**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования  
«Национальный исследовательский университет   
«Высшая школа экономики»**

Факультет Компьютерных наук

Департамент больших данных и информационного поиска

Базовая кафедра Яндекс

УТВЕРЖДАЮ

Академический руководитель

образовательной программы

«Науки о данных»

по направлению 01.04.02

«Прикладная математика и информатика»

С.О. Кузнецов

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2014 г.

**Программа дисциплины** «Алгоритмы обработки больших данных»

для направления 01.04.02 "Прикладная математика и информатика" подготовки магистра

для магистерской программы "Науки о данных"

**Автор программы:**

Бабенко М.А., к.ф.-м.н. (maxim.babenko@gmail.com)

Одобрена на заседании базовой кафедры Яндекс «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2014  г.

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_\_\_\_\_ М.А. Бабенко

Рекомендована Академическим советом образовательной программы

«Науки о данных» «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 2014  г.

Менеджер базовой кафедры Яндекс \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.Ф. Баулин

Москва, 2014

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения разработчика программы.*

Пояснительная записка

## Автор программы

Бабенко М.А., к.ф.-м.н.

## Требования к студентам

## Данный курс предназначен для студентов, успешно освоивших годовой предмет «Алгоритмы и структуры данных для поиска». Для понимания излагаемого материала необходимо иметь представление о RAM модели вычисления во внутренней памяти, знать основные алгоритмы сортировки, поиска, а также стандартные структуры данных (стеки, деки, очереди, кучи, деревья). Полезным также является знакомство с базовыми теоретико-графовыми алгоритмами (DFS, BFS, алгоритм Дейкстры).

## Аннотация

Дисциплина «Алгоритмы обработки больших данных» предназначена для подготовки магистров 01.04.02 – Прикладная математика и информатика.

Данный курс позволит студентам в своей практической деятельности точнее оценивать сложность алгоритмов, манипулирующих с большими объемами данных, а следовательно быстрее и точнее выбирать наиболее эффективное решение. По мере роста объемов данных, с которыми приходится иметь дело, а также усложнения иерархий памяти в современных вычислительных системах, данные навыки становятся с каждым годом все более актуальными.

Программа курса предусматривает лекции (26 часов) и практические занятия (38 часов).

## Учебные задачи курса

Целью данного курса является ознакомление слушателей с иерархиями памяти, как с теоретической, так и с практической точек зрения. Рассказывается про модель вычислений во внешней памяти, а также cache-oblivious вычисления. В рамках указанных моделей изучаются задачи сортировки и поиска (вместе с сопутствующими структурами данных – кучи, буферизированные деревья, биномиальные списки и т.д.) Также внимание уделяется базовым концепциям вычислений на потоках данных с ограниченной памятью. Все темы курса снабжены практическими заданиями, призванными дать возможность сопоставить теорию с практикой.

Тематический план дисциплины «Алгоритмы обработки больших данных»

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№** | **Название темы** | **Всего часов по дисциплине** | **Аудиторные часы** | | **Самосто-ятельная работа** |
| **Лекции** | **Сем. и практика занятия** |
| 1 | Простейшие вычисления во внешней памяти | 16 | 2 | 4 | 10 |
| 2 | Деревья во внешней памяти | 30 | 6 | 8 | 16 |
| 3 | Графы во внешней памяти | 24 | 4 | 6 | 14 |
| 4 | Кеширование (caching) | 30 | 6 | 8 | 16 |
| 5 | Хеширование (hashing) и создание эскизов (sketching) | 22 | 4 | 6 | 12 |
| 6 | Потоковые алгоритмы | 22 | 4 | 6 | 12 |
|  | Итого | 144 | 26 | 38 | 80 |

# Источники информации

## Список литературы

### Основная литература

1. Jeffery Scott Vitter, Algorithms and data structures for external memory
2. U. Meyer, P. Sanders, and J. Sibeyn (eds.), Algorithms for Memory Hierarchies
3. N. Zeh, I/O-Efficient Graph Algorithms
4. Frigo, Leiserson, Prokop, Cache-oblivious algorithms
5. Demaine, Cache-oblivious algorithms and data structures
6. A. Rajaraman, J. Ullman, Mining of Massive Datasets

### Дополнительная литература

1. Ulrich Drepper, What every programmer should know about memory
2. Brodal, Fagerberg, Jacob, Cache-oblivious search trees via binary trees of small height
3. J. Abello, A. Buchsbaum, J. Westbrook, A Functional Approach to External Graph Algorithms
4. P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality
5. A. Broder, M. Charikar, A. Frieze, M. Mitzenmacher, Min-wise independent permutations
6. M. Charikar, Similarity Estimation Techniques from Rounding Algorithms
7. J. Munro, M. Paterson, Selection and sorting with limited storage
8. G.S. Manku, S. Rajagopalan, B.G. Lindsay, Approximate Medians and other Quantiles in One Pass and with Limited Memory
9. G. Cormode, S. Muthukrishnan, An Improved Data Stream Summary: The Count-Min Sketch and its Applications

# Формы контроля и структура итоговой оценки

Текущий контроль - домашняя работа в первом модуле, контрольная работа в первом модуле.

Итоговый контроль – письменный экзамен (120 мин.)

Итоговая оценка вычисляется следующим образом:

0,1\*оценка за домашнюю + 0,2\*оценка за контрольную + 0,7\*оценка за экзамен.

### Таблица соответствия оценок по десятибалльной и системе зачет/незачет

|  |  |
| --- | --- |
| **Оценка по 10-балльной шкале** | **Оценка по 5-балльной шкале** |
| 1 | Незачет |
| 2 |
| 3 |
| 4 | Зачет |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

### Таблица соответствия оценок по десятибалльной и пятибалльной системе

|  |  |
| --- | --- |
| **По десятибалльной шкале** | **По пятибалльной системе** |
| 1 – неудовлетворительно  2 – очень плохо  3 – плохо | неудовлетворительно – 2 |
| 4 – удовлетворительно  5 – весьма удовлетворительно | удовлетворительно – 3 |
| 6 – хорошо  7 – очень хорошо | хорошо – 4 |
| 8 – почти отлично  9 – отлично  10 - блестяще | отлично – 5 |

# Программа дисциплины «Алгоритмы обработки больших данных»

**Тема 1. Простейшие вычисления во внешней памяти**

### Модель вычислений во внешней памяти, измерение сложности алгоритмов

### Оценки сложности сортировки во внешней памяти

### MergeSort как I/O-оптимальная сортировка, ее сложность

### Основная литература

1. Jeffery Scott Vitter, Algorithms and data structures for external memory
2. U. Meyer, P. Sanders, and J. Sibeyn (eds.), Algorithms for Memory Hierarchies

### Дополнительная литература

1. Ulrich Drepper, What every programmer should know about memory

**Тема 2. Деревья во внешней памяти**

B-деревья (B-trees) поиска и их разновидности (B+, B\*)

Использование B-деревьев в DBMS-системах, типы индексных структур

Персистентные (persistent) B-деревья

Слоистые (stratified) B-деревья

Буферизированные (buffered) B-деревья

B-боры (B-tries)

### Основная литература

1. Jeffery Scott Vitter, Algorithms and data structures for external memory
2. U. Meyer, P. Sanders, and J. Sibeyn (eds.), Algorithms for Memory Hierarchies

**Тема 3. Графы во внешней памяти**

Представление графов во внешней памяти

Задача о ранжировании списка (list ranking)

Обходы графов во внешней памяти: DFS, BFS

### Основная литература

1. Jeffery Scott Vitter, Algorithms and data structures for external memory
2. U. Meyer, P. Sanders, and J. Sibeyn (eds.), Algorithms for Memory Hierarchies
3. N. Zeh, I/O-Efficient Graph Algorithms

### Дополнительная литература

1. J. Abello, A. Buchsbaum, J. Westbrook, A Functional Approach to External Graph Algorithms

**Тема 4. Кеширование (caching)**

Кеши, их типы и организация, стратегии замещения и способы оценки их эффективности

Нечувствительные к кешированию (cache-oblivious) алгоритмы

Задача об умножении матриц

Деревья van Emde Boas

Оптимальный статический бинарный поиск

Деревья поиска

Приоритетные очереди

### Основная литература

1. Frigo, Leiserson, Prokop, Cache-oblivious algorithms
2. Demaine, Cache-oblivious algorithms and data structures

### Дополнительная литература

1. Ulrich Drepper, What every programmer should know about memory
2. Brodal, Fagerberg, Jacob, Cache-oblivious search trees via binary trees of small height

**Тема 5. Хеширование (hashing) и создание эскизов (sketching)**

Гипотеза равномерного хеширования, универсальные семейства хеш-функций

Хеш-функции, учитывающие близость (locality-sensitive hashing), эскизы (sketches)

Мера сходства, связь с LSH

Мера сходства Жаккара, система LSH на основе min-wise перестановок

Косинусная близость, система LSH на основе случайных проекций

### Основная литература

1. A. Rajaraman, J. Ullman, Mining of Massive Datasets

### Дополнительная литература

1. P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality
2. A. Broder, M. Charikar, A. Frieze, M. Mitzenmacher, Min-wise independent permutations
3. M. Charikar, Similarity Estimation Techniques from Rounding Algorithms

**Тема 6. Потоковые алгоритмы**

Поиск порядковых статистик: точных (Munro-Paterson) и приближенных (Manku-Rajagopalan-Lindsay)

Приближенный подсчет числа различных элементов

Поиск слонов (elephants, heavy hitters)

### Основная литература

1. A. Rajaraman, J. Ullman, Mining of Massive Datasets

### Дополнительная литература

1. J. Munro, M. Paterson, Selection and sorting with limited storage
2. G.S. Manku, S. Rajagopalan, B.G. Lindsay, Approximate Medians and other Quantiles in One Pass and with Limited Memory
3. G. Cormode, S. Muthukrishnan, An Improved Data Stream Summary: The Count-Min Sketch and its Applicationsxx

# Методические указания студентам

Самостоятельная работа студента предусматривает выполнение теоретических заданий, направленных на овладение техникой построения алгоритмов, оперирующих большими данными, а также практических заданий по программной реализации этих алгоритмов.

Автор программы: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ <Бабенко М.А.> /