

Программа учебной дисциплины «Рекомендательные системы и поиск закономерностей в данных»

Утверждена
Академическим советом ООП
Протокол № 2.3-09/ 2706-01 от «27» июня 2018г.

Автор	Игнатов Дмитрий Игоревич
Число кредитов	6
Контактная работа (час.)	68
Самостоятельная работа (час.)	160
Курс	Рекомендательные системы и поиск закономерностей в данных
Формат изучения дисциплины	без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Рекомендательные системы и поиск закономерностей в данных» являются овладение студентами основными методами рекомендательных систем и подходами к поиску закономерностей в данных, например, таких как базы транзакций и последовательностей.

В результате освоения дисциплины студент должен:

знать:

- основные типы рекомендательных систем и алгоритмические подходы к решению задачи формирования рекомендаций
- методы майнинга данных и поиска закономерностей в различных типах данных, включая транзакции и последовательности событий и транзакций;
- основные методы проверки качества рекомендаций и достоверности выявленных закономерностей

уметь:

- адекватно определять и применять методы рекомендательных систем в зависимости от типа задачи;
- оценивать качество и степень достоверности результатов, полученных с помощью методов майнинга данных и поиска закономерностей в данных;
- работать с современным программным обеспечением и языками программирования для создания и поддержки методов рекомендательных систем и майнинга данных;
- определять возможности применения теоретических концепций из теории множеств, дискретной математики и теории графов, теории вероятности и математической статистики, теории оптимизации для постановки и решения конкретных задач поиска закономерностей в данных и рекомендательных систем;
- ориентироваться в потоке научной информации для использования новых подходов в данной предметной области;

владеть:

- навыками использования методов линейной алгебры, теории вероятностей, математической статистики, теории оптимизации, дискретной математики и теории графов в практических приложениях рекомендательных систем и поиска закономерностей в данных;

- навыками обработки и интерпретации результатов компьютерных экспериментов с реальными и синтетическими данными;
- навыками использования стандартных методов и моделей рекомендательных систем и поиска закономерностей в данных и их применения к решению конкретных практических задач.

Изучение дисциплины «**Рекомендательные системы и поиск закономерностей в данных**» основано на следующих дисциплинах:

- линейная алгебра;
- методы оптимизации;
- теория вероятности и математическая статистика;
- дискретная математика и теория графов;
- теория алгоритмов.

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- знать основные понятия линейной алгебры (вектор, матрица, скалярное произведение, собственный вектор, норма вектора и др.), методы оптимизации (производная и градиент, условия первого порядка, множители Лагранжа), теории вероятностей (распределение случайной величины, математическое ожидание, среднее и дисперсия и др.), дискретной математики и теории графов (операции с множествами, мощность множества, последовательность, двудольный граф и др.), теории алгоритмов (вычислительная сложность, O-нотация и др.).
- знать простейшие методы решения задач линейной алгебры (поиск решения системы линейных уравнений и собственных чисел) и оптимизации (поиск экстремумов при ограничениях);
- обладать навыками работы с программными средствами анализа данных (предобработка данных, настройка параметров моделей и т.п.).

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение в рекомендательные системы (РС). Обзор курса.

В рамках первой вводной темы дается представление о существующих типах рекомендательных системах и возникающих задачах. Методы майнинга данных и поиска закономерной данных лишь упоминаются в контексте задач рекомендательных систем и приводятся в плане изучения дисциплины.

Тема 2. Коллаборативная фильтрация на основе сходства по пользователям и предметам (user-based и item-based подходы).

В рамках этой темы предлагается знакомство с методами коллаборативной фильтрации, одним из самых распространенных методов рекомендательных систем. Входным типом данных, как правило является матрица оценок, в случае явного отклика, или матрица транзакций, в случае неявного отклика клиента. Расстраиваются исторически первые методы формирования рекомендаций на основе сходства по пользователям и предметам рекомендации, а также различные меры сходства на основе расстояния Евклида и Хемминга, а также корреляция Пирсона и косинусное сходство и др. Закрепление темы проходит в рамках Case-study 1.

Case-study 1. Оценка качества РС на основе бимодальной кроссвалидации.

В ходе тематического исследования на примере данных MovieLens рассказывается о применении двух методов коллаборативной фильтрации на основе сходства по пользователям и предметов рекомендации. Особое внимание уделяется оценке их качества на основе бимодального скользящего контроля и доопределенных мерах качества Precision и Recall при варьировании числа соседей, размере выдачи и параметров метода оценки.

Тема 3. Частые множества и ассоциативные правила. Элементы Анализа формальных понятий. Импликация и ассоциативные правила. Основные алгоритмы (Apriori, FP-growth). Меры интересности (support, confidence, lift, stability). Компактное представление частых множеств и ассоциативных правил (замкнутые множества, понятие базиса).

В этой теме дается знакомство с ключевыми методами поиска закономерностей в данных на примере задачи поиска частопокупаемых товаров при анализе покупательских корзин. Особое внимание уделено признаковым зависимостям, записываемым в виде правил вида $A \rightarrow B$, называемых ассоциативными правилами. Прикладные аспекты данных методов раскрываются в рамках Case-study 2.

Case-study 2. Анализ посещаемости сайтов. Рекомендация контекстной рекламы.

В рамках тематического исследования наглядно показано как решетки частых множеств и ассоциативные правила могут быть применены в задаче изучения аудитории целевого-вебсайта. На примере задачи контекстной рекламы рассказывается не только о применении ассоциативных правил для формирования рекомендаций, но и об использовании морфологии рекламных словосочетания для порождения альтернативных рекомендательных метаправил. Рассказывается о пользе таксономий и онтологий в таких задач. Уделено внимание оценке качества правил на основе процедуры скользящего контроля.

Тема 4. Методы на основе матричной факторизации для рекомендательных систем. PureSVD, SVD++, timeSVD. Схемы решения: стохастический градиентный спуск (SGD) и чередующиеся наименьшие квадраты (ALS). Случай неявного отклика. Факторизационные машины.

Case-study 3. Булева матричная факторизация, неотрицательная матричная факторизация (NMF) и разложение по сингулярным числам (SVD).

Тема 5. Спектральная кластеризация. Поиск минимального разреза. Контекстная реклама. Рекомендация музыкальных композиций.

Тема 6. Поиск частых последовательностей. Case-study 4. Анализ демографических последовательностей.

Тема 7. Гибридные рекомендательные системы. Case-study 5. Рекомендация радиостанций.

Тема 8. Бикластеризация. Фолксономии. Трикластеризация и мультимодальная кластеризация парных отношений и тензоров.

Тема 9. Стандартные (*Precision, Recall, F_1-мера, MAE, RMSE*) и дополнительные меры оценки качества рекомендательных систем (*HitRate, Mean Reciprocal Rank, nDCG, diversity, serendipity*). A/B тестирование.

Тема 10. Контекстные рекомендательные системы. Встраивание дополнительной информации в модели.

Тема 11. Ансамбли рекомендательных алгоритмов.

Тема 12. Глубинное обучение для рекомендательных систем.

III. ОЦЕНИВАНИЕ

Формула итоговой оценки

Упрощенная формула:

$$0,4 * [\text{Домашняя работа}] + 0,4 * [\text{проект}] + 0,2 * [\text{защита}]$$

Результирующая оценка по дисциплине рассчитывается по формуле:

$$O_{\text{итог}} = 0,8 * O_{\text{накопл}} + 0,2 * O_{\text{экс/защита}}$$

Накопленная и итоговая оценки, участвующие в этой формуле, округляются арифметически.

Накопленная оценка рассчитывается по формуле:

$$O_{\text{накопл}} = 0,5 * O_{\text{дз}} + 0,5 * O_{\text{проект}}$$

Оценка за домашние задания рассчитывается как среднее значение оценок за все выданные домашние задания.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля студента

Домашняя работа

Домашнее задание 1

[Коллаборативная фильтрация на основе сходства по пользователям \(user-based\) и продуктам \(item-based\)](#)

Домашнее задание 2 (одна из тем)

1. [Частые множества и ассоциативные правила](#)
2. [Спектральная кластеризация](#)
3. [Матричная факторизация](#)
4. [Анализ последовательностей](#)

Оценочные средства для промежуточной аттестации

Проект

Требования к оформлению

В файле по ссылке содержатся требования к отчету. Даны источники данных, указаны типы задач для решения, приведено релевантное программное обеспечение. Указаны сроки выбора задачи и сдачи отчета. Предполагается экзамен в виде защиты проекта.

V. РЕСУРСЫ

5.1 Основная литература

1. Charu C. Aggarwal. Recommender Systems: The Textbook, Springer, April 2016 – Режим доступа: <https://www.springer.com/gp/book/9783319296579>
2. Recommender Systems Handbook. Francesco Ricci, Lior Rokach, Bracha Shapira (Eds.), Springer, 2015 – Режим доступа: <https://www.springer.com/la/book/9781489976369>

5.2 Дополнительная литература

1. Dmitry I. Ignatov, Andrey V. Konstantinov, Sergey I. Nikolenko, Jonas Poelmans, Vasily Zaharchuk: Online Recommender System for Radio Station Hosting. BIR 2012: 1-12
2. Dmitry I. Ignatov, Sergey I. Nikolenko, Taimuraz Abaev, Jonas Poelmans: Online recommender system for radio station hosting based on information fusion and adaptive tag-aware profiling. Expert Syst. Appl. 55: 546-558 (2016) [статья](#)
3. Dmitry I. Ignatov, Sergey I. Nikolenko, Taimuraz Abaev, Natalia Konstantinova: Online Recommender System for Radio Station Hosting: Experimental Results Revisited. WI-IAT (2) 2014: 229-236 [слайды](#)
4. Vladimir Bobrikov, Elena Nenova, Dmitry I. Ignatov: What is a Fair Value of Your Recommendation List? EEML@CLA 2016: 1-12 [Статья](#)
5. Dmitry I. Ignatov, Dmitry V. Gnatyshak, Sergei O. Kuznetsov, Boris G. Mirkin: Triadic Formal Concept Analysis and triclustering: searching for optimal patterns. Machine Learning 101(1-3): 271-302 (2015) [статья](#)
6. Faris Alqadah, Chandan K. Reddy, Junling Hu, Hatim F. Alqadah: Biclustering neighborhood-based collaborative filtering method for top-n recommender systems. Knowl. Inf. Syst. 44(2): 475-491 (2015) [статья](#)
7. Ignatov, D. I., Kuznetsov, S. O., Poelmans, J., & Zhukov, L. E. (2013). Can triconcepts become triclusters? International Journal of General Systems, 42(6), 572–593. [статья](#)
8. Ignatov, D. I., Kuznetsov, S. O., & Poelmans, J. (2012). Concept-based biclustering for internet advertisement. In: IEEE computer society ICDM workshops, pp. 123–130. [статья](#)
9. Nanopoulos, A., Rafailidis, D., Symeonidis, P., & Manolopoulos, Y. (2010). Musicbox: Personalized music recommendation based on cubic analysis of social tags. IEEE Transactions on Audio, Speech & Language Processing, 18(2), 407–412. [статья](#)
10. Cerf, L., Besson, J., Robardet, C., & Boulicaut, J. F. (2009). Closed patterns meet n-ary relations. ACM Transactions on Knowledge Discovery from Data, 3, 3:1–3:36. [статья](#)

11. Symeonidis, P., Nanopoulos, A., Papadopoulos, A. N., & Manolopoulos, Y. (2008). Nearest-biclusters collaborative filtering based on constant and coherent values. *Information Retrieval*, 11(1), 51–75. [статья](#)
12. du Boucher-Ryan, P., & Bridge, D. G. (2006). Collaborative recommending using formal concept analysis. *Knowledge-Based Systems*, 19(5), 309–315. [статья](#)
13. Marat Akhmaturov, Dmitry I. Ignatov: Context-Aware Recommender System Based on Boolean Matrix Factorisation. CLA 2015: 99-110 [слайды](#) [статья](#)
14. Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, Alexander Tuzhilin: Context-Aware Recommender Systems. *AI Magazine*, Vol 32, No 3, 2011 [статья](#)
15. Elena Andreeva, Dmitry I. Ignatov, Artem Grachev, Andrey Savchenko: Extraction of Visual Features for Recommendation of Products via Deep Learning. AIST 2018 [статья](#) [слайды](#)
16. [Deep Learning for RecSys survey](#)

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	<i>Из внутренней сети университета (договор)</i>
2.	Microsoft Office Professional Plus 2010	<i>Из внутренней сети университета (договор)</i>
3.	Язык программирования Python версии 3.6 и выше с библиотеками numpy, pandas, scikit-learn, matplotlib и т.д.	<i>Свободно распространяемое лицензионное соглашение</i>
4.	Библиотека рекомендательных алгоритмов surprise.	http://surpriselib.com (Свободно распространяемое лицензионное соглашение)
5.	Библиотека рекомендательных алгоритмов polara	https://github.com/Evfro/polara (Свободно распространяемое лицензионное соглашение)

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	<i>Профессиональные базы данных, информационно-справочные системы</i>	
1.	Электронные библиотеки: Elsevier, Springer, ACM	<i>Из внутренней сети университета (договор)</i>
	<i>Интернет-ресурсы (электронные образовательные ресурсы)</i>	
1.	Портал machinelearning.ru	URL: https://machinelearning.ru/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических презентаций, соответствующих программе дисциплины в составе:

- компьютер с доступом в Интернет (операционная система, офисные программы, антивирусные программы) и предустановленным программным обеспечением;
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для практических и самостоятельных занятий по дисциплине оснащены компьютерами и предустановленным программным обеспечением, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.