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

# ABSTRACT

There are many benefits to mining software repositories to gather history from software projects. The data can be used to analyze which variables most likely 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 data source to mine, there is also a wide range of data formats that need to be handled and each of these types of repositories are stored in different ways and in different places. In addition, each project stores them differently.

The result is that these historical information sources are very hard to navigate and parse for particular software project, and are 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 project becomes very limited and the scalable of any conclusions is 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 historical repositories from the research analysis process. Currently, data mining is considered part of the analysis process, but needs to be its own step because of the breadth and complexity truly involved in mining repositories.

There are many benefits of this on the research of software development projects. Ways in which this will help include that it will help to produce new “best practices”, it will promote common data formats, tools, schemas, and algorithms for others to pickup and approve upon, and it will enhance the results and conclusions of software repository mining research by separating data mining, something which has to almost be treated as a research process itself because of its complexities, from analysis, making sure that researchers are no hampered by this process and forced to gloss over pertinent parts of the research.

**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.

# Introduction

This section will expand upon the Abstract in all areas.

TODO

# 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

# 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

## iSPARQL

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

TODO

## CVSgrab

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

TODO

## Survival Analysis Models

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

TODO

## TODO

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

TODO

# 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

# Advancements and Techniques for MSR

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

TODO

# 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

# 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

# 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

# Future Work

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

TODO

# TODO

TODO

TODO

# 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.

# 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. 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.
3. “A Framework for Describing and Understanding Mining Tools in Software Development” – FULL CITATION TODO
4. “Mining Software Repositories with iSPARQL and a Software Evolution Ontology” – FULL CITATION TODO
5. “Mining a Change-Based Software Repository” – FULL CITATION TODO
6. “Mining Software Repositories with CVSgrab” – FULL CITATION TODO
7. “Mining Software Code Repositories and Bug Databases using Survival Analysis Models” – FULL CITATION TODO
8. “Fine Grained Indexing of Software Repositories to Support Impact Analysis” – FULL CITATION TODO
9. “Mining Email Social Networks” – FULL CITATION TODO

# APPENDIX

TODO

TODO