Прикладные задачи анализа данных

Игнатов Дмитрий Игоревич и Черняк Екатерина Леонидовна

Национальный исследовательский университет Высшая школа экономики
Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

2017

План

1. Программа курса
   - Оценка по курсу
2. Системы и библиотеки ML&DM
3. Чего бы почитать и посмотреть?

План лекции

1. Программа курса
   - Оценка по курсу
2. Системы и библиотеки ML&DM
3. Чего бы почитать и посмотреть?

Примерная программа курса

1. Введение
2. Введение в автоматическую обработку текстов (text mining) ✓
3. Частые множества признаков (frequent itemsets) и ассоциативные правила
4. Анализ последовательностей (sequence mining)
5. Анализ формальных понятий и его приложения
6. Рекомендательные системы и алгоритмы
7. Бикластеризация. Мультимодальная кластеризация
8. Понижение размерности (SVD, BMF, NMF)
9. Спектральная кластеризация
10. Анализ связей (алгоритм Page Rank)

Text mining. План

Е. Черняк

1. Введение в автоматическую обработку текстов
2. Введение в информационный поиск
3. Вероятностное тематическое моделирование как кластеризация текстов
4. Классификация текстов по теме и по тональности
5. Использование методов классификации последовательностей (sequence labelling) в задачах извлечения информации
6. Методы определения семантической близости (word2vec)

Прикландные задачи

Д. Игнатов

1. Поиск сходства текстовых документов на основе частых множеств
2. Рекомендация контекстной рекламы (частые множества, АФП, ассоциативные правила и бикластеризация)
3. Анализ посещаемости сайтов (АФП)
4. Рекомендация фильмов
5. Гибридные рекомедации радиостанций
6. Поиск сообществ в соцсетях (на основе спектральной кластеризации и бикластеризации)
7. Анализ демографических последовательностей

Итоговая оценка

Типовой сценарий

- Домашние задания
- Проект (индивидуальный или групповой)
- Экзамен (защита проекта)

Итоговая оценка

Формула итоговой оценки (ИО)

$$\text{ИО} = 0.5 \cdot \text{ДЗ} + 0.1 \cdot \text{ТЗ} + 0.2 \cdot \text{ПЗ} + 0.2 \cdot \text{ПП}$$

- ДЗ – домашние задания
- ТЗ – техническое задание к проекту (индивидуальный или групповой (≤ 3 участников))
- ПЗ – пояснительная записка (письменный отчет по проекту)
- ПП – презентация проекта (защита проекта)

Поиск паттернов/зависимостей

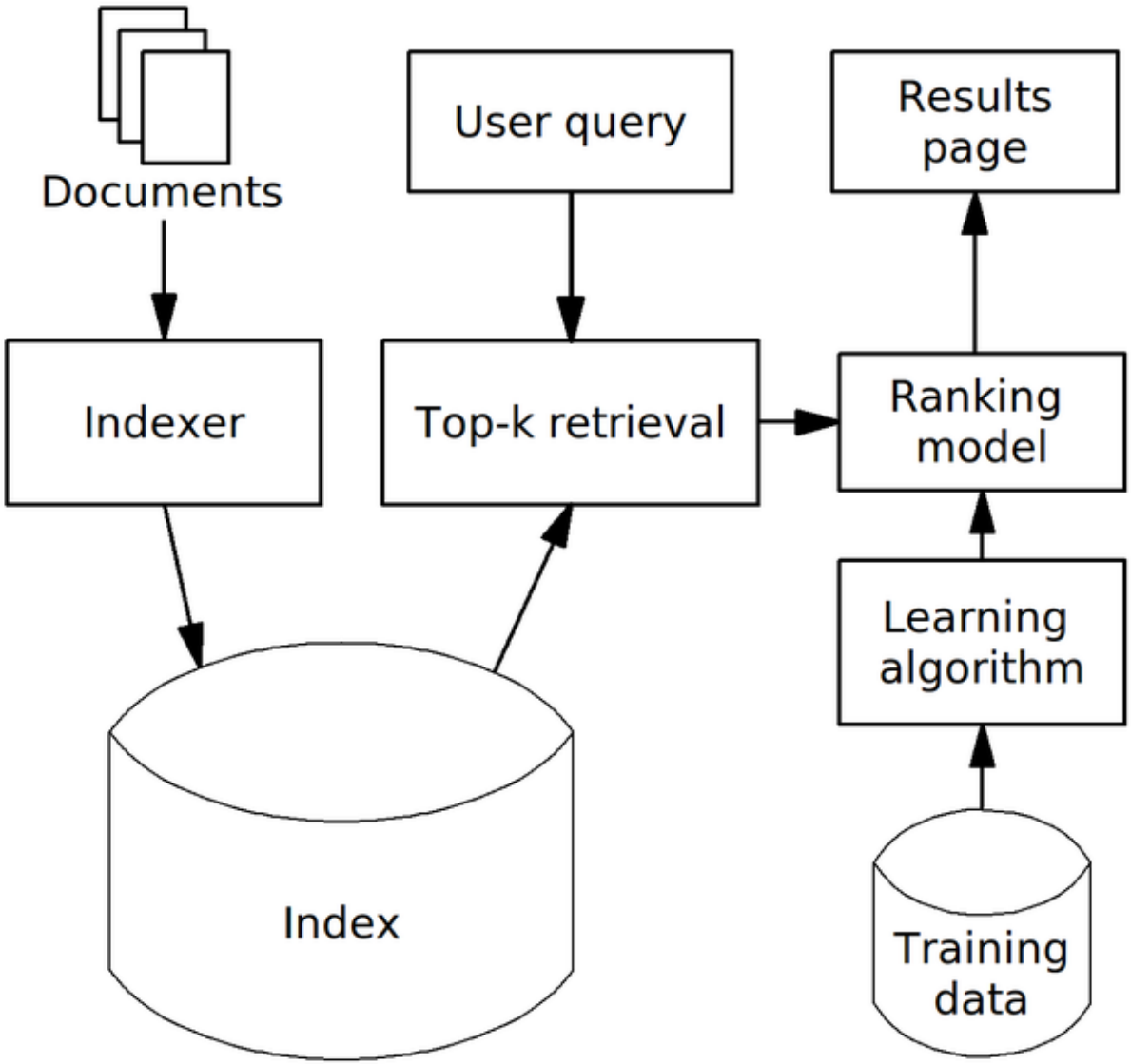
Постановка задачи

- Поиск закономерностей в данных об использовании каких-либо ресурсов. Например, часто используемых вместе ресурсов.
- Пример. $support(\{\text{хлеб, молоко}\}) = 0.7$
- Часто такие закономерности записываются в виде правил $A \longrightarrow B$
- Пример. $\{\text{Студент, Возраст от 16 до 25}\} \longrightarrow \{iPhone, iPad\}$

Поиск паттернов/зависимостей

The FIMI'03 best implementation award was granted to Gosta Grahne and Jianfei Zhu (on the left). The award consisted of the most frequent itemset: $\{diapers, beer\}$.

Ранжирование



Рекомендательные системы

http://Amazon.com

Frequently Bought Together

Price For All Three: $86.01

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

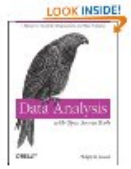
- **This item:** Machine Learning for Hackers by Drew Conway Paperback $33.87
- Machine Learning in Action by Peter Harrington Paperback $25.75
- Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback $26.39

Customers Who Bought This Item Also Bought

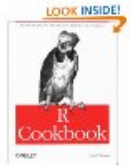
Programming Collective Intelligence: Building ...
› Toby Segaran
(84)
Paperback
$26.39

Machine Learning in Action
› Peter Harrington
(10)
Paperback
$25.75

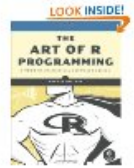
Mining the Social Web: Analyzing Data from ...
› Matthew A. Russell
(19)
Paperback
$26.36



Data Analysis with Open Source Tools
› Philipp K. Janert
(29)
Paperback
$24.05



R Cookbook (O'Reilly Cookbooks)
› Paul Teetor
(18)
Paperback
$32.43



The Art of R Programming: Tour of Statistical ...
Norman Matloff
(29)
Paperback
$25.06

Are any of these items inappropriate for this page? Let us know

Рекомендательные системы

http://lmhonet.ru

Оценки фильма Любопытное стечение обстоятельств

Een Bizarre Samenloop Van Omstandigheden, A Curious Conjunction of Coincidences

Фильмы / Комедии / обстоятельств» /

Да, Вам стоит смотреть фильм «Любопытное стечение обстоятельств»

Людям, с оценками, похожими на Ваши, этот фильм нравится

А ещё они рекомендуют Вам 31 фильм

Ваша прогнозируемая оценка фильма после его просмотра
8.2

Смотрели? Оцените

Не рекомендовать

Про фильм | Онлайн | Скачать | Отзывы | Персоны | Кадры | Оценки | Похожие

Распределение оценок

1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	1	4	4

Кому больше нравится

Кому нравится: 99% 1%

План лекции

1. Программа курса
   - Оценка по курсу
2. Системы и библиотеки ML&DM
3. Чего бы почитать и посмотреть?

Системы машинного обучения и анализа данных

1. Orange (freely available)
2. Conexp(freely available)
3. SPMF (freely available)
4. Weka (freely available)

План лекции

1. Программа курса
   - Оценка по курсу
2. Системы и библиотеки ML&DM
3. Чего бы почитать и посмотреть?

Книги

- M. Zaki et al. *Data Mining and Analysis: Fundamental Concepts and Algorithms*, 2014 (free)
- J. Leskovec et al. Mining of Massive Datasets, 2014 (free)
- Барсегян А. и др. *Анализ данных и процессов*, 2009

Портал machinelearning.ru

http://machinelearning.ru

- Лекции К.В. Воронцова. *Математические методы обучения по прецедентам (машинное обучение)*

Coursera: курсы и специализации

http://www.coursera.org/

- Jiawei Han Pattern Discovery in Data Mining
- Jure Leskovec et al. Mining Massive Datasets

Специализации (платные сертификаты) — состоят из отдельных курсов (участие бесплатно)

- Joseph A Konstan & Michael D. Ekstrand *Recommender Systems Specialization*
- Jiawei Han Data Mining

ИНТУИТ

http://intuit.ru

- Интернет-университет информационных технологий
- К.В. Воронцов Машинное обучение, 2015 (Видео к курсу на сайте ШАД)
- И.А. Чубукова. Data Mining, 2006

Конференции

- IEEE ICDM – IEEE International Conference on Data Mining
- KDD – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- ECML & PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- RecSys – The ACM conference series on Recommender Systems
- ИОИ & ММРО – Серия конференций «Интеллектуализация обработки информации»/«Математические методы распознавания образов»
- АИСТ – International conference on Analysis of Images, Social Networks, and Texts

Вопросы и контакты

www.hse.ru/staff/dima

Спасибо!

dmitrii.ignatov[at]gmail.com