Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

Московский государственный университет имени М.В. Ломоносова

Механико-математический факультет

Кафедра теоретической информатики

УТВЕРЖДАЮ

Заведующий кафедрой

\_\_\_\_\_\_\_\_\_\_\_\_\_\_/А.В.Михалев /

«22» января 2019 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ**

**Наименование дисциплины :**

спецкурс по выбору студента

**Аналитика больших данных: основные алгоритмы**

**Уровень высшего образования:**

**аспирантура**

**Направление подготовки (специальность):**

**01.06.01 Математика и механика**

**Направленность (профиль) ОПОП:**

**Математика**

Форма обучения:

очная

Рабочая программа рассмотрена и одобрена

на заседании кафедры Теоретической информатики

(протокол № 2а, «22 » января 2019 года)

Москва 2019

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по специальности

« Математика и механика » (программы аспирантуры) в редакции приказа МГУ от 30 декабря 2016 г.

Год (годы) приема на обучение\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1.** Место дисциплины в структуре ОПОП ВО. **Вариативная часть ОПОП ВО***.*

**2.** Входные требования для освоения дисциплины, предварительные условия : **знание основ статистики, линейной алгебры и** **программирования.**

**3.** Результаты обучения по дисциплине , соотнесенные с требуемыми компетенциями выпускников*.*

|  |  |
| --- | --- |
| **Компетенции выпускников (коды)** | **Планируемые результаты обучения по дисциплине , соотнесенные с компетенциями** |
| *ПК-1, ПК-5* | ***Знать:*** *стандартные методы обработки больших данных в задачах нахождения схожих элементов, частых наборов и кластеризации.****Уметь*** *ориентироваться в современных методах и алгоритмах компьютерной математики, уметь решать стандартные задачи по изученным темам.****Владеть:*** *прикладными методами решения практических задач по изученным темам.* |
| *ОПК-1, ОПК-2, ОПК-4* | ***Уметь*** *решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий****\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_******Иметь опыт*** *использования на практике математических алгоритмов, в том числе с применением современных вычислительных систем**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
| *УК-1* | ***Уметь*** *формулировать научно обоснованные гипотезы, создавать теоретические модели явлений и процессов, применять методологию научного познания в профессиональной деятельности****\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_******Владеть*** *специальными разделами фундаментальной математики, методами анализа и решения задач**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |

**4.** Формат обучения очный .

**5.** Объем дисциплины составляет **1** з.е., в том числе **50**  академических часов, отведенных на контактную работу обучающихся с преподавателем, **30**  академических часов на самостоятельную работу обучающихся.

**6.** Содержание дисциплины , структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий

|  |  |  |
| --- | --- | --- |
| **Наименование и краткое содержание разделов и тем дисциплины ,****Форма промежуточной аттестации по дисциплине**  | **Всего****(часы**) | В том числе |
| **Контактная работа (работа во взаимодействии с преподавателем)****Виды контактной работы, часы** | **Самостоятельная работа обучающегося,** **часы** *(виды самостоятельной работы – эссе, реферат, контрольная работа и пр. – указываются при необходимости)* |
| Занятия лекционного типа\* | Занятия семинарского типа\* | **Всего** |  |
| Тема 1. Введение в DataMining и аналитику больших данных.  | 4 | 2 |  | 2 | 2 |
| Тема 2. Технология MapReduce распараллеливания вычислений. Алгоритмы, использующие MapReduce. Операции реляционной алгебры. Матричные вычисления. | 6 | 4 |  | 4 | 2 |
| Тема 3. Алгоритмы обнаружения схожих элементов. Сходство по Жаккару. | 4 | 2 |  | 2 | 2 |
| Тема 4. Сходство по Жаккару. Методы сжатия больших файлов. Хеширование, подписи больших файлов. | 6 | 4 |  | 4 | 2 |
| Тема 5. Локально-чувствительное хеширование документов. | 4 | 2 |  | 2 | 2 |
| Тема 6. Техника группировок. Построение семейств функций. | 4 | 2 |  | 2 | 2 |
| Текущий контроль успеваемости | 4 |  | 2 | 2 | 2 |
| Тема 7. Метрики на пространствах данных. | 4 | 2 |  | 2 | 2 |
| Тема 8. Локально-чувствительные семейства функций хеширования в различных метрических пространствах. | 6 | 4 |  | 4 | 2 |
| Тема 9. Методы высоких степеней сходства. Индексация. Использование позиции и длины. | 4 | 2 |  | 2 | 2 |
| Тема 10. Анализ «рыночной корзины». Представление данных и основные алгоритмы. | 6 | 4 |  | 4 | 2 |
| Тема 11. Методы кластеризации в обработке больших данных. | 6 | 4 |  | 4 | 2 |
| Тема 12. Кластеризация в различных метрических пространствах. | 4 | 2 |  | 2 | 2 |
| Консультации | 4 | 2 | 2 |
| Экзамен/зачет | 4 | 4 |  |
| Промежуточная аттестация – контрольная работа | 10 | 8 | 2 |
| **Итого** | 80 |  50 | 30 |

7. Фонд оценочных средств (ФОС) для оценивания результатов обучения по дисциплине

7.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости.

1. Операция реляционной алгебры R(A, B)⊲⊳ B<C S(С, D) создает все кортежи (a, b, c, d) такие, что кортеж (a, b) принадлежит отношению R, кортеж (c, d) принадлежит S, и b < c. Разработайте алгоритм MapReduce этой операции, полагая, что R и S являются множествами.

2. Разработать алгоритм MapReduce для обработки очень большого файла целых чисел и вычисления:

(а) наибольшего числа;

(б) среднего значения всех чисел;

(с) того же множества целых чисел, но где каждое число встречается только один раз;

(d) количества различных чисел.

3. Вычислить двумерные подписи множеств, использую заданные функции хеширования, вычислить получаемые коэффициенты сходства по Жаккару, оценить точность.

3.Универсальное множество U содержит *n* элементов. Случайным образом выбираются подмножества S и T, содержащие *m* элементов каждое. Каково ожидаемое значение коэффициента сходства по Жаккару этих множеств?

4. Дана характеристическая матрица множества *S1, S2, S3* и *S4*.

(a) вычислите матрицу подписей, используя следующие три хеш-функции:

h1(x) = 2x + 1 (mod 6); h2(x) = 3x + 2 (mod 6); h3(x) = 5x + 2 (mod 6).

(b) какие из этих функция являются реальными перестановками строк характеристической матрицы?

(c) насколько близка оценка сходства по Жаккару этих множеств (с помощью вычисленных подписей) к его истинному значению?

5. На пространстве неотрицательных чисел какие их следующих функций задают метрику (доказать справедливость, либо привести пример нарушения свойств).

(a) max(x, y);

(b) diff(x, y) = |x − y|;

(c) sum(x, y) = x + y.

6. Докажите, что расстояние по косинусу между двумя векторами (одной длины), чьи компоненты принимают значения 0 или 1, составляет не более 90°.

7. Пусть эскизы вычисляются следующим набором «случайных» векторов:

v1 = [+1,+1,+1,−1], v2 = [+1,+1,−1,+1], v3 = [+1,−1,+1,+1] и v4 = [−1,+1,+1,+1].

Вычислите эскизы следующих векторов:

(a) [2, 3, 4, 5].

(b) [−2, 3,−4, 5].

(c) [2,−3, 4,−5].

Для каждой из пар этих трех векторов какова оценка расстояния по косинусу и каково его истинное значение?

8. Есть 100 корзин, пронумерованных числами 1,2, ..., 100, и 100 элементов, также являющихся числами от 1 до 100. Элемент i находится в корзине j тогда и только тогда, когда i делит j без остатка. Например, корзина 24 – это набор элементов {1,2,3,4,6,8,12,24}.

Опишите все ассоциативные правила, имеющие достоверность (confidence) = 100%.

Какие из следующих правил имеют 100% достоверность (confidence)?

– {4,6} → 24;

– {8} → 16;

– {3,4,5} → 42;

– {8,10} → 20.

9. Пусть {A,B,C} – часто встречающийся набор элементов, а {B,C,D,E} – не часто встречающийся набор элементов.

Верны ли следующие утверждения:

– {B,C,E} – не часто встречающийся набор элементов.

– {B,C} – часто встречающийся набор элементов.

– {A,B,C,D,E,F} может быть как часто встречающимся набором элементов, так и не часто встречающимся набором элементов.

– {B,C,F} – не часто встречающийся набор элементов.

7.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации.

Матричное представление множеств.

Вычисление хешированных подписей.

Локально-чувствительное хеширование документов.

LSH для хешированных подписей.

Анализ техники группировок (*S*-кривые)

Использование методов хешированных подписей и LSH для определения вероятно схожих документов.

Теория локально-чувствительных функций.

Локально-чувствительные семейства для расстояния по Жаккару

Усиление локально-чувствительного семейства (AND-конструкции и OR-конструкции).

Случайные гиперплоскости и расстояние по косинусу.

LSH-семейства для евклидова расстояния (размерность 2).

Ассоциативные правила. Достоверность и поддержка. Алгоритм поиска ассоциативных правил с заданными достоверностью и поддержкой.

Использование оперативной памяти при поиске частых наборов элементов. Метод треугольной матрицы и метод троек.

Свойство антимонотонности наборов элементов. Масштабируемый алгоритм поиска ассоциативных правил (A-Priori Algorithm)

|  |
| --- |
| **ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине (модулю)**  |
| ОценкаРО исоответствующие виды оценочных средств  | 2 | 3 | 4 | 5 |
| **Знания***(виды оценочных средств: устные и письменные опросы и контрольные работы, тесты, и т.п. )* | Отсутствие знаний | Фрагментарные знания | Общие, но не структурированные знания | Сформированные систематические знания |
| **Умения***(виды оценочных средств: практические контрольные задания, написание и защита рефератов на заданную тему и т.п.)*  | Отсутствие умений | В целом успешное, но не систематическое умение | В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера) | Успешное и систематическое умение |
| **Навыки (владения, опыт деятельности)***(виды оценочных средств: выполнение и защита курсовой работы, отчет по практике, отчет по НИР и т.п.)*  | Отсутствие навыков (владений, опыта) | Наличие отдельных навыков (наличие фрагментарного опыта) | В целом, сформированные навыки (владения), но используемые не в активной форме | Сформированные навыки (владения), применяемые при решении задач |

8. Ресурсное обеспечение:

* Перечень основной и дополнительной литературы,
* А. Основная литература
* 1. Ю. Лесковец, А. Раджарамаран, Дж. Ульман. Анализ больших наборов данных. ДМК, Москва, 2016.
* 2. Wil M. P., van der Aalst. Process Mining Discovery, Conformance and Enhancement of Business Processes. ISBN: 978-3-642-19344-6 (Print) 978-3-642-19345-3 (Online).
* Б. Дополнительная литература

# 1. [Jure Leskovec](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople), [Anand Rajaraman](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople), [Jeffrey D. Ullman](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople). Mining of Massive Datasets. 2nd Edition. Stanford University, California, 2014.

* Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. <http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition#c3bEPha9lWJHRUhW.99>

2. <http://i.stanford.edu/~ullman/mmdsn.html>

* Описание материально-технического обеспечения.

Компьютерный класс

9. Язык преподавания.

Русский

10. Преподаватели

Главацкий С.Т., Бурыкин И.Г., Айдагулов Р.Р.

11. Авторы программы

Главацкий С.Т., Бурыкин И.Г., Айдагулов Р.Р.