ПАРАЛЛЕЛЬНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМА ОБУЧЕНИЯ СИСТЕМЫ ТЕКСТОВОЙ КЛАССИФИКАЦИИ

Е. В. Котельников,

Т. А. Пескишева

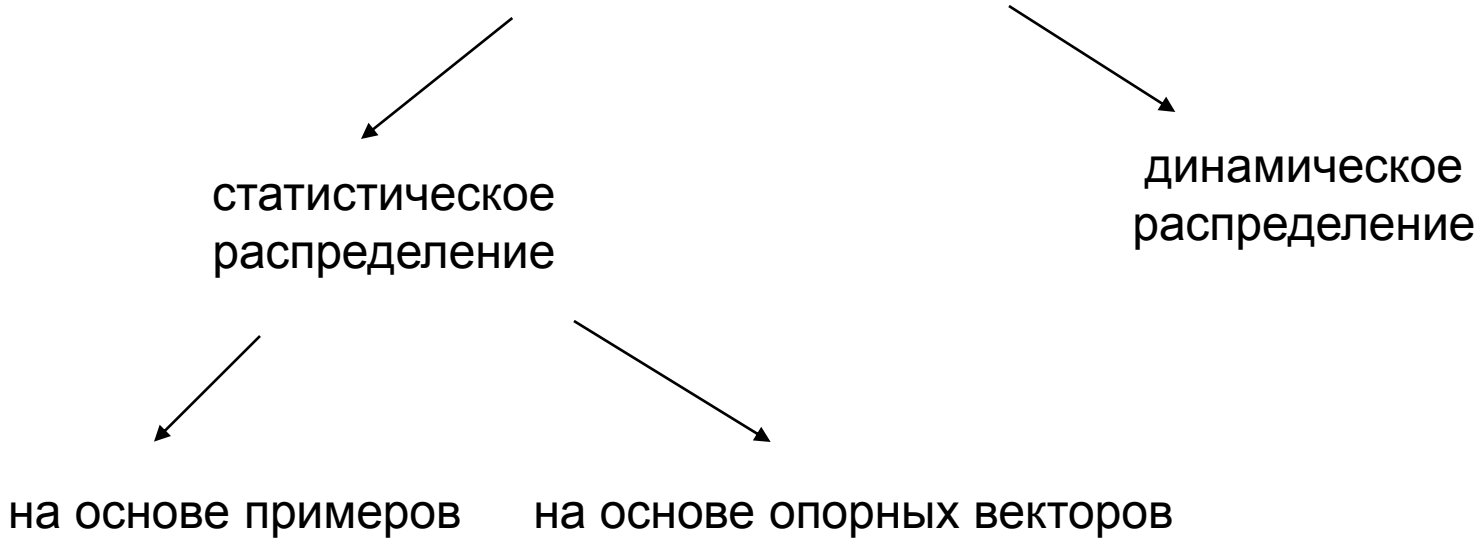
Вятский государственный гуманитарный университет

Введение

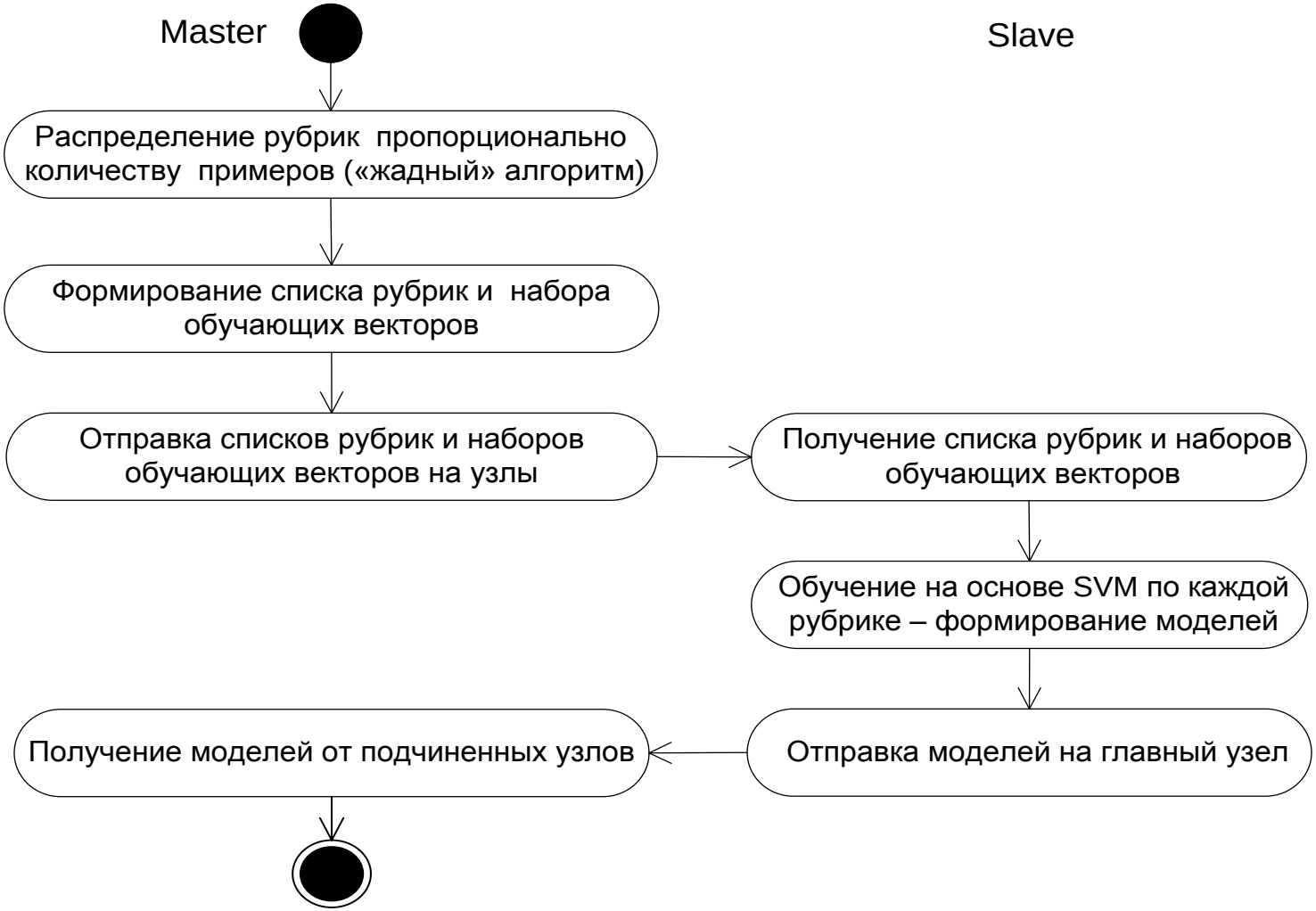
- Текстовая классификация (рубрикация)
- Метод опорных векторов (Support Vector Machines, SVM)
- Применение SVM сводится к двум основным процессам:
  - обучение
  - классификация
- Наиболее трудоемким является процесс обучения

Методы многоклассовой классификации в SVM

Подходы к распределению нагрузки



Статическое распределение на основе примеров



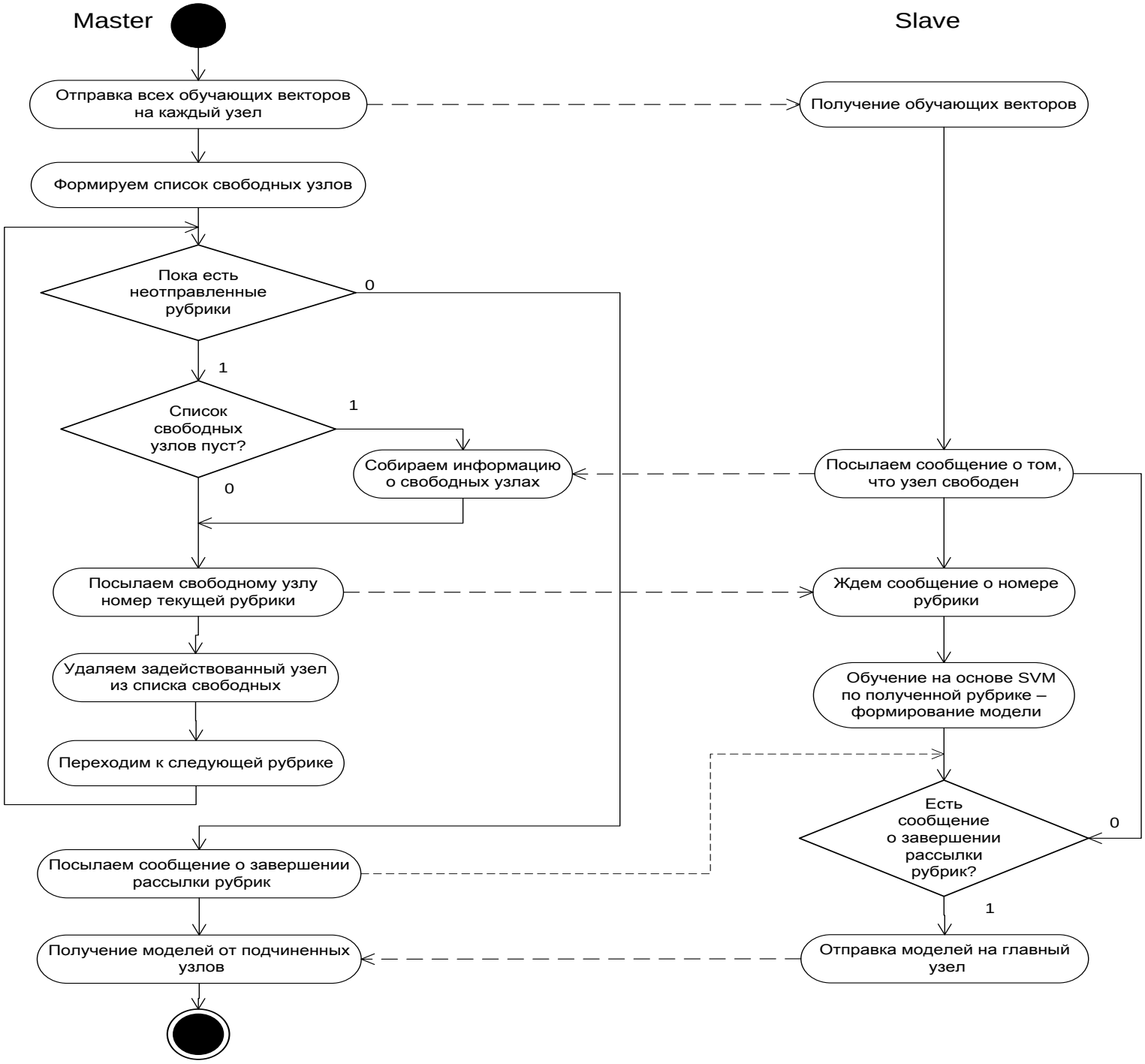
Статическое распределение на основе опорных векторов

- Классы сортируются не по количеству примеров, а по количеству опорных векторов.

- В результате получается более сбалансированная нагрузка на узлы и уменьшается время простоя узлов.

Динамическое распределение

- Узлы загружаются по мере их освобождения
- Не используется априорная информация о количестве обучающих примеров или опорных векторов
- Увеличивается количество обменов информацией между главным и подчиненными узлами кластера
- Увеличивается время ожидания узлов по сравнению со статическими методами
- Эффективность алгоритма в некоторой степени уменьшается


Master
Slave
Отправка всех обучающих векторов на каждый узел
Получение обучающих векторов
Формируем список свободных узлов
Пока есть неотправленные рубрики
0
1
Список свободных узлов пуст?
1
0
Собираем информацию о свободных узлах
Посылаем сообщение о том, что узел свободен
Посылаем свободному узлу номер текущей рубрики
Ждем сообщение о номере рубрики
Удаляем задействованный узел из списка свободных
Обучение на основе SVM по полученной рубрике – формирование модели
Переходим к следующей рубрике
Есть сообщение о завершении рассылки рубрик?
0
1
Посылаем сообщение о завершении рассылки рубрик
Получение моделей от подчиненных узлов
Отправка моделей на главный узел

Программная реализация

- Для кластерной архитектуры параллельная реализация проводилась с использованием библиотеки MPI.NET версии 1.0
- Расчеты проводились на вычислительном кластере ВятГГУ, состоящем из 30 вычислительных узлов.
- Каждый вычислительный узел представляет собой ПК с процессором Intel Core 2 Duo 2 ГГц и 2 Гб оперативной памяти. Узлы связаны сетью Gigabit Ethernet.
- Кластер функционирует на базе ОС Microsoft Windows HPC Server 2008, на каждом узле установлена среда выполнения MPI.NET Runtime версии 1.0.
- Для экспериментального исследования использовалась коллекция финансовых новостей агентства Reuters (Reuters-21578, Distribution 1.0)
- Обучающий набор представлен 7775 документами, каждый из которых относится к одной или нескольким из 115 рубрик.

Результаты экспериментов

- Ускорение

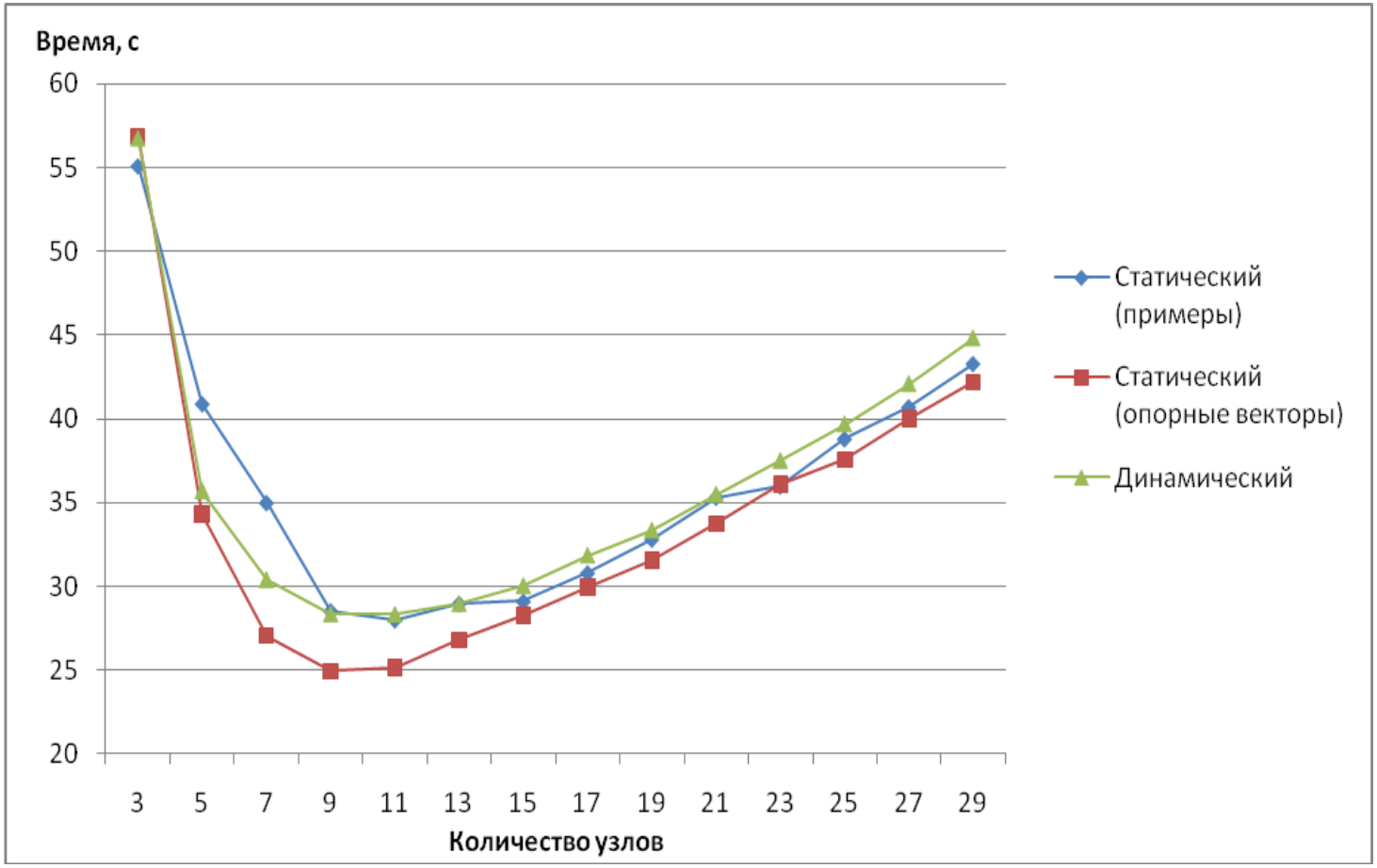
$$S = \frac{T_s}{T_p}$$

- Эффективность

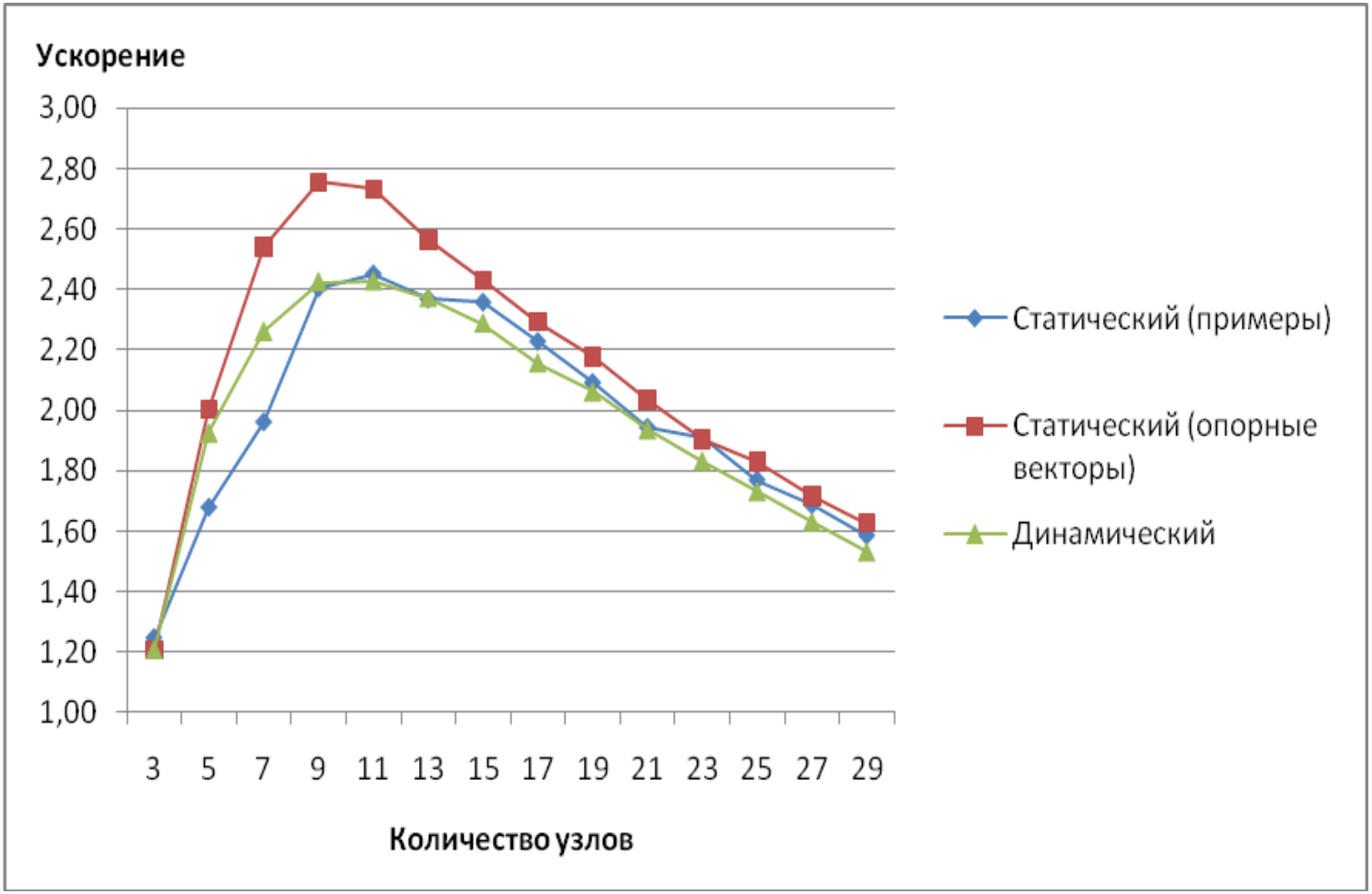
$$E = \frac{S}{p}$$

  - T_s – время решения задачи на одном вычислительном узле
  - T_p – время решения на p идентичных вычислительных узлах
  - p – количество вычислительных узлов

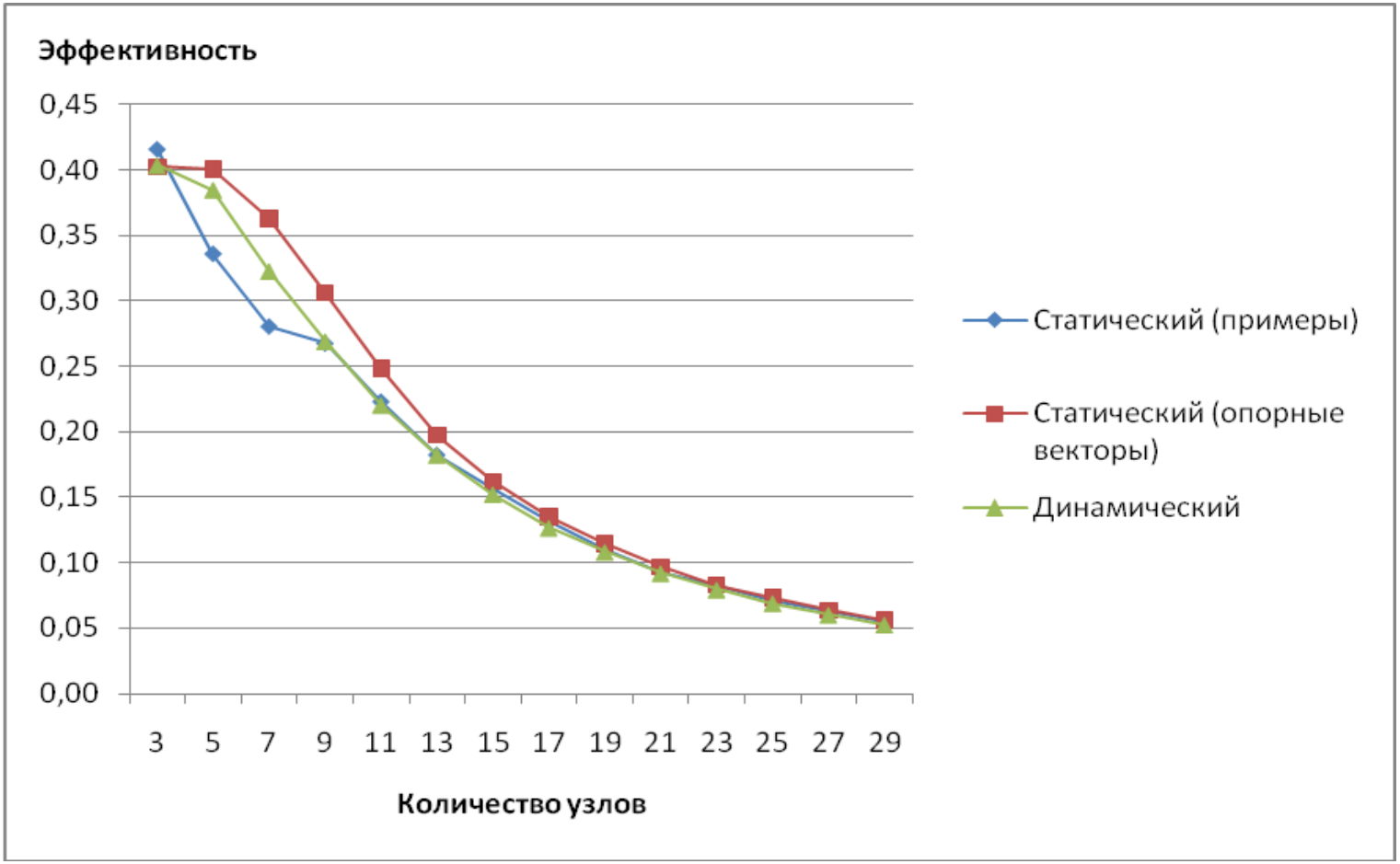
Результаты экспериментов для трех методов



Зависимость ускорения от количества узлов



Зависимость эффективности от количества узлов



Данные по десяти рубрикам коллекции Reuters-21578 с наибольшим количеством примеров

Условные номера рубрик	Количество примеров в рубрике	Время обучении для рубрики, с	Количество опорных векторов
31	2877	8,53	766
1	1650	9,79	857
62	538	6,77	520
38	433	4,96	386
25	389	5,42	414
112	369	6,13	438
46	347	5,40	436
114	212	2,89	211
96	197	4,89	387
17	181	4,12	331

Заключение

- Предлагаемые методы можно применять для повышения эффективности решения задачи текстовой классификации

- Когда априорная информация неизвестна могут применяться методы статического распределения нагрузки на основе примеров и динамического распределения.

  Следует отдавать предпочтение последнему.

- При решении задачи поиска оптимальных параметров классификатора на первом проходе рекомендуется использовать метод динамического распределения, а при последующих – метод статического распределения на основе опорных векторов

Спасибо за внимание!
Вопросы?