The Benefits of Separating the Data Mining of Repositories from the Research Analysis Phase

CS680 Distributed Software Development Position Paper (Spring 2010)

Sammie Stahlback
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
sws28@drexel.edu

Jordan Osecki
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
jmo34@drexel.edu

Michael Kim
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
mk394@drexel.edu

1. ABSTRACT

There are many benefits from collecting history from software projects by mining software repositories. The data can be used to analyze which variables contributed to the success or failures of a particular task or possibly the whole project in a number of different categories. Much can be learned both by the developers of the software and by researchers from mining software repositories. It can give guidance to the developers regarding the current project and help researchers come to conclusions about more general results, benefitting future projects everywhere.

Historical information is typically mined from a wide range of sources. Data sources include code repositories, bug reports, emails, and other communication archives. Because there is such a wide breadth of data sources to choose from, there is also a wide range of data formats that need to be handled. Each type of repository is stored in different ways and in different places. In addition, each project stores each of them differently.

The result is that these historical information sources are very hard to navigate and parse for any particular software project and it is almost impossible to scale research results across multiple data sources. Because of the intense effort required for data mining, the conclusions and types of research that can be performed on software projects becomes very limited and the scalability of any conclusions can be extremely questionable. Research analysis is severely impacted by these constraints.

This position paper proposes that there is value in separating the data mining effort of software repositories from the research analysis process. Currently, data mining is performed alongside the analysis process, but there appears to be a need to develop data extraction techniques outside of the research itself because of the breadth and complexity truly involved in mining software repositories.

This paper will also investigate existing tools and techniques that are currently available for mining software repositories, to see if any can be incorporated into this new separate process. It will analyze what these tools try to accomplish, the effectiveness of them, and what can be learned from them as the need for mining software repository tools grows.

Treating the task of mining software repositories as its own discipline will help expand research in software development projects. Ways in which this will help include the following: creating new "best practices" for extracting data and promoting common data formats, tools, schemas, and algorithms for others to pickup and approve upon. As a result, it will allow researchers to more easily duplicate results of others or expand upon their research without having to reinvent the data extraction process.

By allowing others to more easily build upon common extraction techniques, researchers will be less hampered by this process and no longer forced to gloss over pertinent parts of the research. It will also bring additional researchers into the area of both mining software repositories and analysis.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining.

General Terms

Design, Documentation, Experimentation, Management, Measurement, Reliability, Standardization, Theory.

Keywords

Distributed software systems, co-located software systems, mining software repositories, research analysis, repositories.

2. Introduction

This section will expand upon the Abstract in all areas.

TODO

3. Current Method to Conduct Research

This section will describe exactly how research is currently conducted, including how each research effort has to re-define how to mine each repository, etc.

TODO

4. Available Methods of Data Mining

This section will describe other methods of data mining found in our research that could be utilized by researchers to bypass mining.

TODO

4.1 iSPARQL

This section describes how iSPARQL can be used to mine data.

TODO

4.2 CVSgrab

This section describes how CVSgrab can be used to mine data.

TODO

4.3 Survival Analysis Models

This section describes how Survival Analysis Models can be used to mine data.

TODO

4.4 TODO

This sub-section will describe any other tools found in our research.

TODO

5. Researcher's Method of Data Mining

This section will describe our methods of mining MySQL by putting their mailing lists into a more searchable format and how this has benefitted our analysis of the mailing list. We can conclude by stating how if it had already been in a DB, we could have immediately just started analysis.

TODO

6. Advancements and Techniques for MSR

This section will describe advancements and techniques in the field of mining software repositories.

TODO

7. The Benefits of Separation

This section will be the MAIN BODY of the paper, discussing our position and supporting it with everything just stated.

TODO

8. Business Preparation

This section describes how a business can prepare itself and its projects to implement our position and be able to set-up their stuff to make MSR easier to do.

TODO

9. Conclusion

This section will review the major points of our position and summarize everything else discussed, as well as reflect on any new conclusions or points discovered throughout the process.

TODO

10. Future Work

This section will describe what future work could be done to validate our position.

TODO

11. TODO

TODO

TODO

12. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

The team would also like to thank Professor Valetto for helping to guide this position paper in the right direction and raise issues for the team to consider.

13. BIBLIOGRAPHY

[1] Alonso, O., Devanbu, P. T., Gertz, M., "Database Techniques for the Analysis and Exploration of Software Repositories", *First MSR Workshop*, ICSE 204, Scotland UK.

[2] Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathin, A., "Mining Email Social Networks", *International Conference on Software Engineering, Proceedings of the 2006 international workshop on Mining software repositories*, 2006, ACM.

[3] Canfora, G., Cerulo, L., "Fine Grained Indexing of Software Repositories to Support Impact Analysis", *METRICS, Proceedings of the 11th IEEE International Software Metrics Symposium*, 2005, IEEE Computer Society.

[4] German, D., Cubranić, D., Storey, MA, "A Framework for Describing and Understanding Mining Tools in Software Development", *International Conference on Software Engineering archive, Proceedings of the 2005 international workshop on Mining software repositories*, 2005, ACM.

[5] Hassan, Ahmed e., "The Road Ahead for Mining Software Repositories". *16th Frontiers of Software Maintenance*, FoSM 2008, Sept 30 – Oct 2 2008, IEEE Computer Society, pp. 48-57.

[6] Kiefer, C., Bernstein, A., Tappolet, J., "Mining Software Repositories with iSPARQL and a Software Evolution Ontology", *Proceedings of the 29th International Conference on Software Engineering Workshop*, 2007, IEEE Computer Society.

[7] Robbes, R., "Mining a Change-Based Software Repository", *ICSEW, Proceedings of the 29th International Conference on Software Engineering Workshops*, 2007, IEEE Computer Society.

[8] Voinea, L., Telea, A., "Mining Software Repositories with CVSgrab", *International Conference on Software Engineering, Proceedings of the 2006 international workshop on Mining software repositories*, 2006, ACM .

[9] Wedel, M., Jensen, U., Göhner, P., "Mining Software Code Repositories and Bug Databases using Survival Analysis Models", *Empirical Software Engineering and Measurement*, 2008, ACM.

14. APPENDIX

TODO

TODO