На правах рукописи

КИЖАЕВА Наталья Александровна

ИССЛЕДОВАНИЕ ПАТТЕРНОВ В ТЕКСТАХ НА ОСНОВЕ ДИНАМИЧЕСКИХ МОДЕЛЕЙ

01.01.09 — дискретная математика и
математическая кибернетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург
2018

Работа выполнена в Санкт-Петербургском государственном университете.

Научный руководитель:	доктор физико-математических наук, профессор Граничин Олег Николаевич
Официальные оппоненты:	Хлебников Михаил Владимирович, доктор физико-математических наук, профессор РАН, ФГБУН «Институт проблем управления им. В.А. Трапезникова» РАН, главный научный сотрудник, и.о. заведующего лабораторией адаптивных и робастных систем им. Я.З. Цыпкина
	Петухова Нина Дмитриевна, кандидат физико-математических наук, ФГБОУ ВО «Санкт-Петербургский государственный морской технический университет», доцент
Ведущая организация:	Институт системного анализа Федерального исследовательского центра «Информатика и управление» РАН

Защита состоится «30» мая 2018 года в 18 часов на заседании диссертационного совета Д 212.232.29 на базе Санкт-Петербургского государственного университета по адресу: 199178, Санкт-Петербург, 10 линия В.О., д. 33/35, ауд. 74.

С диссертацией можно ознакомиться в Научной библиотеке им. М. Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб., д. 7/9. и на сайте https://disser.spbu.ru/files/disser2/disser/d0wHQhP77v.pdf

Автореферат разослан «____» _____ 2018 г.

Ученый секретарь
диссертационного совета Д 212.232.29,
доктор физико-математических наук,
профессор В. М. Нежинский

Общая характеристика работы

Актуальность темы. На протяжении последних десятилетий наблюдается значительный рост объема текстовой информации, генерируемой каждый день. Этот огромный объем данных представляется в различных формах, таких, как записи в социальных сетях, записи осмотра пациентов, данные медицинского страхования, статьи новостных агентств, отчеты о работе технических устройств и т. п. Текстовые данные — это пример неструктурированной информации, которая легко обрабатывается и воспринимается человеком, но является гораздо более сложной для понимания компьютером. Задача интеллектуального анализа текстов состоит в извлечении полезной информации из неструктурированных текстов, их автоматической категоризации, классификации и кластеризации. Автоматизированный анализ позволяет исследователям не только собирать и изучать объем материала, анализ которого вручную невозможен, но и выявлять закономерности, незаметные при простом прочтении.

Интеллектуальный анализ текстов является частью более широкого класса задач интеллектуального анализа данных, машинного обучения и теории распознавания образов. Современные алгоритмы машинного обучения (классификации, кластеризации) и теории распознавания образов базируются на работах С. А. Айвазяна, М. А. Айзермана, Э. М. Бравермана, В. Н. Вапника, Ф. Розенблатта, Л. И. Розоноэра, Р. А. Фишера , В. Н. Фомина, К.Фукунаги, Я.З.Цыпкина, А.Я.Червоненкиса, Дж.Хартигана, Дж.Хопфилда, В. А. Якубовича и др. Многие современные системы распознавания образов основаны на принципах нейронных сетей (см. С. Хайкин, Ф. Уоссермен, А. В. Тимофеев, А. И. Галушкин и др.)

Большинство методов интеллектуального анализа текстов рассматривает текст как статический объект, не учитывая процесс его написания или динамику последовательности изложения. В то же время динамика текстового документа может служить его отличительной характеристикой, признаком, по которому в множестве текстов можно выделить группы схожих документов. Это, в свою очередь, открывает множество сфер применения: определение авторства текстов, выявление плагиата, поиск аномалий в отчетах о работе технических устройств.

Перечисленные факторы актуализируют разработки методов классификации текстовых документов, которые кроме статических характеристик текстов учитывали бы связи (корреляции) между последовательностями их фрагментов.

Целью работы является исследование паттернов динамической модели текстовых документов.
Были поставлены и решены следующие задачи:

- Разработать метод построения динамических моделей текстовых документов.
- Исследовать, является ли динамика изменений фрагментов текстового документа его отличительной характеристикой.
- Разработать и обосновать алгоритмы кластеризации динамических моделей.

Методы исследования. В диссертации применяются методы теории оценивания и оптимизации, функционального анализа, теории вероятностей и математической статистики, машинного обучения и компьютерной лингвистики.
Основные результаты. В ходе выполнения работы получены следующие научные результаты:

1. Предложен метод построения динамических моделей текстовых документов.
2. Разработан и теоретически обоснован алгоритм классификации фрагментов текстовых документов, основанный на кластеризации спектрального представления динамических моделей текстовых документов.
3. Разработан и теоретически обоснован алгоритм классификации фрагментов текстовых документов, основанный на кластеризации динамических моделей текстовых документов с помощью расстояний на ядрах.

Научная новизна. Все основные научные результаты диссертации являются новыми.
Теоретическая ценность и практическая значимость. Теоретическая ценность работы состоит в предложенном методе построения динамической модели текста, разработке и обосновании алгоритмов классификации фрагментов текстовых документов.

Предложенные новые методы находят применение в множестве прикладных задач и исследовательских задач. Определение авторства текстов в литературных исследованиях, в криминалистике, при выявлении плагиата. Анализ неструктурированной текстовой информации в отчетах технических устройств с помощью

предложенного алгоритма предоставляет возможность выявления неоднородности стиля, а, значит, и возможного сбоя технического устройства.

Степень достоверности и апробация работы. Достоверность основных утверждений диссертации подтверждается строгостью математических доказательств. Работоспособность предлагаемых методов подтверждена численными экспериментами.

Материалы диссертации докладывались на семинарах кафедр системного программирования и теоретической кибернетики математико-механического факультета СПбГУ, семинарах Лаборатории анализа и моделирования социальных процессов СПбГУ, семинарах факультета интеллектуальной обработки информации колледжа ОРТ им. Брауде (Кармиэль, Израиль), на международных конференциях AINL-ISMW FRUCT Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search (9-14 ноября, 2015, Санкт-Петербург, Россия), 2015 IEEE International Symposium on Intelligent Control (September 21-23, 2015, Sydney, Australia), 8th International Scientific Conference on Physics and Control (PhysCon 2017) (July 17-19, Florence, Italy), 2017 IEEE Conference on Control Technology and Applications (August 27-30, 2017, Coast, Hawaii, USA).

Результаты диссертации были использованы в работах по грантам СПбГУ "Исследование возможностей кластеризации рукописных текстов на арабском языке" 6.37.181.2014, "Определение формальных характеристик арабографических рукописей и их цифровая обработка" 2.37.175.2014.

Публикация результатов. Основные результаты исследований опубликованы в работах [1-7]. Из них четыре [1-4] в периодических рецензируемых изданиях, индексируемых в наукометрических базах данных SCOPUS и Web of Science или включенных в перечень научных журналов, рекомендованных ВАК.

Работы [1-5] написаны в соавторстве. В работах [1-5] Н.А. Кижаевой принадлежат формулировки и доказательства теорем, результаты моделирования, а соавторам — постановки задач и выбор методов решения.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы, включающего 150 источников. Текст занимает 86 страниц и содержит 10 рисунков.

Содержание работы

Во **введении** обосновывается актуальность темы диссертационной работы и кратко излагаются основные результаты.

В **первой главе** "Интеллектуальный анализ текстов" приводится краткий обзор литературы по теме исследования, вводятся основные понятия и обозначения, описываются постановки задач исследований предметной области.

В п. 1.1 рассматриваются основные проблемы и задачи, которые возникают в сфере интеллектуального анализа текстовых данных. Ключевые задачи интеллектуального анализа текстов включают в себя извлечение информации, реферирование, обучение с учителем, обучение без учителя, извлечение мнений, анализ биомедицинских данных и т. п.

В п. 1.2 перечисляются этапы предварительной обработки текстовых документов и дается описание распространенных моделей представления текстовых данных. Предобработка текстов — важный этап большинства алгоритмов. Этап предобработки обычно состоит из токенизации, фильтрации, лемматизации и стемминга. Векторная модель — представление текстов в виде векторов из некоторого общего для всех текстов векторного пространства.

В пп. 1.3 и 1.4 формулируются проблемы классификации и кластеризации и приводятся классические алгоритмы для их решения.

Пусть $\mathcal{Z} = \{\mathbf{z}^j\}_{j=1}^m$, $\rho(\mathbf{z}, \mathbf{z}')$ — метрика. Задача кластеризации заключается в нахождении разбиения множества $\mathcal{Z}$ на k кластеров таких, что

$$\mathcal{T}^k(\mathcal{Z}) = \{C_1, \ldots, C_k\},$$

$$\mathcal{Z} = \bigcup_{i=1}^{k} C_i, \quad C_i \cap C_j = \emptyset,\ i \neq j.$$

Для разбиения $\mathcal{T}^k(\mathcal{Z})$ функция $\gamma_{\mathcal{T}^k} : \mathcal{Z} \to \{1, \ldots, k\}$, соотносящая точки кластерам, определена следующим образом

$$\gamma_{\mathcal{T}^k}(\mathbf{z}) = i \Leftrightarrow \mathbf{z} \in C_i, i = 1, \ldots, k.$$

Таким образом

$$C_i = \{\mathbf{z} \in \mathcal{Z} | \gamma_{\mathcal{T}^k}(\mathbf{z}) = i\}.$$

Для любого k для множества $\mathcal{Z}$ существуют различные разбиения $\mathcal{T}^k(\mathcal{Z})$.

Разбиение должно обладать следующим свойством: объекты, принадлежащие одному кластеру более "похожи" между собой, чем объекты, принадлежащие разным кластерам. Определим q_i — функцию "близости" к кластеру i, для любого $i = 1, \ldots, k$. Рассмотрим задачу минимизации

$$f(\mathcal{T}^k, \mathbf{z}) = \sum_{i=1}^{k} \gamma_{\mathcal{T}^k}(\mathbf{z}) q_i(\mathcal{T}^k, \mathbf{z}) \to \min_{\mathcal{T}^k}. \tag{1}$$

Результат минимизации функции (1) зависит от z. Пусть вероятностное распределение $P(\cdot)$ определено на множестве $\mathcal{Z}$. Тогда можно рассматривать задачу минимизации функции качества

$$F(\mathcal{T}^k) = Ef(\mathcal{T}^k, \mathbf{z}) = \sum_{i=1}^{k} \int_{C_i} q_i(\mathcal{T}^k, \mathbf{z}) P(d\mathbf{z}) \to \min_{\mathcal{T}^k} \tag{2}$$

В некоторых случаях можно ограничиться разбиением $\mathcal{T}^k$, которое полностью определяется множеством k векторов $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathbb{R}^m$, которые формируют $m \times k$ матрицу $C = (\mathbf{c}_1, \ldots, \mathbf{c}_k)$ и для $i = 1, \ldots, k$ и $\mathbf{z} \in \mathcal{Z}$ функции $q_i(\cdot, \mathbf{z})$ зависят только от $\mathbf{c}_i$, то есть $q_i(\cdot, \cdot) : \mathbb{R}^m \times \mathcal{Z} \to \mathbb{R}$. Правило разбиения можно задать следующим образом

$$\begin{aligned} C_i(\mathcal{Z}) = \{\mathbf{z} \in \mathcal{Z} : & q_i(\mathbf{c}_i, \mathbf{z}) < q_j(\mathbf{c}_j, \mathbf{z}),\ j = 1, \ldots, i-1 \\ & q_i(\mathbf{c}_i, \mathbf{z}) \le q_j(\mathbf{c}_j, \mathbf{z}),\ j = i+1, \ldots, k\},\ i = 1, \ldots, k, \end{aligned}$$

которое минимизирует (1). Вектора $\mathbf{z}_i$, $i = 1, \ldots, k$ интерпретируются как центры кластеров, когда $\mathcal{Z}$ — подмножество евклидова пространства $\mathbb{R}^m$. В этом случае функционал качества (2) принимает форму

$$F(\mathcal{T}^k) = \sum_{i=1}^{k} \int_{C_i} q_i(\mathbf{c}_i, \mathbf{z}) P(d\mathbf{z}) \to \min_{\mathcal{T}^k}. \tag{3}$$

и может быть переписан в виде

$$F(C) = \int_{\mathcal{Z}} \langle l(C, \mathbf{z}), q(C, \mathbf{z}) \rangle P(d\mathbf{z}) \to \min_{C}, \tag{4}$$

где $l(C, \mathbf{z})$ и $q(C, \mathbf{z})$ — вектора длины k такие, что первый состоит из значений характеристической функции $\mathbb{1}_{C_i(C)}(C, \mathbf{z})$, а второй из $q_i(\mathbf{c}_i, \mathbf{z})$, $i = 1, \ldots, k$.

Такая формализация имеет простую геометрическую интерпретацию. Пусть распределение $P(\cdot)$ равномерно на $\mathcal{Z}$ и пусть функции $q_i(\mathbf{c}_i, \mathbf{z}) = ||\mathbf{z}_i - \mathbf{z}||^2,\ i = 1, \ldots, k$ представляют расстояние до центров кластеров $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$. Интеграл $\int_{C_i} ||\mathbf{c}_i - \mathbf{z}||^2 dz$ определяет разброс точек $\mathbf{z}$ множества C_i. Функционал (4) принимает вид

$$F(C) = \sum_{i=1}^{l} \int_{C_i} ||\mathbf{c}_i - \mathbf{z}||^2 d\mathbf{z} \to \min_C . \tag{5}$$

Таким образом, задача кластеризации свелась к задаче нахождения такого множества центров $\{\mathbf{c}_1^*, \ldots, \mathbf{c}_k^*\}$, для которых общий разброс точек минимален.

В п. 1.5 даны определения мер сходства и различия, приведены примеры широко используемых функций расстояния и схожести. Пусть $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, $P > 0$, $|\mathcal{X}| = N$. Обозначение $\mathbf{x}_{ik}$ означает k-й элемент $\mathbf{x}_i$.

В численных экспериментах в работе были использованы следующие функции расстояния:

- Корреляция Спирмена: $d_{Spearman} := 1 - \dfrac{6 \sum_{i=1}^{N} (R(\mathbf{x}_i) - R(\mathbf{y}_i))^2}{N(N^2 - 1)}$, где $R(x_i), R(y_i)$ – ранги элементов $\mathbf{x}_i, \mathbf{y}_i$ в последовательностях $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ и $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ соответственно.
- Расстояние Канберра: $d_{Canberra} := \sum_{i=1}^{P} \dfrac{|\mathbf{x}_{ik} - \mathbf{x}_{jk}|}{|\mathbf{x}_{ik}| + |\mathbf{x}_{jk}|}$.

Во **второй главе** "Динамическая модель процесса эволюции текстовых документов" предложен один из возможных методов построения динамической модели текста. На основе предложенной динамической модели были разработаны и обоснованы два метода классификации документов и их фрагментов. Первый метод основан на кластеризации периодограмм, второй использует кластеризацию с помощью расстояния, основанного на некоторых ядрах. Сформулированы теоремы об однозначности и корректности построенных процедур классификации.

В п. 2.1 описывается метод построения динамической модели текстовых документов, исследованы свойства модели.

Пусть $\{X_i\}_{i=1}^n$ — множество текстовых документов. Под текстовым документом будет понимать упорядоченное множество символов.

$\forall i = 1, \ldots, n$ разделим документ X_i на m_i последовательных фрагментов:

$$X_i = x_i^1 + \ldots + x_i^{m_i}, \tag{6}$$

где "+" — операция конкатенации строк. Рассмотрим множество всех фрагментов $\overline{X} = \{x_i^j\}_{i \in 1..n, j \in 1..m_i}$.

Введем отображение V, которое сопоставляет фрагменту $x_i^j \in \overline{X}$ некоторое вероятностное распределение $P \in \mathcal{P}_M$ из множества вероятностных распределений на $\{1, \ldots, M\}$:

$$V : \overline{X} \to \mathcal{P}_M,$$

$$P \in \mathcal{P}_M : \quad P = \{p_i\}_{i=1}^M, \quad p_i \geq 0, \; \sum_{i=1}^M p_i = 1.$$

Таким образом

$$\mathbf{x}_i^j = V(x_i^j) \in \mathbb{R}^M. \tag{7}$$

Обозначим $\mathcal{X} = \{\mathbf{x}_i^j\}_{i \in 1..n, j \in 1..m_i}$ — множество всех фрагментов в векторном представлении.

Значение параметра M определяется выбранной векторной моделью. Примеры распространенных векторных моделей приведены в п. 1.2.2 диссертации. Пусть $\mathcal{V} = \{v_1, \ldots, v_A\}$ — множество всех термов в коллекции документов, называемое словарем. В случае модели "мешка слов" $M = |\mathcal{V}|$, текст представляется в виде распределения частот появления в нем всех термов из словаря. Модель ключевых слов является частным случаем предыдущей, текст представляется распределением частот появления слов из некоторого подмножества $\mathcal{V}' \subset \mathcal{V}$, таким образом $M = |\mathcal{V}'|$. В модели N-грамм, строится словарь всех N-грамм $\mathcal{V}_N$, встречающихся в документах из множества документов, в этом случае $M = |\mathcal{V}_N|$.

Будем считать, что на множестве $\mathbb{R}^M \times \mathbb{R}^M$ определена некоторая функция похожести двух фрагментов:

$$r : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}. \tag{8}$$

Пусть $T > 0$. Для $i \in 1..n$, $j > T$, $\mathbf{x}_i^j \in \mathcal{X}$ обозначим через $\Delta_{x_i^j}$ множество предшествующих ему векторов-фрагментов: $\Delta_{\mathbf{x}_i^j} = \{\mathbf{x}_i^{j-T}, \ldots, \mathbf{x}_i^{j-1}\}$.

Каждая последовательность векторов-фрагментов $\Delta_{\mathbf{x}}$ с помощью описанной вы-

ше функции (8) порождает функцию $s_{\mathbf{x}}(\cdot) : \mathbb{R}^M \to \mathbb{R}$:

$$s_{\mathbf{x}}(\mathbf{y}) = \frac{1}{T} \sum_{\mathbf{x}' \in \Delta_{\mathbf{x}}} r(\mathbf{x}', \mathbf{y}), \tag{9}$$

которую будем называть *динамической моделью*.

Значения функции $s_{\mathbf{x}}(\mathbf{y})$ соответствуют средней похожести вектора-фрагмента $\mathbf{y}$ с каждым из векторов-фрагментов из $\Delta_{\mathbf{x}}$.

Таким образом, введено отображение

$$\psi : \mathbf{x}_i^j \to s_{\mathbf{x}}(\cdot). \tag{10}$$

В п. 2.2.1 сформулирован алгоритм кластеризации с помощью спектрального представления и правило классификации на его основе. Сформулирована теорема о корректности описанной ниже процедуры.

Каждый документ из множества $\{X_i\}_{i=1}^n$ разделим на одинаковое количество последовательных фрагментов $\bar{m}$. Для каждого фрагмента получим его векторное представление согласно (7). Сопоставим документу последовательность векторов-фрагментов:

$$X_i \mapsto \{\mathbf{x}_i^j\}_{j \in 1..\bar{m}}. \tag{11}$$

Пусть $T > 0$. Для $j > T$, для каждого $\mathbf{x}_i^j \in \{\mathbf{x}_i^j\}_{j \in T+1..\bar{m}}$ построим динамическую модель $s_{\mathbf{x}_i^j}(\cdot)$. Рассмотрим последовательность выходов динамической модели:

$$\{\mathbf{x}_i^j\}_{j \in T+1..\bar{m}} \mapsto \{s_{\mathbf{x}_i^j}(\mathbf{x}_i^j)\}_{j \in T+1..\bar{m}}. \tag{12}$$

Последовательность (12) представляет собой временной ряд, соответствующий i-му документу.

Введем следующие обозначения:

- $\mathrm{s}_i^j = s_{\mathbf{x}_i^j}(\mathbf{x}_i^j)$, $j \in T+1..\bar{m}$, $i \in 1..n$ — средняя мера похожести фрагмента $\mathbf{x}_i^j$ и предшествующих ему фрагментов.
- $\mathcal{S}_i = \{\mathrm{s}_i^j\}_{i \in 1..n, j \in T+1..\bar{m}}$ — последовательность средних мер похожести, временной ряд.
- $\mathbb{S} = \{\mathcal{S}_i\}_{i \in 1..n}$ — множество последовательностей – временных рядов, соответствующих разным документам коллекции.

Периодограммой называется оценка спектральной плотности мощности сигнала, ее вычисление основано на подсчете коэффициентов преобразования Фурье с последующим усреднением.

Для каждого временного ряда $\mathcal{S}_i$ вычислим его периодограмму.

$$\mathcal{S}_i \mapsto \mathrm{PG}(\mathcal{S}_i). \tag{13}$$

Обозначим $\mathbb{F} = \{\mathrm{PG}(\mathcal{S}_i)\}_{i\in 1..n}$ — множество всех периодограмм документов. Заметим, элементы множества $\mathbb{F}$ являются векторами из $\mathbb{R}^{\bar{m}}$, будем называть $\mathbb{F}$ — пространством коэффициентов Фурье.

Будем кластеризовать элементы множества помощью алгоритма кластеризации Cl, минимизирующего функционал (5).

Количество кластеров определяется значением индекса алгоритма валидации кластеризации.

Описанную процедуру можно сформулировать в виде следующего алгоритма.

Алгоритм 1

- X — множество текстов
- T — параметр задержки
- $k^\star$ — максимальное количество кластеров
- Cl — алгоритм кластеризации
- CLV — индекс алгоритма валидации кластеризации

1. Преобразовать документ $\mathcal{X}_i \in X$ во временной ряд $\mathcal{S}_i$ последовательно применив (11) и (12).
2. Для каждого временного ряда вычислить периодограмму $\mathrm{PG}(\mathcal{S}_i)$.
3. **for** $k = 2$ *to* k^* **do**
4. $\mathcal{T} = Cl(\{\mathrm{PG}(S_i)\}_{i\in 1..n}, k)$;
5. $ind_k = CLV(\mathcal{T})$;
6. **end for**

7. Количество кластеров соответствует оптимальному числу кластеров, согласно значению индекса ind_k $\{k = 2, .., k^*\}$.

Пусть в результате работы Алгоритма 1 периодограммы документов разделились на k кластеров $L_1, \ldots, L_k$. Тогда в пространстве временных рядов можно определить следующее правило классификации, относящее документ к одному из классов $l_1, \ldots, l_k$:

Правило классификации 1

Два документа X_i и X_j относятся к одному классу l_k, если соответствующие им периодограммы $\mathrm{PG}(\mathcal{S}_i)$ и $\mathrm{PG}(\mathcal{S}_j)$ попали в один кластер k.

Теорема 1. *Кластеризация в пространстве $\mathbb{F}$ обеспечивает однозначность и корректность Правила классификации 1.*

В п. 2.2.2 сформулирован алгоритм кластеризации по расстояниям, основанным на ядрах, и правило классификации на его основе. Сформулирована теорема о корректности описанной ниже процедуры.

Каждый документ из множества $\{X_i\}_{i=1}^n$ разделим на последовательные фрагменты одинаковой длины. Далее для каждого фрагмента получим его векторное представление согласно (7). Сопоставим документу последовательность векторов-фрагментов:

$$X_i \mapsto \{\mathbf{x}_i^j\}_{j \in 1..m_i}.$$

Пусть $T > 0$, $\mathbb{X} = \{\mathbf{x}_i^j\}_{i \in 1..n, j \in T+1..\bar{m}}$ — множество векторов-фрагментов, для которых $j > T$, $\bar{m} = m_1 + \ldots + m_n$.

По формуле (9) $\forall \mathbf{x}_i^j$ построим динамическую модель:

$$\mathbf{x}_i^j \mapsto s_{\mathbf{x}_i^j}(\cdot).$$

Для строгого теоретического обоснования дальнейших выкладок предположим выполнение следующего условия для $s_{\mathbf{x}}(\cdot)$:

Предположение 1

$$s_{\mathbf{x}}(\mathbf{x}) \leq s_{\mathbf{x}}(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbb{X}.$$

То есть каждый вектор-фрагмент наиболее тесно связан только со своими T предшественниками.

Введем функцию $D : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$

$$D(\mathbf{x}, \mathbf{y}) = s_{\mathbf{x}}(\mathbf{x}) + s_{\mathbf{y}}(\mathbf{y}) - s_{\mathbf{x}}(\mathbf{y}) - s_{\mathbf{y}}(\mathbf{x}). \tag{14}$$

Произведем вложение пространства $(\mathbb{X}, D(\mathbf{x}, \mathbf{y}))$ в пространство R^m, где $m = m_1 + \ldots + m_n - nT = |\mathbb{X}|$:

$$F : (\mathbb{X}, D(\mathbf{x}, \mathbf{y})) \to (\mathbb{R}^m, \|\cdot\|)$$

по следующему правилу: каждому вектору-фрагменту $\mathbf{x}_i^j$ сопоставим вектор $F \in R^m$ по следующему правилу:

$$F(\mathbf{x}_i^j) = \begin{pmatrix} D(\mathbf{x}_i^j, \mathbf{x}_1^{T+1}) \\ D(\mathbf{x}_i^j, \mathbf{x}_1^{T+2}) \\ \ldots \\ 0 \\ \ldots \\ D(\mathbf{x}_i^j, \mathbf{x}_n^{m_{n-1}}) \\ D(\mathbf{x}_i^j, \mathbf{x}_n^{m_n}) \end{pmatrix}. \tag{15}$$

Таким образом, $\forall j > T, i \in 1..n$ координаты вектора $F(\mathbf{x}_i^j)$ соответствуют расстояниям от вектора-фрагмента $\mathbf{x}_i^j$ до всех векторов-фрагментов из множества $\mathbb{X}$.

Рассмотрим пример вложения. Пусть $\mathbb{X} = \{\mathbf{x}_1^{t_1}, \mathbf{x}_2^{t_1}, \mathbf{x}_3^{t_1}\}$ и

- $D(\mathbf{x}_1^{t_1}, \mathbf{x}_2^{t_1}) = 0.5$,
- $D(\mathbf{x}_1^{t_1}, \mathbf{x}_3^{t_1}) = 1$,
- $D(\mathbf{x}_2^{t_1}, \mathbf{x}_3^{t_1}) = 0.2$.

Тогда соответствующие вектора F равны

$$F(\mathbf{x}_1^{t_1}) = \begin{pmatrix} 0 \\ 0.5 \\ 1 \end{pmatrix}, F(\mathbf{x}_2^{t_1}) = \begin{pmatrix} 0.5 \\ 0 \\ 0.2 \end{pmatrix}, F(\mathbf{x}_3^{t_1}) = \begin{pmatrix} 1 \\ 0.2 \\ 0 \end{pmatrix}.$$

Обозначим $\mathcal{F} = \{F(\mathbf{x}_i^j\}_{\mathbf{x}_i^j \in \mathbb{X}}$. Будем кластеризовать элементы множества $\mathcal{F}$ с помощью алгоритма кластеризации Cl, минимизирующего функционал (5).

Описанную процедуру можно сформулировать в виде следующего алгоритма.

Алгоритм 2

- $\mathcal{X}$ — коллекция текстов
- T — параметр задержки
- k — число групп

1. Построить $\mathbb{X} = \{\mathbf{x}_i^j\}_{j=T+1}^m$.
2. Для каждого $\mathbf{x}$ построить динамическую модель $s_{\mathbf{x}}$ по (9).
3. Вычислить $F(\mathbf{x})$ для каждого $\mathbf{x}$ по (15).
4. Разделить множество $\mathcal{F}$ на k кластеров с помощью алгоритма кластеризации Cl.

Пусть в результате работы Алгоритма 2 вектора $F(\mathbf{x})$ разделились на k кластеров $L_1, \ldots, L_k$. Тогда в пространстве фрагментов можно определить следующее правило классификации, относящее фрагмент к одному из классов $l_1, \ldots, l_k$:

Правило классификации 2

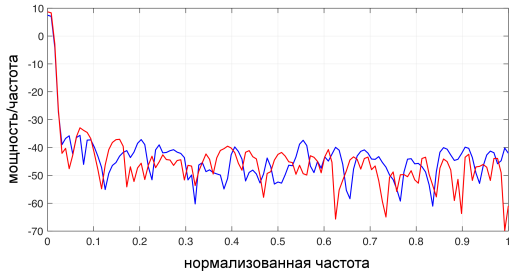
Два фрагмента $\mathbf{x}_i$ и $\mathbf{x}_j$ относятся к одному классу l_k, если соответствующие им вектора $F(\mathbf{x}_i)$ и $F(\mathbf{x}_j)$ попали в один кластер k.

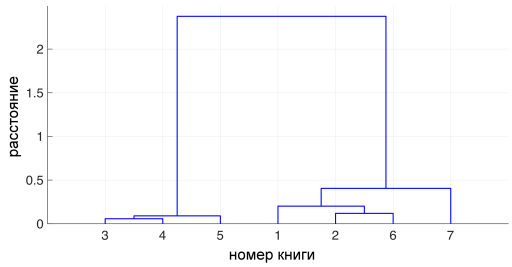
Теорема 2. *Если $r(\mathbf{x}, \mathbf{y})$ — положительно определенное ядра и выполнено Предположение 1, то кластеризация в пространстве $\mathcal{F}$ обеспечивает однозначность и корректность Правила классификации 2.*

В **третьей главе** "Экспериментальные результаты" представлены результаты применения предложенных алгоритмов кластеризации к задаче определения авторского стиля текстов нескольких серий популярных книг.

В п. 3.1 дается определение задачи определения авторства, описываются основные алгоритмы решения этой задачи.

В п. 3.2 приводится результат применения алгоритма классификации текстов на основе кластеризации с помощью спектрального представления к задаче определения авторского стиля в трех коллекциях книг. Ниже представлены примеры графика периодограмм для двух книг А. Азимова из цикла "Основание" и результат иерархической кластеризации всех книг из цикла (7 романов).

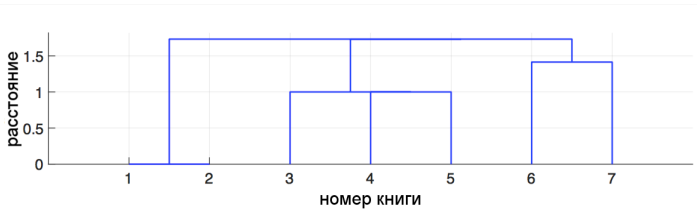




В п. 3.3 приводится результат применения алгоритма классификации текстов на основе кластеризации с помощью расстояния, основанного на ядрах. Ниже в виде таблицы представлен результат сравнения стилей книг А. Азимова из цикла “Основание”, полученный с помощью расстояния $d_{Spearman}$. Здесь ‘1’ обозначает, что для соответствующей пары книг найдено различие в стилях:

	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$
$F1$	0	0	1	1	1	1	1
$F2$	0	0	1	1	1	1	1
$F3$	1	1	0	1	0	1	1
$F4$	1	1	1	0	0	1	1
$F5$	1	1	0	0	0	1	1
$F6$	1	1	1	1	1	0	1
$F7$	1	1	1	1	1	1	0

Результатом иерархической кластеризации книг из цикла является следующая дендрограмма:



Эксперименты показали, что рассмотренные в диссертации новые динамические модели текстов действительно оказались уникальными характеристиками авторского стиля.

В **заключении** формулируются основные результаты диссертации.

Работы автора по теме диссертации

Статьи в периодических рецензируемых изданиях, индексируемых в наукометрических базах данных SCOPUS и Web of Science или включенных в перечень научных журналов, рекомендованных ВАК:

[1] ***Amelin K., Granichin O., Kizhaeva N., Volkovich Z.*** **Patterning of writing style evolution by means of dynamic similarity** // Pattern Recognition, 2017, https://doi.org/10.1016/j.patcog.2017.12.011

[2] ***Granichin O., Kizhaeva N., Shalymov D., Volkovich Z.*** **Writing style determination using the KNN text model** // Proceedings of the 2015 IEEE International Symposium on Intelligent Control. — Sydney, Australia, 2015. — September 21–23. — P. 900–905.

[3] ***Kizhaeva N., Volkovich Z., Granichin O., Granichina O., Kiyaev V.*** **Spectral profiling of writing process** // Proceedings of the 2017 IEEE Conference on Control Technology and Applications. — Coast, Hawaii, USA, 2017. — August 27–30. — P. 2063–2068.

[4] ***Кижаева Н.А., Шалымов Д.С.*** **Определение авторского стиля текстов на основе статистического подхода двухвыборочного тестирования и метода K-ближайших соседей** // Компьютерные инструменты в образовании, 2015. — №5. — С.14–23.

Другие научные публикации:

[5] *Kizhaeva N., Shalymov D., Granichin O., Volkovich Z.* Studying of KNN two-sample test approach applications for writing style comparison of English and Russian text collections // Proceedings of the AINL-ISMW FRUCT (Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search). — ITMO University, FRUCT Oy, Finland. — Saint-Petersburg, Russia, 2015. — November 9–14. — P. 163–166.

[6] *Кижаева Н.А.* Тематическое моделирование и кластеризация текстов на арабском языке // Стохастическая оптимизация в информатике, 2013. — Т. 9, — №2. — С. 33–40

[7] *Кижаева Н.А.* Динамическая модель процесса эволюции текстовых документов // Стохастическая оптимизация в информатике, 2018. — Т. 14. — №1. — С. 31–45.