Автономная некоммерческая организация высшего образования
«Университет Иннополис»
(АНО ВО «Университет Иннополис»)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
(МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)
по направлению подготовки
09.04.01 – «Информатика и вычислительная техника»

GRADUATION THESIS
(MASTER GRADUATE THESIS)
Field of Study
09.04.01 – «Computer Science»

Направленность (профиль) образовательной программы
«Анализ данных и искусственный интеллект»
Area of Specialization / Academic Program Title:
«Data Analysis and Artificial Intelligence»

Тема / Topic	**Простая система рекомендации коррективных действий, основанная на измерениях репозиториев программного обеспечения / A simple framework to recommend corrective actions based on the measurements of software repositories**

Работу выполнил / Thesis is executed by	**Данякин Кирилл Дмитриевич / Daniakin Kirill Dmitrievich,** **Жолха Фирас / Jolha Firas**	подпись / signature
Руководитель выпускной квалификационной работы / Graduation Thesis Supervisor	**Суччи Джианкарло / Giancarlo Succi**	подпись / signature

Иннополис, Innopolis, 2022

Abstract

Several works attempted to establish procedures to individuate bugs, defects, or anomalies during the different phases of software development, especially in the implementation phase. The mere detection of anomalies is not sufficient, though, at least until they get fixed. Corrective actions can be formulated to remove anomalies and enhance the software quality. To know whether an anomaly exists in a software, one must measure the software quality attributes related to it using specific software metrics. The main aim of this work is to highlight the industrial challenge in managing software development issues and find out and explain how to meaningfully attribute software metrics to useful corrective actions.

We have conducted a systematic literature review, where we have collected three kinds of data (metrics, anomalies, actions), which helped us individuate the dimensions of the problem. We found 384 software metrics, which are used to detect 374 anomalies related to 494 corrective and preventive actions.

Our findings show the need to formulate remedial strategies and build tools to automate the process of determining actions from abnormal metric values. Therefore, we propose a simple framework for detecting anomalies in software projects by using the measurements of the corresponding GitHub repositories and recommending corrective actions where needed. In this framework, we use clustering of software repositories, graph neural networks, and topic modelling.

Appendix E

Software Metrics Used in This Work

Table XX: GitHub API metrics used in this study.

Number	Metric Name	Metric Description
1	[Commits] Average additions	Average number of lines added through all the commits
2	[Commits] Average deletions	Average number of lines deleted through all the commits
3	[Commits] Average files changed	Average number of files changed through all the commits
4	[Commits] Average message length (chars)	Average length of messages within a commit (in chars)
5	[Commits] Count	Number of commits to this repository

6	[Commits] Days since first	Number of days from the first commit
7	[Commits] Days since last	Number of days from the last commit
8	[Commits] Maximum per day	Maximum number of commits per day
9	[Commits] Per day (True)	Average number of commits (counted only days with commits)
10	[Commits] Total lines added	Total number of lines added through all the commits
11	[Commits] Total lines deleted	Total number of lines deleted through all the commits
12	[Contributors Top-100] Average additions	Average number of lines added through all the commits by top-100 contributors
13	[Contributors Top-100] Average commits	Average number of commits made by top-100 contributors
14	[Contributors Top-100] Average deletions	Average number of lines deleted through all the commits by top-100 contributors
15	[Contributors Top-100] Average participation weeks	Number of participation weeks of top-100 contributors
16	[Contributors] Count	Number of contributors
17	[Forks] Count	Number of forks
18	[Forks] Max per day	Maximum number of forks per day
19	[Issues] Average body len (chars)	Average time to close an issue
20	[Issues] Average comment len (chars)	Average length of issue message (in chars)
21	[Issues] Average comments	Average number of comments
22	[Issues] Average labels	Average number of labels in issues

23	[Issues] Average title len (chars)	Average length of issue title (in chars)
24	[Issues] Count	Number of issues
25	[Issues] Labels	Number of issue labels
26	[Issues] Maximum per day	Maximum number of issues per day
27	[Issues] Open	Number of open issues
28	[Issues] Per day	Average number of issues per day (counted all days)
29	[Issues] Per day (True)	Average number of issues (counted only days with issues)
30	[Issues] Total comments	Total number of issue comments
31	[Pulls] Average body len (chars)	Average length of a body within a pull (in chars)
32	[Pulls] Average comments	Average number of comments in a pull
33	[Pulls] Average commits	Average number of commits in a pull
34	[Pulls] Average files changed	Average number of files changed through all the pulls
35	[Pulls] Average labels	Average number of labels within a pull
36	[Pulls] Average lines added	Average number of lines deleted through all pulls
37	[Pulls] Average lines deleted	Average number of lines added through all pulls
38	[Pulls] Average review comments	Average number of review comments in a pull
39	[Pulls] Average title len (chars)	Average length of a title within a pull (in chars)
40	[Pulls] Count	Number of pulls
41	[Pulls] Created per day (True)	Average number of created pull per day (counted only days with pulls)

42	[Pulls] Maximum created per day	Maximum number of created pulls per day
43	[Pulls] Total lines added	Total number of lines added through all pulls
44	[Pulls] Total lines deleted	Total number of lines deleted through all pulls
45	[Releases] Average asset downloads	Average number of asset downloads
46	[Releases] Average asset size	Average asset size
47	[Releases] Average assets	Average number of assets
48	[Releases] Average body len (chars)	Average length of a body within a release (in chars)
49	[Releases] Average title len (chars)	Average length of a title within a release (in chars)
50	[Releases] Count	Number of releases
51	[Releases] Tags	Number of tags
52	[Releases] Total downloads	Total number of downloads
53	[Repo] Age (days)	Age of a repository (in days)
54	[Repo] Branches	Number of branches
55	[Repo] Deployments	Number of deployments
56	[Repo] Milestones	Number of milestones
57	[Repo] Network members	Number of network members
58	[Repo] Programming Languages	Number of programming languages used in a repo
59	[Repo] Readme length (chars)	Number of readme length (in chars)
60	[Repo] Size	Size of repository
61	[Repo] Topics	Number topics of repository
62	[Repo] Watchers	Number of watchers
63	[Repo] Workflows	Number of workflows

64	[Stars] Count	Number of stars
65	[Stars] Maximum per day	Maximum number of stars per day
66	[Stars] Per day (True)	Average number of stars
67	[Workflow Runs] Average duration (ms)	Average duration of workflow runs (ms)
68	[Workflow Runs] Average fails per day	Average number of failed workflow runs per day (counted all days)
69	[Workflow Runs] Average failure duration (ms)	Average duration of failure workflow runs (ms)
70	[Workflow Runs] Average success duration (ms)	Average duration of success workflow runs (ms)
71	[Workflow Runs] Average successes per day (True)	Average number of success workflow runs per day (counted days with workflows)
72	[Workflow Runs] Count	Number of workflow runs

Appendix G

Topic Modeling Results

Here we present the results we obtained from topic modeling on issues and commits. The tables in this chapter show the topics represented by the top words and the number of words for each topic in the topic model. The labels of the topics are generated from the top words using the open source tool "keytotext" [159]. This tool uses pre-trained transformers to generate sentences from the input keywords and it is usually used for topic labeling and fine tuning the outputs of topic modeling but the drawback of this tool is that it is trained only on non-technical text whereas in our case we are dealing with technical text. In order to improve the meaningfulness of the generated labels, we manually paraphrased them but it is better to have an automatic way of labeling the technical text and indeed this method needs a lot of labeled technical text which is out of our concentration in this thesis.

Table XXIII: Topics extracted from “fixing” commits with corresponding 5-top words.

#	Commit topic	Word count per topic
1	['remov', 'test', 'file', 'non', 'api']	108
2	['renam', 'instead', 'provid', 'failur', 'ignor']	68
3	['resolv', 'issu', 'thi', 'work', 'delet']	111
4	['chang', 'sort', 'function', 'packet', 'read']	86
5	['flag', 'us', 'function', 'user', 'vip']	70
6	['return', 'object', 'address', 'maximum', 'superus']	55
7	['test', 'case', 'regress', 'npe', 'move']	74
8	['use', 'string', 'json', 'instead', 'version']	100
9	['display', 'format', 'featur', 'onli', 'version']	95
10	['debug', 'prefer', 'root', 'asdf', 'wifi_ssid']	57
11	['element', 'text', 'html', 'creat', 'document']	84
12	['connect', 'reset', 'client', 'cannot', 'respons']	85
13	['setstr', 'meshtast', 'run', 'configur', 'wifi']	69
14	['result', 'wrap', 'strategi', 'around', 'empti']	69
15	['onli', 'thi', 'user', 'allow', 'end']	162
16	['option', 'compil', 'resolv', 'execut', 'madskristensen']	59
17	['type', 'temperatur', 'field', 'rang', 'content']	85
18	['thank', 'transact', 'properti', 'object', 'error']	79
19	['chang', 'alarm', 'mode', 'class', 'see']	85
20	['handl', 'data', 'forc', 'includ', 'correct']	101
21	['structur', 'devic', 'serial', 'first', 'creat']	76
22	['task', 'claus', 'bf', 'queue', 'main']	42
23	['support', 'name', 'event', 'call', 'function']	98
24	['support', 'search', 'index', 'implement', 'dot']	73
25	['request', 'merg', 'pull', 'defin', 'synonym']	63
26	['messag', 'improv', 'client', 'text', 'bodi']	64
27	['check', 'code', 'error', 'miss', 'null']	119
28	['properli', 'valu', 'valid', 'select', 'disabl']	121
29	['bug', 'get', 'fail', 'charact', 'incorrect']	91
30	['make', 'use', 'order', 'doc', 'thi']	132
31	['issu', 'depend', 'script', 'replac', 'jar']	98
32	['updat', 'link', 'readm', 'version', 'instal']	92

Table XXIV: Commit topic labels generated from top 10 words and manually paraphrased.

#	Commit topic label
1	Remove test file and use different API
2	Fix the failure of a package by renaming or changing the ID
3	Resolve issue related to the work report
4	Change sort function and test the new API
5	When using meta queries, you need to take the size of the tables into consideration.
6	Update a configuration parameter to activate some command
7	Perform a regression test
8	Use a different json version
9	Refactor the display format of the app
10	Debug Reset button in the app
11	Use null-safe variables
12	Handle client connection to the server
13	Configure wifi module
14	Handle the authentication strategy with the client
15	Support online users
16	Add option to resolve the app build
17	Fix the content field in mgrid in python
18	Update local transaction
19	Change build mode of the project
20	Handle issue related to updating the data
21	Create specific volume for the device
22	Update task dialog
23	Set position:fixed for fullscreen mode
24	Implement support for misplaced dot on input
25	Uninstall the new installed package
26	Improve client notification
27	Check the error code of the null output
28	Properly add select validation to the form
29	Prevent the incorrect character number in the window of the android app
30	Fix ordering of list items
31	Fix script dependencies
32	Update readme file

Table XXV: Topics extracted from “bug” issues with corresponding 5-top words.

#	Issue topic	Word count per topic
1	[’animation’, ’route’, ’page’, ’app’, ’src’]	5937
2	[’server’, ’error’, ’client’, ’player’, ’reproduce’]	6767
3	[’component’, ’entity’, ’gree’, ’homeassistant’, ’py’]	4223
4	[’search’, ’name’, ’query’, ’str’, ’result’]	4862
5	[’mongodb’, ’connect’, ’kafka’, ’org’, ’converter’]	3205
6	[’time’, ’start’, ’service’, ’second’, ’stop’]	8979
7	[’dll’, ’php’, ’address’, ’vendor’, ’thread’]	4130
8	[’lib’, ’module’, ’node’, ’ghost’, ’logger’]	4857
9	[’data’, ’model’, ’train’, ’py’, ’input’]	4712
10	[’java’, ’com’, ’lang’, ’run’, ’util’]	10499
11	[’map’, ’rest’, ’metadata’, ’row’, ’swagger’]	2862
12	[’file’, ’line’, ’py’, ’lib’, ’self’]	15103
13	[’lua’, ’framexml’, ’bagnon’, ’interface’, ’component’]	4152
14	[’behavior’, ’reproduce’, ’expected’, ’bug’, ’step’]	8294
15	[’jar’, ’xml’, ’scala’, ’user’, ’play’]	4316
16	[’npm’, ’err’, ’node’, ’http’, ’react’]	4469
17	[’none’, ’highlight’, ’logback’, ’development’, ’exe’]	3215
18	[’build’, ’cpp’, ’lib’, ’library’, ’src’]	6593
19	[’string’, ’type’, ’data’, ’json’, ’value’]	11148
20	[’hie’, ’bios’, ’ghc’, ’haskell’, ’cabal’]	2903
21	[’angular’, ’cli’, ’ember’, ’mocha’, ’bower’]	2630
22	[’html’, ’template’, ’class’, ’href’, ’div’]	5925
23	[’module’, ’node’, ’lib’, ’user’, ’webpack’]	7942
24	[’product’, ’subscription’, ’cart’, ’order’, ’price’]	5285
25	[’airflow’, ’docker’, ’info’, ’compose’, ’postgresql’]	4437
26	[’link’, ’page’, ’menu’, ’click’, ’browser’]	9247
27	[’file’, ’directory’, ’root’, ’rw’, ’user’]	4544
28	[’go’, ’transaction’, ’git’, ’github’, ’com’]	3131
29	[’flutter’, ’src’, ’dart’, ’org’, ’springframework’]	3029
30	[’this’, ’bug’, ’product’, ’jet’, ’widget’]	8392
31	[’test’, ’download’, ’py’, ’error’, ’exception’]	3532
32	[’lua’, ’function’, ’defined’, ’bagnon’, ’addons’]	6146
33	[’date’, ’value’, ’status’, ’end’, ’summary’]	3228
34	[’io’, ’client’, ’netty’, ’vertx’, ’connection’]	4624
35	[’table’, ’mysql’, ’id’, ’db’, ’data’]	4603
36	[’item’, ’bag’, ’bank’, ’bagnon’, ’character’]	10712
37	[’log’, ’numjobs’, ’iodepth’, ’randread’, ’iop’]	7070
38	[’php’, ’woocommerce’, ’wp’, ’plugins’, ’index’]	7162
39	[’public’, ’new’, ’void’, ’string’, ’import’]	7590
40	[’run’, ’install’, ’command’, ’build’, ’sh’]	10533
41	[’java’, ’org’, ’junit’, ’hoverfly’, ’engine’]	3891
42	[’openid’, ’app’, ’cluster’, ’exporter’, ’google’]	4776
43	[’this’, ’issue’, ’work’, ’problem’, ’working’]	57062
44	[’this’, ’set’, ’value’, ’result’, ’using’]	25246
45	[’request’, ’response’, ’error’, ’url’, ’api’]	12986
46	[’atom’, ’app’, ’package’, ’remote’, ’edit’]	6747
47	[’user’, ’email’, ’password’, ’account’, ’permission’]	7112
48	[’field’, ’form’, ’value’, ’post’, ’type’]	9412
49	[’import’, ’python’, ’py’, ’module’, ’file’]	7068
50	[’system’, ’microsoft’, ’window’, ’runtime’, ’aspnetcore’]	6470
51	[’meshtastic’, ’serial’, ’debug’, ’gpio’, ’arduino’]	5322
52	[’android’, ’view’, ’script’, ’com’, ’app’]	5155
53	[’error’, ’context’, ’addon’, ’running’, ’software’]	6479
54	[’active’, ’language’, ’inactive’, ’syntax’, ’autocomplete’]	3240
55	[’div’, ’class’, ’de’, ’md’, ’col’]	5272
56	[’version’, ’docker’, ’issue’, ’false’, ’information’]	8347
57	[’version’, ’latest’, ’package’, ’dependency’, ’update’]	9705
58	[’image’, ’color’, ’style’, ’text’, ’font’]	6266
59	[’filter’, ’listing’, ’post’, ’page’, ’grid’]	10507
60	[’java’, ’samczsun’, ’sun’, ’net’, ’ssl’]	4183
61	[’error’, ’this’, ’file’, ’get’, ’any’]	27735
62	[’function’, ’error’, ’this’, ’console’, ’undefined’]	11047
63	[’rb’, ’gem’, ’ruby’, ’lib’, ’redmine’]	3383
64	[’embedded’, ’swimmingseadragon’, ’bagnon’, ’dev’, ’sanctimoniousswamprat’]	8110

Table XXVI: Issue topic labels generated from top 10 words and manually paraphrased.

#	Issue topic label
1	issue in the animation image in the web page.
2	bug in the version of the client player app
3	bug in climate component of the home assistant app
4	bug in search result of the query
5	issue in connection to mongodb
6	service crashes due to memory issues
7	issue in php deployment using magento
8	issue in logger module of node.js app
9	issue in input/output layer of the model during training
10	bug in concurrent thread of java app
11	issue in column/row display in swagger typescript rest api
12	issue in one of the python libraries
13	error in bagnon addon in lua component
14	unexpected behaviour of the app
15	bug in the the player written in scala
16	error in the installation of the npm modules react and node-gyp
17	error in gui development in aragon
18	unknown error in library while building a cpp app
19	issue in data type of json key/value content
20	build issue of haskell project
21	issue in installing the ember-mocha module of npm.
22	issue in html template.
23	issue in one of the react modules the in node.js app
24	issue in payment page of woocommerce app
25	bug in docker compose file related to postgresql
26	issue in open button of the page
27	illegal usage of the directory group permission
28	bug in transaction of the website's cpanel
29	issue in one bean of the flutter app
30	filter problem in woobuilder plugin
31	failure in integration with travis ci
32	issue in one of the lua addons
33	error in summary description of the app
34	client connect issue to the database
35	issue in backup of the database
36	issue in one of the addons
37	errors in configuration parameters in json files
38	plugin issue in woocommerece app
39	issue in access modifiers of a variable
40	docker build failure
41	issues in testing the app with a testing module
42	issue in database exporter
43	some unspecified issue
44	bug in setting values of some fields
45	error in request/response of the api client related to access token
46	issue in one of addons of atom editor
47	access permission bug in user login windows
48	bug in form fields while doing post
49	bug in one of the python libraries related to sharepoint services
50	object exception in aspnet app
51	bug in access port in arduino app
52	issue in using ajax in android app
53	error in one of the addons of the software patch
54	issue in themes of the app
55	bugs in html elements of the webpage
56	issue in docker version used
57	deprecation issue in one of the dependencies
58	bugs in css files
59	bugs in grid widgets used in the webpage
60	security issues in the app
61	file errors
62	type error in the console
63	issue in one of the ruby packages
64	issue in lua addons used for embedded modules

Discussion and Evaluation

Evaluation

Average precision@5 = 0.8
Average recall@5 = 0.512

Mean Average precision@2 = 0.5
Mean Average recall@2 = 0.08

Repository	Predicted issues	Predicted Relevance	Actual Relevance	Predicted commits	Predicted Relevance	Actual Relevance
smira/ txZMQ	42	0.56	Relevant	29	0.507	Relevant
				17	0.5068	Relevant
	56	0.554	Relevant	7	0.534	Not relevant
				28	0.5335	Not relevant
	35	0.553	Not relevant	30	0.5442	Not relevant
				29	0.5442	Relevant
	43	0.552	Relevant	23	0.544	Relevant
				10	0.5438	Relevant
	44	0.5513	Relevant	7	0.535	Not relevant
				22	0.535	Relevant

8 actual relevant issues

Evaluation

Average precision@5 = 0.54
Average recall@5 = 0.11

Mean Average precision@2 = 0.8
Mean Average recall@2 = 0.067

Repository	Predicted issues	Predicted Relevance	Actual Relevance	Predicted commits	Predicted Relevance	Actual Relevance
notanumber/xapian-haystack	57	0.506	Not relevant	28	0.55	Relevant
				14	0.54	Not relevant
	42	0.504	Relevant	29	0.508	Relevant
				17	0.503	Relevant
	43	0.503	Relevant	23	0.52	Relevant
				10	0.511	Relevant
	61	0.5029	Relevant	10	0.522	Relevant
				8	0.516	Not relevant
	18	0.5024	Relevant	10	0.53	Relevant
				9	0.525	Relevant

17 actual relevant issues

Evaluation

Average precision@5 = 0.125
Average recall@5 = 0.0625

Mean Average precision@2 = 0.35
Mean Average recall@2 = 0.1

Repository	Predicted issues	Predicted Relevance	Actual Relevance	Predicted commits	Predicted Relevance	Actual Relevance
denjones/hexo-theme-chan	44	0.564	Not relevant	7	0.53	Not relevant
				22	0.52	Relevant
	57	0.5631	Relevant	28	0.55	Relevant
				14	0.54	Not relevant
	46	0.5601	Not relevant	29	0.51	Not relevant
				22	0.5042	Not relevant
	37	0.5587	Not relevant	29	0.52	Not relevant
				23	0.513	Not relevant
	23	0.5586	Not relevant	30	0.506	Relevant
				29	0.504	Relevant

4 actual relevant issues

issue_topic	commit_topics
0	"[29, 2, 17]"
1	"[16, 29, 18, 3, 26, 2, 14]"
2	"[6, 22, 19, 29]"
3	"[31, 23, 7, 6, 29, 0, 1, 21, 8, 22, 26]"
5	"[1, 29, 22, 2, 7, 28, 30]"
6	[28]
7	"[18, 29, 31, 22, 17, 13]"
8	[29]
9	"[6, 12, 25, 29]"
10	"[29, 13]"
11	"[26, 24, 30, 18, 27, 28, 16]"
12	[14]
13	[29]
14	"[0, 29, 10]"
15	"[31, 29]"
17	"[26, 13, 15]"
18	"[14, 23, 28, 21, 27, 29, 2, 8, 22, 7, 10, 1, 19, 0, 24, 18]"
19	"[7, 17, 4, 0, 30]"
20	"[23, 31]"
21	"[30, 10, 3, 27, 31, 28, 15]"
22	"[29, 20, 7]"
23	"[29, 7, 30, 1]"
24	"[29, 21]"
25	"[31, 3, 2, 29, 27]"
27	"[5, 31]"
28	[7]
29	[30]
30	"[26, 8]"
31	[29]
33	"[11, 25]"
34	"[2, 8, 29]"
35	"[29, 8]"
37	"[4, 0, 30, 26, 28, 9]"
38	"[29, 17, 26, 22, 7, 6, 27, 2]"

issue_topic	commit_topics
39	"[28, 7, 22, 29, 18, 31]"
40	[29]
41	"[0, 29]"
42	"[25, 17, 30, 0, 19, 21, 22, 4, 2, 24, 7, 28, 20, 1, 16, 23, 31, 8, 29, 18, 27, 26, 6, 13, 15, 10, 3]"
43	"[29, 7, 23, 26, 28, 14, 4, 27, 17, 0, 25, 10, 6, 22, 24, 16, 1, 19, 8, 13, 3, 11, 9, 5, 30]"
44	"[2, 17, 14, 13, 22, 26, 1, 31, 8, 23, 29]"
45	[29]
46	"[25, 5]"
47	"[27, 29, 4]"
48	"[22, 29, 17, 5, 19]"
49	"[31, 19, 26, 15, 22, 8]"
50	"[9, 3, 6, 24, 29, 26, 18, 5, 12, 7, 1]"
52	[29]
54	"[8, 27, 15, 22, 16, 20, 31]"
55	[31]
56	"[31, 22, 25, 0, 30]"
57	"[19, 29, 27, 3, 28]"
58	"[3, 4, 17]"
59	"[29, 19, 28, 10, 21]"
60	"[26, 25, 29, 20, 30, 8, 18, 31, 4, 27, 1, 19, 16, 10, 15, 28, 2]"
61	"[31, 22, 24, 10, 23, 14, 29, 26, 27, 28, 30, 7, 11, 3, 0, 16, 19, 15, 17]"
63	"[28, 29]"