MSR 2012 @ ICSE

目錄

Mining Software Repository 2012 @ ICSE

參加了今年的 MSR，會場在 University of Zurich。一大早來到大學，註冊有點 小插曲，顯然瑞士人搞不清楚中國人的名字，3 個楊(Yang)姓的中國人的名牌 被搞錯了。然後堀田學長的所屬被寫作了“Japan, Japan”，成爲了全日本的代表。

MSR(MicroSoft Research) talk @ MSR(Mining Software Repositories)

首先是來自微軟亞洲研究院(MicroSoft Research @ Asia, MSR Asia）的 Keynots，於是就變成了 MSR 在 MSR 的演講。MSR 的張冬梅(Dongmei Zhang)女士的演講 分爲關於 Software Analysis 和 XIAO 的兩部分。XIAO 是 MSRA 開發的 Code Clone Detector，似乎我要給井上研做的就是這個。想更多瞭解 Xiao 的細節，不過張女士 演講結束的時候的鼓掌導致了話筒的小故障。

Towards Improving BTS with Game Mechanisms

感覺這篇的內容基本上就是關於

http://www.joelonsoftware.com/items/2008/09/15.html

這裏寫到的東西，然後說同樣的理論是否可以用於 Issue Tracking 之類的事情上。 個人感覺這個意義不大，stackoverflow 之所以成功是因爲它把開源社區本身就 具有的名譽體系具現化了，本着大家都喜歡被別人奉爲大牛的心態，就如同 wikipedia 一樣。同樣的理論如果用於公司內部的 Issue Tracking 系統上，會得到 完全不同的東西吧。就像 MSDN

的組織方式雖然和 wikipedia 是一樣的，但是在 MSDN 裏找信息的感覺和在 wikipedia 完全不一樣。個人不太看好這個方向。

GHTorrent

這篇的 slide 在這裏可以看到：http://www.slideshare.net/gousiosg/ghtorrent-githubs-data-from-a-firehose-13184524

Data exporter for github. Github 的主要數據，代碼，已經可以通過 git 接口 獲得了，wiki 是 git 的形式保存的。所以這個項目的目的就是暴露別的數據，主要 是 issue tracking，code comments，這種。代碼 訪問 github api，然後用分佈式 實現以克服 api 的限制，然後提供 torrents 形式的 history 下載。github api 獲得 的 json 數據以 bson 的形式保存在 MongoDB 裏，解析過的有了 Schema 之後的數據保存 在 MySQL 裏並可以導出 SQL。

個人的想法，覺得數據如果能夠更統一，全部存在 Git 裏或許更好，像 Wiki 一樣。同樣是要暴露全部歷史記錄的目的，用 Torrent 自己實現的歷史遠不如用 Git 的 接口實現的歷史記錄方便吧，git blame 之類的也更方便追蹤 code comment 之類的 作者信息。當然對 git 的 raw date 直接讀寫，需要對 git 的內部原理有足夠的理解，或許只有 github 的人有這種能力了。

Topic Mining

用得兩個參數，DE 和 AIC，完全不能理解，過後研究。實驗針對了 Firefox, Mylyn, Eclipse 三個軟件。試圖從 Repo 中分析源代碼的 identifier 和 comments，找到 topic 和 bug 之間的關係，比如怎樣的 topic 更容易導致 bug。得出的結論似乎 也很曖昧，只是說核心功能被報告的 bug 更多，但是不知道原因。這只能表示核心 功能受到更多關注和

更多測試吧，並不能說明核心功能就容易產生 bug。

不過這個的 Slide 做得很漂亮，很容易理解。

SeCold

A linked data platform for mining software repositories

沒聽懂這個項目的目的。

The evolution of software

第二天的 Keynotes，關於將 Social Media 和 Software Development 相結合的想法。或許就是 Github 賴以成功的基礎。講到代碼中的 comment, Tags, uBlog, blog 之類 的 social 的特性和 IDE 的融合的趨勢。

Do Faster Releases Imporve Software Quality?

使用 Firefox 作爲例子。

結論是快速發佈導致 bug 更多，更容易 crash，但是 bug 更快得到修復，並且用戶 更快轉向新的發佈。

Security vs Performance Bugs in Firefox

Performance bugs are regression, blocks release.

一些感想

基於自然語義分析的 commit 分割

經常工具（比如 git）的使用者並沒有按照工具設計者的意圖使用工具，這給 MSR 帶來很多困難。舉個例子，git 有非常完美的 branch 系統，通常期望 git 的使用者 能夠在一次 commit 裏 commit 一個功能，比如一個 bug 的修復，或者一個 feature 的 添加，但是事實上經常有很多邏輯上的 commit 被合併在一個裏面了。

或許這不是使用者的錯，而是工具仍然不夠人性的表現。或許我們可以自動把 一次的 commit 按照語義分割成多個。

分割之後，可以更容易地把 issue 和 commit 關聯，也更容易組織更多的研究。

關於這次發表中大家用的 slides 系統

題目爲``Incorporating Version Histories in Information Retrieval Based Bug Localization'' 的人用的 slide 是 beamer 的。公式很多，overlay 很多，列表 很多，圖片很少，典型的 beamer 做出的 slide。思維導圖用得很不錯。今天一天 有至少 3 個 slide 是用 beamer 做的。

題目爲``Towards Improving Bug Tracking Systems with Game Mechanisms'' 的人用了 prezi，圖片很多，過度很多。但是比如沒有頁號沒有頁眉頁腳，正式 會議的場合不太方便。

至少有六個以上用了 Apple Keynotes，Keynotes 做出來的東西真的和 Powerpoint 做出來的很難區別，其中兩個人用了初始的主題所以才看出來。

剩下的自然是 PPT。MSRA 的張女士做的雖然是 PPT，倒是有很多 beamer 的感覺，比如頁眉頁腳和 overlay 的用法。這些如果都是 PPT 做出來的，會多很多額外的 人力吧。

值得一提的是有一個題目爲``Green Mining: A Methodology of Relating Software Change to Power Consumption''的人的 slide 全是``劣質'' 的手繪漫畫，效果意外地好,很低碳很環保很綠色很可愛。具體效果可以參考下面的動畫，雖然 現場看到的不是一個版本：

http://softwareprocess.es/a/greenmining-presentatation-at-queens-20120522.ogv

微軟是個腹黑娘！

嘛雖然這也不是什麼新聞了。MSR2012 的 Mining Challenge 的贊助商是微軟,管理 組織者來自微軟研究院,獎品是 Xbox 和 Kinect。然後今年的題目是：

> **Mining Android Bug**

我看到了微軟滿滿的怨氣……