Программа “Машинное обучение и майнинг данных”

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

Тема: *Частые множества признаков и ассоциативные правила.*

Домашнее задание №1

Автор: Д.И. Игнатов Срок сдачи: 12 октября 2017

Задание высылается в виде отчета в формате PDF или DOC по адресу dmitrii.ignatov@gmail.com и ассистенту курса Данилу Гиздатулину (gizdatullindanil@gmail.com) с темой письма [ML2017-HSE-BIG5-HW1-FIM]-<Фамилия Имя>.

Задание 1 (30 баллов). Поиск частых множеств

а) Для массива данных о контекстной рекламе размером 2000 компаний × 3000 словосочетаний найти частые множества для минимальной поддержки minsupp=35[1]. Необходимо указать число таких множеств.

Пример в SPMF

б) Повторить подзадание а) для частых замкнутых множеств.

Пример в SPMF

в) Повторить поздание а) для частых максимальных множеств.

Пример в SPMF

г) Среди множеств, найденных в заданиях а), б), в) указать около 10 размером более 10 словосочетаний и провести их интерпретацию как рынков.

Данные.

Данные в формате с разделителями табуляцией.

Списки пар словосочетание-фирма в виде индентификаторов и словосочетания.

```
3000 2000 92345
% размеры данных: число словосочетаний, число фирм, число пар

0 23 1
0 96 1
0 188 1
```

[1]поддержка дана в абсолютных единицах, а не процентах

```
0 328 1
0 556 1
0 632 1
```

Рекомнедуемое программное средство: SPMF.

Задание 2 (30 баллов). Поиск ассоциативных правил

а) Для массива данных о контекстной рекламе 2000 компаний $\times$ 3000 словосочетаний найти ассоциативные правила для минимальной поддержки $minsupp = 35$ и $minconf = 1$. Необходимо указать число таких правил.

Пример в SPMF

б) Для исходного массива данных найти замкнутые ассоциативные правила для минимальной поддержки minsupp=35 и minconf=1. Необходимо указать число таких правил.

Пример в SPMF

в) Для исходного массива данных найти 5 самых частых правил при минимальной достоверности $minconf = 0,8$. Необходимо указать эти правила и дать интерпретацию.

Пример в SPMF

Задание 3 (40 баллов). Анализ посещаемости сайтов на основе решеток формальных понятий

Для трех контекстов о посещаемости сайта Высшей школы экономики в терминах посещений сайтов новостной, образовательной и финансовой тематики необходимо выполнить пункты задания ниже.

а) Удалением некоторого числа сайтов (признаков) или пользователей (объектов) добиться числа формальных понятий не менее 100, но не сильно превышающего это значение.

б) Для контекстов, полученных удалением объектов или признаков в пункте а), построить диаграммы решеток понятий.

в) Привести 3–5 примеров понятий в виде пары $\langle$ размер объема понятия, содержание понятия $\rangle$ для размера содержания 2 и более сайта. Дать содержательную интерпретацию найденных понятий.

г) Привести пример импликации вида $A \to B$, найденной по диаграмме решетки понятий с указанием ее поддержки.

Рекомендуемое программное средство: Concept Explorer.

Дополнительная информация может быть найдена в статьях Ignatov and Kuznetsov [2008], Ignatov et al. [2012], Kuznetsov and Ignatov [2007], Yevtushenko [2006], Zaki and Hsiao [2005], Zhukov [2004], Ignatov [2014], Zaki and Wagner Meira [2014].

Список литературы

Dmitry I. Ignatov and Sergei O. Kuznetsov. Concept-based Recommendations for Internet Advertisement. In R. Belohlavek and S. O. Kuznetsov, editors, *Proc. CLA 2008*, volume Vol. 433 of *CEUR WS*, pages 157–166. Palacký University, Olomouc, 2008, 2008. ISBN 978–80–244–2111–7. URL `http://ceur-ws.org/Vol-433/paper13.pdf`.

Dmitry I. Ignatov, Sergei O. Kuznetsov, and Jonas Poelmans. Concept-Based Biclustering for Internet Advertisement. In *ICDM Workshops*, pages 123–130, 2012. URL `https://www.researchgate.net/publication/268288810_Concept-based_Biclustering_for_Internet_Advertisement`.

Sergei O. Kuznetsov and Dmitry I. Ignatov. Concept Stability for Constructing Taxonomies of Web-site Users. *in Proc. Social Network Analysis and Conceptual Structures: Exploring Opportunities, S. Obiedkov, C. Roth (Eds.), Clermont-Ferrand (France), February 16, 2007*, 2007. URL `http://arxiv.org/abs/0905.1424`.

Serhiy A. Yevtushenko. *Concept Explorer. The User Guide*, September 12 2006. URL `http://www.comp.dit.ie/pbrowne/compfund2/UserGuide.pdf`.

Mohammed Javeed Zaki and Ching-Jui Hsiao. Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. *IEEE Trans. Knowl. Data Eng.*, 17(4):462–478, 2005. URL `http://www.cs.rpi.edu/~zaki/PaperDir/TKDE05-charm.pdf`.

L. E. Zhukov. Spectral Clustering of Large Advertiser Datasets. Technical report, Overture R&D, April 2004. URL `http://leonidzhukov.ru/papers/spectral_clustering-zhukov.pdf`.

Dmitry I. Ignatov. Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields. In *Information Retrieval - 8th Russian Summer School, RuSSIR 2014, Nizhniy, Novgorod, Russia, August 18-22, 2014, Revised Selected Papers*, pages 42–141, 2014. URL `http://bit.ly/2lpTbH2`.

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`.