SIGMOD 2010 RWE Review on Paper “Towards Proximity Pattern Mining in Large Graphs”

Zhenjie Zhang,

National University of Singapore
zhenjie@comp.nus.edu.sg

1 Experiment Overview

In this report, we will provide some analysis on the repeatability and workability on the source codes provided by the authors of SIGMOD 2010 paper “Towards Proximity Pattern Mining on Large Graphs”. Since the primary reviewer on the paper has identified some problems on part of the experimental results, we will focus on these experimental items in this report.

Based on the source codes and running scripts provided by the authors, we have set up some experimental environment on a Red hat Linux Operating system (CentOS 5.0) equipped on IBM x255 server with four Intel Xeon MP 3.0 GHz CPU, 18G DDR memory and six 73.4GB Ultra320 SCSI hard disks. All the programs are compiled with GCC 4.4.3 and each process is handled by a single core at any time.

2 Summary from Primary Reviewer

In this section, we give a brief summary on the RWE results by the primary reviewer on this paper.

The primary reviewer is able to reproduce the results in Table 3, Table 4, Table 8, Table 9, Table 10 and Figure 11.

Table 5 and Table 6 cannot be repeated, due to the lack of the Intrusion Network data set. Table 7 is not fully verified, because the author selected some of the results to present in their original paper. The primary review also found some problem with the program when the memory consumption exceeded some threshold. This causes the inconsistency to the original results in Figure 9. One of the curve in Figure 10, when depth is 3, is significantly higher than the reported results. And finally, Figure 12 is unavailable for test, for the missing of appropriate program to generate the data with the specific number of labels.

Given the observations from the primary reviewer, we decided to concentrate on Figure 9, Figure 10, since the other inconsistency results cannot be dissolved directly. In the following section, some new results in our experiments are presented and discussed.

3 New Results

3.1 Figure 9(a)

The goal of the experiments with respect to Figure 9(a) is investigating the impact of graph cardinality on the computation time of NmPA. In particular, the authors used different number of nodes from the DBLP graph data and test with their NmPA method. While the primary reviewer discovered some memory problem when too large memory space is allocated, the program on our server is capable of completing all the experiments without any runtime error. The detailed results are listed in the following figure. Despite of some vibration on the curves, the general trends of the results match those of Figure 9(a) in the original paper. Therefore, we believe there might be some system configuration on compilation problem with the experimental environment with the primary reviewer.

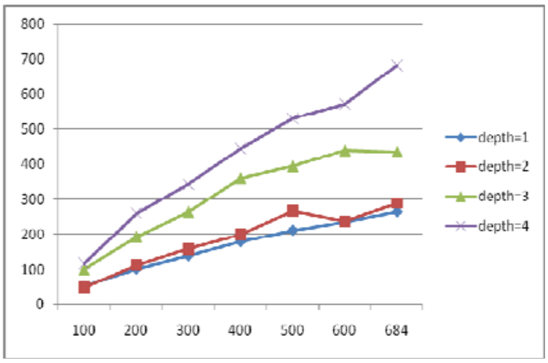

Fig. 9(a). On the x-axis are the numbers of nodes (K) in the DBLP data set. The y-axis shows the corresponding running time (seconds) of NmPA algorithm. Different curves apply different parameters on the depth of the propagation.

2.2 Figure 9(b)

Similar to Figure 9(a), the experiments in Figure 9(b) tests the impact of propagation depth. On my experimental setting, we have observed a much lower increase on the CPU time when the depth grows from 3 to 4. Again, we believe this is because of the different experimental setting on both hardware and software.

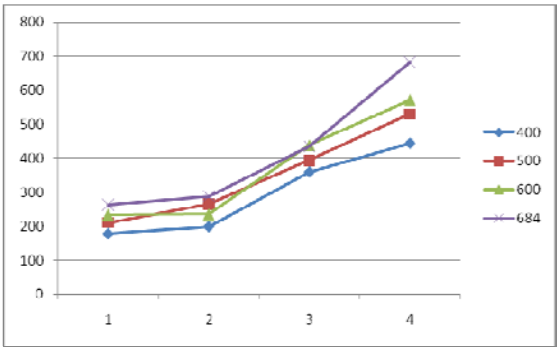

Fig. 9(b). On the x-axis are the propagation depth in the DBLP data set. The y-axis shows the corresponding running time (seconds) of NmPA algorithm. The curves denote data sets with different numbers of nodes (K).

2.3 Figure 10

In Figure 10, the authors tested on varying number of labels. After repeating the instructions in the program package, we found that the curves with depth=1 and depth=2 are very different to separate, which is similar to the results from the primary reviewer. On the other hand, the curve when depth=3 is close to the reports in their original paper.

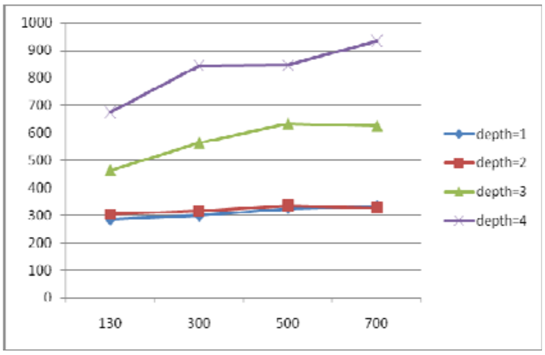

Fig. 10. On the x-axis are the number of labels in DBLP data set. The y-axis shows the corresponding running time (seconds) of NmPA algorithm. Different curves apply different parameters on the depth of the propagation.

3 Conclusion

After some tests on the experiments of the paper titled “Towards Proximity Pattern Mining on Large Graphs”, we found most of the experiments are generally repeatable. However, some of the results cannot be successfully tested by the reviewers, for some reasons including 1) the authors failed to provide the data and the processing programs, 2) the results are not stable on different machines. One of the major problem of the experiment is the huge amount of memory consumption, leading to some potential problems on runtime errors.