Поиск закономерностей в последовательностях (Sequence Mining)

Игнатов Дмитрий Игоревич[♢]

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

PM 2017

Содержание

Анализ темпоральных данных

[J. Han & M. Kamber, Data Mining: Concepts and Techniques,2^{nd} Edition, Chapter 8, 2006]

Задачи

- Анализ потоков данных (Mining Data Streams)
- Анализ временных рядов (Time-Series Analysis)
- Анализ последовательностей (Mining Sequence Data)

Содержание

[Agrawal and Srikant, 1995]

"Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of a set of items, and given a user-specified minimum support threshold of min sup, sequential pattern mining finds all frequent subsequences, that is, the subsequences whose occurrence frequency in the set of sequences is no less than min sup."

Поиск частых последовательностей

Основные понятия

- Множество элементов (Itemset):
  a и $\{a, b, c\}$ для краткости a и abc
- Последовательность (Sequence):
  $s = \langle a(abc)(ac)d(cf)\rangle$
- Событие (Event):
  $s = \langle e_1 e_2 e_3 e_4 e_5\rangle$
  $e_1 = a$ и $e_2 = (abc)$ с точностью до интервала наступления
- Длина последовательности – число вхождений всех элементов:
  длина s равна 9
- l-последовательность:
  s – 9-последовательность

Поиск частых последовательностей

Han et al., Zaki & Meira, Aggarwal et al., и др.

- $s = \langle s_1, \ldots, s_k \rangle$ – подпоследовательность $r = \langle r_1, \ldots, r_l \rangle$ ($s \sqsubseteq r$) если $k \leq l$ и существуют $1 \leq t_1 < t_2 < \ldots < t_k \leq l$, такие что $s_j = r_{t_j}$ для всех $1 \leq j \leq k$[1].
- $support(s, D)$ – поддержка последовательности s в D, т.е. число последовательностей в D, таких что s является их подпоследовательностью.

$$support(s, D) = |\{r | r \in D, s \sqsubseteq r\}|$$

[1] если элементы последовательности s_i – множества (sequences of itemsets), то требуют $s_j \subseteq r_{t_j}$

Поиск частых (под)последовательностей

База данных последовательностей

Номер последовательности (Sequence_ID)	Последовательность (Sequence)
1	$\langle a(abc)(ac)d(cf)\rangle$
2	$\langle (ad)c(bc)(ae)\rangle$
3	$\langle (ef)(ab)(df)cb\rangle$
4	$\langle eg(af)cbc\rangle$

- $\langle af\rangle$ является подпоследовательностью 1 и 3
- $support(\langle af\rangle) = 2$

Содержание

Алгоритмы поиска частых последовательностей

GSP: [Srikant & Agrawal, 1996]

- Generalized Sequential Pattern (GSP)
- Расширение алгоритма Apriori

Алгоритмы поиска частых последовательностей

SPADE: [Zaki, 2001]

- Sequential PAttern Discovery using Equivalence classes
- Использует вертикальный формат данных
- Действует по принципу “порождай и проверяй”

Алгоритмы поиска частых последовательностей

SPADE: [Zaki, 2001]

SID	Sequence
1	$\langle a(abc)(ac)d(cf)\rangle$
2	$\langle (ad)c(bc)(ae)\rangle$
3	$\langle (ef)(ab)(df)cb\rangle$
4	$\langle eg(af)cbc\rangle$

SID	*EID*	*itemset*
1	1	a
1	2	abc
1	3	ac
1	4	d
1	5	cf
2	1	ad
2	2	c
2	3	bc
2	4	ae
3	1	ef
3	2	ab
3	3	df
3	4	c
3	5	b
4	1	e
4	2	g
4	3	af
4	4	c
4	5	b
4	6	c

(a) vertical format database

a		b		···
SID	*EID*	*SID*	*EID*	···
1	1	1	2	
1	2	2	3	
1	3	3	2	
2	1	3	5	
2	4	4	5	
3	2			
4	3			

(b) ID_lists for some 1-sequences

ab			ba			···
SID	*EID*(*a*)	*EID*(*b*)	*SID*	*EID*(*b*)	*EID*(*a*)	···
1	1	2	1	2	3	
2	1	3	2	3	4	
3	2	5				
4	3	5				

(c) ID_lists for some 2-sequences

aba				···
SID	*EID*(*a*)	*EID*(*b*)	*EID*(*a*)	···
1	1	2	3	
2	1	3	4	

(d) ID_lists for some 3-sequences

The SPADE mining process: (a) vertical format database; (b) to (d) show fragments of the ID_lists for 1-sequences, 2-sequences, and 3-sequences, respectively.

Алгоритмы поиска частых последовательностей

PrefixSpan: [Jian P. & Jiawei H. et al., 2001]

- Использует FP-деревья (Frequent Pattern Trees) и проекции баз данных (projected dabases)
- Понятие суффикса и префикса последовательности

Алгоритмы поиска частых последовательностей

PrefixSpan: [Jian P. & Jiawei H. et al., 2001]

prefix	*projected database*	*sequential patterns*
$\langle a \rangle$	$\langle (abc)(ac)d(cf) \rangle$, $\langle (\_d)c(bc)(ae) \rangle$, $\langle (\_b)(df)eb \rangle$, $\langle (\_f)cbc \rangle$	$\langle a \rangle$, $\langle aa \rangle$, $\langle ab \rangle$, $\langle a(bc) \rangle$, $\langle a(bc)a \rangle$, $\langle aba \rangle$, $\langle abc \rangle$, $\langle (ab) \rangle$, $\langle (ab)c \rangle$, $\langle (ab)d \rangle$, $\langle (ab)f \rangle$, $\langle (ab)dc \rangle$, $\langle ac \rangle$, $\langle aca \rangle$, $\langle acb \rangle$, $\langle acc \rangle$, $\langle ad \rangle$, $\langle adc \rangle$, $\langle af \rangle$
$\langle b \rangle$	$\langle (\_c)(ac)d(cf) \rangle$, $\langle (\_c)(ae) \rangle$, $\langle (df)cb \rangle$, $\langle c \rangle$	$\langle b \rangle$, $\langle ba \rangle$, $\langle bc \rangle$, $\langle (bc) \rangle$, $\langle (bc)a \rangle$, $\langle bd \rangle$, $\langle bdc \rangle$, $\langle bf \rangle$
$\langle c \rangle$	$\langle (ac)d(cf) \rangle$, $\langle (bc)(ae) \rangle$, $\langle b \rangle$, $\langle bc \rangle$	$\langle c \rangle$, $\langle ca \rangle$, $\langle cb \rangle$, $\langle cc \rangle$
$\langle d \rangle$	$\langle (cf) \rangle$, $\langle c(bc)(ae) \rangle$, $\langle (\_f)cb \rangle$	$\langle d \rangle$, $\langle db \rangle$, $\langle dc \rangle$, $\langle dcb \rangle$
$\langle e \rangle$	$\langle (\_f)(ab)(df)cb \rangle$, $\langle (af)cbc \rangle$	$\langle e \rangle$, $\langle ea \rangle$, $\langle eab \rangle$, $\langle eac \rangle$, $\langle eacb \rangle$, $\langle eb \rangle$, $\langle ebc \rangle$, $\langle ec \rangle$, $\langle ecb \rangle$, $\langle ef \rangle$, $\langle efb \rangle$, $\langle efc \rangle$, $\langle efcb \rangle$.
$\langle f \rangle$	$\langle (ab)(df)cb \rangle$, $\langle cbc \rangle$	$\langle f \rangle$, $\langle fb \rangle$, $\langle fbc \rangle$, $\langle fc \rangle$, $\langle fcb \rangle$

Замкнутые (под)последовательности

Определение

Определение

Замкнутая подпоследовательность (closed subqequence) – последовательность, которая не содержит надпоследовательности с такой же поддержкой.

(Частые) замкнутые (под)последовательности

Пример

Номер последовательности (Sequence_ID)	Последовательность (Sequence)
1	$\langle a(abc)(ac)d(cf)\rangle$
2	$\langle (ad)c(bc)(ae)\rangle$
3	$\langle (ef)(ab)(df)cb\rangle$
4	$\langle eg(af)cbc\rangle$

- $minsupp = 3$
- Замкнуто ли $\langle a\rangle$?
- Замкнуто ли $\langle ab\rangle$?
- Замкнуто ли $\langle af\rangle$?

Замкнутые (под)последовательности

Алгоритмы

- CloSpan [X. Yan, J. Han, and R. Afshar, 2003]
- BIDE [J. Wang and J. Han, 2004]

Содержание

Доступные реализации
SPMF

- SPMF A Sequential Pattern Mining Framework
  >100 алгоритмов анализа данных:
  - association rule mining
  - itemset mining
  - sequential pattern mining
  - sequential rule mining
  - sequence prediction
  - high-utility pattern mining
  - clustering and classification
- Репозиторий М.Заки
- TraMiner – библиотека для анализа последовательностей в R:
  - Extracting frequent event subsequences
  - Identifying most discriminating event subsequences
  - Association rules between subsequences

Содержание

Демографические последовательности

Данные опроса РиДМиЖ

- Источник – обследование “Родители и дети, мужчины и женщины в семье и обществе”
- Предобработанная выборка Github
  - 2300 опрошенных
  - Социально-демографические признаки: пол, возраст, образование, место жительства (город/село) и т.п.
  - Демографические события: отделение от семьи, партнер, вступление в брак, рождение ребенка, устройство на работу.

Демографические последовательности

Статьи и слайды

- H. Blockeel, J. Fürnkranz, A. Prskawetz, F.C. Billari: Detecting Temporal Change in Event Sequences: An Application to Demographic Data. PKDD 2001: 29-41
- F. C. Billari, J. Fürnkranz, A. Prskawetz. Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. European Journal of Population, 2006, 22(1), pp 37–65

Научно-учебная группа “Модели и методы анализа демографических последовательностей”

- D.I. Ignatov, E. Mitrofanova, A. Muratova, D. Gizdatullin: Pattern Mining and Machine Learning for Demographic Sequences. KESW 2015: 225-239
  [Статья] [Слайды]
- Pattern-based classification of demographic sequences [Слайды]

Содержание

Домашнее задание

Данные опроса РиДМиЖ

На основе имеющихся опросных данных (см. Github) определить:

- Какое событие чаще всего является первым у каждого поколения? (можно построить гистограмму)
- Какое событие чаще всего является последним у каждого поколения? (можно построить гистограмму)
- Какая последовательность событий наиболее частая у каждого поколения (длины 2, 3, 4 и 5)?
- Аналогичные вопросы для признака “пол” и признаков “поколение” и “пол”

Содержание

Что почитать?

Книги и статьи

- Учебник Zaki & Meira (Глава 10)
- Учебник Han & Kamber, 2-е издание (слайды) (Глава 8.3), нет в 3-м издании
- Frequent Pattern Mining. Editors: Charu C. Aggarwal, Jiawei Han
- Поиск частых последовательностей (статьи – мой архив): PrefixSpan, CloSpan, GSP, SPADE, BIDE и др.
- Подборка Fournier-Viger'a
  - Поиск частых последовательностей (в т.ч. многоуровневые/многомерные последовательности)
  - Поиск правил на последовательностях (sequential rules mining)
  - Предсказание следующего события (sequence prediction)

Recommender Systems Challenge 2015

- RecSys Challenge 2015
- Статьи победителей в моём DropBox
- Наше решение

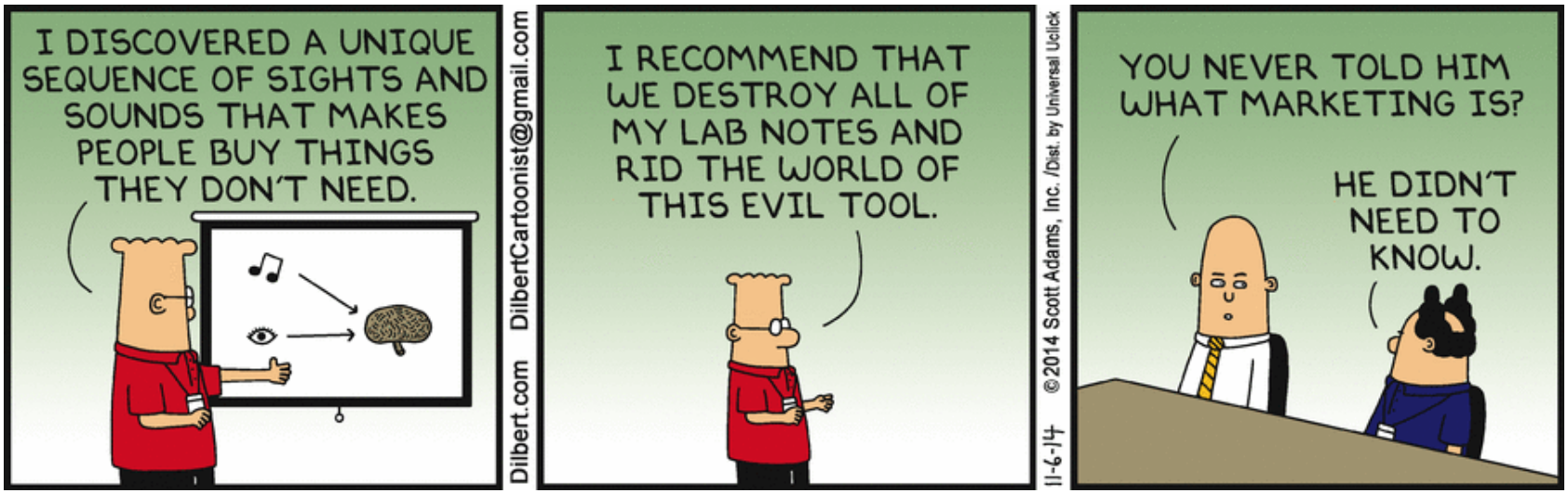
Задача

Given a sequence of click events performed by some user during a typical session in an e-commerce website, the goal is to predict whether the user is going to buy something or not, and if he is buying, what would be the items he is going to buy. The task could therefore be divided into two sub goals:

1. Is the user going to buy items in this session? Yes|No
2. If yes, what are the items that are going to be bought?

Just for fun или шутки ради

http://dilbert.com/strip/2014-11-06



Вопросы и контакты

www.hse.ru/staff/dima

Спасибо!

dmitrii.ignatov[at]gmail.com