А.В. Кравченко

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МЕТОДОВ ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТОВ

(Самарский национальный исследовательский университет
имени академика С.П. Королева)

Постановка задачи

Сравнить количество ошибок первого и второго рода при определении автора анонимного произведения используя метод «Топ 100» и метод опорных слов.

Методы

Для определения авторства текстов используются методы «Топ 100», метод опорных слов. Подобные методы могут найти своё важное применение в работе историков, литературоведов, юристов.

Метод «Топ 100» был описан в предыдущей работе [1].

Метод опорных слов заключается в следующем. На основе произведений известного автора высчитывается средняя частота использования служебных слов. Для того, чтобы определить автора анонимного произведения, для него находится частота использования служебных слов. Из средних частот известных авторов ищется средняя частота с наименьшей разницей относительно частоты анонимного произведения. Пример используемых служебных слов в методе опорных слов: 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, и, что, но, а, да, хотя, когда, чтобы ,если, тоже, или, то есть, зато, будто, не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, уже, либо [2].

Описание и результаты эксперимента

Проведем два эксперимента, в которых определения количества ошибок первого и второго рода для методов «Топ 100» и опорных слов. Ошибка первого рода называется ошибка, состоящая в опровержении верной гипотезы, в текущих экспериментах это случай, когда автор не распознает свой собственный текст. Ошибкой второго рода называется ошибка, состоящая в принятии ложной гипотезы, в текущих экспериментах это случай, когда автор распознает чужой текст как свой.

Эксперименты будут проводиться с помощью разработанной автоматизированной системы, которая позволяет рассчитывать характеристики для авторов и определять на основе этих характеристик авторов анонимных произведений (рисунок 1).

Проведем первый эксперимент для определения количества ошибок первого и второго рода для метода «Топ 100».

Взято множество произведений тридцати авторов. Тексты произведений были получены из электронной библиотеки Максима Мошкова [3]. Были взяты произведения известных русских писателей таких

как, А.С. Пушкин, А.П. Чехов, Н.В. Гоголь, Л.Н. Толстой, Ф.М. Достоевский, И.С. Тургенев, М. Горький, М.А. Булгаков, М.Ю. Лермонтов, И.А. Гончаров, И.А. Бунин, В.В. Набоков, А.И. Солженицын, М.Е. Салтыков-Щедрин, В.Я. Брюсов, А.Н. Островский, Н.С. Лесков, М.М. Пришвин, Е.И. Замятин, М.А. Шолохов, А.А. Фадеев, А.И. Герцен, А.Н. Толстой, В.Г. Короленко, Л.Н. Андреев, М.П. Арцыбашев, В.В. Вересаев, А.И. Эртель, Ф.М. Решетников, А.Ф. Писемский.

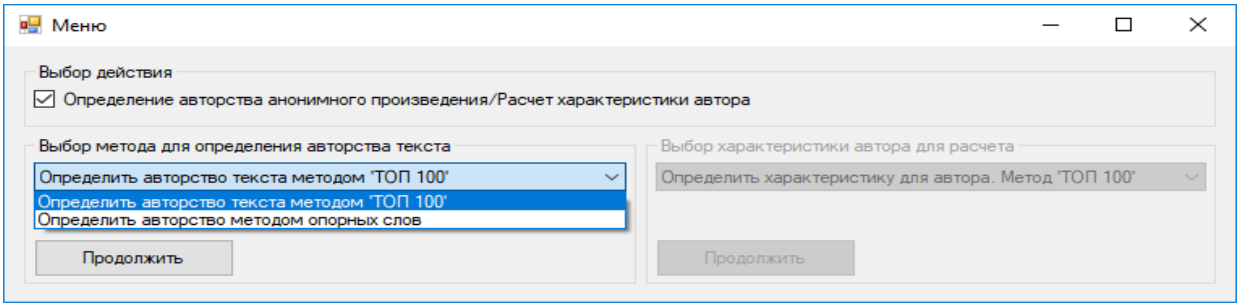

Рисунок 1 – Настройка параметров программы

Сто самых часто используемых слов языка было взято с электронного ресурса «Национальный корпус русского языка» [4]. Пример таких слов это: и, в, не, на, с, что, я, а, он, как, к, по, но, его, это, из, все, у, за, от, то, о, же, так, для, было, она, только, мы, бы, мне, был, ее, или, еще, меня, их, они, до, когда, уже, ты, если, да, вы, вот, при, ни, ему, чтобы, нет, есть, даже, может, быть, во, время, очень, были, была, сказал, ли, под, со, себя, нас, где, него, чем, того, без, будет, этого, теперь, после, там, можно, этом, раз, себе, тем, этот, ну, том, потом, более, них, которые, всех, человек, ничего, надо, тут, тогда, здесь, потому, один, кто, через, который. Для каждого автора на основе текстов его произведений был получен вектор, характеризующий частоту встречаемости ста самых часто используемых слов языка. Вектора составлялись на основе данных, полученных в предыдущей работе, для составления векторов каждого автора использовалось по 60000 – 70000 слов [1]. Для анонимных произведений были составлены аналогичные вектора. Вектора авторов и вектора анонимных произведений были сравнены по формуле среднего абсолютного отклонения. Пример определения автора анонимного произведения, в котором в качестве анонимного произведения выступает третья книга «Тихий дон» представлен на рисунке 2. Как видно, минимальное отклонение относительно анонимного произведения имеет М.А. Шолохов.

В результате проведенных экспериментов было получено, что при определении авторов ста пятидесяти анонимных произведений метод «Топ 100» дает 39 ошибок первого рода и 34 ошибок второго рода.

Проведем эксперимент для метода опорных. Для метода будем использовать те же самые известные произведения авторов и те же анонимные произведения. Для каждого автора на основе его текстов была получена средняя частота использования служебных слов. Для анонимных произведений была рассчитана частота служебных слов. В результате проведенного эксперимента было

получено, что при определении авторов ста пятидесяти анонимных произведений метод опорных слов дает 63 ошибок первого рода и 47 ошибок второго рода.

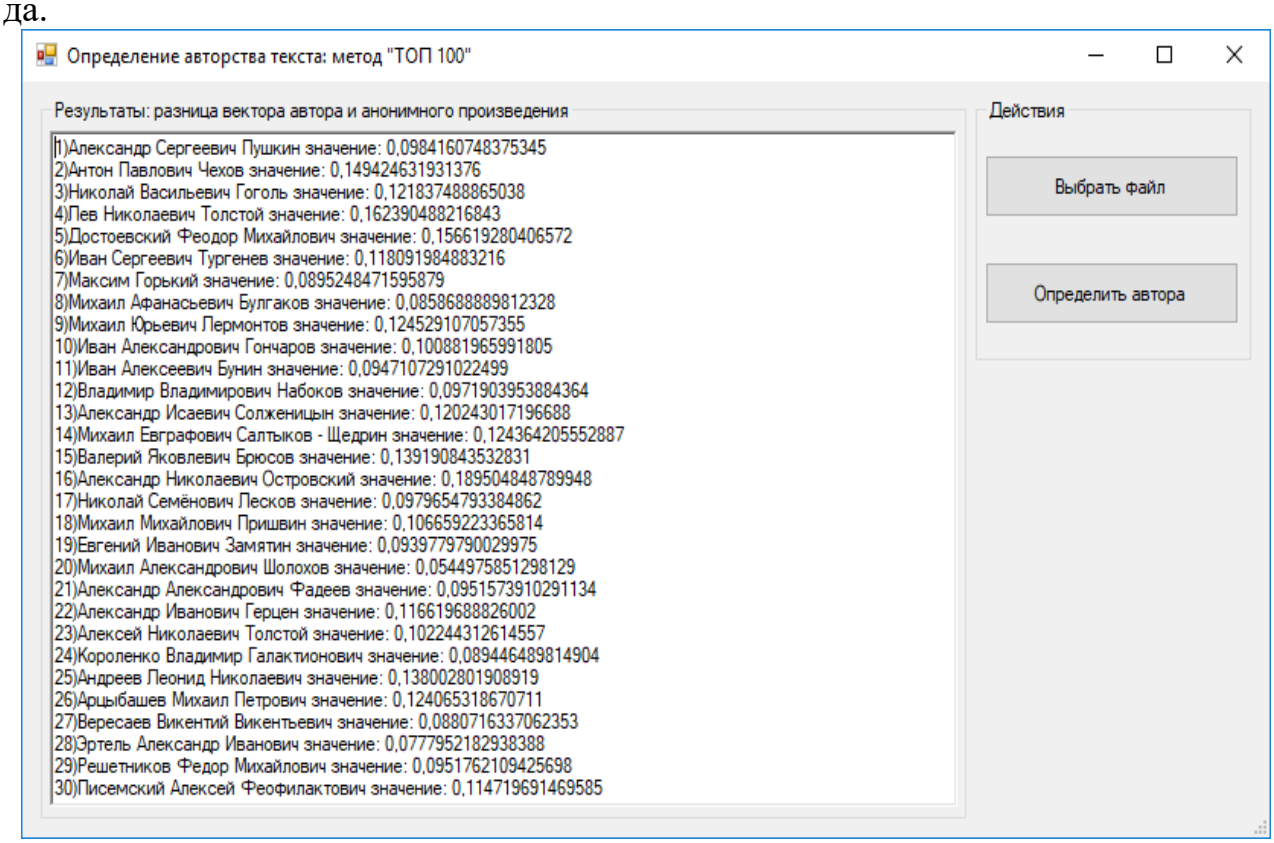

Рисунок 2 – Определение авторства анонимного произведения методом «Топ 100»

Заключение

Таким образом метод «Топ 100» позволяет определить авторство произведения с достаточной долей для большинства практических задач. В результате проведенных экспериментов при использовании формулы среднего абсолютного отклонения метод «Топ 100» дает на выходе 39 ошибок первого рода и 34 ошибок второго рода для ста пятидесяти анонимных произведений.

В результате второго эксперимента при использовании формулы среднего абсолютного отклонения метод опорных слов дает на выходе 63 ошибок первого рода и 47 ошибок второго рода для ста пятидесяти анонимных произведений.

Литература

1 Кравченко, А.В Исследование методов определения авторства текстов [Текст] / А.В Кравченко // Перспективные информационные технологии (ПИТ 2017): труды Международный научной – технической конференции / под ред. С.А. Прохорова. – Самара: Издательство Самарского научного центра РАН, 2017. – с. 102…105.

2 В.П. Фоменко, Т.Г. Фоменко Дополнение 3. Авторский инвариант русских литературных текстов. Приложение: Кто был автором «Тихого дона»? [Электронный ресурс]. – http://chronologia.org/seven2_2/add3.html (дата обращения 01.02.2018 г.).

3 Библиотека Максима Мошкова [Электронный ресурс]. – http://www.lib.ru/ (дата обращения 01.02.2018 г.);

4 Национальный корпус русского языка [Электронный ресурс]. – http://ruscorpora.ru/1grams.top.html (дата обращения 01.02.2018 г.).

Т. Г. Кудрявцева

ВИЗУАЛИЗАЦИЯ ГРАФОВЫХ МОДЕЛЕЙ НА ОСНОВЕ АНАЛИЗА ТОПОЛОГИИ ГРАФОВ

(Самарский университет)

Графы являются универсальным средством представления структурированной информации во многих областях науки и применимы для отображения любой информации, которая может быть представлена в виде совокупности объектов и связей между ними. Поэтому визуализация графов является ключевой компонентой во многих приложениях науки и техники.

Для удобного анализа графовых моделей широко применяются различные средства визуализации. Все системы визуализации предоставляют возможность загрузить граф из файла, рассчитать какие-либо его характеристики и построить изображение. Но ни одна из них не позволяет осуществить определение топологии графа, что позволило бы строить более качественные изображения.

Целью работы является решение проблемы визуализации графовых моделей, которая возникает при обработке больших объемов информации и сложных структур данных. Значительное внимание уделяется способам отображения графов в зависимости от их топологии. Анализ топологии графа позволяет определить наиболее оптимальный метод визуализации, учитывающий эстетические критерии для конкретного типа графа.

В работе анализируются следующие основные топологии графов:

6. безмасштабный граф (scale-free) (рисунок 1);
7. геометрический граф (рисунок 2);
8. граф Эрдеша-Реньи (рисунок 3).

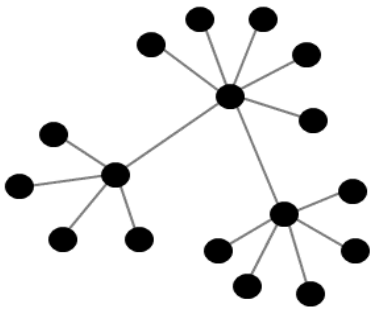

Рисунок 1 – Безмасштабный граф