Opponent review of dissertation work of Ing. Nour Easa Oweis

As requested, I had the honor of reviewing and evaluating the dissertation work of Ing. Nour Easa Oweis as part of his PhD study at the Technical University of Ostrava. His thesis “Parallel Association Rule Mining Algorithm Based on MapReduce by Using Lift Interestingness Measure for Big Data” clearly belongs to the field of his study. The work is divided into six main chapters including the objectives of the work and the conclusions. Load-bearing part of the work is represented in chapters 4 and 5. The dissertation work contains 93 pages. This final report is prepared based on thorough review of the dissertation work.

1. The fulfillment of the dissertation objectives:

The author proposed to use two algorithms for data mining. The first is the use of Lift Interestingness Measure (LIM) and integrates it with parallel algorithm under MapReduce approach. This is very useful as it could handle massive datasets with a large number of nodes. This will help discover and describe the type of correlation(s) between Left Hand Side (LHS) and Right Hand Side (RHS) in association rules. The second algorithm is for dimensionality reduction using the Singular Value Decomposition (SVD) approach, applied to reduce the data into fewer dimensions as pre-processing technique. This is a good approach as it has been implemented as an optimization processes for the dataset.

The dissertation objectives were novel and unique and they were fulfilled during the dissertation work presented in the results section and the pseudocodes presented during the work. It was also clear upon applying the algorithms on various case studies. The results and the discussion has provided sample evidence that the objectives were fulfilled in the following manner: the first proposed algorithm is a novel algorithm that designed for parallelism using LIM that allowed the data mining processes to be used with higher efficiency and better performance. The second proposed algorithm provided pre-processing techniques to reduce the dimensionality of the dataset, therefore, reduce the data capacity, the required time and memory to handle, easy visualization of data, and eliminate irrelevant features and noise reduction, thus reducing overall costs and effort related to Big Data.

2. Contribution to the practice and further development of scientific field

The Big Data had, for sure, helped the advancement of the world technology, industries, healthcare and scientific knowledge. However, it is a package that came with a lot of challenges that needed to be addressed. Mr. Oveis has tackled an important part of these problems related to data management with high scalability of the data and finding meaningful association rules. His approach by the two algorithms mentioned above have introduced a novel approach for looking
at the association rules as well as offered a pre-processing tool for data management to reduce the amount of memory and time needed to process this data.

In this thesis project, mining association rule by using LIM based on MapReduce have been developed for extracting relationship between the itemset instead of using confidence. Reduce dimensionality was also utilized to reduce the attributes into fewer dimensions as a pre-processing technique deployed for association rule mining.

Overall, I strongly feel that Mr. Oweis had contributed to the Big data mining and management both in terms of scientific knowledge and practical applications.

3. Student publication

I strongly feel that Mr. Nour Oweis has identified and proved by publications significant avenues of scientific knowledge. The author published his work in 07 national and international conferences and journals related to the topic, and 03 publications in different theme.

In addition, Ing, Oweis is a well-rounded researcher that he liked to investigate more in the science and the minor details of other topics and made some other publications not necessarily directly related to his dissertation work, but definitely add to the scientific body in computer engineering and networks.

Summary:

During his time in this thesis work, Mr. Oweis has clearly met or exceeded expectations with respect to scientific approach, documentation and exploring the field of Big Data and its relation to the parallel Association Rule Mining Algorithm and the Pre-processing of the data. He shows strong performance as a researcher, with every promise of continued development, and has provided a high level of development in the Computer Sciences. His professional productivity and dissertation work are substantial, and he has shown considerable initiative as well as the ability to make unique contributions, as exemplified by the novel approach that he offered. His publications at the international level is already demonstrating his ability to have an impact at the scientific knowledge and its applications aspects in computer sciences. In general, I feel that Mr. Oweis has identified appropriate venues for his development, and I strongly encourage him to continue and expand his current trajectory of activities.
Other notes:

- The references number must be included BEFORE full stop. On p. 15, (i.e. ... journal [17]).
- Unify the use of each:
  - the proposed algorithm name (MRLAR), the author uses it on several ways (i.e. MapReduce-Based for Lift Association Rule on p. 16, and MapReduce Lift Association Rule on p. 15).
  - MapReduce approach, the author uses it on several ways such as (MapReduce functionality On p. 16, MapReduce approach on p. 16, 17, 27, and MapReduce process on p. 50).
  - LBA the author uses it on several ways such as (Lift Base Algorithm on p. 17, Lift Based Algorithm on p. 47)
- Repetitive definition for Big Data, the author defined the term of Big Data on the top of p. 18, and again at the end of the same page again wrote ‘This is known as Big Data’.
- The use of some references should be more concrete. On p. 19, the authors state: ‘For more information about big data, please refer to very useful articles [60-68]’ This should be broken down to details.
- The fifth Big Data analytics techniques is missing (MATLAB programing Language on p. 28), the author list with details five analytics techniques while he mentioned only four on p. 26.
- Reference is missing on p. 29 in the first paragraph of the Big Data Application and its Notable Uses.
- Abbreviation missing on p. 30 for (SaaS).
- Reference is missing on p. 31 in the third paragraph of the Data Mining section (i.e. Data mining techniques contain a variety of applications and notable uses...).
- The variable definition of $\Sigma$ in SVD is wrong on p. 37, the author mentioned that $\Sigma$ is an orthogonal matrix while it’s a diagonal matrix with positive or zero elements.
- Some of the attributes in the dataset are not clear on p. 48 such as (C, D, and F).
- I suggest to discriminate the accepted rules > 1 and rejected rules < 1 on a separate two tables on p. 60 to be more clear to the reader.
- On the first attribute’s table on p. 63 (Data Size), the author mentioned that it’s a number of transaction and write it as a data size.
- References needed on the first paragraph of the proposed algorithm FSPCA ‘The main purpose of using... on p. 65.
- In the summery and wrap up of the thesis, first paragraph ‘This work came into two parts: ... on p. 74, the author miss to mentioned the second part of the thesis (data reduction) while its explain later.
- In the conclusion Chapter on p. 77 there is uncomplete sentence at the end of first paragraph ‘This thesis process’

To summarize, the author proved his ability to work systematically on a given problem using the scientific method, and showed a good potential for further work in research. The results presented in this thesis are interesting and practically usable for researchers and practitioners world-wide, which is exemplified by the numerous publications of the author on various international conferences and journal.
Given all other requirements has been fulfilled, I recommend to submit the dissertation work to defense.

Additional Questions:

- In the section of (Big data analytics) on p. 26, 27 and 28, the author mentioned that the MATLAB can easily run the same MapReduce algorithm in different execution environments by minimal changes (one line of your code), could you provide more details with example?
- In the section of (Motivation and objective on the study) on p. 16 and 17, the author mentioned that the proposed algorithm require less computational resources, could you explain this point?
- In the section of (Association rule, Lift of the rule) on p. 35, the author mentioned that if Lift = 1 then independent correlation, could you provide more details with example?

Respectfully submitted,

Ostrava 29.05.2016

Ing. Nabil Oudrar, Ph.D

[Signature]