Разработка параллельного алгоритма кластеризации текстовых документов FRiS-Tax на основе технологии MPI

М.Е. Мансурова[1], В.Б. Барахнин[2, 3], С.С. Аубакиров[1], Е. Хибатханулы[1], Мусина А.Б.[1]

Казахский национальный университет имени ал-Фараби[1], Институт вычислительных технологий СО РАН[2], Новосибирский государственный университет[3]

Введение

- Большой объем данных
- Кластеризация – инструмент анализа данных
- Процесс разбиения множества документов электронной базы на классы (кластеры), при котором элементы, объединяемые в один класс, имеют большее сходство

Задача

- Кластеризация алгоритмом FRiS-Tax
- Мера близости – мера конкурентного сходства

$$F_{b/c}(a) = \frac{m(a,b) - m(a,c)}{m(a,b) + m(a,c)}$$

- Процесс кластеризации является ресурсоемким, с ростом объема обрабатываемой информации задача еще больше усложняется
- Подбор атрибутов и весовых коэффициентов требует участия человека

Цель

- Выбор атрибутов и настройка весовых коэффициентов при помощи генетического алгоритма (ГА)
- Параллельная реализация на основе Message Passing Interface (MPI)

FRiS-Tax алгоритм

- Разбиение всего множества объектов выборки *A* на линейно разделимые кластеры похожих между собой объектов
- Под похожестью понимается конкурентное сходство с центральным объектом кластера, столпом.
- S - множество столпов
- Сходство атрибутов документа (авторы, аннотация, ключевые слова и т.д.)

Мера сходства

$$m(d_1, d_2) = \sum a_i m_i(d_1, d_2), \tag{1}$$

где i — номер элемента (атрибута) библиографического описания, a_i — весовые коэффициенты, причём $\sum a_i = 1$, $m(d_1, d_2)$ — мера сходства по i-му элементу (иными словами, по i-й шкале).

Мера сходства

$$F^*_{s_{a1}}(a) = \frac{m(a, s_{a1}) - m^*}{m(a, s_{a1}) + m^*}$$

Естественно, что в задаче таксономии множество столпов S заранее не задано. Выбираться оно будет таким образом, чтобы средняя величина конкурентного сходства каждого объекта выборки A с ближайшим к нему столпом из множества S была максимальна:

$$\overline{F}(S) = \sum_{a \in A} F^*_{s_{a1}}(a) \to \max_S . \qquad (2)$$

FRiS-Tax алгоритм

1. Поочередно перебирая все объекты выборки A, выделяем объект a^*, для которого величина $\overline{F}(\{a^*\})$ максимальна, и назначаем его на роль первого столпа s_1.

2. После того как первый столп зафиксирован, на роль второго столпа поочередно назначаются все объекты выборки, не совпадающие с s_1. В качестве второго столпа s_2 выбирается тот объект b^*, который в паре с s_1 обеспечивает максимальное суммарное конкурентное сходство $\overline{F}(\{b^*, s_1\})$.

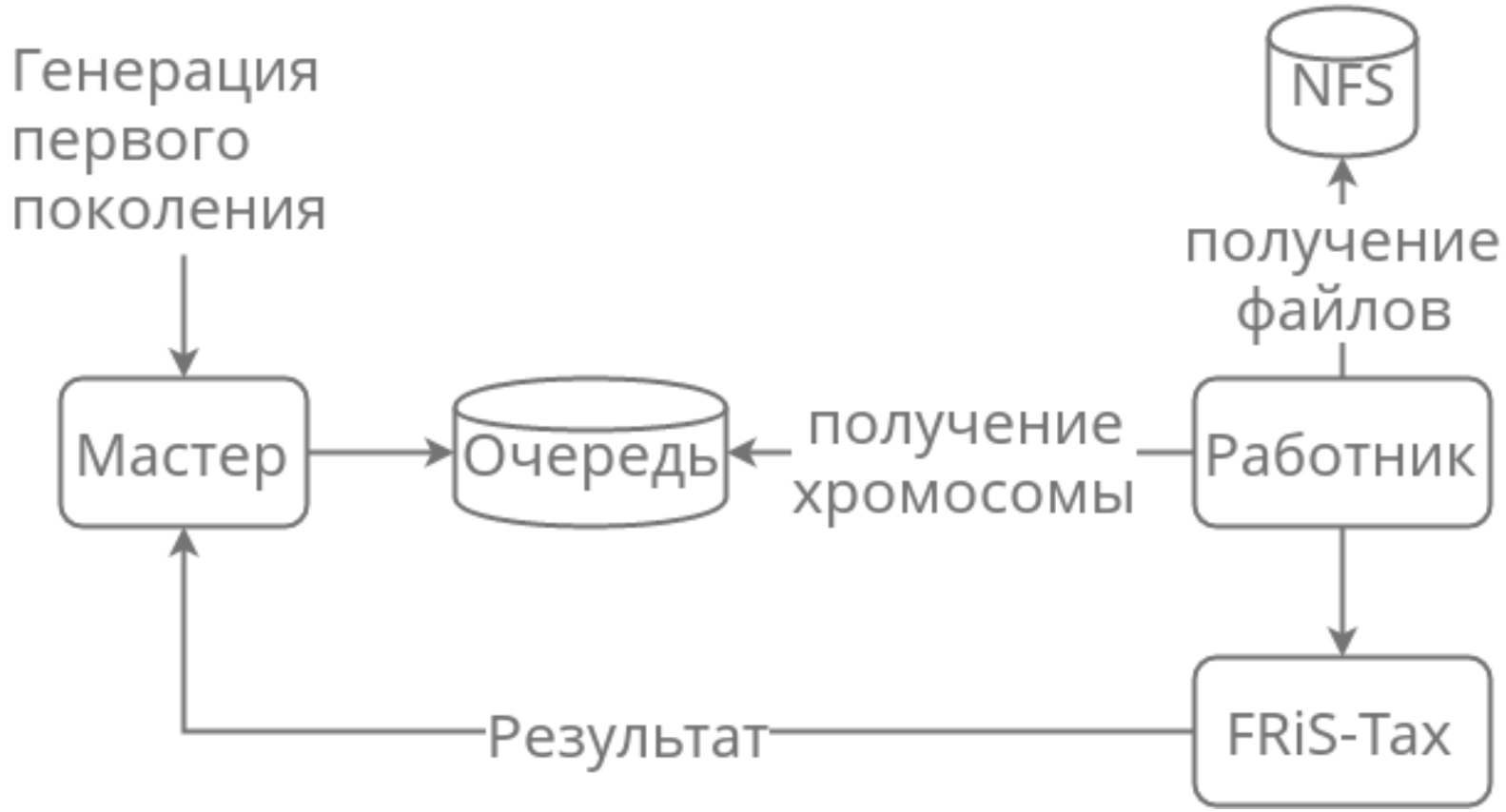
3. Наращивание числа столпов продолжается по этому же принципу. Если первые i столпов $\{s_1, s_2, \ldots, s_i\}$ уже определены, то на роль s_{i+1}-го столпа выбирается объект $z^* \in A/\{s_1, s_2, \ldots, s_i\}$, обеспечивающий максимум функционалу $\overline{F}(\{z^*, s_1, s_2, \ldots, s_i\})$. Процесс продолжается до тех пор, пока не будет набрано заданное число столпов k.

4. После того как было найдено множество столпов $\{s_1, s_2, \ldots, s_k\}$, вся выборка распределяется между ними. Объект относится к тому кластеру, сходство со столпом которого максимально. Объекты, присоединённые к первому столпу s_1, образуют кластер A_1, объекты, ближайшим для которых оказался столп s_2, образуют кластер A_2 и т. д. В результате получаем разбиение множества объектов A на k кластеров A_1, A_2, $\ldots$, A_k.

FRiS-Tax


Инициализация параметров
Парсинг PDF файлов
Нахождение первого столпа
Нахождение следующего столпа
Количество столпов меньше K
Да
Нет
Распределение остальных объектов по кластерам
Обновление столпов
Конец
Нет

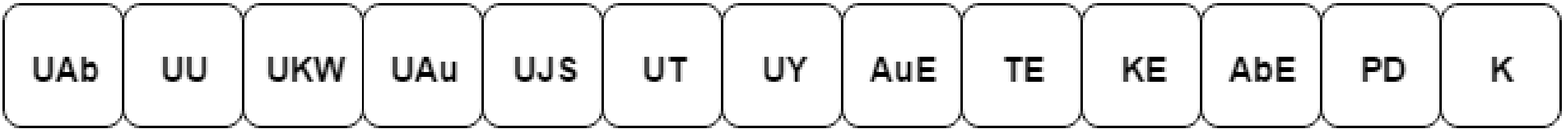
Архитектура



Набор генов

№	Полное название гена	Сокращенное название гена	Возможные значения
1	POSSIBLE_DIFFERENCES	PD	0-3
2	UseAbstract	UAb	0-3
3	UseUdk	UU	0-1
4	UseKeyWords	UKW	0-3
5	UseAuthors	UAu	0-3
6	UseJournalSeria	UJS	0-1
7	UseTitle	UT	0-3
8	UseYear	UY	0-1
9	AuthorEquality	AuE	0-1
10	TitleEquality	TE	0-1
11	KeywordsEquality	KE	0-1
12	AbstractEquality	AbE	0-1
13	K (количество кластеров)	K	2-15

Хромосома

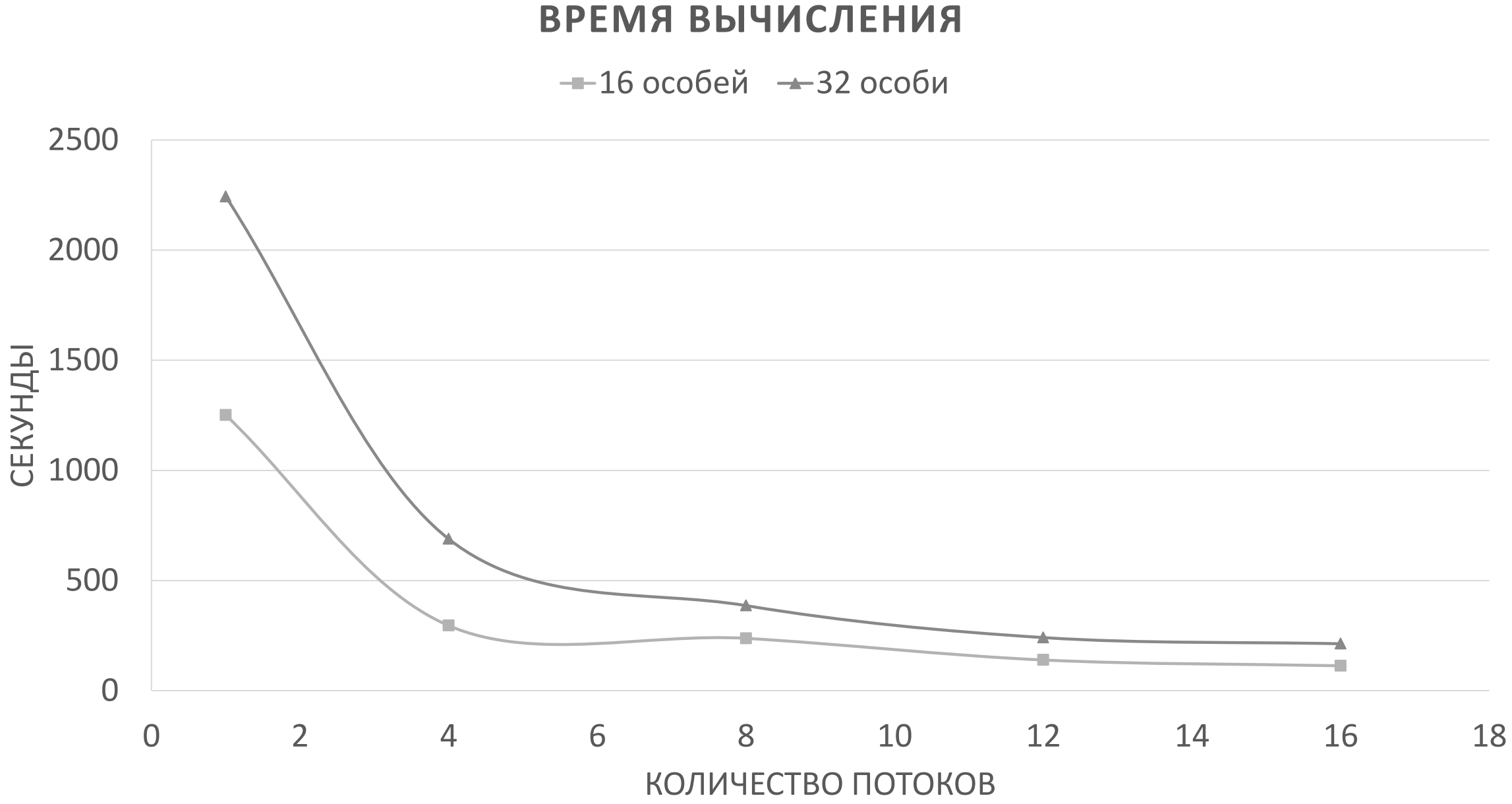


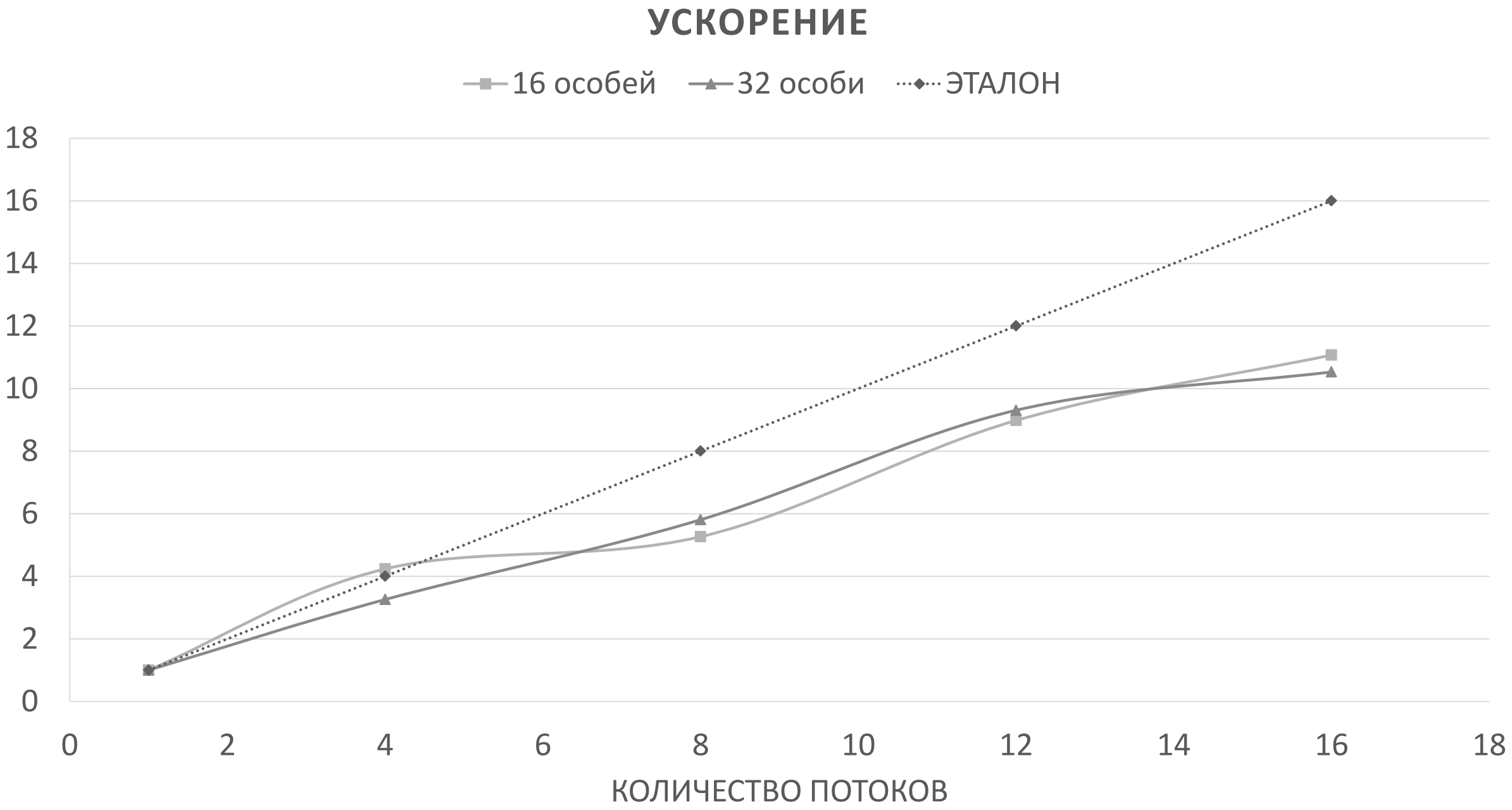
Эксперименты

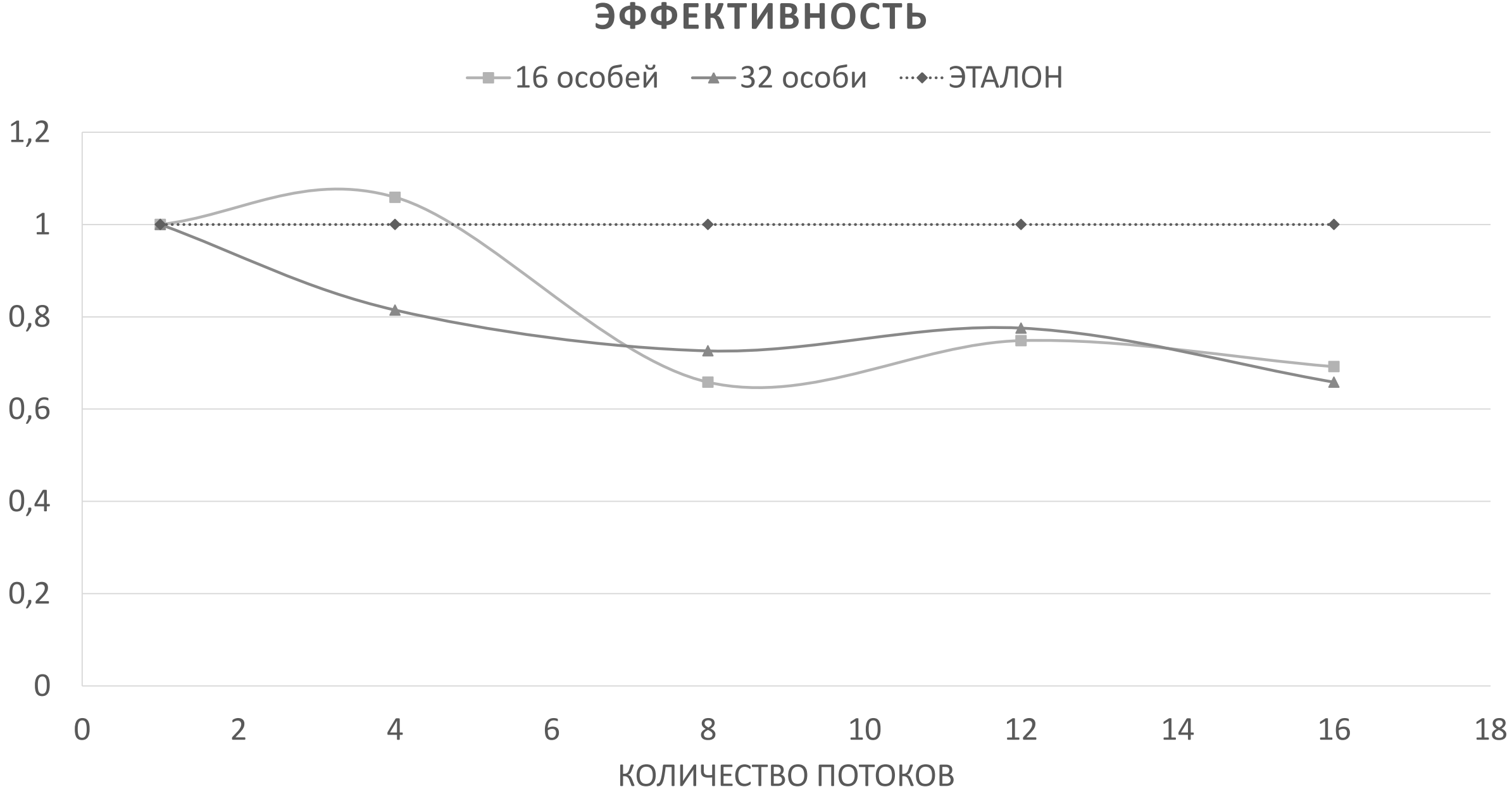
- Лаборатория НИИ ММ при КазНУ имени аль-Фараби
  - 16 машин
- Характеристики машин
  - RAM: 16Gb
  - Архитектура: x86_64
  - CPU(s): Intel Core i5-2500 CPU 3.30GHz
  - Сеть: 1Gbit/s

Корпус

- http://www.mathnet.ru «Сибирский математический журнал»
- Атрибуты
  - год выпуска
  - код Mathematical Subject Classification (MSC)
  - ключевые слова
  - авторы
  - серия
  - аннотация
  - заголовок


ВРЕМЯ ВЫЧИСЛЕНИЯ
16 особей
32 особи
2500
2000
1500
1000
500
0
СЕКУНДЫ
0
2
4
6
8
10
12
14
16
18
КОЛИЧЕСТВО ПОТОКОВ


УСКОРЕНИЕ
16 особей
32 особи
ЭТАЛОН
18
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18
КОЛИЧЕСТВО ПОТОКОВ


ЭФФЕКТИВНОСТЬ
16 особей
32 особи
ЭТАЛОН
1,2
1
0,8
0,6
0,4
0,2
0
0
2
4
6
8
10
12
14
16
18
КОЛИЧЕСТВО ПОТОКОВ

К зафиксирована

- Целевая функция – Purity
- Селекция – турнирная
- Мутация – равномерная (Uniform)
- Скрещивание – равномерное (Uniform)

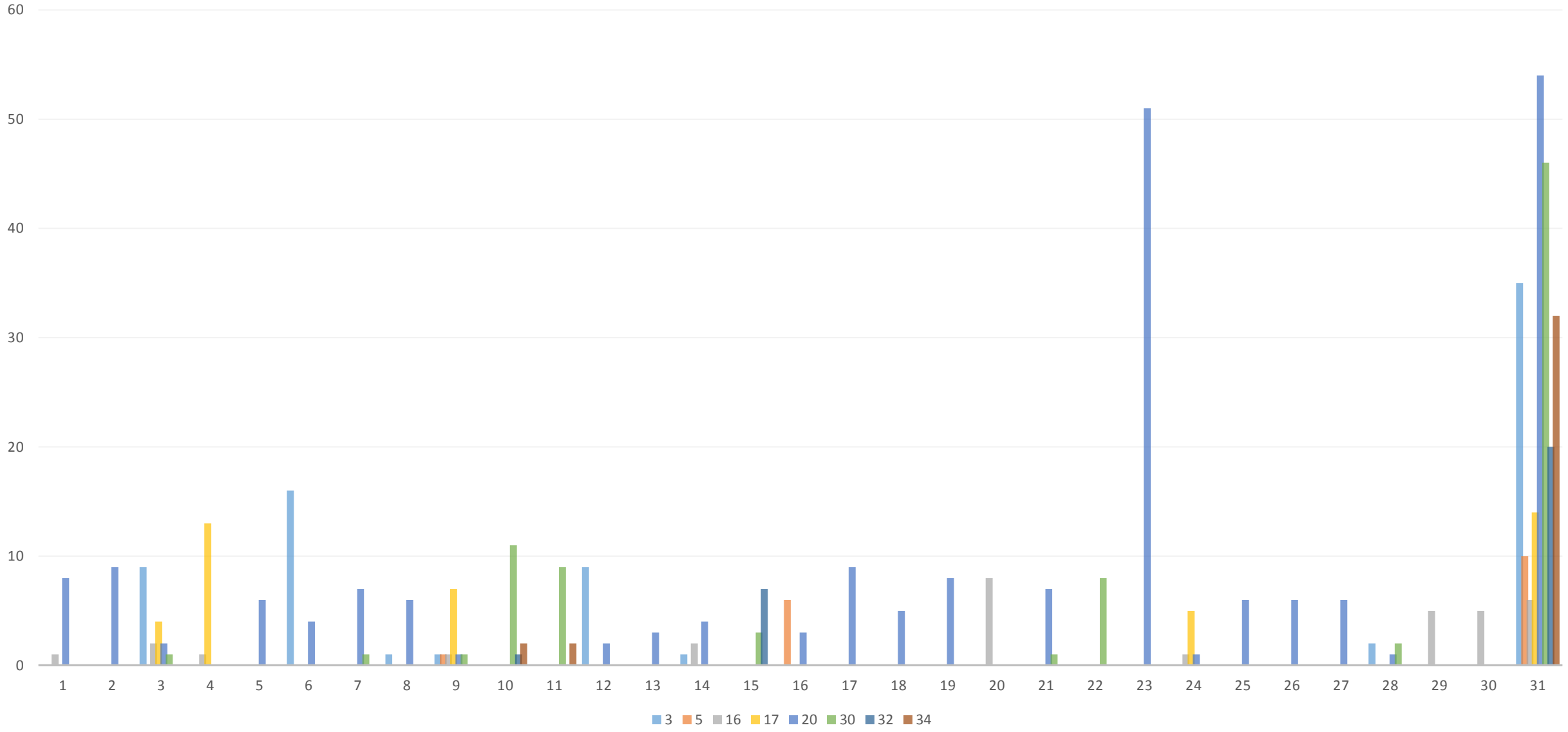
К не зафиксирована

- Целевая функция – среднеквадратическое отклонение
  - уменьшение радиусов кластеров
  - увеличение разброса кластеров
- Селекция – турнирная
- Мутация – равномерная (Uniform)
- Скрещивание – равномерное (Uniform)

Результаты

- К зафиксирована
  - К=8
  - Абстракт, Ключевые слова, Авторы и Заголовок
  - Purity – 83%
- К не зафиксирована
  - К=31
  - Абстракт, Ключевые слова и Заголовок

“К” НЕ ЗАФИКСИРОВАНА



Обсуждение

Как видно из представленных графиков, несмотря на очевидную близость документов внутри узкой тематики, алгоритм успешно разбил выборки на части.

Кластер с номером центроида, равным **31**, включает в себя документы, которые не удалось отнести ни к одной из формируемых групп.

Количество таких документов в зависимости от количества кластеров варьируется в интервале **7.3–12.7 %**, что является приемлемым для работы алгоритмов кластеризации.

Будущая работа

- Подобрать лучшие методы ГА
- Последний кластер
- Кэширование функции m() в общем хранилище
- Добавить семантику, онтологии предметных областей

Вопросы?