Курс “Практические задачи анализа данных”

Раздел курса: *Поиск закономерностей в данных (Pattern Mining).*

Тема: *Частые (под)последовтельности.*

Домашнее задание №5

Автор: Д.И. Игнатов Срок сдачи: 13 апреля 2017

Задание высылается в виде отчета в формате PDF или DOC по адресу dmitrii.ignatov@gmail.com с темой письма [MLDM2017-HW2-SEQ]-<Фамилия Имя> с копией ассистенту – Баранецкой Дарье <dabaranetskaya@edu.hse.ru>.

Задание 1 (2 балла). Поиск частых событий

На основе данных опроса РиДМиЖ (см. файл socAttrAndSeqFusion(full).txt в репозитории Github) определить:

1. Какое событие чаще всего является первым у каждого поколения? (можно построить гистограмму)
2. Какое событие чаще всего является последним у каждого поколения? (можно построить гистограмму)
3. Аналогичные вопросы для признака “пол” и комбинаций признаков “поколение” и “пол”.

Рекомендация: для подсчета частот можно пользоваться табличным процессором или средствами, доступными в Питоне.

Задание 2 (5 баллов).

1. Определите число частых и частых замкнутых последовательностей при минимальной поддержке 1 для всего набора данных.

   Примеры в SPMF: поиск частых последовательностей с PrefixSpan и поиск частых замкнутых последовательностей с BIDE+ (не забывайте про модификатор в конце имени алогоритма string, иначе нельзя будет работать с последовательностями из названий событий).

2. Какая последовательность событий наиболее частая в группе мужчин, а какая – в группе женщин (длины 2, 3, 4 и 5)?

3. Аналогичные пунктам 1 и 2 вопросы для комбинаций признаков “поколение” и “пол” (без ограничения на длину).

   Рекомендация: можно воспользоваться идеей добавления в начало каждой последовательностей комбинаций двух признаков в начало

```
generation=1 -1 gender=0 -1 separation -1 work -1 -2
```

   из файла SocialAttributes.csv (оставить только столбцы gender и generation).

   Пример см. скрипт reader.py.

```
def exp1():
    ds= DataListToSequenceList2(dataList, names)
     fl=AttrAndSeqFusion(ds,'SocialAttributes.csv')
    SequenceToSPMF(fl, 'gener_gender_seq.txt')
```

   Альтернативно можно использовать т.н. многомерные/многоуровневые последовательности (опция со строками не доступна): пример в SPMF.

4. Постройте все последовательностные ассоциативные правила для женщин и для мужчин ($minsupp = 1$, $minconf = 0,5$) и приведите примеры трех самых достоверных для этих двух типов правил.

Задание 3 (3 балла)

Приведите пример постановки задачи и фрагмента или описания данных, в которой можно было бы использовать

1. поиск частых множеств и ассоциативных правил,
2. поиск частых (под)последовательностей,
3. поиск частых (под)графов (данные для графов можно не приводить, достаточно описания задачи).

Дополнительная информация может быть найдена в статье Ignatov et al. [2015] и учебниках Zaki and Wagner Meira [2014] (часть 2), Han and Kamber [2006] (Главы 5, 8.3, 9.1).

Список литературы

Dmitry I. Ignatov, Ekaterina Mitrofanova, Anna Muratova, and Danil Gizdatullin. Pattern mining and machine learning for demographic sequences. In *Knowledge Engineering and Semantic Web - 6th International Conference, KESW 2015, Moscow, Russia, September 30 - October 2, 2015, Proceedings*, pages 225–239, 2015. doi: 10.1007/978-3-319-24543-0_17. URL https://www.researchgate.net/publication/312307491_Pattern_Mining_and_Machine_Learning_for_Demographic_Sequences.

Mohammed J. Zaki and Jr. Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, May 2014. ISBN 9780521766333. URL http://www.dataminingbook.info/pmwiki.php/Main/BookDownload.

Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2 edition, 2006.