Patterns for cleaning and analyzing bug data

Rodrigo Souza*†, Christina Chavez*‡ and Roberto Bittencourt*§
*Dept. of Computer Science, Federal University of Bahia, Brazil
†Data Processing Center, Federal University of Bahia, Brazil
‡Fraunhofer Project Center, Federal University of Bahia, Brazil
§Dept. of Exact Sciences, State University of Feira de Santana, Brazil
Email: rodrigo@dcc.ufba.br, flach@dcc.ufba.br, roberto@uefs.br

***Abstract*—Bug reports provide insight about the quality of an evolving software and its development process. Such data, however, is often incomplete and inaccurate, and thus must be analyzed with care. In this paper, we present patterns that help both novice and experienced data scientists to extract useful information from bug data, while avoiding common mistakes.**

***Index Terms*—data analysis, mining software repositories, patterns, bugs.**

I. Introduction

Bug tracking systems store bug reports for a software project and keep record of discussion and progress changes for each bug. This data can be used not only to assess quality attributes of the software, but also to reason about its development process. Such richness of information makes bug reports an invaluable source for data scientists interested in software engineering.

However, bug tracking systems often contain data that is inaccurate [1], biased [2], or incomplete [3]. For example, changing the status of a bug report to `VERIFIED` usually means that, after a resolution was found to the bug, some kind of software verification (source code inspection, testing etc.) was performed and the resolution was considered appropriate. Sometimes, however, old bug reports are marked as verified just to help users and developers keep track of current bug reports [4].

Without proper guidance, it is easy to overlook pitfalls in the data and draw wrong conclusions. In this paper, we provide best practices and step-by-step solutions to recurring problems from cleaning up bug data up to deriving meaningful conclusions from bug reports.

Each solution is presented in a structured form called pattern. A pattern consists of 7 sections: (1) a short *name*; (2) the *problem* being solved; (3) a *context* in which the pattern can be applied; (4) one or more *solutions* to the problem; (5) a *discussion* of trade-offs and common mistakes to consider when using the pattern; (6) *examples* of the pattern in use; (7) *related patterns*.

II. Data Set

Each pattern contains an *Example* section with code snippets showing how to apply the pattern on real data. The snippets are written in R, a programming language for data analysis[1]. The data used are the bug reports from the project NetBeans/Platform, made available for the 2011 edition of the Mining Challenge[2].

The NetBeans project uses Bugzilla[3] as its bug tracking system, and stores all data in a MySQL database. Although the source code presented here refers to database tables and columns used by Bugzilla, it should work with any bug tracking system with minor modifications.

Although the full NetBeans data set contains 57 database tables, in this paper only three are used: `bugs`, `changes` (originally `bugs_activity`), and `comments` (originally `longdescs`). Notice that some tables and columns were renamed for clarity purposes.

The `bugs` table contains general information about each bug report, which is identified by a unique number (column `bug`). Each bug report has a `severity`, a `priority`, and two timestamps: the time of creation (`creation.time`), and the time of the last modification (`modif.time`). Table I shows a sample of the `bugs` table.

bug	severity	priority	creation.time	modif.time
397	normal	P4	1998-07-31	2008-12-23
427	normal	P4	1998-08-08	2008-12-23
479	normal	P2	1998-08-19	2008-12-23

TABLE I
Sample of the `bugs` table. The hour part of the times is ommitted.

The `comments` table contains comments that each `user` added to a `bug` report at some point in `time`. To reduce the file size, the text for the comment was replaced by its MD5 hash (column `comment.md5`). With high probability, two comments are represented by the same hash number if and only if they contain the same text. Table II shows a sample of the `comments` table.

The `changes` table contains all modifications made by users on bug reports over time. This includes changes in priority, status, resolution, and virtually any other field in a bug report. Each row contains the `new value` of a

[1]http://www.r-project.org/
[2]http://2011.msrconf.org/msr-challenge.html
[3]http://www.bugzilla.org/

bug	comment.md5	user	time
397	bafe80ea4ba2b73b6883243d8718ae3b	1887	2000-11-01
427	64d38f33f8848525a81b415406d064df	46	1998-08-08
427	264e61d606f0a4c0c8d477de00fba347	124	2000-07-25

TABLE II
SAMPLE OF THE `comments` TABLE. THE HOUR PART OF THE TIME IS OMMITTED.

`field` that was modified by a `user`[4] at some point in `time`. Table III shows a sample of the `bugs table`.

bug	user	time	field	new.value
427	17822	2009-10-30	resolution	WONTFIX
500	182	2002-04-12	bug_status	CLOSED
755	182	2002-01-11	bug_status	REOPENED

TABLE III
SAMPLE OF THE `changes` TABLE. THE HOUR PART OF THE TIME IS OMMITTED.

The *status* field is used to keep track of the progress of the bug fixing activity. A bug report is created with status `NEW` or `UNCONFIRMED`. Then, its status may be changed to `ASSIGNED`, to denote that a user has taken responsibility on the bug. After that, the bug is `RESOLVED`, then optionally `VERIFIED` by the quality assurance team and `CLOSED`, after the next software release comes out. If, after resolving the bug, someone finds that the resolution was not appropriate, the status is changed to `REOPENED`.

There are many forms of resolving a bug. To reflect that, when a bug status is changed to `RESOLVED`, the *resolution* field is changed either to `FIXED`—if the software was changed to solve the issue—, `WORKSFORME`—if developers were not able to reproduce the problem—, `DUPLICATE`—if a previous bug report describes the same problem—, among other resolutions.

All the data and code used in this paper is available online[5].

Write a short story referencing the patterns, that shows how one typically uses them

III. Look Out For Mass Updates

Problem

Discover changes to bug reports that were the result of a mass update.

Context

Changes to a bug report are often the result of an effort made by developers to triage, fix or verify a bug. Sometimes, however, hundreds or thousands of bug reports are changed almost simultaneously. Such changes are not caused by a burst of productivity; instead, they are the result of a mass update, often done with the purpose of cleaning up the bug tracking system.

[4]In this context, user denotes a user of the bug tracking system, which can be a developer or a final user.

[5]https://github.com/rodrigorgs/dapse13-bugpatterns

Mass updates can be motivated by a policy change. In Eclipse Modeling Framework, for instance, developers decided that bug reports containing fixes that were already published on their website should have the status `VERIFIED`. A mass status update was needed to make previous bug reports conform to the new policy.

Mass updates are characterized by a large number of changes of the *same type* (e.g., marking a bug report as `VERIFIED` or changing a target milestone) made by a single developer in a short period of time. Often such changes are accompanied by a comment that is the same for all changes in the mass update.

Solution

First, choose the type of change that you wish to analyze (e.g., changing a bug status to `VERIFIED`). Then apply one of the following solutions.

Solution 1. Select only the changes of the chosen type, along with the time of the change. Sort the changes by time and then plot the accumulated number of changes over time as a line chart.

The line is always increasing, but periods with many changes will stand out as steep ascents. Examine your vertical axis to assess whether such ascents represent a large number of changes (e.g., thousands). If this is the case, then it is likely that the changes were caused by a mass update.

Solution 2. Select only the changes of the chosen type, along with the date of the change, the user who made the change and the comment. Then, for each triple $\langle date, user, comment \rangle$, count the number of occurrences in the data set. Then, sort the triples by number of occurences. Triples with highest frequencies are good candidates for mass updates. You may analyse the comment text to see if they refer to cleanup, policy change or mass update in general.

Discussion

The first solution to find mass updates is more visual, but less accurate, as it does not take into account the user who made the changes. It is useful for a quick exploratory assessment of mass updates. The second solution is numeric and takes into account the users and their comments, but is more computationally intensive.

Examples

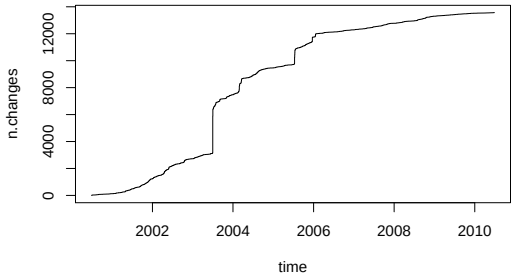
For example, Souza et al. [4] used the first solution (see Figure 2 in their paper) and a variation of the second without taking comments into account to detect mass verifications (i.e., changes in which the status is set to `VERIFIED`). They discarded all changes that were part of a mass update which updated at least 50 bug reports.

Solution 1. Here is a sample R code to plot the accumulated number of verifications over time.

Center figure

```
> ver <- subset(changes, field == "bug_status"
+   & new.value == "VERIFIED")
> ver <- ver[order(ver$time), ]
```

```
> ver$n.changes <- 1:nrow(ver)
> with(ver, plot(n.changes ~ time, type="l"))
```



In this chart, some line segments are almost vertical (e.g., the line between 2002 and 2004). Such segments mark dates when there were mass updates.

Solution 2. Here is a sample R code to count updates by user, date, and comment. First, select only verifications, and then use the `merge` operation to associate them with their respective comments. After that, do the counting as usual. Here, the 6 records with higher counts are shown.

```
> library(plyr)
> ver <- subset(changes, field == "bug_status"
+   & new.value == "VERIFIED")
> ver$date <- as.Date(ver$time)
> full <- merge(ver, comments)
> cnt <- count(full, c("date", "user",
+                      "comment.md5"))
> cnt <- cnt[order(cnt$freq, decreasing=T), ]
> head(cnt[, c("date", "user", "freq")])
```

date	user	freq
2003-07-01	17822	1703
2003-07-01	17822	972
2003-07-02	17822	447
2005-07-12	182	437
2005-12-20	182	209
2005-07-13	182	181

Related Patterns

How many changes in a day are considered normal and how many indicate mass updates? Be sure to Choose a Suitable Threshold (Section V).

IV. Old Wine Tastes Better

aka Let them mature / give 'em some time / a long future ahead / the benefit of the doubt

Problem

Discard bug reports that are work in progress.

Context

Temporal logic: bugs that will, *eventually*, be reopened.

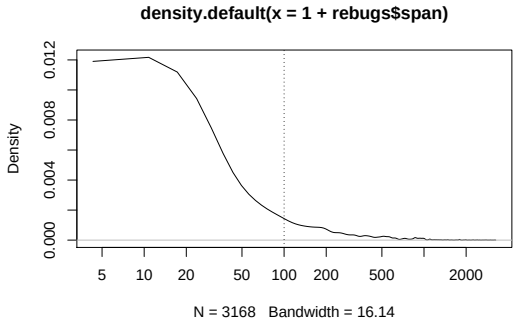
Bug reports change over time. Sometimes, one needs to classify a bug report according to the occurence of some change. For example, has a bug been fixed? Is this bug report a duplicate? A negative answer may turn into a positive one with the arrival of new data.

Therefore, for recent bugs, the outcome may still be unknown. For instance, when training a model to predict which bugs get reopened and which do not, it is likely that recent bugs have not been reopened—*yet*, because some of them may be in the future. The question is: how long to wait for the bug to be reopened before classifying it as not reopened?

Solution

Analyze the time spans of bug reports from their creation to the outcome being studied. For example, let's consider bug reopenings as the outcome. Select only reopened bugs and use a box plot or a density plot to visualize the distribution of time spans from creation to reopening. Then choose a threshold. It should be a time span that is large enough so that most reopened bug reports are reopened within that span. R code:

```
> library(data.table)
> changes <- data.table(changes)
> reopenings <- changes[
+   field == "bug_status" & new.value == "REOPENED",
+   list(first.reopen = min(time)),
+   by=bug]
> rebugs <- merge(bugs, reopenings)
> rebugs$span <- as.numeric(
+         rebugs$first.reopen - rebugs$creation.time,
+         units="days")
> plot(density(1 + rebugs$span), log="x")
> abline(v=100, lty=3)
```



Suppose 100 days was chosen as threshold. Now, compute the life span of the bug reports, from their creation to the time of the last change in the data set. Discard all bugs whose life span is less than the threshold. Such bugs are too recent: they do not have enough data to conclude if they will be reopened or not.

```
> threshold <- 100
> present <- max(changes$time)
> bugs$life.span <- present - bugs$creation.time
> valid.bugs <- subset(bugs, life.span >= threshold)
```

After discarding recent bugs according to this criteria, assume that bugs that were not reopened will also never be reopened in the future.

Discussion

It is a common mistake to keep recent temporal data in analyses. This is equivalent to choosing a threshold of 0, i.e., even if data is limited to a few days, current information is used to predict the future.

The choice of a threshold can be based on the desired confidence that current data can be extrapolated to the future. For example, consider a confidence level of 0.95 that bugs that were not reopened within a time span will not ever be reopened. The confidence can be approximated as the probability that a bug will never reopen divided by the probability that a bug will not reopen within the given time span[6].

Examples

Related Patterns

Be sure to [Choose a Suitable Threshold].

V. Choose a Suitable Threshold

Problem

Find a threshold for a metric so that data is separated into two distinct groups.

Context

Metrics are often used to help classify a data point into one of two groups. For example, they may help to determine if a data point is an outlier or not. When the mapping between data and groups is known for a subset of the data, then standard data mining techniques, such as cross-validation, can be leveraged to yield a threshold. The problem is more difficult if there is no information about the composition of the groups.

Solution

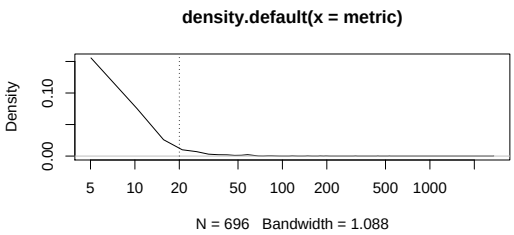
In fact, this is a hard, highly context-dependant problem, and it is related to finding outliers. Usually, the distribution of a metric is analyzed by means of a density plot, a box plot, or a bean plot (that mixes concepts from the later two). For heavily skewed data, logarithmic scales are used.

For example, consider the metric "number of verifications per developer and day in a bug tracking system" ("verification" means that the bug status was changed to `VERIFIED`. We would like to find a threshold such as, when the metric is greater than the threshold, it suggests that the verifications were actually just a mass verification for cleanup purposes.

It does not make sense to talk about a mass verification with just one or two verifications, so let's consider only greater values, then plot the density estimates. Because the distribution is heavily skewed, a logarithmic scale is used:

```
> metric <- metric[metric > 2]
> plot(density(metric), log="x")
> abline(v=20, lty=3)
```

[6]The proof is a direct application of Bayes' theorem.



A value around 20 appears to be a boundary between frequent and unfrequent values for the metric, as shown with a vertical line, and thus may be a good choice for a threshold to distinguish between usual and unusual numbers of verifications per developer and day.

Discussion

Looking for a suitable threshold can be highly subjective, and depends on the problem at hand. For example, if the problem requires an answer with high precision, it is better to choose a more restrictive threshold. Sometimes, thresholds can be inferred from natural human limits. For example, how many bugs can a developer fix in a single day? A reasonable answer can be used as a threshold to filter out invalid data.

It can be that the metric chosen is not sufficient to distinguish between the two groups. In that case, choose another metric or a combination of metrics.

Examples

Souza et al. [4] used density plots of the ratio between verifications and fixes by developers in order to find a threshold to help identify quality engineers.

Related Patterns

This pattern can be used together with several other patterns, e.g., [Mass Edits], [Know Your Subjects] etc.

References

[1] G. Antoniol, K. Ayari, M. Di Penta, F. Khomh, and Y.-G. Guéhéneuc, "Is it a bug or an enhancement?: a text-based approach to classify change requests," in *Proc. of the 2008 Conf. of the Center for Adv. Studies on Collaborative Research*, ser. CASCON '08. ACM, 2008, pp. 23:304–23:318.

[2] C. Bird, A. Bachmann, E. Aune, J. Duffy, A. Bernstein, V. Filkov, and P. Devanbu, "Fair and balanced?: bias in bug-fix datasets," in *European Soft. Eng. Conf. and Symposium on the Foundations of Soft. Eng.*, ser. ESEC/FSE '09. ACM, 2009.

[3] J. Aranda and G. Venolia, "The secret life of bugs: Going past the errors and omissions in software repositories," in *Proc. of the 31st Int. Conf. on Soft. Engineering*, 2009, pp. 298–308.

[4] R. Souza and C. Chavez, "Characterizing verification of bug fixes in two open source ides." in *MSR*, M. Lanza, M. D. Pent, and T. Xi, Eds. IEEE, 2012, pp. 70–73. [Online]. Available: http://dblp.uni-trier.de/db/conf/msr/msr2012.html#SouzaC12