Course "Machine Learning and Data Mining"

Course chapter:*Pattern Mining.*

Topic: *Frequent Itemset Mining and Association Rules.*

Homework 1

Author: D.I. Ignatov Deadline: June 4, 2017

The report should be send in PDF or DOC to <dmitrii.ignatov@gmail.com> with the email's topic [MLDM2017m-HW1-FIM]-<Name Family Name> with a CC to TA, Anna Muratova <anyamuratova@yandex.ru> and Daniil Stepenskiy <reinkarn@gmail.com>.

Task 1 (30 points). Frequent Itemset Mining

a) Given a contextual advertising dataset of 2000 companies × 3000 terms, find all frequent itemsets with minsupp=35. Report the number of such itemsets.

Example for SPMF

b) Repeat subtask a) for frequent closed itemsets.

Example for SPMF

c) Repeat subtask a) for maximal frequent itemsets.

Example in SPMF

d) Among the resulting itemsets for a), b), and c), indicate 10 itemsets composed of 10 terms or greater and interpret them as "markets".

Data.

Tab-separated data.

Pairs of term-firm as respective IDs.

```
3000 2000 92345
% dataset size: the number of terms, the number of firms, the
   number of pairs

0 23 1
0 96 1
0 188 1
0 328 1
0 556 1
```

```
0 632 1
```

The recommended software: SPMF.

Task 2 (30 points). Association Rules Mining

a) For advertising dataset with 2000 firms $\times$ 3000 terms find association rules with $minsupp = 35$ and $minconf = 1$. Indicate the number of such rules.

Example for SPMF

b) For the input dataset find closed association rules with minsupp=35 and minconf=1. Indicate the number of such rules.

Example for SPMF

c) For the input datset find the top-5 frequent rules with $minconf = 0,8$. Provide all the found rules and their interpretation (at least for a couple them) in the report.

Example for SPMF

Task 3 (40 points). Analysis of web site visitors' behavior based on concept lattices

For three input context about visitors of Higher School of Economics in terms of their visits of news websites, education-related and finance-related websites perform the subtasks below.

a) By removal of certain websites (attributes) or visitors (objects) in the input dataset make sure that the number of formal concepts is about 100.

b) For the context from subtask a) that obtained by object/attribute removal build the corresponding lattice diagrams.

c) Provide 3–5 examples of concepts as pairs $\langle$ extent size, intent $\rangle$ for the intent size greater than 2. Give the interpretation of the found concepts.

d) Provide the examples of implications $A \to B$ found by lattice diagram with the indication of their support and confidence.

The recommended software: Concept Explorer.

Auxiliary information can be found in 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].

References

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`.