

**Федеральное государственное автономное образовательное учреждение
высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"**

Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

**Рабочая программа дисциплины «Упорядоченные множества для
анализа данных»**

для образовательной программы «Науки о данных»
направления подготовки 01.04.02. Прикладная математика и информатика
уровень - магистр

Разработчик программы
Кузнецов С.О., д.ф.-м.н, профессор, skuznetsov@hse.ru

Одобрена на заседании департамента анализа данных и искусственного интеллекта
«__»_____ 2015 г.

Руководитель департамента анализа данных и искусственного интеллекта Школы
С.О. Кузнецов _____

Утверждена Академическим советом образовательной программы
«__»_____ 2015 г., № протокола _____

Академический руководитель образовательной программы
С.О. Кузнецов _____

Москва, 2015

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

I. Пояснительная записка

Автор программы: Доктор физико-математических наук С.О. Кузнецов

Требования к студентам: Изучение курса «Упорядоченные множества для анализа данных» требует предварительных знаний по элементарной теории множеств и отношений, булевой алгебре и теории алгоритмов (курс «Дискретная математика»).

Аннотация. Дисциплина «Упорядоченные множества для анализа данных» предназначена для подготовки магистров по направлению 010500.68 (магистерская программа «математическое моделирование»)

Теория решеток замкнутых множеств и зависимостей предоставляет математические основы современных методов поиска зависимостей в данных – импликаций и ассоциативных правил на множествах признаков. Поиск ассоциативных правил находится в центре внимания методов разработки данных (data mining). Изложение курса начинается с повторения основных понятий теории отношений, теории графов, теории упорядоченных множеств и решеток. Одним из разделов современной теории решеток является анализ формальных понятий, исходным объектом которого служит бинарное отношение на множествах объектов и их свойств (признаков). На основе отношения определяется соответствие Галуа и оператор замыкания. Замкнутые множества объектов (признаков) образуют решетку (понятий), которая, с одной стороны, позволяет наглядно представлять иерархию классов объектов, а с другой – зависимости на признаках, определяемых в терминах импликаций и ассоциативных правил (частичных импликаций). Решетки формальных понятий дают удобный формализм для описания ряда моделей машинного обучения, таких как пространства версий, деревья решений, ДСМ-гипотезы, а также онтологий – современного средства представления знаний. Серьезным ограничением в применении решеток понятий является сложность вычислительных задач, связанных с алгоритмическим порождением решеток и базисов импликаций. В связи с этим важно изучение сложности задач и сложности различных алгоритмов порождения решеток и базисов импликаций.

Учебные задачи курса.

Данный курс позволит студентам овладеть математическими основами важнейшей области разработки данных (Data mining) - построения иерархий классов объектов, импликаций, ассоциативных правил и зависимостей других типов на признаках. Студенты получат навыками автоматического построения иерархическую модель предметной области, и находить зависимости в данных, а также анализировать алгоритмическую сложность такого рода задач и строить эффективные алгоритмы порождения иерархий классов объектов и систем зависимостей на множествах признаков объектов.

II. Тематический план курса "«Упорядоченные множества для анализа данных»"

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практика	
1	Введение. Отношения и графы	8	1	1	6
2	Порядки и графы	8	1	1	6
3	Решетки и полурешетки	12	2	2	8
4	Бинарные отношения и соответствия Галуа	12	2	2	8
5	Анализ формальных понятий (АФП)	12	2	2	8
6	Импlications и функциональные зависимости	10	2	2	6
7	Ассоциативные правила через соответствия Галуа и решетки понятий	10	2	2	6
8	Алгоритмические проблемы построения решеток замкнутых множеств и базисов импликаций	10	2	2	6
9	Кластеризация и устойчивость	10	2	2	6
10	Модели машинного обучения через соответствия Галуа и решетки понятий	12	2	2	8
11	Представление знаний с помощью решеток понятий	10	2	2	6
	Итого	190	20	20	74

Базовый учебник по курсу – ридер «Дискретные структуры», составленный по следующим источникам:

1. Биркгоф Г., Теория решеток. - М.: Наука, 1984. - 568 с.
2. Биркгоф Г., Барти Т., Современная прикладная алгебра, М., Лань, 2005 – 400 с.
3. Гретцер Г., Общая теория решеток. - М.: Мир, 1982. - 452 с.
4. В. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
5. Ф.Т. Алескеров, Э.Л. Хабина, Д.А. Шварц, Бинарные отношения, графы и коллективные решения, М., ГУ-ВШЭ, 2006.
6. А.А. Зыков, Основы теории графов, М., Наука, 1987.
7. О. Оре, Теория графов, М., Мир, 1965.
8. Ф. Харари, Теория графов, М., Мир, 1973.

Дополнительная литература по курсу

1. V. Duquenne and J.-L. Guigues, Familles minimales d'implications informatives resultant d'un tableau de donnees binaires, Math. Sci. Humaines, vol. 95, pp. 5-18, 1986.
2. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
3. S.O. Kuznetsov, On Computing the Size of a Lattice and Related Decision Problems, Order, 2001, vol. 18 (4), pp. 313-321.
4. S.O. Kuznetsov, S.A. Obiedkov, Comparing performance of algorithms for generating concept lattices, J. Exp. Theor. Artif. Intell., 2002, vol. 14, 2-3, pp. 189-216.
5. M. Luxenburger, Implications partielle dans un contexte, Math. Sci. Hum., 1991.
6. Кузнецов С.О. Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. 2001. - N 10. - с. 3-27.
7. T.S. Blyth, M.F. Janowitz, Residuation Theory, Pergamon Press, 1972.
8. P. Buitelaar, P. Cimiano, B. Magnini, Eds., Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.
9. C. Carpineto and G. Romano, Concept Data Analysis: Theory and Applications, Wiley, 2004.
10. B. Ganter, G. Stumme, R. Wille, Eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, State-of-the Art Series (2005), vol. 3626, pp. 196-225.
11. T. Mitchell, Machine Learning, Mc Graw Hill, 1997.
12. B. A. Davey and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 1990.

Формы контроля и структура итоговой оценки.

- 1 модуль: 0,3* домашнее задание + 0,7* контрольная
2 модуль: 0,2* работа на занятиях + 0,3* домашнее задание + 0,5* контрольная

Итоговая оценка: среднее по модулям

Программа курса

Упорядоченные множества в анализе данных.

Тема 1. Введение: обзор курса. Отношения и графы.

Бинарные отношения. Графы, подграфы, части, циклы, клики, деревья, двудольные графы. Графы бинарных отношений. Свойства бинарных отношений (рефлексивность, симметричность, асимметричность, антисимметричность, транзитивность, связность, ацикличность, полнота) и их теоретико-графовое выражение. Важные виды бинарных отношений: эквивалентность, толерантность, частичный порядок. Дополнительное отношение, обратное (дуальное) отношение, кодуальное отношение, симметрическое дополнение.

Свойства бинарных отношений: рефлексивность, транзитивность, симметричность, асимметричность, антисимметричность. Иллюстрация свойств на графе отношения.
Важные виды отношений: эквивалентность (классы эквивалентности), толерантность (классы толерантности),

Основная литература

1. Биркгоф Г., Барти Т., Современная прикладная алгебра, М., Лань, 2005 – 400 с.
2. Ф.Т. Алескеров, Э.Л. Хабина, Д.А. Шварц, Бинарные отношения, графы и коллективные решения, М., ГУ-ВШЭ, 2006.
3. А.А. Зыков, Основы теории графов, М., Наука, 1987.
4. О. Оре, Теория графов, М., Мир, 1965.
5. Ф. Харари, Теория графов, М., Мир, 1973.

Тема 2. Частично-упорядоченные множества и графы.

Частичный порядок, строгий порядок, квазипорядок, линейный порядок, отношение покрытия (доминирования), ориентированный граф порядка, диаграмма (Хассе) порядка. Частичный порядок как транзитивное замыкание отношения покрытия (также через произведение матриц), квазипорядок, отношение несравнимости, частичный порядок на элементах фактор-множества по отношению эквивалентности в квазипорядке. Примеры порядков в математике и приложениях: порядок на мультимножествах, порядок на разбиениях, (квази)порядок на помеченных (раскрашенных) графах. Размерность упорядоченного множества (порядковая и мультипликативная). Топологическая сортировка

Основная литература

1. Биркгоф Г., Теория решеток. - М.: Наука, 1984. - 568 с.
2. Биркгоф Г., Барти Т., Современная прикладная алгебра, М., Лань, 2005 – 400 с.
3. Гретцер Г., Общая теория решеток. - М.: Мир, 1982. - 452 с.
4. Ф.Т. Алескеров, Э.Л. Хабина, Д.А. Шварц, Бинарные отношения, графы и коллективные решения, М., ГУ-ВШЭ, 2006.
5. А.А. Зыков, Основы теории графов, М., Наука, 1987.
6. О. Оре, Теория графов, М., Мир, 1965.

Дополнительная литература

1. B. A. Davey and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 1990.
2. T.S. Blyth, M.F. Janowitz, Residuation Theory, Pergamon Press, 1972.

Тема 3. Решетки и полурешетки.

Инфимум, супремум, полурешетки, квазирешетки, два определения решеток. Диаграммы полурешеток решеток. Виды решеток (полные, модулярные, матроиды, дистрибутивные, булевы) и их диаграммы. (Порядковые) фильтры и идеалы решеток. Пополнения частичных порядков до решеток (пополнение Дедекинда-Макнила) и дистрибутивных решеток (Теорема Биркгофа).

Основная литература

1. Биркгоф Г., Теория решеток. - М.: Наука, 1984. - 568 с.
2. Биркгоф Г., Барти Т., Современная прикладная алгебра, М., Лань, 2005 – 400 с.
3. Гретцер Г., Общая теория решеток. - М.: Мир, 1982. - 452 с.
4. О. Оре, Теория графов, М., Мир, 1965.

Дополнительная литература

1. B. A. Davey and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 1990.
2. T.S. Blyth, M.F. Janowitz, Residuation Theory, Pergamon Press, 1972.

Тема 4. Бинарные отношения и соответствия Галуа.

Соответствия Галуа и их свойства. Соответствие Галуа, основанное на бинарном отношении. Оператор замыкания и система замыканий (семейство Мура). Замкнутые множества, решетка замкнутых множеств.

Основная литература

1. Биркгоф Г., Теория решеток. - М.: Наука, 1984. - 568 с.
2. Биркгоф Г., Барти Т., Современная прикладная алгебра, М., Лань, 2005 – 400 с.
3. Гретцер Г., Общая теория решеток. - М.: Мир, 1982. - 452 с.
4. О. Оре, Теория графов, М., Мир, 1965.

Дополнительная литература

1. B. A. Davey and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 1990.
2. T.S. Blyth, M.F. Janowitz, Residuation Theory, Pergamon Press, 1972.

Тема 5. Анализ формальных понятий (АФП).

Формальный контекст, формальное понятие, частичный порядок на формальных понятиях, решетка формальных понятий. Супремум и инфимум-неразложимые элементы решетки. Основная теорема АФП (Р. Вилле) о представимости полной решетки решеткой формальных понятий. Число формальных понятий контекста. Характеризация решеток через бинарное отношение. Отношение «стрелка». Характеризация дистрибутивных решеток через отношения «стрелок». Многочленные контексты, шкалирование.

Основная литература

1. B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.

Дополнительная литература

1. B. A. Davey and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 1990.
2. T.S. Blyth, M.F. Janowitz, Residuation Theory, Pergamon Press, 1972.

Тема 6. Импликации и зависимости.

Системы импликаций, правила Армстронга, связь с функциональными зависимостями. Базисы импликаций: прямой базис, минимальный базис (Дюкенна-Гига). Псевдосодержания: определения Дюкенна-Гига и Гантера. Характеризация типов решеток

по виду импликаций в минимальном базисе (дистрибутивность, к-дистрибутивность, и т.д.). Размеры базисов.

Основная литература

1. B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.

Дополнительная литература

1. B. A. Davey and H. A. Priestley, Introduction to Lattices and Order, Cambridge University Press, 1990.

1. V. Duquenne and J.-L. Guigues, Familles minimales d'implications informatives resultant d'un tableau de donnees binaires, Math. Sci. Humaines, vol. 95, pp. 5-18, 1986.

Тема 7. Ассоциативные правила.

Ассоциативные правила в разработке данных (Data mining), их поддержка (support) и степень уверенность (confidence). Ассоциативные правила и решетки формальных понятий. Базис Люксембургера для ассоциативных правил. Базис, основанный на основном дереве диаграммы решетки понятий.

Основная литература

1. B. Ganter, G. Stumme, R. Wille, Eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, State-of-the Art Series (2005), vol. 3626, pp. 196-225.

2. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

5. M. Luxenburger, Implications partielle dans un contexte, Math. Sci. Hum., 1991.

Тема 8. Алгоритмические проблемы построения решеток замкнутых множеств и базисов импликаций.

Теоретические оценки временной сложности в худшем случае. Классы сложности P, NP, co-NP и #P. #P-полнота задач подсчета размера решетки замкнутых множеств и размера минимального базиса. NP-полнота некоторых задач о понятиях: Задача определения псевдозамкнутости и co-NP.

Алгоритмы построения решеток: Норриса, Гантера, Замыкай-по-Одному, Нурина и др. Алгоритмы построения минимального базиса импликаций и базиса ассоциативных правил. Программная система ConExp построения решеток понятий, базисов импликаций и ассоциативных правил.

Основная литература

1. B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.

2. S.O. Kuznetsov, On Computing the Size of a Lattice and Related Decision Problems, Order, 2001, vol. 18 (4), pp. 313-321.

3. S.O. Kuznetsov, S.A. Obiedkov, Comparing performance of algorithms for generating concept lattices, J. Exp. Theor. Artif. Intell., 2002, vol. 14, 2-3, pp. 189-216.

4. М. Гэри, Д. Джонсон, Вычислительные машины и труднорешаемые задачи, М., Мир, 1982

Тема 9. Кластеризация и устойчивость понятий

Классические методы кластеризации, основанные на отношении и метриках сходства. Определение кластера как замкнутого множества объектов с «большим» общим числом признаков. Устойчивость понятия как мера качества кластера. Уровневые и интегральный индексы устойчивости. Устойчивость и дисперсия. Устойчивость и импликации. Устойчивость и свойства решетки понятий. Соотношение между уровневыми индексами устойчивости. Динамика устойчивости при росте числа примеров. Трудновычислимость устойчивости. Алгоритм с полиномиальной задержкой для вычисления индексов устойчивости. Приближенное вычисление устойчивости. Устойчивость в анализе сообществ.

Основная литература

1. B. Ganter, G. Stumme, R. Wille, Eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, State-of-the Art Series (2005), vol. 3626, pp. 196-225.
2. S.O. Kuznetsov, Stability of a Formal Concept, Proc. 4th Journee d'Informatique Messine (JIM'03), E. San-Juan, Ed., Metz, 2003.
3. Кузнецов С.О. Устойчивость как оценка обоснованности гипотез, получаемых на основе операционального сходства// НТИ. Сер.2 - 1990. - N12. - С.21-29.

Тема 10. Модели машинного обучения через соответствия Галуа и решетки понятий.

Пространство версий через соответствия Галуа. Пространства версий с полурешеточным упорядочением классификаторов. ДСМ-метод порождения гипотез, гипотезы как содержания решетки понятий положительного контекста. Импликации и ДСМ-гипотезы. Гипотезы и пространства версий. Деревья решений и их погружение в решетку полупроизведения шкал. Узорные структуры и их проекции, обучение на узорных структурах. Импликации и ассоциативные правила на узорных структурах.

Основная литература

1. Кузнецов С.О. Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. 2001. - N 10. - с. 3-27.
2. B. Ganter and S.O. Kuznetsov, Pattern Structures and Their Projections, Proc. 9th Int. Conf. on Conceptual Structures, ICCS'01, G. Stumme and H. Delugach, Eds., Lecture Notes in Artificial Intelligence, vol. 2120 (2001), pp.129-142.
3. S.O. Kuznetsov, Machine Learning and Formal Concept Analysis, Proc. 2nd Int. Conf. on Formal Concept Analysis, ICFCA'04, P. Eklund, Ed., Lecture Notes in Artificial Intelligence, vol. 2961 (2004), pp. 287-312.
4. B. Ganter and S.O. Kuznetsov, Hypotheses and Version Spaces, Proc. 10th Int. Conf. on Conceptual Structures, ICCS'03, A. de Moor, W. Lex, and B.Ganter, Eds., Lecture Notes in Artificial Intelligence, vol. 2746 (2003), pp. 83-95.

Тема 11. Представление знаний с помощью решеток понятий.

Решетки понятий как средство для построения таксономий и мерономий (системы классов, связанных отношением «быть частью»).

Определения онтологий. Онтология как частично-упорядоченное множество с дополнительным отношением на элементах. Программные средства построения

онтологий. Автоматическое построение онтологий по объектно-признаковым таблицам как решеток понятий.

Основная литература

1. B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
2. P. Buitelaar, P. Cimiano, B. Magnini, Eds., Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.
3. B. Ganter, G. Stumme, Creation and Merging Ontology Top-levels, Proc. 13th Int. Conf. on Conceptual Structures, ICCS'06, P. Hitzler, F. Sharfe, Eds., Lecture Notes in Artificial Intelligence, (2006).
4. N.F. Noy, R. Fergerson, M. Musen, The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility, Proc. EKAW 2000, LNCS 1937, Springer, Heidelberg 2000, pp.17-32.

Тематика заданий по различным формам текущего контроля:

1. Соотношение между графовым и табличным заданием отношений
2. Соотношение между заданием частичных порядков с помощью графов, диаграмм и таблиц.
3. Размерности порядков
4. Свойства, выполняющиеся в различных типах решеток,
5. Решетки понятий контекстов и основная теорема анализа формальных понятий
6. Взаимная переводимость импликаций в контекстах и функциональных зависимостей в реляционных базах данных
7. Ассоциативные правила через решетки понятий, базисы правил через остовные деревья диаграммы
8. Полиномиальная задержка эффективных алгоритмов вычисления множества всех понятий и их решеток.
9. Эффективные алгоритмы вычисления устойчивости
10. Пространства версий для классификаторов, задаваемых узорными структурами.
11. Онтологии как решетки с дополнительными отношениями на элементах

Вопросы для оценки качества освоения дисциплины

Тема 1. Для отношения, заданного следующей таблицей, определить, является ли оно

	a	b	c	d
a	x	x	x	
b		x	x	
c	x			
d			x	

симметричным,
асимметричным,
антисимметричным,

рефлексивным,
 транзитивным,
 и если нет, то объяснить почему.

Тема 2. Для отношения частичного порядка, заданного следующей таблицей

	1	2	3	4	5	6	7
1	x		x	x	x	x	x
2		x	x	x	x	x	x
3			x	x	x	x	x
4				x	x	x	x
5					x		
6						x	
7							x

- а) построить ориентированный граф и диаграмму
- б) определить размерность частичного порядка двумя способами (порядковым и мультипликативным)

Тема 3. Доказать, что для любых элементов решетки имеет место неравенство

$$x \wedge (y \vee z) \geq (x \wedge y) \vee (x \wedge z).$$

Темы 4, 5. Для контекста, представленного таблицей

	a	b	c	d
1	x	x	x	
2		x	x	
3	x			
4			x	

- а) построить решетку понятий
- б) определить (объяснив ответ), имеют ли место признаковые импликации $ac \rightarrow b$, $cb \rightarrow a$, $bd \rightarrow c$
- в) привести еще как минимум три нетривиальные импликации, выполняющиеся в контексте (импликация $A \rightarrow B$ называется тривиальной если $B \subseteq A$).

Темы 5, 6

1. Для контекста, представленного таблицей

	a	b	c	d
1		x	x	x
2	x		x	x
3	x	x		x
4	x	x		
5			x	x

Построить множество всех понятий, диаграмму решетки понятий
 минимальный базис импликаций (базис Дюкенна-Гига), прямой базис.

2. Для множества импликаций $a \rightarrow b$, $b \rightarrow cd$, $d \rightarrow e$ построить контекст с множеством признаков $\{a,b,c,d,e\}$, в котором выполняются только эти импликации (и те, которые следуют по ним по правилам Армстронга).

3. По многозначному контексту

	a	b	c	d
1	r	s	t	t
2	s	r	t	t
3	s	r	s	s
4	t	t	r	r

построить бинарный контекст, в котором импликации синтаксически совпадают с функциональными зависимостями в исходном многозначном контексте.

4. По контексту, представленному таблицей

	a	b	c	d
1	x	x	x	
2		x	x	
3	x			
4			x	

построить многозначный контекст, для которого множество функциональных зависимостей синтаксически совпадает с множеством импликаций в исходном контексте, с использованием всех значений из множества натуральных чисел от 1 до 7.

Тема 7. По контексту, представленному таблицей

	a	b	c	d
1		x	x	x
2	x		x	x
3	x	x		x
4	x	x		
5			x	x

построить все ассоциативные правила с поддержкой не менее $1/3$ и степенью уверенности не менее $1/2$

Тема 8.

Какова временная сложность в худшем случае следующей задачи:

УСЛОВИЕ Дан контекст (G, M, I) , множество $N \subseteq M$ и натуральное число k

ВОПРОС Существует ли подмножество $S \subseteq N$ такое, что $S' = N$ и $|N| \leq k$

Сколько операций пересечения произведет алгоритм Замыкай-по-Одному, вычисляя множество всех понятий контекста представленного следующей таблицей?

	a	b	c	d
1		x	x	x
2	x		x	x

3	x	x		x
4	x	x		
5			x	x

Тема 9. Для контекста, представленного таблицей

	a	b	c	d
1		x	x	x
2	x		x	x
3	x	x		x
4	x	x		
5			x	x

определить понятия с максимальной интегральной устойчивостью.

Тема 10.

1. Считая, что в следующей таблице объекты 1-5 – положительные примеры, а объекты 6, 7 – отрицательные,

	a	b	c	d
1		x	x	x
2	x		x	x
3	x	x		x
4	x	x		
5			x	x
6	x	x	x	
7		x	x	

построив решетку положительного контекста, найти множество положительных ДСМ-гипотез с запретом на контрпример, а также построить множество отрицательных ДСМ-гипотез с запретом на контрпример.

Тема 11.

Построить онтологию области «классы химических элементов», на основе описания с помощью бинарных признаков, с помощью алгоритма «Замыкай-по-Одному».

Автор программы: _____ / С.О. Кузнецов/