Understanding Trends in Global Terrorism

Konrad Cybulski, Julian Kardis, Matteo Batelic

Final report
FIT2083 Research Project

Contents

1 Introduction

1.1 Hypothesis

1.2 Data

<u>Control Dataset</u>

Figure 1: Lottery Data supplied from Geoff Webb.

It is evident that the newer version of OPUS MINER finds more associations due to the new relaxed statistical tests, as seen in Figure 2, 3 and 4.

Figure 2: Number of associations found in chess.fimi, retail.fimi and mushroom.fimi

Figure 3: Number of associations found in connect.fimi

Time

When OPUS MINER is being run on datasets such as Mushroom.fimi or Chess.fimi, as more associations were found, the time in leverage mode decreased while the time in lift mode increased. Whilst in retail.fimi and multistore.txt dataset, the time taken in lift mode was similar to that of mushroom.fimi and chess.fimi datasets, but in this instance, time taken in leverage mode increased. In connect.fimi we can see that the second version was the fastest, the first version was slower but the latest version in leverage mode takes the longest. Unlike any of the other datasets we can uniquely see that in lift mode, the time in the second version increases from the first version but decreases in the current version.

Figure 4: Number of associations found in multistore.txt

1.3 Discussion

With all of the datasets, it is clear to see that the number of associations found had increased over the three versions, however the relationship between the amount of associations found and computational time is a very complicated connection. As previously predicted within the hypothesis, unlike the old version of OPUS MINER the newest and second version were able to remove untestable hypothesis from the search area, allowing the program to run faster. However, because of the subsequent relaxed Bonferroni correction, there will be a large variance in time between each individual dataset as the new versions will have to spend more time compensating for these relaxed statistical tests to remain accurate. The results produced, reveal the complex relationship between time and the associations found. In the chess.fimi and mushroom.fimi dataset running in leverage mode, it is seen that the newer versions are faster than the previous version, but in lift mode, the newer versions are slower. Whereas, in the retail.fimi dataset, the program in lift mode has the same characteristics as the chess.fimi and mushroom.fimi dataset but the second version is the fastest. This is unexpected as the newest version and second version are similar, both filtering the untestable hypothesis before the Bonferroni correction. This discrepancy could be due to error in the results as it was only run five times. Unlike any of the other datasets, it can be seen in multistore.txt that the time taken in leverage mode is faster than the time taken in lift mode and in both of the

Figure 5: Time taken for Mushroom.fimi

datasets the current version is slower than the prior versions.

The most unique results came from the connect dataset where in leverage mode there was a large difference in time between the second version and the newest version, with the oldest version surprisingly being faster than the latest version and second being the fastest. With the computational time in lift mode being unlike any of the other datasets this result was surprising.

As the hypothesis stated, it is a very complex relationship between the associations found and computational time, as each of the datasets produced unique results. If we were focusing on the datasets that found little associations such as chess.fimi and mushroom.fimi, it is clear that the average running time in leverage mode decreased while in lift mode increased. Further the datasets multistore.txt and retail.fimi that found few associations but more than chess.fimi and mushroom.fimi, the time in both leverage and lift mode increased for the newer versions. Whereas, in the connect dataset where a large number of associations were found, OPUS MINER'S computational time to search in leverage mode increased while the time to search in lift mode decreased.

With this select amount of datasets it was found that the datasets with fewer associations within leverage mode ran faster in the newer versions than the older version. Whilst in lift mode the newer versions take a longer time to find the associations when compared to the older version. The opposite occurs in the dataset with a great number of associations where in leverage mode it takes a longer time to find the associations in the newest versions compared to the

Figure 6: Time taken for Chess.fimi

older version while within lift mode the newer versions run faster than he older version. Surprisingly the datasets with not a lot of associations has the opposite relationship between associations found and computation time to the dataset with a lot of associations.

To gain a more accurate view of the results, more datasets with a greater range of associations will have to be run through OPUS MINER as the selected 5 datasets are so uniquely different from each other nothing can be said with great accuracy.

2   Conclusion

The results found clearly show that the newer version is more powerful than the previous as it finds more associations but the relationship of associations found and computational time between versions is inconclusive. The connection seems to be case by case, not in general along all the cases.

It is seen that some datasets with a small amount of associations run faster in leverage mode and slower in lift mode in the newer versions, while others with few found associations take a longer time to compute for both leverage and lift mode, while the dataset with a large amount of associations behave the opposite, with the newest version taking a longer time in leverage mode and a shorter time in lift mode. Due to these contrasting results within this

Figure 7: Time taken for Retail.fimi

report, there seems to be no indication of a clear relationship between associations found, computational time.

3   Further work

- More datasets with associations between 20,000 and one million (between multistore.txt and connect.fimi) will have to be tested so a more detailed picture can be built.

- More than 5 runs will have to be averaged, so the results are more accurate.

Figure 8: Time taken for Connect.fimi

References

[1] Geoffrey I. Webb *Discovering Significant Patterns*. Mach Learn (2007) 68: 1–33
DOI: 10.1007/s10994-007-5006-x

[2] Geoffrey I. Webb *Layered critical values: a powerful direct-adjustment approach to discovering significant patterns*. Mach Learn (2008) 71: 307–323
DOI: 10.1007/s10994-008-5046-x

[3] Geoffrey I. Webb, Jilles Veerken *Efficient Discovery of the Most Interesting Associations*. Trans. Knowl. Discov. Data 8, 3, Article 15 (May 2014), 31 pages.
DOI: http://dx.doi.org/10.1145/2601433

[4] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, Jun Sese *Statistical significance of combinatorial regulations*. Publication date: 2013/8/6, Volume: 110, Issue: 32, Pages: 12996-13001, Publisher: National Acad Sciences

[5] Geoffrey I. Webb *Read Layered Critical Values: A Powerful Direct-Adjustment Approach to Discovering Significant Patterns*. Machine Learning, 71(2-3), 307-323 [Technical Note], 2008
http://dx.doi.org/10.1007/s10994-008-5046-x

Figure 9: Time taken for multistore.txt.fimi

[6] *Frequent Itemset Mining Implementations Repository.* http://fimi.ua.ac.be/