**Спецкурсы кафедры математических методов прогнозирования для аспирантов**

на 2015/16 учебный год

|  |  |  |  |
| --- | --- | --- | --- |
| № | **Спецкурс** | **Семестр** | **Лектор** |
| 1 | **«Непрерывные морфологические модели и алгоритмы»**Continuous morphological models and algorithms | Осень | д.т.н., профессор Местецкий Леонид Моисеевич  |
| 2 | **«Задачи и алгоритмы вычислительной геометрии»**Problems and algorithms of computational geometry | Весна | д.т.н., профессор Местецкий Леонид Моисеевич  |
| 3 | **«Логический анализ данных в распознавании»** Logical data analysis in pattern recognition | Осень | д.ф.-м.н, доцент Дюкова Елена Всеволодовна |
| 4 | **«Методы машинного обучения и поиск достоверных закономерностей в данных»** Machine learning and knowledge discovery | Осень | д.ф.-м.н. Сенько Олег Валентинович |
| 5 | **«Нестатистические методы анализа данных и классификации»,** Non-statistical methods of data mining and classification | Осень | д.ф.-м.н., профессор Рязанов Владимир Васильевич |
| 6 | **«Нестатистический анализ данных»,** Non-statistical data mining | Весна | д.ф.-м.н., профессор Рязанов Владимир Васильевич |
| 7 | **«Метрические методы интеллектуального анализа данных»**Metric methods of machine learning | Весна | к.ф.-м.н., доцент Майсурадзе Арчил Ивериевич |
| 8 | **«Вероятностное тематическое моделирование»**Probabilistic topic modelling  | Весна | д.ф.-м.н.Воронцов Константин Вячеславович |

**Курс «Непрерывные морфологические модели и алгоритмы»**

**Местецкий Л.М.**

**Аннотация**

В компьютере изображения представляются прямоугольными матрицами точек, обладающих определенным цветом и яркостью. Такое дискретное представление является удобным для ввода, запоминания, обработки в компьютере. Однако для анализа формы объектов такое представление неудобно. Человеку привычнее и проще при описании формы объектов оперировать непрерывными геометрическими фигурами. Основные преимущества непрерывного представления: адекватность его с физической сущностью «сплошных» объектов реального мира, возможность использования для анализа, преобразований и распознавания формы методов «непрерывной» математики.

В курсе рассматриваются вопросы анализа формы плоских фигур и связанные с этим приложения в области распознавания изображений, компьютерной графики и геоинформатики. Рассматриваются темы, относящиеся к вычислительной геометрии – геометрический поиск, выпуклые оболочки, пересечения и близость объектов. Вычислительная геометрия является теоретической основой для разработки эффективных алгоритмов работы с изображениями.

**Программа**

* Основные понятия вычислительной геометрии. Мера сложности вычислений. Асимптотический анализ сложности. Рекурсивные алгоритмы. Оценка вычислительной сложности задачи.
* Алгоритмическая парадигма «разделяй и властвуй». Геометрический поиск. Локализация точки в простом и в выпуклом многоугольнике при уникальном и массовом запросе. Локализация точки в планарном подразбиении.
* Построение выпуклых оболочек. Алгоритмы Джарвиса, Грэхема. Слияние выпуклых оболочек. Метод редукции для оценки сложности задачи.
* Пересечения геометрических объектов. Пересечения конечного множества отрезков. Алгоритмическая парадигма плоского заметания. Структуры данных в алгоритме заметания.
* Близость геометрических объектов. Разбиения Вороного и триангуляции Делоне. Преобразования двойственных графов. Алгоритмы построения триангуляции Делоне: наивные, жадные, инкрементные, флип-флоп, рекурсивные. Обобщенные диаграммы Вороного для различных множеств сайтов – кругов, отрезков, многоугольников.
* Обобщенная диаграмма Вороного и скелет многоугольной фигуры.

**Литература**

1. Препарата Ф., М.Шеймос М. Вычислительная геометрия: введение. Москва: «Мир», 1989.
2. Местецкий Л.М. [Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры.](http://www.machinelearning.ru/wiki/index.php?title=%D0%A3%D1%87%D0%B0%D1%81%D1%82%D0%BD%D0%B8%D0%BA:Mest#.D0.9C.D0.BE.D0.BD.D0.BE.D0.B3.D1.80.D0.B0.D1.84.D0.B8.D1.8F) Москва: Физматлит. 2009.
3. Местецкий Л.М. Лекции по вычислительной геометрии. Учебное пособие. ВМК МГУ, 2013. [Конспект [1.46 Мб]](http://www.machinelearning.ru/wiki/images/8/86/Lectures_CG.rar)
4. Местецкий Л.М. Лекции по непрерывным морфологическим моделям [Презентации PowerPoint [3.8 Мб]](http://www.machinelearning.ru/wiki/images/9/9d/Lectures_Mestetskiy.rar)

**Курс «Задачи и алгоритмы вычислительной геометрии»**

**Местецкий Л.М.**

**Аннотация**

Эффективные алгоритмы работы с геометрической информацией являются непременным атрибутом всех современных систем компьютерной графики, анализа и распознавания изображений, машинного зрения, геоинформационных систем. Геометрические алгоритмы предоставляют хорошее поле для развития алгоритмического мышления, необходимого в прикладной математике.

В этом спецкурсе будут рассмотрены классические темы вычислительной геометрии: геометрический поиск, выпуклые оболочки, пересечение и близость объектов, диаграммы Вороного, триангуляции Делоне, скелеты многоугольников.

**Программа**

* Бинарное изображение как результат сегментации формы. Фигура, как модель формы, открытое множество, связное множество, область, односвязная и многосвязная области, замкнутая область, жорданова кривая
* Непрерывные и дискретные модели формы. Граничное и скелетное описание фигуры, их достоинства и недостатки. Дискретные модели границы и скелета фигуры. Эквивалентность непрерывной и дискретной моделей формы.
* Непрерывные границы дискретной фигуры. Связность в дискретном пространстве, дискретная фигура. Треугольные структуры соседства, гексагональные структуры соседства (6-смежность), объектная и компонентная структуры соседства. Граничные пары, граничные треугольники, граничные коридоры дискретной сцены.
* Поиск и прослеживание границы. Постановка задачи прослеживания границ дискретной сцены. Симплексное прослеживание. Прослеживание методом подвижного моста.
* Аппроксимация граничного коридора разделяющей фигурой минимального периметра. Алгоритм вытягивания нити. Аппроксимация границы составными кривыми Безье.
* Обобщения диаграммы Вороного и триангуляции Делоне: в манхэттенской и в шахматной метриках, для сайтов-окружностей в метрике Лагерра, для разделённых сайтов-сегментов.
* Диаграмма Вороного многоугольной фигуры. Связь диаграммы Вороного многоугольника и его скелета, алгоритм построения скелета многоугольника по его диаграмме Вороного. Обобщённая триангуляция Делоне для коллекции сайтов.
* Регуляризация скелета методом стрижки. Силуэт скелетного графа, силуэт подграфа, расстояние между силуэтами. Последовательная стрижка скелета с контролем точности. Стрижка тонких элементов.
* Циркулярные фигуры и жирные линии. Преобразования формы изображений в компьютерной графике и в распознавании образов. Жирная кривая, циркулярная фигура, осевой граф, функция ширины, силуэт. Деформация циркуляра и задача подгонки, аппроксимация скелета циркуляром.

**Курс «Логический анализ данных в распознавании»**

**Дюкова Е.В.**

**Аннотация**

В спецкурсе будут изложены общие принципы, лежащие в основе дискретных методов анализа информации в задачах распознавания, классификации и прогнозирования. Будут рассмотрены подходы к конструированию процедур распознавания на основе использования аппарата логических функций и методов построения покрытий булевых и целочисленных матриц. Будут изучены основные модели и рассмотрены вопросы, связанные с исседованием сложности их реализации и качества решения прикладных задач.

**Программа**

Постановка задачи распознавания. Сущность дискретного подхода к задачам распознавания. Общие принципы построения дискретных (логических) процедур распознавания. Понятие элементарного классификатора. Основные модели дискретных (логических) процедур распознавания и используемые в них семейства элементарных классификаторов.

Модели алгоритмов голосования по представительным наборам. Детерминированные и стохастические модели тестовых алгоритмов. Модели алгоритмов голосования, основанные на построении покрытий классов. Модели, основанные на построении решающих деревьев. Методы построения элементарных классификаторов.

Основные понятия теории дизъюнктивных нормальных форм. Задача построения сокращенной дизъюнктивной нормальной формы двузначной логи-ческой функции, заданной конъюнктивной нормальной формой. Критерии допустимости, неприводимости и максимальности элементарной конъюнкции для двузначной логической функции, заданной конъюнктивной нормальной формой. Способы построения сокращенной дизъюнктивной нормальной формы для всюду определенной двузначной логической функции и частичной двузначной логической функции. Построение элементарных классификаторов на основе преобразования нормальных форм двузначных логических функций.

Понятие неприводимого покрытия булевой матрицы. Формулировка задачи построения неприводимых покрытий булевой матрицы как задачи преобразования конъюнктивной нормальной формы монотонной булевой функции в дизъюнктивную нормальную форму. Понятие тупикового покрытия целочисленной матрицы. Формулировка задачи построения тупиковых покрытий целочисленной матрицы как задачи преобразования конъюнктивной нормальной формы двузначной логической функции в дизъюнктивную нормальную форму. Построение элементарных классификаторов на основе поиска покрытий булевых и целочисленных матриц. Геометрическая интерпретация понятий покрытия и тупикового покрытия целочисленной матрицы. Примеры алгоритмов поиска покрытий и тупиковых покрытий булевых и целочисленных матриц, используемых на практике.

Сложность реализации дискретных (логических) процедур распознавания. Метрические свойства множества покрытий целочисленной матрицы.

Асимптотические оценки типичных значений числа покрытий и длины покрытия целочисленной матрицы. Асимптотические оценки типичных значений числа тупиковых покрытий и длины тупикового покрытия целочисленной матрицы в случае, когда число строк матрицы существенно превосходит число столбцов. Асимптотические оценки типичных значений числа (тупиковых) покрытий с длиной близкой к минимальной. Подходы к оценке сложности трудно решаемых перечислительных задач. Асимптотически оптимальный алгоритм поиска тупиковых покрытий целочисленной матрицы с полиномиальной задержкой.

Методы повышения эффективности дискретных процедур распознавания. Методика предварительного анализа обучающей информации. Выделение типичных для классов объектов и объектов, лежащих на границе между классами. Снижение влияния шумящих признаков. Применение дискретных процедур распознавания для обработки вещественнозначной информации. Проблема понижения значности исходной информации. Полные и частичные перекодировки.

Проблема построения корректных процедур распознавания. Алгебро-логический синтез корректных процедур распознавания на базе элементарных классификаторов. Задача классификации (таксономии, кластерного анализа). Процедуры классификации, основанные на построении покрытий классов.

**Литература**

**Основная литература**

1. Дискретная математика и математические вопросы кибернетики / Под ред. С.Б. Яблонского, О.Б. Лупанова. М.: Наука. 1974.
2. Дюкова Е.В. Дискретные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели. Учебное пособие для студентов математических факультетов педвузов. М: МПГУ. 2003. [Пособие (PDF)](http://www.ccas.ru/frc/papers/djukova03mp.pdf), [Приложение к пособию (PDF)](http://www.ccas.ru/frc/papers/djukova03appendix.pdf).
3. Дюкова Е.В. Алгоритмы распознавания типа Кора: сложность реализации и метрические свойства / Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука. 1989. Вып. 2. С. 99-125.
4. Дюкова Е.В. О числе тупиковых покрытий целочисленной матицы // Ж. вычисл. матем. и матем. физ. 2005. Т. 27, № 1. С. 114-127.
5. Дюкова Е.В., Журавлёв Ю.И. Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. 2000. Т. 40, №8. С. 1264-1278.
6. Дюкова Е.В., Журавлев Ю.М., Рудаков К.В. Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // Ж. вычисл. матем. и матем. физ. 1996. Т. 36, № 8. С. 217-225.
7. Дюкова Е.В., Инякин А.С. О процедурах классификации, основанных на построении покрытий классов // Ж. вычисл. матем. и матем. физ. 2003. Т. 43, № 12. С. 1910-1921.
8. Дюкова Е.В., Песков Н.В. Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Ж. вычисл. матем. и ма-тем. физ. 2002. Том 42, № 5. С. 741-753.

**Дополнительная литература**

1. Баскакова Л.В., Журавлёв Ю.И. Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ. 1981. Т. 21, №5. С. 1264-1275.
2. Вайнцвайг М.Н. Алгоритм обучения распознаванию образов «Кора» / Ал-горитмы обучения распознаванию образов. М.: Сов. Радио. 1973. С. 82-91.
3. Дмитриев А.И., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификации предметов или явлений / Дискретный анализ. Новоси-бирск: ИМ СО АН СССР. 1966. Вып. 7. С. 3-17.
4. Донской В.И. Алгоритмы обучения, основанные на построении решающих деревьев // ЖВМиМФ. 1982. Т. 22, № 4. С. 963-974.
5. Дюкова Е.В. Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. 1977. 233, № 4. С. 527-530.
6. Дюкова Е.В. Асимптотически оптимальные тестовые алгоритмы в задачах распознавания / Пробл. кибернетики. М.: Наука. 1982. Вып. 39. С. 165-199.
7. Дюкова Е.В. О сложности реализации некоторых процедур распознавания // Ж. вычисл. матем. и матем. физ. 1987. Т. 27, №1. С.114-127.
8. Дюкова Е.В. Асимптотические оценки некоторых характеристик множества представительных наборов и задача об устойчивости //Ж. вычисл. матем. и матем. физ. 1995. Т. 35, № 1. С. 122-134.
9. Дюкова Е.В. О сложности реализации дискретных (логических) процедур распознавания // Ж. вычисл. матем. и матем. физ. 2004. Т. 44, № 3. С. 550-572.
10. Журавлев Ю.И. Избранные научные труды. М.: Изд. Магистр. 1998. 416 с.
11. Кузнецов В.Е. Об одном стохастическом алгоритме вычисления информативных характеристик таблиц по методу тестов / Дискретный анализ. Новоси-бирск: ИМ СО АН СССР. 1973. Вып. 23. С. 8-23.
12. Рязанов В.В. О построении оптимальных алгоритмов распознавания и таксономии (классификации) при решении прикладных задач / Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука. 1988. Вып. 1. С. 229-279.
13. Чегис И.А., Яблонский С.В. Логические способы контроля электрических схем // Тр. МИАН СССР. М.: 1958.
14. Яблонский С.В. Введение в дискретную математику. М.: Наука. 1986.

**Курс «Методы машинного обучения и поиск достоверных закономерностей в данных»**

**Сенько О.В.**

**Аннотация**

В курсе обсуждаются основные проблемы, возникающие при использовании методов обучения по прецедентам (машинного обучения). Даётся краткий обзор существующих методов распознавания и регрессионного анализа. Рассказывается о способах оценки точности на генеральной совокупности (обобщающей способности). Обсуждаются различные способы повышения обобщающей способности методов машинного обучения.

Рассказывается о тесно связанной с задачей машинного обучения задаче поиска закономерностей в данных. Под закономерностью понимается некоторое относительно простое правило, связывающее прогнозируемую величину с другими переменными. Рассказывается о современных методах статистической верификации закономерностей, основанных на перестановочных тестах. Обсуждаются методы распознавания и регрессионного анализа, основанные методы, основанные на вычислении коллективных решений по системам закономерностей. Рассказывается также о различных методах кластерного анализа и способах их применения.

Приведены примеры успешно решённых задач диагностики, прогнозирования, интеллектульного анализа данных в различных областях, включая медицину, антропологию, социологию, химию др.

**Программа**

1. Область использования и основные задачи методов машинного обучения (ММО)
2. Основные проблемы, возникающие при использовании ММО
3. Методы точности прогнозирования для обученных алгоритмов
4. Краткий обзор методов линейного регрессионного анализа, включая методы, основанные на регуляризации (гребневая регрессия, Лассо, эластичная сеть)
5. Краткий обзор методов распознавания, включая
	* статистические методы;
	* нейросетевые методы;
	* решающие деревья и леса;
	* комбинаторно-логические методы;
	* метод опорных векторов;
	* методы, основанные на голосовании по системам закономерностей,
	* метод мультимодельных статистически взвешенных синдромов.
6. Анализ ROC-кривых как способ оценки эффективности методов распознавания
7. Понятие обобщённой ошибки прогнозирования в машинном обучении.
8. Структура обобщённой ошибки (трёх компонентное разложение )
9. Краткий обзор коллективных методов в машинном обучении, включая
	* простейшие комитеты;
	* логическая коррекция;
	* алгебраическая коррекция.
10. Выпуклые комбинации (ВК) прогностических алгоритмов
	* структура ошибки ВК;
	* структура составляющих обобщённой ошибки для ВК,
	* методы поиска оптимальных ВК, основанные на концепции несократимых и нерасширяемых наборов;
	* метод выпуклой регрессии, сравнение с Лассо и эластичной сетью.
11. Методы верификации закономерностей, основанные на перестановочных тестах, включая
	* верификацию простейших моделей;
	* оценка обоснованности использования более сложной модели по отношению к простой (лезвие Оккама).
12. Метод интеллектуального анализа данных (ИАД), основанный на оптимальных достоверных разбиениях.
13. Примеры использования в ММО и ИАД для решения медицинских задач.

**Литература**

1. Журавлёв Ю.И. Рязанов В.В. Сенько О.В. РАСПОЗНАВАНИЕ. Математические методы. Программная система. Применения. - Москва: Фазис, 2006
2. Сенько О.В. Перестановочный тест в методе оптимальных разбиений. // Ж. выч. матем. и матем. физ. N9, 2003, с.1438-1447.
3. А.В. Кузнецова, И.В. Костомарова, Н.Н. Водолагина, Н.А. Малыгина, О.В. Сенько. [Изучение влияния клинико-генетических факторов на течение дисциркуляторной энцефалопатии с использованием методов распознавания](http://mi.mathnet.ru/rus/mbb/v6/i1/p115) // Матем. биолог. и биоинформ., 2011, том 6, выпуск 1, страницы 115–146.
4. Oleg V.Senko and Anna V. Kuznetsova. [The Optimal Valid Partitioning Procedures](http://interstat.statjournals.net/YEAR/2006/abstracts/0604002.php). InterStat. Statistics on the Internet, April, 2006.
5. Сенько О.В., Докукин А.А. Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности. ЖВМиМФ, Т. 51, №9 с.1751-1760, 2011.
6. Senko O., Kuznetsova A. A recognition method based on collective decision making using systems of regularities of various types // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. Vol. 20, No. 2, 2010, pp. 152-162.

**Курс «Нестатистические методы анализа данных и классификации»**

**Рязанов В.В.**

**Аннотация**

Основная цель спецкурса состоит в изложении основанных на оптимизационных, дискретных и эвристических подходах методов анализа данных. Будут рассмотрены логические модели распознавания (классификации с учителем) и анализа разнотипных многомерных данных, методы оптимизации моделей распознавания, алгоритмы поиска скрытых логических закономерностей и связей по признаковым описаниям, алгоритмы обработки множеств логических закономерностей, методы создания качественных моделей объектов, ситуаций, явлений или процессов. Будут рассмотрены практические численные методы решения данных задач, и их применения в медицине, бизнесе, химии, технике и других областях.

**Программа**

1. Задача распознавания (классификации) по прецедентам. Тупиковые тесты и их вычисление, случаи вещественных признаков, стохастический тестовый алгоритм.
2. Логические модели распознавания для разнотипных признаков (алгоритмы вычисления оценок, алгоритм «Кора», модели с представительными наборами, практические реализации).
3. Эффективное вычисление оценок для различных типов признаков.
4. Оптимизация моделей распознавания. Релаксационный и комбинаторный алгоритмы поиска максимальной совместной подсистемы системы линейных неравенств.
5. Логические закономерности классов, эвристический критерий качества логических закономерностей.
6. Сведение задачи поиска логических закономерностей классов к задаче целочисленного линейного программирования.
7. Логические закономерности классов, стандартный критерий качества логических закономерностей и его оптимизация.
8. Генетические методы поиска, генетический алгоритм поиска логических закономерностей классов.
9. Веса признаков и прецедентов, логические корреляции. Минимизация признакового пространства.
10. Логические описания классов. Обработка множеств логических закономерностей. Минимальные и кратчайшие описания классов.
11. Алгоритмы распознавания, основанные на голосовании по системам логических закономерностей, построение устойчивых на обучении оценок.
12. Построение минимальных по сложности логических закономерностей классов.
13. Бинарные решающие деревья. Допустимые разбиения единичного куба.
14. Алгоритм построения допустимого разбиения, представление допустимого разбиения бинарным решающим деревом.
15. Прямые методы построения бинарных решающих деревьев, критерии ветвления.

**Литература**

1. Журавлев Ю.И. Избранные научные труды. М.: Магистр. 1998.
2. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации. Проблемы кибернетики. М.: Наука. 1978. Вып. 33. С.5-68.
3. Донской В.И., Башта А.И. Дискретные модели принятия решений при неполной информации. Симферополь: Таврия. 1992.
4. Рязанов В.В. Логические закономерности в задачах распознавания (параметрический подход) // Журнал вычислительной математики и математической физики. Т.47, № 10. 2007. С. 1793-1808.
5. Ковшов Н.В., Моисеев В.Л., Рязанов В.В. Алгоритмы поиска логических закономерностей в задачах распознавания // Журнал вычислительной математики и математической физики. 2008. Т.48, № 2. С. 329-344.

**Курс «Нестатистический анализ данных»**

**Рязанов В.В.**

**Аннотация**

В спецкурсе будут рассмотрены проблемы и методы кластерного анализа (подходы и алгоритмы кластеризации с известным и неизвестным числом кластеров, критерии кластеризации, вопросы устойчивости, построение оптимальных решений), дискретные методы для решения задач классификации множествами алгоритмов, новые подходы и алгоритмы в регрессионном анализе и анализе данных (решение задач восстановления зависимостей на основе решения задач классификации, задачи классификации с большим числом классов). Будут рассмотрены практические задачи классификации и поиска зависимостей по прецедентам, применения в медицине, бизнесе и технике.

**Программа**

1. Задачи кластерного анализа, меры близости, функции подобия.
2. Критерии качества кластеризации при заданном числе кластеров: дисперсионный и родственные критерии, основанные на матрицах рассеяния, след в качестве критерия, основанные на матрицах рассеяния, определитель матрицы внутригруппового рассеяния.
3. Критерии кластеризации при неизвестном числе кластеров, меры концентрации.
4. Итеративная оптимизация критериев кластеризации на примере дисперсионного критерия.
5. Алгоритмы иерархической группировки.
6. Критерии устойчивости кластеризаций и их вычисление.
7. Алгоритмы кластеризации, основанные на поиске центров сгущений (алгоритмы к-средних, нечетких к-средних, ФОРЕЛЬ, построения оптимальных покрытий).
8. Эвристические алгоритмы кластеризации, алгоритмы «к-эталонов» и «взаимного поглощения».
9. Восстановление компонент смеси по заданной обучающей выборке.
10. Нейросетевые алгоритмы обучения и самообучения. Метод встречного распространения, сеть Хопфильда, алгоритмы Хэбба.
11. Логические корректоры для решения задач классификации.
12. Решение задач кластерного анализа коллективами алгоритмов, метод коллективных к-средних.
13. Комитетный синтез коллективных кластеризаций, критерии качества коллективных кластеризаций.
14. Нахождение оптимальных коллективных решений задачи кластерного анализа.
15. Эвристические методы построения коллективных кластеризаций и практические алгоритмы.
16. Методы восстановления регрессионных зависимостей по прецедентам, основанные на решении задач распознавания и дискретной оптимизации.

**Курс «Метрические методы интеллектуального анализа данных»**

**Майсурадзе А.И.**

**Аннотация**

В курсе рассматриваются фундаментальные и прикладные задачи интеллектуального анализа данных, на которых иллюстрируется применение различных приемов и методов, базирующихся на понятиях сходства, близости, аналогии. Представлена теоретическая основа для построения, реализации и анализа широкого спектра методов, которые в настоящее время принято называть метрическими. Примеры из прикладных областей демонстрируют возникновение рассматриваемых идей или их использование. Многие примеры будут из области компьютерного зрения и обработки текстов.

Идея сходства свойственна человеческому мышлению, соответственно, для всех фундаментальных задач ИАД возник целый комплекс подходов и методов, опирающихся на эту идею. Специфика различных предметных областей добавляет к фундаментальным задачам новые требования или заставляет решать новые задачи. В курсе особое внимание уделено задачам идентификации, верификации, кластеризации. Рассмотрены методы построения и вычисления функций сходства, согласование сходства на различных множествах объектов, синтез новых способов сравнения объектов на базе уже имеющихся. Рассмотрен комплекс технологий, предназначенный для эффективного представления и обработки метрической информации вычислительными системами.

Проводятся аналогии с задачами, которые возникают в традиционной статистике многих переменных. Такое сопоставление позволяет одновременно улучшить освоение материала спецкурса и материала по статистике.

**Программа**

**Основные подходы к заданию сходства.** Функциональный подход: двуместные функции, удовлетворяющие аксиомам. Геометрический подход: определение в пространстве множеств точек. Табличный подход: матрицы попарного сходства над конечными множествами.

**Классическое определение метрики и метрического пространства.** Аксиоматическое задание метрики. Построение топологии по метрике. Пространства сходящихся последовательностей. Фундаментальные последовательности и полные пространства. Роль аксиомы треугольника и непрерывность метрики. Роль аксиомы сепарабельности и единственность предела сходящейся последовательности. Сопоставление метрик и отношений эквивалентности, 0,1-метрики. Различные модификации системы аксиом метрики и их интерпретация: расстояние, полуметрика, ультра-метрика, квази-метрика, неравенство Птолемея.

**Локальные метрики и их продолжение на всё пространство.** Формализация понятия «между» в метрическом пространстве. Выпуклость метрического пространства по Менгеру. Аксиомы существования и единственности точек между заданными точками. Аксиомы существования и единственности продолжения луча. Теорема о единственности продолжения локально совпадающих метрик. Практический пример проверки аксиом и использования локального продолжения метрики.

**Геометрические подмножества общих метрических пространств.** Понятия открытого и замкнутого шара, их согласованность с топологией метрического пространства. Понятия открытого и замкнутого обобщенного эллипсоида. Клетки Дирихле («сферы влияния»), автоматическое исправление ошибок. Геометрическое место точек, равноудаленных от заданных точек, проблема меры указанного подмножества. Понятие кривой в метрическом пространстве, длина кривой. Геодезическая линия, кривая наименьшей длины, сегмент. Свойство совпадения геодезических с множествами равноудаленных точек в обобщенных евклидовых пространствах.

**Примеры метрических пространств.** Пространство изолированных точек, дискретная топология. Метрики  (городских кварталов),  (евклидова),  (Чебышёва). Их физический смысл. Метрика  (Минковского). Форма шаров, вложенность единичных шаров. Зависимость объема шара от размерности пространства. Проблема сопоставления объема шаров в разных метриках с ростом размерности. Проблема единственности кратчайшего пути. Хаусдорфова метрика и другие метрики между подмножествами метрического пространства, индуцированные исходной метрикой между точками. Расстояния между функциями (графиками). Метрики на декартовом произведении метрических пространств. Случай конечного и бесконечного числа сомножителей, метрики на последовательностях.

**Классификация функций сходства.** Сопоставление значений: номинальные, порядковые, арифметические (интервальные, относительные, разностные, абсолютные) шкалы. Понятие о граничных объектах. Аксиомы сходства, главный и вспомогательный аргументы. Классификация мер сходства по одному свойству (признаку). Функции сходства на декартовом произведении пространств со значениями в различных шкалах.

**Характеристики метрик.** Инвариантность расстояния относительно сдвига, поворота. Инвариантность формы шаров относительно положения центра и направления на центр. Инвариантность объема шаров относительно положения центра и направления на центр. Ограниченность метрики. Ограниченность шаров. Понятие полностью абсолютных и полностью относительных метрик, промежуточные метрики. Выпуклость шаров. Односвязность шаров. Существование и единственность сегментов, непрерывность сегментов.

**Преобразования метрик.** Изометрические преобразования пространств. Преобразования функций, сохраняющие метрические свойства. Некоторые достаточные условия преобразований, сохраняющих метрические свойства. Ограничение значений метрики (range companders). Примеры универсальных компандеров. Возможность монотонного преобразования произвольной функции в метрику. Возможность линейного преобразования произвольной ограниченной функции в метрику. Нормализация метрик, зависимость от точки отсчета. Переход от булеанов конечных множеств к пространствам бинарных векторов, соответствие мощности множества и длины вектора.

**Реализация метрик.** Реализация конечных метрик точками ЛВП, точечные конфигурации. Алгоритмическая сложность решения задачи точного вложения в линейные пространства с метриками. Примеры МК, имеющих или не имеющих точную реализацию. Задача поиска оптимальной точечной конфигурации в пространстве малой размерности, методы метрического и неметрического многомерного шкалирования. Реализация многомерных данных элементами функциональных пространств. Методы визуализации многомерных данных: параллельные координатные оси, графики Эндрюса, шкалирование и иерархии, таблицы проекций, параметризованные глифы (звезды, лица Чернова).

**Принцип самоорганизации.** Принцип самоорганизации при построении эвристических информационных моделей. Понятие представителей, мера сходства между объектами и представителями. Функции представительства и назначений, структура метода. Самоорганизация в задаче кластеризации. Самоорганизация и задача факторного анализа, самоорганизация и задача дискриминантного анализа. Модификация прецедентной информации, понятие типологического дискриминантного анализа. Самоорганизация и задача восстановления пропусков.

**Метрики на конечных множествах.** Представление метрик таблицами попарных расстояний. Метрическая конфигурация (МК). Специальное линейное пространство метрических конфигураций. Система неравенств треугольника как определение полиэдрального конуса полуметрик. Грани и экстремальные лучи полуметрического конуса, проблема их определения. Векторное представление метрических конфигураций. Достаточные условия сохранения метрических свойств покомпонентными корректорами метрических конфигураций. Примеры использования достаточных условий. Несовместимость метрических свойств и ортогональности метрических конфигураций.

**Разложение МК по конечным системам МК.** Полные системы, базисы МК. Проблема использования переполненных систем МК. Гомогенные базисы, интерпретация коэффициентов разложения. Ранг МК. Ранговые и полуметрические ранговые базисы. Неполные системы, оптимальная аппроксимация МК. Разложение по системе «отдельных объектов», метрика попарных сумм, эффективное вычисление признака «общая удаленность» для индивидуальных объектов.

**Литература**

### Основная литература

1. Воронин Ю.А. Начала теории сходства. Новосибирск: Наука. СО. 1991.
2. Деза М., Лоран М. Геометрия разрезов и метрик. М.: МЦНМО. 2001.
3. Майсурадзе А.И. Гомогенные и ранговые базисы в пространствах метрических конфигураций // Ж. вычисл. матем. и матем. физ. (ЖВМиМФ). 2006. Т.46, № 2. С.344-361.
4. Basalaj W. Proximity Visualization of Abstract Data. Dissertation work. 2001.

### Дополнительная литература

1. Буземан Г. Геометрия геодезических. М.: Физматгиз. 1962.
2. Воронин Ю.А. Теория классификации и её приложения. Новосибирск: Наука. СО. 1985.
3. Дидэ Э. Методы анализа данных. М.: Финансы и статистика. 1985.
4. Дэйвисон М. Многомерное шкалирование. М.: Финансы и статистика. 1988.
5. Кочетков Д.В. О функциях близости. Сообщения по прикл. матем. ВЦ АН СССР. 1978.
6. Кочетков Д.В. Построение алгоритма вычисления расстояний для одного класса метрических пространств. Сообщения по прикл. матем. ВЦ АН СССР. 1978.
7. Майсурадзе А.И. О поиске оптимального коллективного слагаемого для набора метрических конфигураций // Искусственный интеллект (ИИ). 2006. №2. С.183-187.
8. Майсурадзе А.И. О свойствах оптимальных точечных конфигураций для одного семейства функционалов сравнения метрических конфигураций // ЖВМиМФ. 2005. Т. 45, № 9. С. 1741-1748.
9. Майсурадзе А.И. Об оптимальных разложениях конечных метрических конфигураций в задачах распознавания образов // ЖВМиМФ. 2004. Т. 44, № 9. С. 1697-1707.
10. Скворцов В.А. Примеры метрических пространств. М.: МЦНМО. 2002.
11. Тылкин М.Е. О геометрии Хэмминга единичных кубов // Доклады АН СССР. 1960. Т.134. С. 1037-1040.
12. Тылкин М.Е. О реализуемости матриц расстояний в единичных кубах // Проблемы кибернетики. 1962. Т. 7. С. 31-42.
13. Шрейдер Ю.А. Что такое расстояние? М.: Физматгиз. 1963.
14. Yianilos P.N. Normalized Forms for Two Common Metrics. Princeton: NEC Re-search Institute. 2002.

**Курс «Вероятностное тематическое моделирование»**

**Воронцов К.В.**

**Аннотация**

В спецкурсе изучаются методы построения вероятностных тематических моделей (topic modeling) коллекций текстовых документов. Развивается многокритериальный подход к решению некорректно поставленной задачи стохастического матричного разложения — аддитивная регуляризации тематических моделей. Особое внимание будет уделено комбинированию статистических и лингвистических методов анализа текстов. Рассматриваются прикладные задачи классификации и категоризации текстов, информационного поиска, персонализации и рекомендательных систем, а также задачи анализа и классификации дискретизированных биомедицинских сигналов. Предполагается проведение студентами численных экспериментов на модельных и реальных данных. Обзорная видео-лекция на странице семинара «[Стохастический анализ в задачах](http://www.mathnet.ru/php/conference.phtml?option_lang=rus&eventID=25&confid=394)», НМУ, 19 апреля 2014: <http://www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=8980>

[http://www.machinelearning.ru](http://www.machinelearning.ru/wiki/index.php?title=%D0%92%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%BD%D1%8B%D0%B5_%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B5_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D0%B8_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29): «Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

**Программа**

### Задачи анализа текстов. Вероятностные модели коллекций текстов

**Задачи классификации текстов.**

* Коллекция текстовых документов. Векторное представление документа.
* Эмпирические законы Ципфа, Ципфа-Мандельброта, Хипса.
* Постановка задачи классификации текстов. Объекты, признаки, классы, обучающая выборка.
* Частоты слов (терминов) как признаки. Распознавание текстов заданной тематики. Анализ тональности.
* Линейный классификатор. Наивный байесовский классификатор.
* Задача распознавание жанра текстов. Распознавание научных текстов. Примеры признаков.
* Задача категоризации текстов, сведение к последовательности задач классификации.

**Задачи предварительной обработки текстов.**

* Очистка, удаление номеров страниц (колонтитулов), переносов, опечаток, оглавлений, таблиц, рисунков, нетекстовой информации.
* Лемматизация и стемминг. Сравнение готовых инструментальных средств.
* Выделение и удаление стоп-слов и редких слов.

**Задачи информационного поиска.**

* Задача поиска документов по запросу. Инвертированный индекс.
* Меры сходства векторов частот. Косинусная мера сходства. Расстояние Хеллингера.
* Критерий текстовой релевантности TF-IDF. Вероятностная модель и вывод формулы TF-IDF.
* Задача ранжирования. Примеры признаков. Формирование асессорских обучающих выборок.

**Униграммная модель документов и коллекции.**

* Вероятностное пространство. Гипотезы «мешка слов» и «мешка документов». Текст как простая выборка, порождаемая вероятностным распределением. Векторное представление документа как эмпирическое распределение.
* Понятие параметрической порождающей модели. Принцип максимума правдоподобия.
* Униграммная модель документов и коллекции.
* *Ликбез.* Теорема Куна-Таккера.
* Аналитическое решение задачи о стационарной точке функции Лагранжа. Частотные оценки условных вероятностей.

**Литература:** [Маннинг 2011].

**Вероятностный латентный семантический анализ**

* *Напоминания:* Коллекция текстовых документов. Векторное представление документа. Задачи информационного поиска и классификации текстов.

**Мотивации вероятностного тематического моделирования**

* Идея перехода от вектора (терминов) к вектору тем.
* Цели тематического моделирования: поиск научной информации, агрегирование и анализ новостных потоков, формирование сжатых признаковых описаний документов для классификации и категоризации текстовых документов, обход проблем синонимии и омонимии.

**Задача тематического моделирования.**

* Вероятностное пространство. Тема как латентная (ненаблюдаемая) переменная. Гипотеза условной независимости. Порождающая модель документа как вероятностной смеси тем.
* Постановка обратной задачи восстановления параметров модели по данным.

**Вероятностный латентный семантический анализ (PLSA).**

* Частотные оценки условных вероятностей терминов тем и тем документов. Формула Байеса для апостериорной вероятности темы. Элементарное обоснование ЕМ-алгоритма.
* Принцип максимума правдоподобия, аналитическое решение задачи о стационарной точке функции Лагранжа, формулы M-шага.
* Рациональный ЕМ-алгоритм (встраивание Е-шага внутрь М-шага).

**Проведение экспериментов на модельных данных.**

* Процесс порождения терминов в документе. Генератор модельных (синтетических) данных. Генерация случайной величины из заданного дискретного распределения.
* Распределение Дирихле. Генерация разреженных и сглаженных векторов дискретных распределений из распределения Дирихле.
* Оценивание точности восстановления модельных данных. Расстояние между дискретными распределениями. Проблема перестановки тем, венгерский алгоритм.
* Проблема неединственности и неустойчивости матричного разложения. Экспериментальное оценивание устойчивости решения.

**Литература:** [Hofmann 1999].

**Модификации алгоритма обучения модели PLSA**

* *Напоминания.* Задача тематического моделирования коллекции текстовых документов. Модель PLSA, формулы Е-шага и М-шага.

**Латентное размещение Дирихле (LDA)**

* Сглаженные байесовские оценки условных вероятностей.

**Робастный ЕМ-алгоритм (REM).**

* Робастная модель с шумом и фоном.
* Упрощённая робастная модель.

**Стохастический ЕМ-алгоритм (SEM).**

* Гипотеза разреженности апоcтериорного распределения тем p(t|d,w).
* Алгоритм сэмплирования Гиббса.

**Онлайновый ЕМ-алгоритм (OEM).**

* Проблема больших данных.
* Эвристика разделения М-шага.
* Эвристика разделения коллекции на пачки документов.
* Добавление новых документов (folding-in).

**Способы формирования начальных приближений.**

* Случайная инициализация.
* Инициализация по документам.
* Поиск якорных слов. Алгоритм Ароры.

**Литература:** [Hoffman 2010, Asuncion 2009].

**Аддитивная регуляризация тематических моделей**

* *Напоминания*. Вероятностная тематическая модель. Принцип максимума правдоподобия. PLSA. EM-алгоритм.
* *Ликбез*. KL-дивергенция.

**Многокритериальная регуляризация.**

* Некорректность постановки задачи тематического моделирования.
* Аддитивная регуляризация.
* Общая формула M-шага для регуляризованного ЕМ-алгоритма.

**Регуляризатор разреживания.**

* Гипотеза разреженности распределений терминов тем и тем документов.
* Энтропийный регуляризатор и максимизация KL-дивергенции.
* Связь разреживания с L0-регуляризацией и методом разреживания нейронных сетей [Optimal Brain Damage](http://www.machinelearning.ru/wiki/index.php?title=OBD).
* Связь разреженности и единственности неотрицательного матричного разложения.

**Регуляризатор сглаживания.**

* Модель латентного размещения Дирихле LDA.
* Обоснование LDA через минимизацию KL–дивергенции. Виды сглаживающих распределений.
* Свойства распределения Дирихле, сопряжённость с мультиномиальным распределением.
* Байесовский вывод. Сглаженные частотные оценки условных вероятностей.
* Оценки максимума апостериорной вероятности.
* Численные методы оптимизации гиперпараметров.

**Комбинирование разреживания и сглаживания.**

* Разреживание предметных тем и сглаживание фоновых тем. Автоматическое выделение стоп-слов.
* Частичное обучение как выборочное сглаживание.

**Ковариационные регуляризаторы.**

* Антиковариация тем.
* Выявление корреляций между темами, модель CTM. Оценивание параметров модели (матрицы ковариаций).
* Корреляция документов.
* Тематические модели цитирования.

**Разреживание и сглаживание**

**Сглаживание**

* Сравнение LDA и PLSA. Экспериментальные факты: LDA скорее улучшает оценки редких слов, чем снижает переобучение.
* Дилемма разреживания и сглаживания.

**Частичное обучение (Semi-supervised EM).**

* Виды частично размеченных данных: привязка документа к темам, привязка термина к темам, нерелевантность, переранжирование списков терминов тем и тем документов, виртуальные документы.
* Использование частично размеченных данных для инициализации.
* Использование частично размеченных данных в качестве поправок на М-шаге ЕМ-алгоритма.

**Литература:** [Blei, 2003].

**Внутренние методы оценивания качества**

**Реальные данные.**

* Текстовые коллекции, библиотеки алгоритмов, источники информации.
* Внутренние и внешние критерии качества.
* Дополнительные данные для построения внешних критериев качества.

**Перплексия и правдоподобие.**

* Определение и интерпретация перплекcии.
* Перплексия контрольной коллекции. Проблема новых слов в контрольной коллекции.

**Когерентность.**

* Определение когерентности.
* Эксперименты, показывающие связь когерентности и интерпретируемости.
* Способы оценивания совместной встречаемости слов.

**Оценивание качества темы.**

* Контрастность темы (число типичных документов темы, число типичных терминов темы).
* Пиковость темы.
* Однородность (радиус) темы.
* Конфликтность темы (близость темы к другим темам).

**Статистические тесты условной независимости.**

* Методология проверки статистических гипотез. Критерий согласия хи-квадрат Пирсона. Матрица кросс-табуляции «термины–документы» для заданной темы.
* Проблема разреженности распределения. Эксперименты, показывающие неадекватность асимптотического распределения статистики хи-квадрат.
* Статистики модифицированного хи-квадрат, Кульбака-Лейблера, Хеллингера.
* Обобщённое семейство статистик Кресси-Рида.
* Алгоритм вычисления квантилей распределения статистики Кресси-Рида.
* Рекуррентное вычисление статистики Кресси-Рида.

**Литература:** [Newman, 2009–2011].

**Внешние методы оценивания качества**

**Оценивание интерпретируемости тематических моделей.**

* Корректность определения асессорами лишних терминов в темах и лишних тем в документах.
* Визуализация тематических моделей.

**Критерии качества классификации и ранжирования.**

* Полнота, точность и F-мера в задачах классификации и ранжирования.
* Критерии качества ранжирования: MAP, DCG, NDCG.
* Оценка качества тематического поиска документов по их длинным фрагментам.

**Литература:** [Blei, 2003].

**Робастные тематические модели**

*Робастность* — устойчивость модели к нарушениям исходных предпосылок, заложенных в основу модели.

**Робастная тематическая модель с фоном и шумом**

* Аналитическое решение задачи о стационарной точке функции Лагранжа, формулы M-шага.
* Аддитивный и мультипликативный М-шаг.
* Оценки тематичности слов.
* Эксперименты: робастная модель не нуждается в регуляризации и более устойчива к разреживанию.

**Разреженная робастная тематическая модель с шумом**

* Максимизация правдоподобия для упрощённой робастной модели.
* Вычисление перплексии для упрощённой робастной модели.

**Робастная тематическая модель с усечёнными распределениями**

* Явления синонимии, взаимной заменяемости терминов, эффект burstiness.
* Гипотеза об усечённых распределениях терминов тем в документах как ослабление гипотезы условной независимости.
* Аналитическое решение задачи о стационарной точке функции Лагранжа. Модификация ЕМ-алгоритма.

**Литература:** [Chemudugunta, 2006].

### Синтаксические тематические модели

**Энграммные модели.**

* Задача выделения терминов как ключевых фраз (словосочетаний). Словари терминов.
* Морфологический анализ текста.
* Синтаксический анализ текста. Выявление подчинительных связей.
* Статистические методы поиска коллокаций. Критерий C-Value.
* Совмещённый статистический критерий TF-IDF & CValue.
* Энграммный онлайновый алгоритм на основе синтаксического анализа и фильтрации терминов путём разреживания.
* Влияние выделения ключевых фраз на качество модели и интерпретируемость тем.

**Марковские модели синтаксиса.**

* Коллокации
* Оценивание матрицы переходных вероятностей.

**Регуляризация для задач классификации**

* *Напоминания*. Аддитивная регуляризация тематических моделей.

**Простейшие модели.**

* Примеры классов: годы, авторы, категории, и т.д.
* Моделирование классов темами.
* Моделирование классов распределениями тем.
* Автор-тематическая модель.
* Многоклассовые задачи. Частотный регуляризатор.

**Тематическая модель классификации.**

* Тематическая модель распределения классов документа. Вероятностная интерпретация.
* Тематическая модель цитирования документов.
* Тематическая модель цитирования авторов.
* Тематическая модель категоризации. Ковариационный регуляризатор.

**Динамические тематические модели**

**Модели с дискретным временем.**

* Модель с фиксированной тематикой.
* Модель с медленно меняющейся тематикой.

**Модели с непрерывным временем.**

**Иерархические тематические модели**

* Задачи категоризации текстов. Стандартный метод решения — сведение к последовательности задач классификации.

**Тематическая модель с фиксированной иерархией.**

* Вероятностная формализация отношения «тема–подтема». Тождества, связывающие распределения тем и подтем
* Задача построения иерархического тематического профиля документа.
* Задача построения одного уровня иерархии. Аналитическое решение задачи максимизации правдоподобия, формулы M-шага.
* Онлайновый иерархический EM-алгоритм.
* Необходимость частичного обучения для задачи категоризации.
* Необходимость разреживания для построения иерархического тематического профиля документа.

**Сетевые иерархические модели.**

* Возможность для темы иметь несколько родительских тем.
* Дивергенция Кульбака–Лейблера. Свойства KL-дивергенции.
* Интерпретация KL-дивергенции как степени вложенности распределений. Оценивание силы связей «тема-подтема» KL-дивергенцией.
* Дополнение тематического дерева до тематической сети.

**Иерархические процессы Дирихле.**

* Оптимизация числа тем в плоской модели.
* Создание новых тем в иерархических моделях.
* Нисходящие и восходящие иерархические модели.

### Многоязычные тематические модели

* Параллельные тексты.
* Сопоставимые тексты.
* Регуляризация матрицы переводов слов.

### Многомодальные тематические модели

* Коллаборативная фильтрация.
* Модель научной социальной сети.
* Персонализация рекламы в Интернете.

### Распараллеливание алгоритмов обучения тематических моделей

* Основы Map-Reduce
* Распределённое хранение коллекции.

**Литература**

**Основная литература**

1. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
2. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China.— 2010.— Vol. 4, no. 2. — pp. 280–301.
3. Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. — 2009.

**Дополнительная литература**

1. Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование 2012 Т. 4, №12. С 693–706.
2. Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'2014, Analysis of Images, Social networks and Texts. — Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2014, Vol. CCIS 439. pp. 28–45.
3. Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning Journal, Special Issue «Data Analysis and Intelligent Optimization», Springer, 2014.
4. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. —pp. 993–1022.
5. Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. — MIT Press, 2006. — Vol. 19. — pp. 241–248.
6. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society, Series B. — 1977. — no. 34. — pp. 1–38.
7. Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — pp. 50–57.
8. Hoffman M. D., Blei D. M., Bach F. R. Online Learning for Latent Dirichlet Allocation // NIPS, 2010. pp. 856–864.
9. Lu Y., Mei Q., Zhai C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA // Information Retrieval. — 2011. — Vol.14, no.2. — pp. 178–203.
10. Wallach H., Mimno D., McCallum A. Rethinking LDA: Why priors matter // Advances in Neural Information Processing Systems 22 / Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta. — 2009. — pp. 1973–1981.
11. Zavitsanos E., Paliouras G., Vouros G. A. Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // Journal of Machine Learning Research. — 2011. — Vol. 12. — pp. 2749–2775.