1. Спецкурс программы аспирантуры, полугодовой: **Аналитика больших данных, основные алгоритмы**.

2. Преподаватели: доц. С.Т.Главацкий, н.с. И.Г.Бурыкин, с.н.с. Р.Р.Айдагулов.

3. Аннотация курса: распределенные файловые системы и Map-Reduce как инструмент для создания параллельных алгоритмов; обработка наборов документов, поиск по сходству; методы ключевого хеширования (minhashing) и локально чувствительного хеширования; метрики на пространствах данных; методы поиска высоких степеней сходства, индексация; алгоритмы кластеризации многомерных массивов данных.

4. Тематическое содержание курса:

|  |  |
| --- | --- |
| **Тема 1.** | Введение в DataMining и аналитику больших данных. |
| **Тема 2.** | Технология MapReduce распараллеливания вычислений. Алгоритмы, использующие MapReduce. Операции реляционной алгебры. Матричные вычисления. |
| **Тема 3.** | Алгоритмы обнаружения схожих элементов. Сходство по Жаккару. |
| **Тема 4.** | Сходство по Жаккару. Методы сжатия больших файлов. Хеширование, подписи больших файлов. |
| **Тема 5.** | Локально-чувствительное хеширование документов. |
| **Тема 6.** | Техника группировок. Построение семейств функций. |
| **Тема 7.** | Метрики на пространствах данных. |
| **Тема 8.** | Локально-чувствительные семейства функций хеширования в различных метрических пространствах. |
| **Тема 9.** | Методы высоких степеней сходства. Индексация. Использование позиции и длины. |
| **Тема 10.** | Анализ «рыночной корзины». Представление данных и основные алгоритмы. |
| **Тема 11.** | Методы кластеризации в обработке больших данных. |
| **Тема 12.** | Кластеризация в различных метрических пространствах. |

5. Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения, характеризующих этапы формирования компетенций.

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)

Примеры задач для самостоятельного решения.

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;

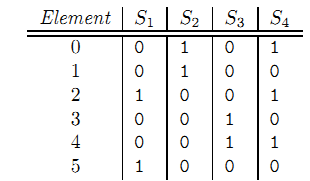
(с) того же множества целых чисел, но где каждое число встречается только один раз;

(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.

6. Перечень основной и дополнительной учебной 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.

# 7. Перечень ресурсов информационно-коммуникационной 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

**Программа утверждена на заседании кафедры теоретической информатики**

**Протокол №**