Why mine software repositories?
- Understand software development and evolution
- Improve software design and reuse
- Use knowledge to plan future development
- Make predictions about software development
- Empirically validate research ideas

Why mine open source repositories?
- Easy to use
- Abstract details of how to mine software repositories
- Efficient and scalable
- Reproducible research results

Boa Goals

Boa Language
- MapReduce*
- Quantifiers*
- User Functions*
- Domain-specific Types
- Protocol Buffers
- Input reader

Boa’s Compiler
- MapReduce*
- Quantifiers
- User Functions
- Domain-specific Types
- Protocol Buffer
- Input reader

Runtime

Domain-specific Types
- Project: id, name, homepage_url, maintainers, code_repositories, etc
- CodeRepository: url, repository_type, revisions
- Revision: id, author, committer, commit_date, log, files
- File: name
- Person: email, real_name, username

Output Aggregators
- mean: An arithmetic mean of all data.
- sum: An arithmetic sum of all data.
- top(N): A sample that records the top N elements.
- maximum(N): A sample of the N highest weighted elements.
- minimum(N): A sample of the N lowest weighted elements.
- set(N): A set of at most N unique elements.
- collection: Collection of all data.

Data Infrastructure
- Replicates the data offline, for faster access
- Translates the data into a common format
- Provides mapping from raw data to domain-specific types in the language

Future work
- Other repositories (GitHub, etc)
- Other VCS (cvs, git, bzr, etc)

Boa: Analyzing Ultra-Large-Scale Software Repositories

Why mine open source repositories?

Number of words in commit log messages

The 10 most used programming languages


For more information about the Boa project contact Hridesh Rajan at hridesh@iastate.edu

http://boa.cs.iastate.edu/

Robert Dyer
Hridesh Rajan
Hoan Nguyen
Tien N. Nguyen