https://saf.petrsu.ru
ISSN 2500-140X

https://petrsu.ru

Студенческий научный электронный журнал

StudArctic Forum

№ 4 (24), 2021

Издатель
ФГБОУ «Петрозаводский государственный университет»
Российская Федерация, г. Петрозаводск, пр. Ленина, 33

Адрес редакции
185910, Республика Карелия, г. Петрозаводск, ул. Ленина, 33.
E-mail: saf@petrsu.ru
http://saf.petrsu.ru

https://saf.petrsu.ru
ISSN 2500-140X

Scientific journal

StudArctic Forum

№ 4 (24), 2021

Publisher
© Petrozavodsk State University, 2012—2021

Address
33, Lenin av., 185910 Petrozavodsk, Republic of Karelia, Russia
E-mail: saf@petrsu.ru
http://saf.petrsu.ru

УДК 81`32

Компьютерные и информационные науки

ХАРИЧЕВ
Евгений Сергеевич

прикладной бакалавриат, Петрозаводский государственный университет (Петрозаводск, Российская Федерация),
E-mail: ge01nia85@gmail.com

Установление авторства текстов с помощью методики сравнения размеченных ориентированных графов

Научный руководитель:
Кулаков Кирилл Александрович
Статья поступила: 11.10.2021;
Принята к публикации: 31.10.2021;

Аннотация. В данной статье автор описывает исследование методик анализа текстов на основе сравнения размеченных ориентированных графов, а также разработку инструмента визуализации для изучения графовых моделей.
Ключевые слова: ИС «СМАЛТ»; атрибуция текстов; установление авторства; сравнение размеченных ориентированных графов; визуализация графовых моделей.

Для цитирования: Харичев Е. С. Установление авторства текстов с помощью методики сравнения размеченных ориентированных графов / // StudArctic Forum. 2021. № 4 (24). С. 13—15.

В литературе одной из самых актуальных филологических задач является проблема установления авторства (атрибуции). Ручная обработка текстов подвергалась критике из-за неполного и субъективного анализа - как следствие, стали появляться разнообразные решения для автоматизации данного процесса. Одним из таких решений является информационная система «Статистические методы анализа литературного текста» (ИС «СМАЛТ») на базе ПетрГУ. Несмотря на многообразие методик установления авторства, проблема до сих пор не имеет однозначного решения, поэтому существует потребность в анализе имеющихся способов, а также в разработке новых.

В качестве основного направления исследования был выбран анализ текстов путем сравнения размеченных ориентированных графов, а именно группы и одного. Потребность в таком анализе возникает в случае, когда имеется ряд текстов определенного автора, и необходимо оценить близость документа неизвестного происхождения к данному набору текстов.

Теоретико-графовая модель строится на уровне лексики. Вершины графа — набор грамматических форм, которые встречаются в текстах. Дуги образуют пары грамматических форм с весом равным относительной частоте встречаемости данной пары в тексте. Расстояние (степень близости) между группой графов и искомым графом определяется по методам ближайшего соседа и вероятностного классификатора [Проблема атрибуции: 225].

Метод ближайшего соседа предполагает оценку расстояний между заданным графом и каждым графом группы, а затем выбор минимального из них. В качестве меры близости используются евклидово расстояние и метрика городских кварталов. После вычислений происходит анализ близости цветом на основе расчёта минимального и максимального расстояния в группе графов.

Пусть A — группа графов, B — матрица для искомого графа. Вторая методика использует меру близости равную (1), где pij — относительная частота встречаемости данной связи в графах их группы. Если она не встретилась, то положим $pij = \alpha$, где α — настроечный коэффициент больший нуля,bij — элемент матрицы B.

$$[P(A,B) = \prod_{i=1}^{n} \prod_{j=1}^{n} p_{ij}^{b_{ij}}]$$

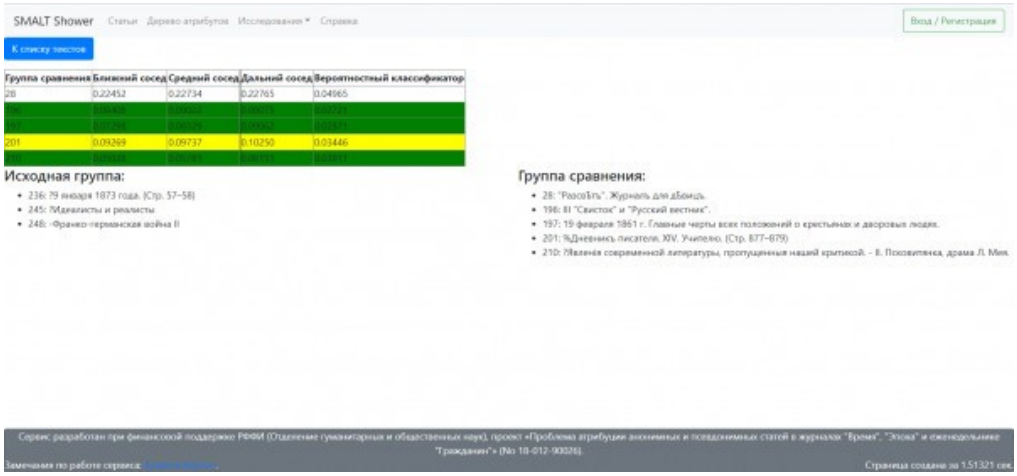

Рисунок 1. Реализация методик в системе.

После реализации предложенных методик в системе была проведена их апробация на текстах Ф. М. Достоевского (36 документов) и В. П. Мещерского (7 документов) [Проблема атрибуции: 348—354]. Результаты исследования показали, что данные методики необходимо совершенствовать, так как процент правильной оценки авторства не удовлетворяет необходимым требованиям (32% — метод ближайшего соседа, 7% — вероятностный классификатор).

Дополнительно к основному исследованию возникла потребность в программе визуализации графовых моделей для одной из первых версий ИС «СМАЛТ», написанной на VBA (Visual Basic for Applications) в Microsoft Excel. Были сформулированы основные требования к программе визуализации:

- считывание матриц смежности графов из файлов Excel;
- обработка порогового (минимальный вес дуги, при котором она не удаляется) и узлового (минимальная полустепень захода вершины, при которой она не удаляется) значений;
- вывод визуального представления отдельного графа, а также наложения двух графов с выделение цветом общих и отличающихся частей;
- некоторые дополнительные функции (вывод легенды, двойные стрелки и др.).

В результате было создано оконное приложение Windows Forms на языке C# в связке с комплексом утилит для автоматической визуализации графов (Graphviz). Чтобы получить изображение, программе на основе матрицы необходимо правильно составить описание графа на специальном языке DOT и запустить обработчик.

Программа успешно протестирована заказчиками [Проблема атрибуции: 225—229]. В дальнейшем планируется встроить её функционал в ИС «СМАЛТ» для удобного изучения графовых моделей текстов.

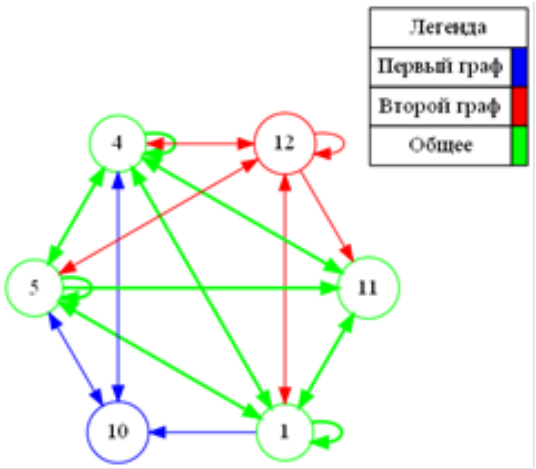

Рисунок 2. Пример результата работы программы визуализации

СПИСОК ЛИТЕРАТУРЫ

Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин» : [монография] / А. А. Рогов, Р. В. Абрамов, Д. Д. Бучнева, О. В. Захарова, К. А. Кулаков, А. А. Лебедев, Н. Д. Москин, А. В. Отливанчик, Е. Д. Савинов, Ю. В. Сидоров. — Петрозаводск : Издательство «Острова», 2021. 391 с.

Computer and Information Sciences

KHARICHEV Evgeniy

applied bachelor, Petrozavodsk State University
(Petrozavodsk, Russian Federation)
E-mail: ge01nia85@gmail.com

Authorship determination of texts using the method of comparison marked oriented graphs

Scientific adviser: Kirill Kulakov
Paper submitted on: 10/11/2021;
Accepted on: 10/31/2021;

Abstract. In this article, the author describes a study of techniques for text analysis based on comparison of marked oriented graphs, as well as the development of a visualization tool for studying graph models.

Keywords: «SMALT» information system; text attribution; identification of authorship; comparison of weighted directed graphs; graph model visualization.

REFERENCES

The problem of attribution in the magazines "Time", "Epoch" and the weekly "Citizen" : [monograph] / A. A. Rogov, R. V. Abramov, D. D. Buchneva, O. V. Zakharova, K. A. Kulakov, A. A. Lebedev, N. D. Moskin, A.V. Castanchik, E. D. Savinov, Yu. V. Sidorov. - Petrozavodsk : Publishing house "Islands", 2021. 391 p.