Virheraportoijien virhemäärien jakaumat virhetietokannassa

(Valmiin työn esittely)

Jari Alahuhta

13.9.2010

Ohjaaja: TkT Mika Mäntylä

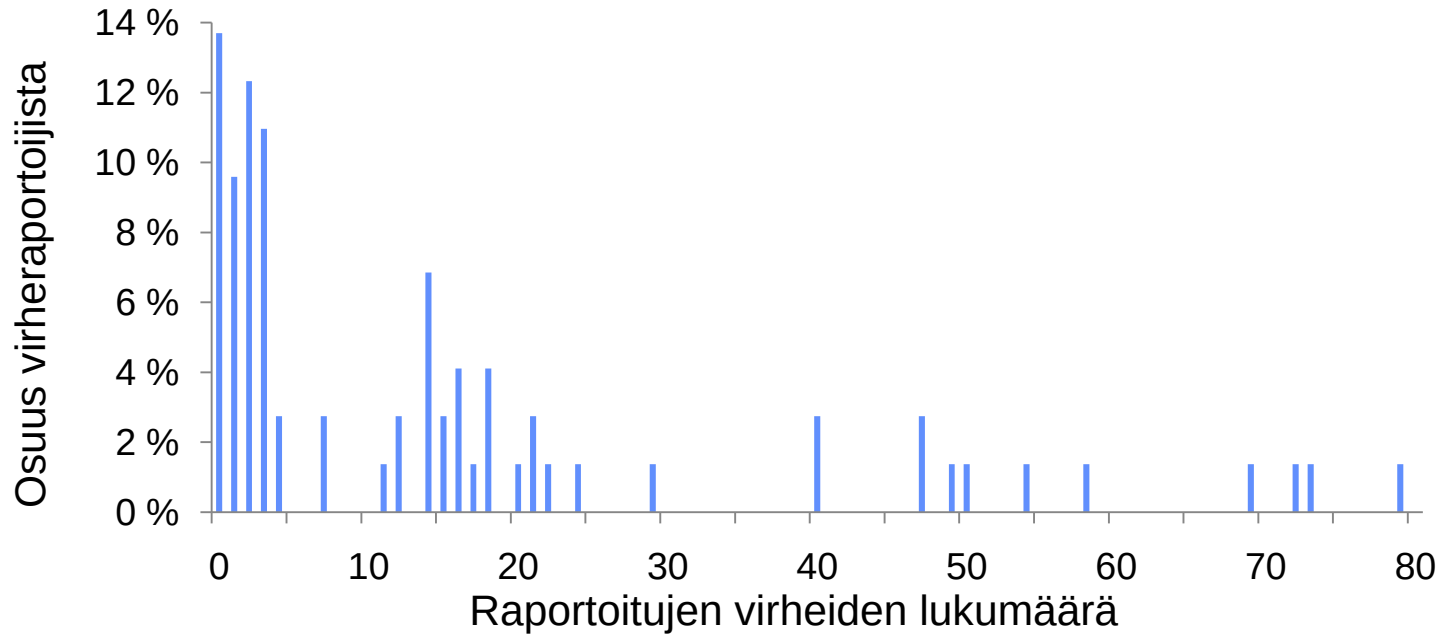
Valvoja: prof. Harri Ehtamo

Yleistä

- ohjelmistoissa virheitä, jotka estävät ohjelmistojen halutunlaisen toiminnan
- tieto virheestä siirrettävä virheen löytäjältä ohjelmistokehittäjälle
- virhetietokanta virheraportoijien ja ohjelmistokehittäjien kommunikaatioväline

Taustaa

- Mäntylä ym. (2010) tutkivat virheiden jakautumista virheraportoijien välille



Tavoitteet

- tavoitteena selvittää

    1. Voidaanko virhejakaumien katsoa noudattavan Pareto-jakaumaa tai jotain muuta teoreettista jakaumaa?

    2. Mistä raportointimäärien erot yksittäisten virheraportoijien välillä johtuvat?

- tavoitteita lähestyttiin kahden tutkimuskysymyksen avulla, joihin haettiin vastauksia tutkimalla empiiristä aineistoa ja olemassa olevaa kirjallisuutta

Tutkimuskysymykset

- **Tutkimuskysymys 1:** Noudattavatko virheraportoijien virhetietokantaan raportoimien virheiden jakaumat Pareto-jakaumaa tai jotain muuta tunnettua teoreettista jakaumaa tarkasteltavassa empiirisessä tutkimusaineistossa?
  - muut tutkitut teoreettiset jakaumat: eksponenttijakauma, lognormaalijakauma, Poisson-jakauma, Weibull-jakauma, Yulen jakauma ja Pareto-eksponenttijakauma
  - tutkittiin niin virhejakaumia kokonaisuudessaan kuin jakaumien pelkkiä häntiäkin

Tutkimuskysymykset

- **Tutkimuskysymys 2:** Mistä erot raportointimäärissä yksittäisten virheraportoijien välillä johtuvat?
  - **Osakysymys 1:** Onko virheraportoijien raportoimien virheiden lukumäärillä ja korjausprosenteilla jonkinlaista yhteyttä tarkasteltavassa tutkimusaineistossa?
  - **Osakysymys 2:** Keskittyvätkö runsaasti virheitä raportoivat virheraportoijat tyypillisesti raportoimaan virheitä erityisesti niistä komponenteista, joista ylipäätään on raportoitu eniten virheitä?

Aineisto

- dataa kolmen suomalaisen ohjelmistoyrityksen virhetietokannasta
- dataa kolmen avoimen lähdekoodin ohjelmiston virhetietokannasta
  - Apachen HTTP-palvelin, Linux-käyttöjärjestelmä, Mozillan Firefox-Internet-selain

	Apache	Linux	Mozilla	Yritys A	Yritys B	Yritys C
Virheraportteja	2389	6713	13610	981	790	1368
Korjausprosentti	35%	61%	24%	84%	91%	67%
Virheraportoijia	1763	3241	7435	69	39	45

Jakaumantutkimismenetelmä

- Clausetin ym. (2009) esittelemä menetelmä
  1. sovitetaan empiiriseen jakaumaan Pareto-jakauma suurimman uskottavuuden menetelmällä
  2. määritetään empiirisen jakauman ja sovitetun Pareto-jakauman yhteensopivuuden aste ja selvitetään, voidaanko empiirisen virhejakauman katsoa noudattavan Pareto-jakaumaa valitulla riskitasolla
  3. tutkitaan, sopiiko jokin muista tarkasteltavista teoreettisista jakaumista empiiriseen jakaumaan Pareto-jakaumaa paremmin

Tulokset

- Pareto-jakauma hyvä malli erityisesti
  - Apachen ja Linuxin kokonaisille virhejakaumille
  - Apachen, Linuxin, Mozillan ja yrityksen B virhejakaumien hännille
- lognormaalijakauma myös hyvä malli
- yritysaineistoista hankalaa tehdä luotettavia johtopäätöksiä pienten raportoijamäärien vuoksi

Tulokset

Virheraportoijien raportoimien virheiden lukumäärien ja korjausprosenttien korrelaatiot kussakin aineistossa

	Apache	Linux	Mozilla	Yritys A	Yritys B	Yritys C
Korrelaatio	0.07	0.02	0.18	0.02	0.20	0.07
p-arvo	**0.00**	0.29	**0.00**	0.89	0.22	0.65

- virheraportoijien raportoimien virheiden ja korjausprosenttien välillä ei ainakaan mitään selkeää lineaarista tilastollista riippuvuutta, vaikkakin lievää positiivista korrelaatiota nähtävissä

Tulokset

- jonkin verran näyttöä siitä, että eniten virheitä raportoineet virheraportoijat keskittyneet niihin komponentteihin, joissa ylipäätään eniten virheitä
  - Mozillalla eniten virheitä raportoinut prosentti virheraportoijista voimakkaasti keskittynyt siihen komponenttiin, josta ylipäätään raportoitu eniten virheitä
  - Apachella samantyyppistä, joskaan ei yhtä voimakasta keskittymistä
  - Linuxilla eniten virheitä raportoinut prosentti virheraportoijista keskittynyt muihin kuin niihin neljään komponenttiin, joista ylipäätään raportoitu eniten virheitä

Rajoitukset

- virhetietokannan datan luotettavuus
  - osa ongelmista suoraan kehittäjille?
  - raportoidaan toisten löytämiä virheitä?
- korjausprosentin merkitys
  - roolin tai aseman vaikutus korjausprosenttiin?
- testausasetelman monimutkaisuus
  - löytyykö virhe juuri testatusta komponentista?

Yhteenveto

- työssä saatiin näyttöä, että
  - Pareto-jakauma melko hyvä malli erityisesti virhejakaumien häntien käyttäytymiselle
  - raportoitujen virheiden lukumäärien ja korjausprosenttien välillä ei voimakasta yhteyttä
  - runsaasti virheitä raportoivat virheraportoijat jonkin verran keskittyvät komponentteihin, joissa eniten virheitä

Lähteet

Clauset, A., Shalizi, C. R. ja Newman M., E., J. 2009. Power-law Distributions in Empirical Data. SIAM Review. Vol. 51:4. S. 661–703. ISSN 0036-1445.

Mäntylä M., Iivonen J., Itkonen J. Who Tested My Software – an Industrial Case Study of Organization, Values, and Distribution of Testing. Tarkastuksessa, ESEM 2010.