On the Efficacy of Keyword Searches to Find Meaningful Architectural Knowledge in Open-Source Software Mailing Lists

Bachelor Thesis for Computing Science

Andrew Lalis andrewlalisofficial@gmail.com

Supervised by **Dr. Mohamed Soliman** m.a.m.soliman@rug.nl

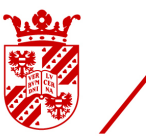
May 11, 2022

university of groningen

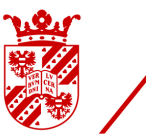
faculty of science and engineering

Abstract

This section will contain a brief overview of the research and conclusions. Test



Contents

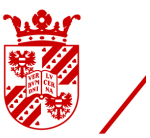


Introduction

In this paper, we'll explore the efficacy of using targeted keyword search queries to find architectural knowledge in mailing lists for open-source software projects. More simply put, we'll build tools, collect data, and analyze that data to qualitatively determine how effective certain keyword-based search queries are at finding useful information in large sets of emails, sent in mailing lists for developers to communicate about large open-source software projects.

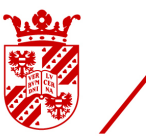
It is important to explore different avenues for acquiring knowledge, especially as the body of information grows exponentially with time. It is difficult for developers and software architects to make informed decisions about their own projects, because the source of their knowledge is distributed in a variety of disparate sources. If we can reliably glean information about software architecture and the successful (and unsuccessful) decisions that other field experts have made, we can make this knowledge more accessible for all.

For the purposes of this research, we will focus on analyzing the contents of mailing lists from three major open-source projects from the Apache Software Foundation: Hadoop, Cassandra, and Tajo. Mailing list data will be obtained from lists.apache.org, and this data will be indexed and searched over using Apache Lucene. We will categorize a subset of emails from these mailing lists, based on the type of architectural design decisions they contain.



Related Work

This section will contain an overview of lots of different sources and what they've done, and how what I'm doing is different.



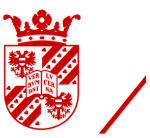
Research Questions

The main research question that this paper attempts to answer is summarized in the following question:

> What is the effectiveness of using keyword search queries to find architectural knowledge in open-source software mailing lists?

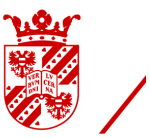
In addition to the main question we're attempting to answer, this paper will also discuss several other possible questions and answers that may be obtained using the data originally gathered for the main question.

1. Does there exist a relationship in the order in which architectural decisions are discussed, chronologically in an email thread? Is there significance in the order in which architectural decisions are made?
2. Is there a relationship between the content of discussion in emails, and related issues in issue/ticket boards such as JIRA and GitHub issues?
3. How can we use data gathered in this research to improve our search queries?



Methodology

This section will provide a detailed overview of how our data was obtained, and how one can replicate it.



Results

This section will show visualizations and aggregate data for results.

Conclusion

This section will answer the research questions and discuss further research.

References

[1] Tom den Boon. “Exploring the effectiveness of search engines for finding architectural knowledge in open source repositories”. In: *University of Groningen Student Theses* (), pp. 3–30. URL: https://fse.studenttheses.ub.rug.nl/25813/.

[2] Philippe Kruchten, Patricia Lago, and Hans van Vliet. “Building Up and Reasoning About Architectural Knowledge”. In: *Quality of Software Architectures* (), pp. 43–58. DOI: 10.1007/11921998_8.

Appendix

This section will contain larger bits of text or code or figures that aren't well suited to being placed inside the body of the paper.