (МИР), являющийся одним из наиболее эффективных инструментов для решения подобных задач. Он является интерпретацией классического критерия максимального правдоподобия в задачах распознавания дискретных объектов, имеющей ряд преимуществ. Это было показано, в частности, в работе [1] для дискретного стационарного случая.

Суть критерия минимума информационного рассогласования заключается в следующем: если задача распознавания объекта X сводится к проверке R статистических гипотез о его законе распределения: W_r :

$P = P^{(r)}$, $r = \overline{1, R}$, то решение в пользу гипотезы v принимается по следующему принципу:

$$W_v(X) : \sum_{i=1}^N w_i \log \left(w_i / p_i^{(r)} \right) \Big|_{r=v} = \min_r$$

$r = \overline{1, R}, v \leq R$, где $\hat{P} = \{w_i\}$ определяет эм-

пирический закон распределения состояний рассматриваемого объекта по результатам n независимых наблюдений. Критерий МИР справедлив для задач любой размерности $m > 1$. Это значит, что при решении задач статистической классификации объектов с помощью данного критерия можно одновременно рассматривать несколько параметров в целях получения более точных результатов. Однако на практике данный подход к созданию автоматизированной системы для решения многомерных задач сопряжен с проблемой значительного увеличения вычислительных затрат, а также с проблемой малости выборок. В самом деле, при увеличении числа исследуемых параметров происходит увеличение размерности исследуемой задачи и, соответственно, увеличивается количество возможных состояний изучаемой системы дискретных случайных величин. При этом объем необходимых выборок, дающих достоверные результаты, экспоненциально возрастает. В данной работе предлагается решение этой проблемы для случаев, когда принцип МИР рассматривается в предположении о статистической независимости всех m анализируемых признаков.

При рассмотрении X как системы m независимых случайных величин был получен следующий результат:

критерий минимума информационного рассогласования, реализующий оптимальный алгоритм распознавания R случайных дискретных объектов в случае изучения системы m независимых дискретных случайных величин сводится к проверке условия минимальности суммы информационных рассогласований, полученных по каждой из m независимых дискретных случайных величин, входящих в систему.

$$W_v(X): \sum_{i=1}^N w_i \log(w_i / p_i^{(r)}) \Big|_{r=v} =$$

$$\sum_{l=1}^m \sum_{i_l=1}^{N_{i_l}} w_{i_l} \log(w_{i_l} / p_{i_l}^{(r)}) \Big|_{r=v} =$$

$$\min_r, r = \overline{1, R}, v \leq R,$$

Предложенный принцип был применен в следующем эксперименте. Были взяты два фрагмента текстов: из романа «Новь» И.Тургенева (2314 предложения) и романа «Соборяне» Лескова (1872 предложения). Затем был взят еще один отрывок из романа «Новь» (2250 предложений), не пересекающийся с обучающим фрагментом. В качестве показателей стилистических особенностей этих текстов были выбраны количество запятых и размах длины слова в каждом отдельно взятом предложении. Были выдвинуты две гипотезы – о принадлежности третьего отрывка каждому из двух рассматриваемых произведений соответственно. Для принятия решения были вычислены величины информационных рассогласований тестируемого отрывка с каждым из обучающих фрагментов. Рассогласования вычислялись по двум признакам в совокупности, а также по каждому из признаков в отдельности. Результаты эксперимента представлены в таблице:

Таблица 1 – Результаты вычислительного эксперимента

	Сумма рассогласований, полученных по каждому из признаков	Рассогласование по двум признакам в совокупности
Тестовый отрывок и обучающий отрывок из романа «Новь»	0.0910451223	0.102776973
Тестовый отрывок и обучающий отрывок из романа «Соборяне»	0.3726505178	0.3593721217

Рассматривая величину информационного рассогласования по двум признакам в совокупности, делаем вывод, что третий отрывок взят из романа И.Тургенева «Новь» (что, как мы знаем, в нашем случае верно). Если рассматривать сумму рассогласований

по каждому из признаков, то решение и здесь принимается в пользу первой гипотезы, а именно: текста романа «Новь». Таким образом, оба подхода приводят к одному и тому же результату.

Кроме того, если рассматривать для каждого текста величину информационного рассогласования, полученного по двум параметрам в совокупности, и сумму информационных рассогласований, полученных по каждому из параметров в отдельности, видим, что эти величины достаточно близки (отличаются на малую величину). Анализ степени зависимости рассматриваемых признаков показал, что коэффициент корреляции между ними в каждом из текстов оказался близок к 0, т.е. данный пример подтверждает справедливость критерия минимальности суммы информационных рассогласований.

Выводы

Таким образом, решение задачи статистической классификации объектов с помощью критерия минимума информационного рассогласования значительно упрощается в случае исследования независимых параметров и сводится к вычислению суммы информационных рассогласований, полученных по каждому из параметров в отдельности.

В частности, при разработке автоматизированной системы авторизации текстов необходимо использовать алгоритм формирования наборов информативных параметров, по которым будет производиться сравнение изучаемого текста с набором текстов различных писателей, имеющимся в базе данных. При этом все измеряемые параметры разбиваются на группы таким образом, что в каждую из них попадают параметры, сильно коррелированные между собой, а слабо связанные параметры попадают в разные группы. Затем в каждой группе находится наиболее информативный параметр. В конечный набор параметров включается по одному наиболее информативному параметру из каждой группы, т.е. происходит «свертывание» параметрического пространства до размерности m , равной числу групп [2]. Таким образом, решение задачи распознавания образов будет производиться по набору из m слабо коррелированных между собой параметров. Тогда, руководствуясь критерием минимума суммы информационных рассогласований, для принятия решения будет необходимо рассчитать информационные рассогласования по каждому из m параметров между исследуемым текстом и каждым из эталонных текстов, а затем просуммировать полу-

ОБ ОДНОМ ПОДХОДЕ К РАЗРАБОТКЕ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ АВТОРИЗАЦИИ ТЕКСТОВ

ченные значения для каждого эталона. Решение будет принято в пользу того автора, которому будет соответствовать минимальная сумма информационных параметрических рассогласований.

Список литературы

1. Савченко В.В., Савченко А.В. Принцип минимального информационного рассогласования в задаче распознавания дискретных объектов / В.В Савченко, А.В. Савченко // Радиоэлектроника. Труды высших учебных заведений России. – 2005. – № 1. – С. 61- 71.
2. Марусенко М.А. Атрибуция анонимных и псевдонимных произведений методами теории распознавания образов / М.А. Марусенко – Л.: Издательство ленинградского университета, 1990.