Dissertation thesis review

The dissertation thesis titled Parallel Association Rule Mining Algorithm Based on MapReduce by Using Lift Interestingness Measure for Big Data, submitted by Nour Easa Owais, describes author’s research in the field of (distributed) association rule mining and big data processing. The author proposes, prototypes, and tests a map-reduce implementation of an association rule mining algorithm utilizing lift as rule interestingness measure. Next, he develops and tests a PCA-based feature selection algorithm.

The thesis is logically structured and presents the research topics, selected algorithms, tools, and frameworks. The first part of the thesis provides motivation of the research, introductory-level information about data mining, its types, and selected techniques, and in general terms defines association rule mining. It also points out the importance of dimensionality reduction for practical data processing. The manuscript also features a popular description and history of the phenomenon of Big Data and mentions selected analytical strategies and tools including NoSQL, map-reduce, R, and Matlab. Section 2.1.5 is dedicated to a description of Big Data applications. However, ref. [55] (and some other) describe regular analytical problems and have in fact nothing to do with Big Data. A motivation for the selection of algorithms and applications, described in this part of the manuscript, is, however, not very clear. Most of them are not related to the rest of the work.

The rest of the thesis first describes the design and experimental evaluation of a map-reduce based implementation of an association rule mining algorithm (MRLAR) and then a PCA-based feature selection method (FSPCA). It describes proposed concepts (sometimes quite vaguely) and their experimental evaluation. The quality of the text varies greatly. The part dedicated to association rule mining and MRLAR algorithm is more mature and reads reasonably well. The second topic is processed less well and it appears that the work was compiled together in a hurry. It is also not clear why are the two concepts (FSPCA and MRLAR) not tied together for a complex experimental evaluation. Instead, the author used FP-Growth to mine association rules from a data set reduced by FSPCA.

The writing of the paper could be improved. For example, the abbreviations MRLAR (and many other) are defined multiple times. Actually, so many times that it becomes annoying. Some parts of the manuscript are not easy to read (e.g. 5.1.2: 'For this part of the thesis, the algorithm used the open source operator Rapidminer (6.5) Weight by PCA for its well-known ability to support big data mining’ - Meaning of the sentence is not clear at all). The ratio between background information and original work is not well balanced. Some sections (e.g. history of big data, simple applications of SDD, SVD) are unnecessary while some other desperately need more technical details and experiments. For example, in the dimensionality reduction part, no comparison of the proposed concept with relevant work is provided. Actually, no related work dealing with PCA-based feature selection, such as [1-4], is mentioned at all. The FSPCA algorithm is poorly described and it is not clear how it works ('In the second phase, we passed the dataset to the weight by PCA operator which used the PCA to assign weights to the attributes’ – how? Technical details are severely missing here!). Some results are not well explained. Tables 8 and 9 are confusing (what is k? It appears that it is not the number of selected features/reduced dimension). In table 8, training time increased with k = 13, it is higher than with the original dataset. The author should definitely explain why this happens. Moreover, the methodology of the
time measurements is not clear (how many times were experiments repeated? Is the presented time average? What was the standard deviation etc).

Experiment results show that the proposed map-reduce implementation of LBA works and that FSPCA can be useful. This and author’s publication portfolio of accepted research papers leads me to the final decision to recommend the thesis for defence, despite its weak sides, summarized in part in the text of this review.

References


doc. Ing. Pavel Krömer, Ph.D.
Hlučín, 30. 5. 2016