МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет» (национальный исследовательский университет)
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

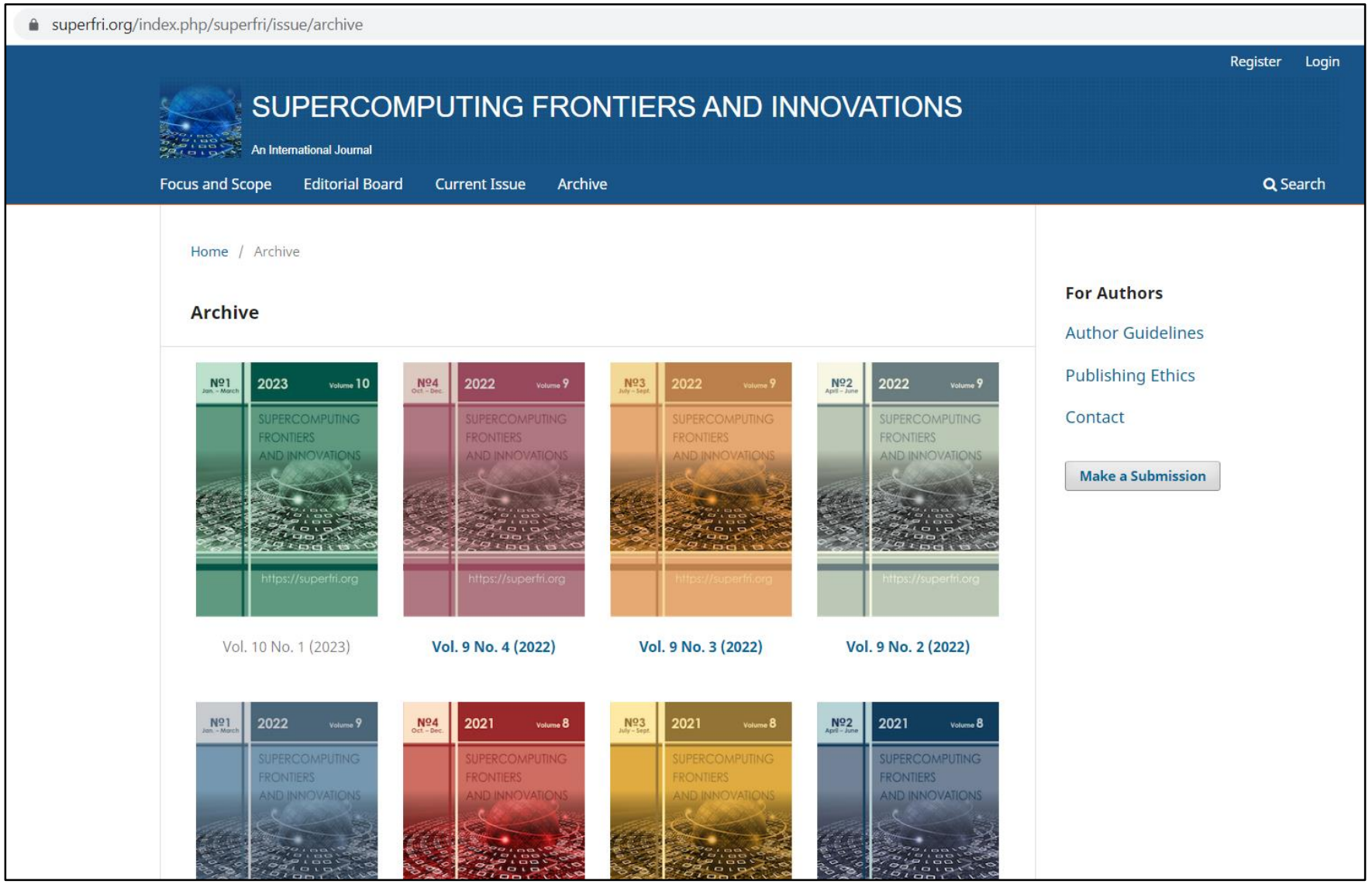
РАЗРАБОТКА АЛГОРИТМА ДЛЯ АВТОМАТИЗИРОВАННОЙ РУБРИКАЦИИ СТАТЕЙ НАУЧНОГО ЖУРНАЛА

Научный руководитель:
Доктор физико-математических наук,
профессор кафедры СП
М.Л. Цымблер

Автор работы:
студент группы КЭ-401
Ю.В. Циммерман

Челябинск, 2023 г.

Актуальность



Актуальность

SUPERCOMPUTING FRONTIERS AND INNOVATIONS
An International Journal

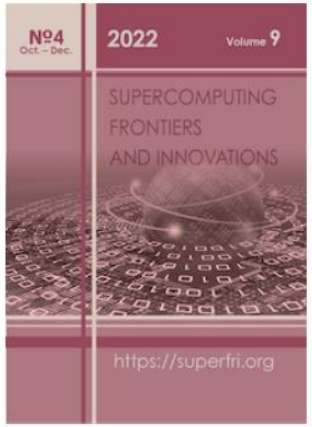
Functional Programming Libraries for Graphics Accelerators

Mikhail M. Krasnov
Keldysh Institute of Applied Mathematics, Moscow, Russian Federation
https://orcid.org/0000-0001-7988-6323

Olga B. Feodoritova
Keldysh Institute of Applied Mathematics, Moscow, Russian Federation
https://orcid.org/0000-0002-2792-9376

DOI: https://doi.org/10.14529/jsfi220403

Keywords: C , functional programming library, CUDA, OpenMP, OpenCL, OpenACC



PDF

Published
2022-12-30

For Authors
Author Guidelines
Publishing Ethics
Contact
Make a Submission

Abstract

Modern graphics accelerators (GPUs) can significantly speed up the execution of numerical tasks. However, porting programs to graphics accelerators is not an easy task, sometimes requiring their almost complete rewriting. CUDA graphics accelerators, thanks to the technology

Пример рубрикации

Scientific Topic

- [ ] Biodiversity and Conservation (82)
- [ ] Biology (60)
- [ ] Chemistry (12)
- [ ] Energy and Climate (56)
- [ ] Food and Agriculture (24)
- [ ] Health and Medicine (75)
- [ ] Paleoscience (14)
- [ ] Physical Science (15)
- [ ] Pollution (30)
- [ ] Social Science (49)
- [ ] Technology (17)
- [ ] Water Resources (47)

Цель и задачи исследования

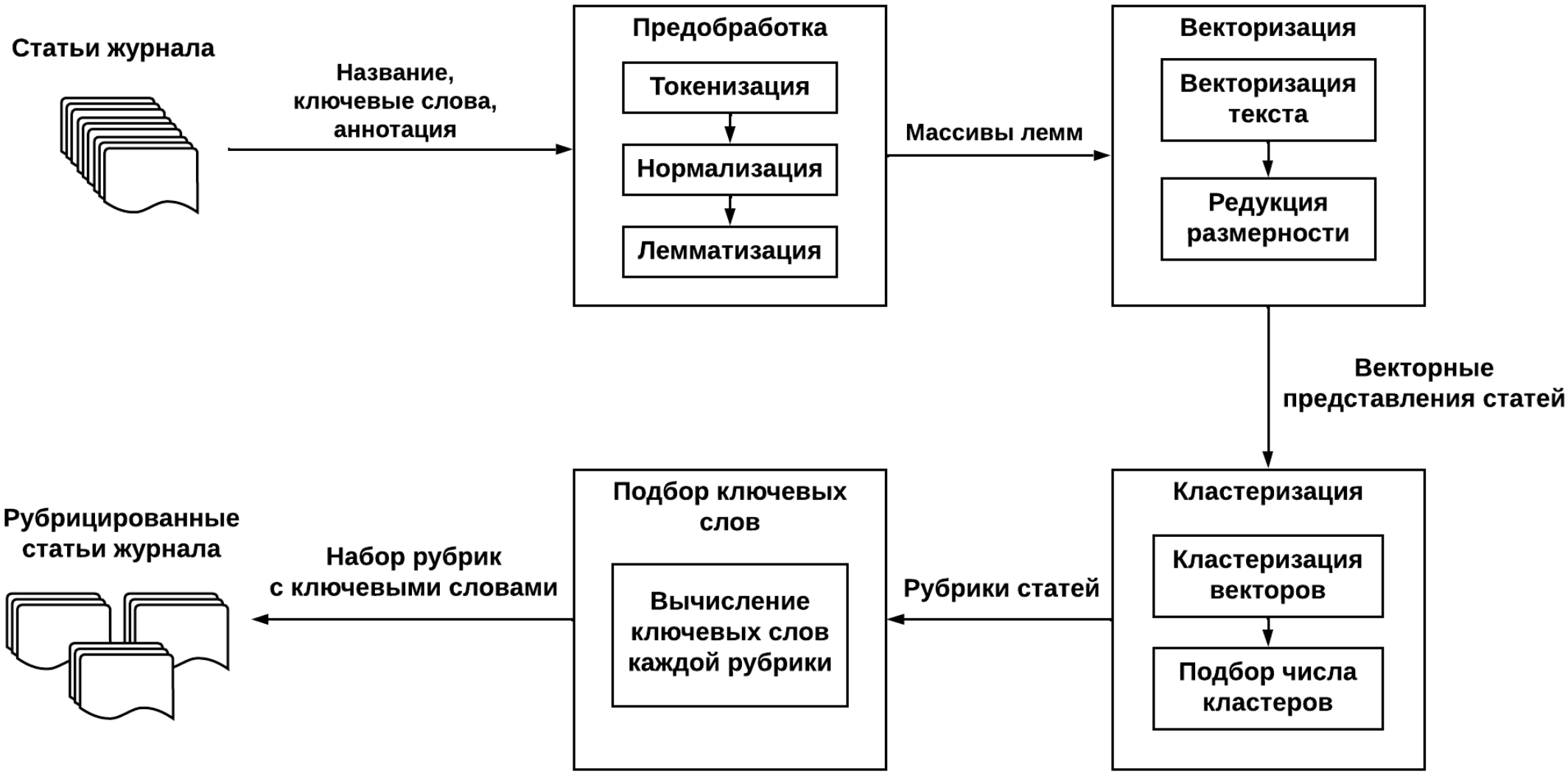
Цель:

Разработка алгоритма, выполняющего автоматизированную рубрикацию статей научного журнала

Задачи:

1. Провести анализ предметной области и обзор существующих решений
2. Выполнить разработку алгоритма рубрикации научных статей
3. Разработать приложение для администратора журнала, позволяющее выполнять подготовку, рубрикацию и визуализацию данных
4. Разработать тестовые наборы и провести тестирование алгоритма

Алгоритм рубрикации



Предобработка метаданных статьи

Исходный текст:

A Review of Supercomputer Performance Monitoring Systems
monitoring, supercomputers, performance monitoring, review
High Performance Computing is now one of the emerging fields in computer science and its applications.

Нормализованный текст:

review supercomputer performance monitoring systems
monitoring supercomputers performance monitoring review
high performance computing emerging fields
computer science applications

Лемматизованный текст, очищенный от дубликатов:

review supercomputer performance monitor system
high compute emerge field
science application

– шумовые слова, подлежащие удалению

– суффиксы, подлежащие удалению/изменению

Векторизация текстовых данных

Леммы
application
compute
emerge
…
system

1	2	3	…	50
0.37	-0.22	0.12	…	0.09
0.63	-0.38	1.15	…	0.38
1.12	-1.61	3.67	…	-3.6
0.33	-0.23	0.77	…	0.11

Корпус слов «glove-wiki-gigaword-50»

	Слово	1	2	3	…	50
1	ability	1.35	-2.41	5.84	…	2.03
	…					
…	apple	0.52	-0.83	0.49	…	0.26
	application	0.37	-0.22	-0.06	…	0.09
	…					
400 000	zygote	0.78	-0.49	0.03	…	-0.41

Редукция размерности векторов

Леммы статьи 1	**1**	**2**	**3**	**…**	**50**
application	0.37	-0.22	0.12	…	0.09
compute	0.63	-0.38	1.15	…	0.38
emerge	1.12	-1.61	3.67	…	-3.6
…					
system	0.33	-0.23	0.77	…	0.11

Леммы статьи N	**1**	**2**	**3**	**…**	**50**
compute	0.63	-0.38	1.15	…	0.38
cluster	1.16	0.17	0.82	…	0.24
…					
network	0.94	0.32	1.33	…	0.58

μ_1	μ_2	μ_3	…	μ_{50}
0.62	-0.22	1.23	…	0.29

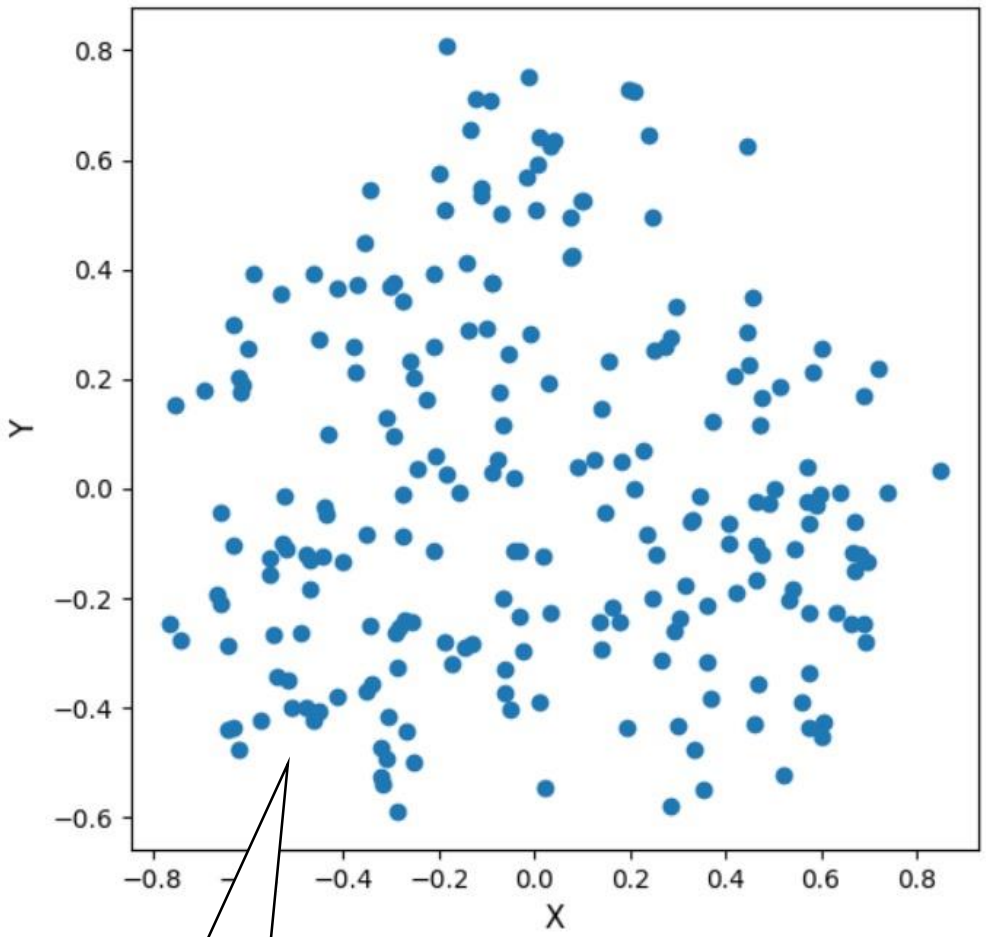
μ_1	μ_2	μ_3	…	μ_{50}
0.88	0.15	0.97	…	0.36

Метод главных компонент

X	**Y**
0.31	-0.14

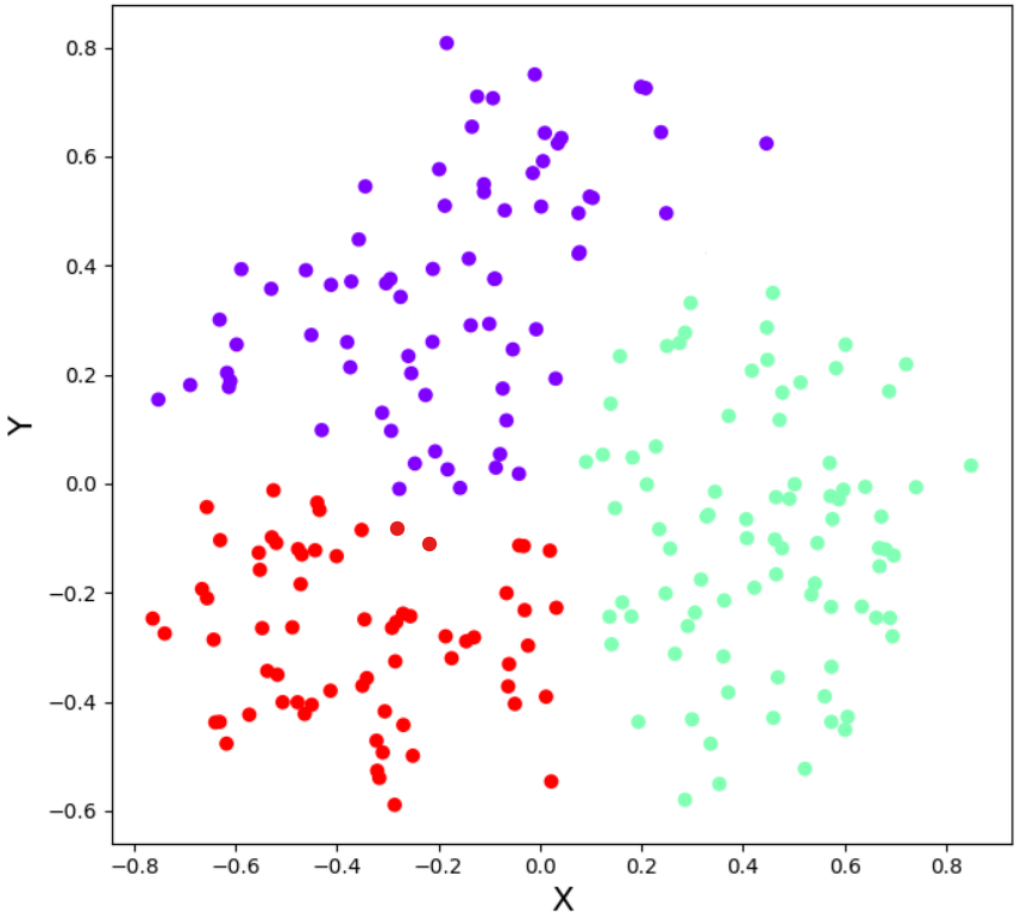
X	**Y**
0.71	0.42

Кластеризация статей



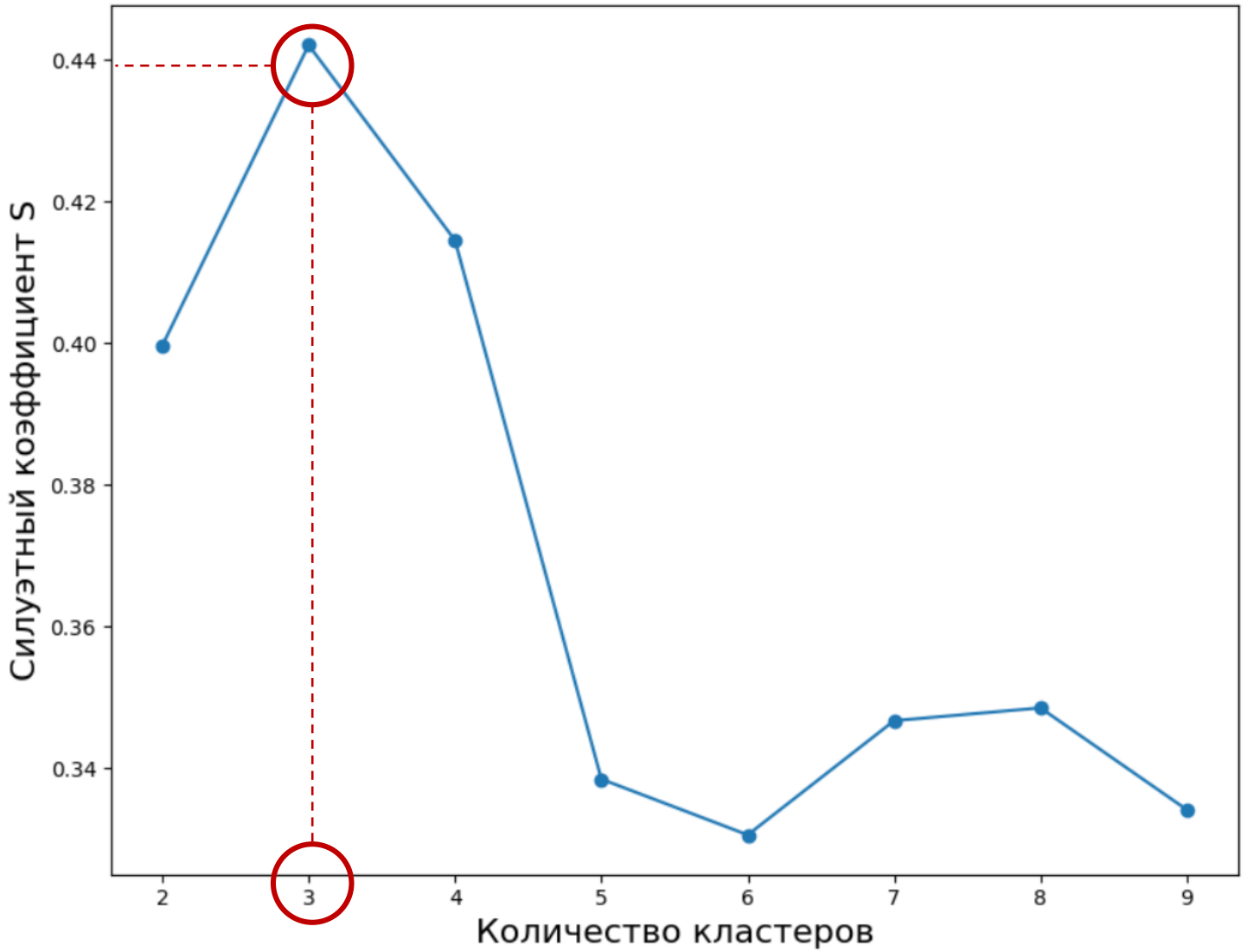
Статья журнала

Агломеративная кластеризация



– рубрика №1
– рубрика №2
– рубрика №3

Подбор числа рубрик



$$S = \frac{b - a}{\max(a, b)}$$

a – среднее расстояние между точками внутри кластера

b – среднее расстояние от a до точек ближайшего кластера

Подбор ключевых слов

№ рубрики	Слово	Вес TF-IDF
1	data	0.09
	hpc	0.88
	…	
	supercomputer	0.68
2	algorithm	0.15
	graphics	0.76
	…	
	visualization	0.82
…		
N	algorithm	0.15
	analysis	0.23
	…	
	tubulin	0.74

$$weight = TF * IDF$$

TF – число появлений слова в рубрике

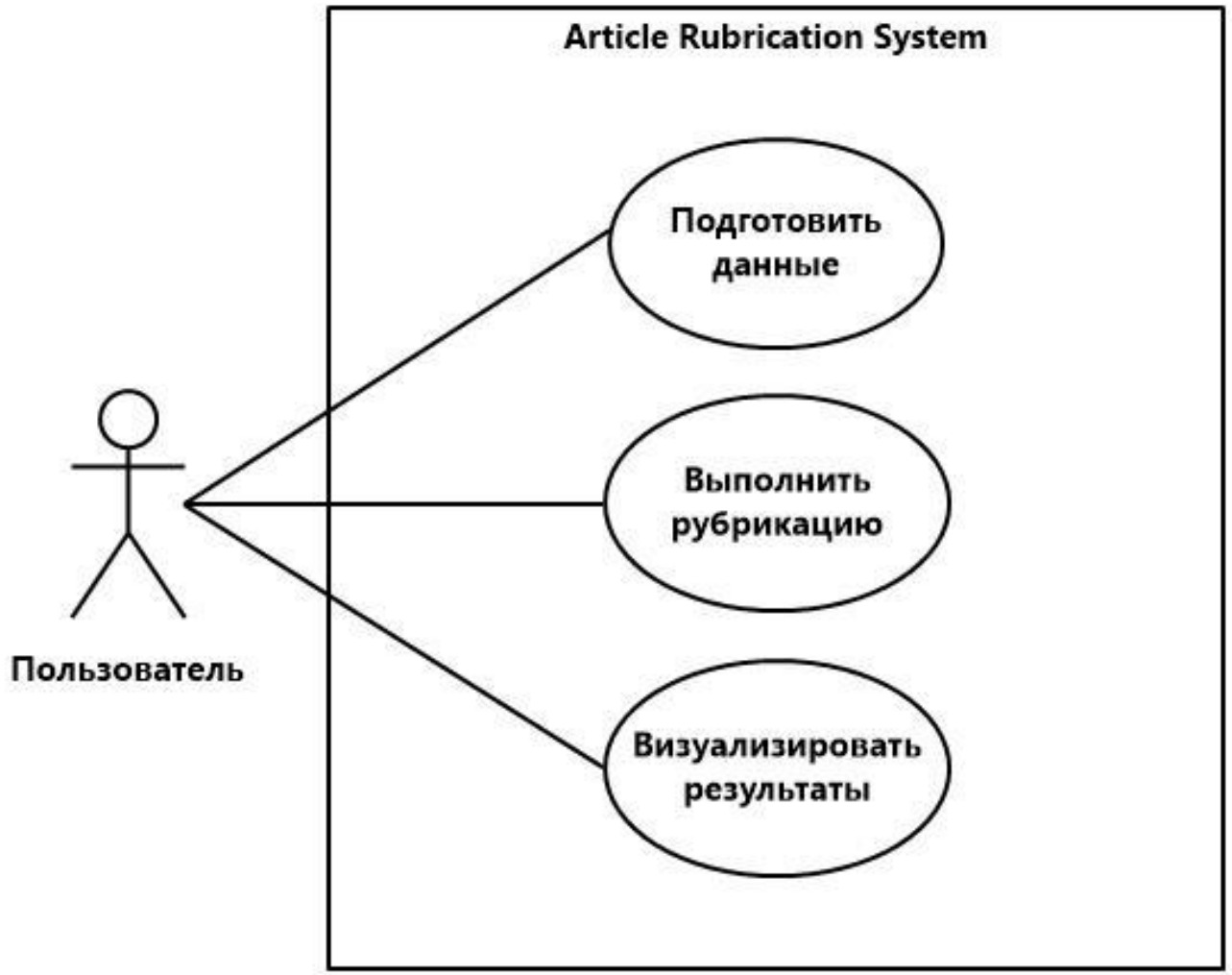
IDF – обратная частота рубрики

$$IDF = log \frac{N}{DF}$$

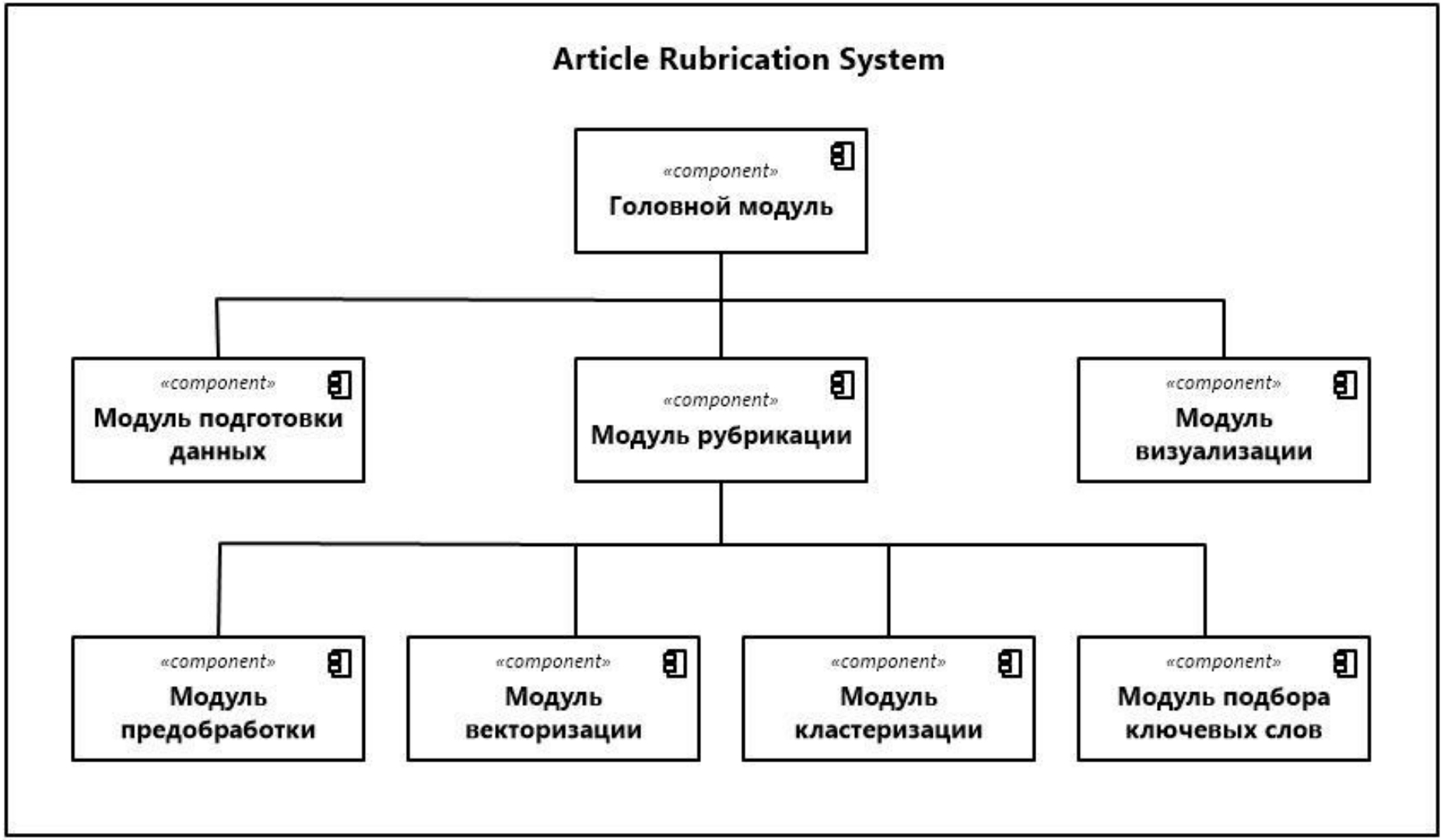
N – общее число рубрик

DF – число рубрик, содержащих слово

Диаграмма вариантов использования



Архитектура приложения



Средства разработки

Среда разработки: JetBrains PyCharm Community 2020.3.3

Языки программирования: Python 3.9.1, JavaScript 1.8.5

Верстка: HTML5, CSS3

Основные библиотеки:

- scikit-learn
- nltk
- plotly
- beautifulsoup4
- PyPDF2
- tkinter
- pandas

Репозиторий: https://github.com/ytsmm/Article-Rubrication-System

Визуализация рубрик

Visualization

Keywords: Perfomance analysis, workflow, Virtualization, graph algorithms, hybrid computing, hardware accelerators, runtime system, memory wall, in-situ visualization, Perfomance evaluation
Quantity of articles: 76

Molecular dynamics

Keywords: parallel algorithms, quantum chemistry, ultrasound tomography, tubulin, CUDA, GPU, floating point, unum computing, computer arithmetic, valid arithmetic
Quantity of articles: 84

Machine learning

Keywords: resilience, HPC benchmarks, drug discovery, deep learning, ultrascale computing, supercomputer, compression, data reduction, climate data, load balancing
Quantity of articles: 65

Visualization

Many-Core Approaches to Combinatorial Problems: case of the Langford Problem

Michael Krajecki, Julien Loiseau, François Alin, Christophe Jaillet

DOI: https://doi.org/10.14529/jsfi160202

Hybrid CPU + Xeon Phi implementation of the Particle-in-Cell method for plasma simulation

Iosif B. Meyerov, Sergey I. Bastrakov, Igor A. Surmin, Alexey V. Bashinov, Evgeny S. Efimenko, Artem V. Korzhimanov, Alexander A. Muraviev, Arkady A. Gonoskov

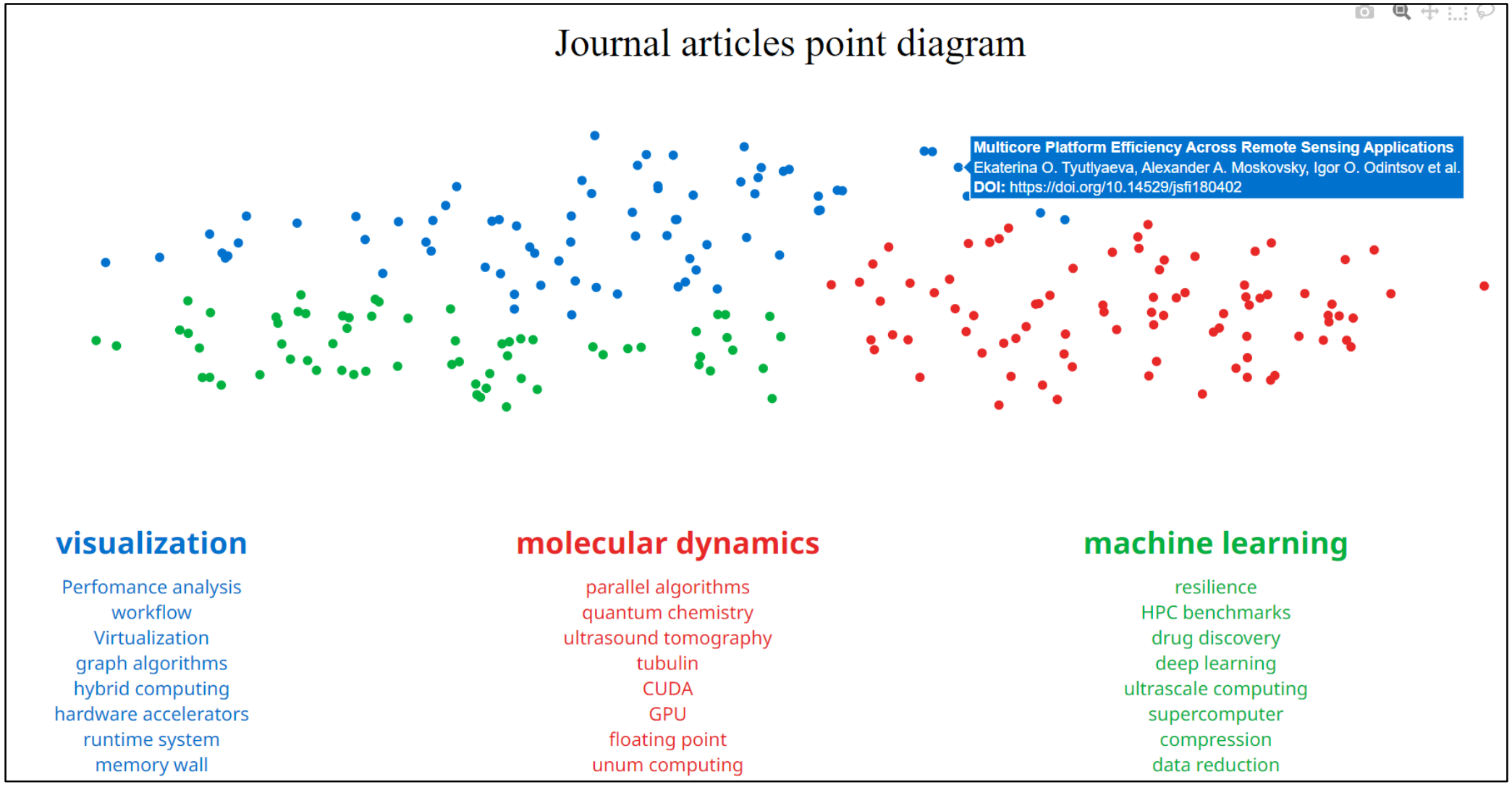
DOI: https://doi.org/10.14529/jsfi160301

Easy Access to HPC Resources through the Application GUI

Matthijs van Waveren, Ahmed Seif El Nawasany, Nasr Hassanein, David Moon, Niall O'Byrnes, Alain Clo, Karthikeyan Murugan, Antonio Arena

DOI: https://doi.org/10.14529/jsfi160302

Визуализация рубрик в виде диаграммы



Поиск по ключевому слову

visualization	molecular dynamics	machine learning
Perfomance analysis	parallel algorithms	resilience
workflow	quantum chemistry	HPC benchmarks
Virtualization	ultrasound tomography	drug discovery
graph algorithms	tubulin	deep learning
hybrid computing	CUDA	ultrascale computing
hardware accelerators	GPU	supercomputer
runtime system	floating point	compression
memory wall	unum computing	data reduction
in-situ visualization	computer arithmetic	climate data
Perfomance evaluation	valid arithmetic	load balancing

Filter by GPU:

Spectral Domain Decomposition Using Local Fourier Basis: Application to Ultrasound Simulation on a Cluster of GPUs
Jiri Jaros, Filip Vaverka, Bradley E. Treeby
https://doi.org/10.14529/jsfi160305

On the Inversion of Multiple Matrices on GPU in Batched Mode
Nikolay M. Evstigneev, Oleg I. Ryabkov, Eugene A. Tsatsorin
https://doi.org/10.14529/jsfi180203

Parallel GPU-based Implementation of One-Way Wave Equation Migration
Alexander L. Pleshkevich, Vadim V. Lisitsa, Dmitry M. Vishnevsky, Vadim D. Levchenko, Boris M. Moroz
https://doi.org/10.14529/jsfi180304

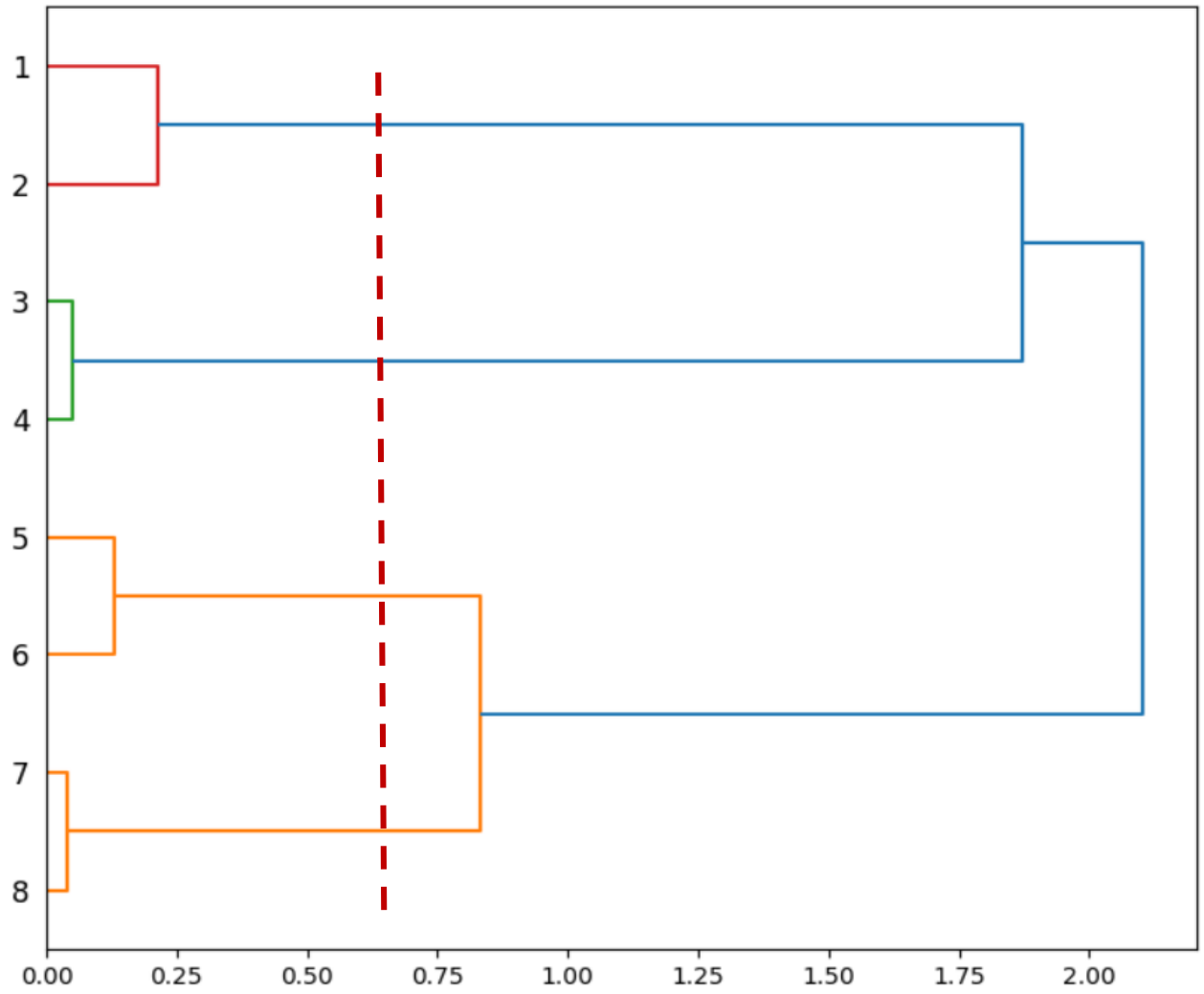
Тестирование

В рамках тестирования были выполнены:

1. Функциональное тестирование разработанного приложения
2. Вычислительные эксперименты на тестовых наборах данных

Вычислительные эксперименты

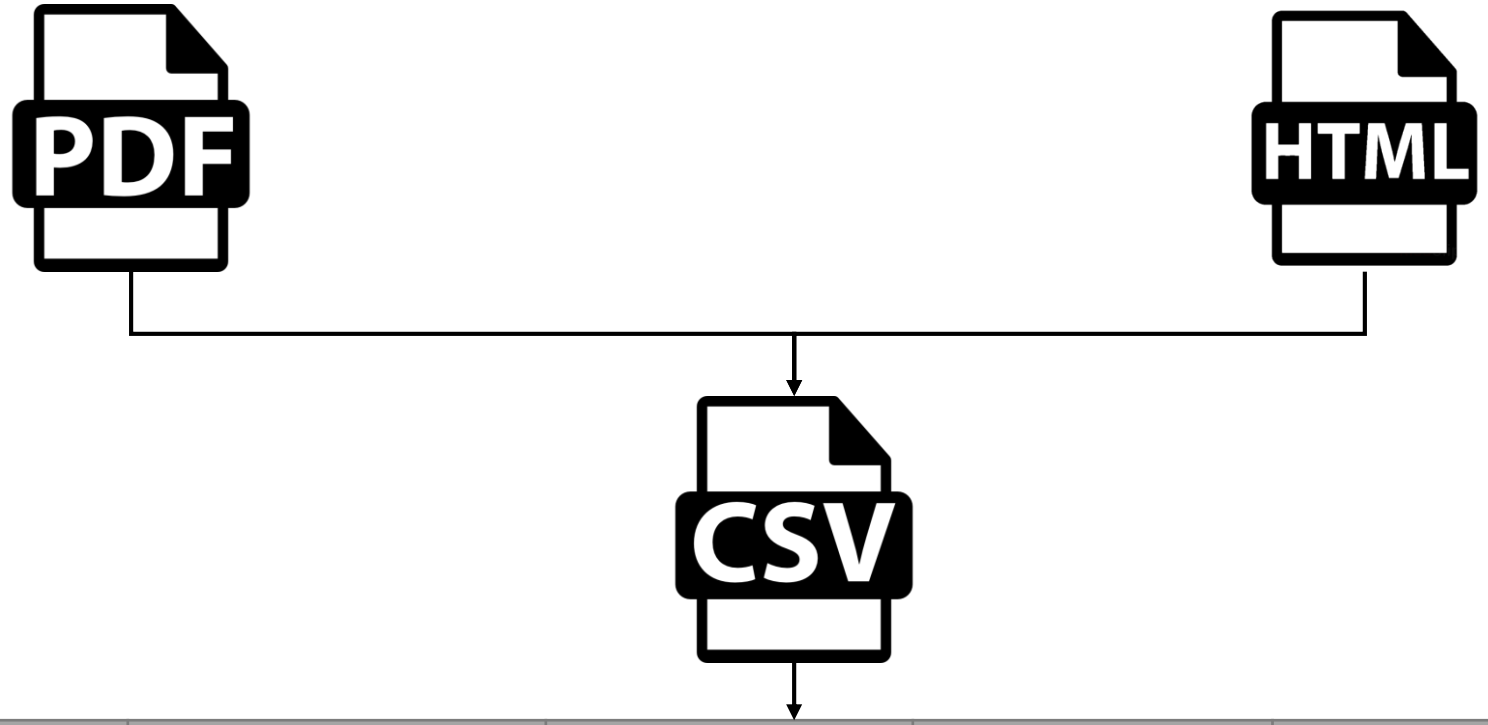
№	Ключевые слова
1	docking, protein-ligand, global minimum, force field, quantum-chemical method
2	docking, protein-ligand, global optimization, tensor train, quantum-chemical method
3	numerical weather prediction, global atmosphere model, computational efficiency, I/O optimization
4	numerical weather prediction, climate modeling, hybrid computing, programming models
5	AlgoWiki, parallel structure, algorithms properties, Top500, problems, methods, implementations, computing platforms
6	scalability, AlgoWiki, parallel structure, problems, methods, algorithms, implementations, computing platforms
7	vector computers, NVIDIA GPUs, graph algorithms, graph framework, VGL, CUDA
8	graph algorithms, NEC SX-ACE, vector computing, data-intensive applications



Основные результаты

1. Выполнен анализ предметной области и проведен обзор существующих решений
2. Выполнена разработка алгоритма рубрикации научных статей
3. Разработано приложение для администратора журнала, позволяющее выполнять подготовку, рубрикацию и визуализацию данных
4. Разработаны тестовые наборы, и проведено тестирование алгоритма

Подготовка данных



Doi	Title	Authors	Keywords	Abstract
https://doi.org/10.14529/jsfi220403	Functional Programming Libraries for Graphics Accelerators	Mikhail M. Krasnov, Olga B. Feodoritova	functional programming library, CUDA, OpenMP, OpenCL, OpenACC	Modern graphics accelerators (GPUs) can significantly speed up the execution of numerical tasks…

Пользовательский интерфейс