Kuinka ohjelmoijat etsivät koodia

Jarmo Isotalo

Referaatti
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 9. syyskuuta 2015

Sisältö

1 Johdanto

Projektien koon kasvaessa ja yhä kehittyneempien haku mahdollisuuksien myötä koodin etsimisestä on muodostunut yhä oleellisempi vaihe sekä ohjelmoitaessa että koodia ylläpidettäessä. Sadowski, Stolee ja Elbaum esittelevä artikkelissaan "How Developers Search for Code: A Case Study" [1], kuinka ohjelmoijat etsivät koodia. He keskittyvät tapaustutkimuksessaan erityisesti tarkastelemaan miksi, miten ja milloin ohjelmoija etsii koodia käyttäen siihen tarkoitettuja työkaluja.

Käyttäen apuna logi analyysiä sekä juuri ennen koodin hakua tehtyjä kyselyitä he selvittivät vastauksia tutkimuskysymyksiinsä.

2 Tutkimusasetelma

Sadowski, Stolee ja Elbaum kertovat artikkelissaan, että Googlella suurin osa ohjelmakoodista on yhdesä isossa repositoriossa, jossa kaikki pääsevät katsomaan ja halutessaan hyödyntämään toistensa koodia. Googlella on käytössä sisäinen koodin haku työkalu, jossa hakua voi rajata mm. kielen ja sijainnin perusteella. Hakutulosten näyttämisen lisäksi koodihaku mahdollistaa koodin läpi navigoinnin kansiorakenteen sekä viittausten perusteella.

Tutkimuksdata kerättiin ennen koodin hakua aukeavalla kyselyllä sekä logi analyysillä. Kyselyä varten he loivat selain pluginin, joka tarjoaa tutkittaville kyselyn, kun he avasivat koodihaun. He kuitenkin rajoittivat sekä kyselyiden määrän enintään kymmeneen kyselyyn päivässä sekä kyselyiden väliajaksi vähintään 10 minuutin tauon. Tällä he pyrkivät vähentämään kyselyn tutkimuksen tuottamaa lisävaikaa tutkittaville, ja siten saamaan laadukkaampia vastauksia.

Toinen tutkimuksen datalähteistä on hakutyökalun logit, minne kaikki interaktio koodihaun kanssa tallentuu. Logeista he saavat selville anonymisoidun käyttäjän, interaktion kellonajan, hakutermit ja jokaisen käytetyn selaimen välilehden idn, sekä mitä tuloksia on klikattu. He korreloivat kyselytulokset sekä logimerkinnät kellonaikojen perusteella. Logeista eritettiin haku sessioita, siten että joka session välissä on vähintään kuuden minuutin tauko. Jokainen sessio koostuu siis koodihaun avaamisesta, hakemisesta, tulosten katselmoinnista ja mahdollisista jatkohauista.

Sadowski, Stolee ja Elbaum esittelevät seuraavat tutkimuskysymykset:

1. Miksi ohjelmoijat etsivät koodia
2. Missä kontekstissa haku suoritettiin
3. Millaiset haku kyselyn ominaisuudet
4. Mitä tyypillinen hakusessio tuottaa tuokseksi

5. Miten eri konteksti vaikuttaa hakutuloksiin

Tapaustutkimusesta he saivat seuraavat tulokset ja näin he hakivat:

1. **Miksi ohjelmoijat etsivät koodia**

   He kartoittivat syitä koodin etsimiseen lajittelemalla kyselyiden vapaamuotoiset vastaukset omiin ryhmiinsä. Tämä tehtiin siten, että vapaamuotoiset kyselylomakkeen vastaukset kirjoitettiin lapuille ja lajiteltiin omiin kategorioihinsa, kunnes kaikki 3 lajittelijaa (artikkelin kirjoittajat) olivat tyytyväisiä lajitteluut. Näin he saivat aikaan korkean luokan kategorioita. tämä tuotti tuloksiksi seuraavat lajit: miten tehdä jotakin, mitä koodi tekee, miksi koodi toimii kuten toimii, koodin paikantamiseen ja kuka muutti ja milloin.

   Lajittelun tuloksista kävi ilmi, että 33.5% hauista oli vastaamaan kysymykseen, miten jokin asia tulisi tehdä. 26% puolestaan vastasivat kysymykseen mitä, eli hakutapahtuma koostui pääosin koodin lukemistesta ja toteutuksen tarkastelusta sekä yleisen koodityylin selvittämisestä. 16% hauista vastaa kysymykseen missä, eli mistä tiettyä koodia käytetään ja mitä koodia se käyttää sekä hauista missä tavoitteena oli selvittää tietyn koodin sijainti repositoriossa. 16% hauista keskittyy vastaamaan kysymykseen miksi tämä koostuu pääosin selvityksitä, kuten miksi jokin toimii eritavalla kuin miten ohjelmoija ajatteli, sekä mahdollisen muutoksen sivuvaikutuksia. 8.5% hauista keskittyy vastaamaan kysymykseen kuka ja milloin; pääosin selvittääkseen kuka muutti koodia ja milloin, sekä koodin omistajuutta, mikä halutaan selvittää mm. jotta oikea henkilö reviewaisi tulevat muutokset.

2. **Missä kontekstissa haku suoritetaan**

   Koodia haetaan monista eri syistä Googlella; 39% hauista tehdään silloin kun työskentelee muutoksen parissa. Myös sekä koodin katselmus vaiheessa että ongelman ratkaisussa koodia tyypillisesti haetaan. Kuitenkin suurin osa hauista tehdään jo osin tuttuun koodiin. Hakuja tehdään myös selvittääkseen miten koodi toimii, miten sitä käytetään.

3. **Hakukyselyn ominaisuudet**

   Hakukyselyn ominaisuuksia selvitettäessä hakukyselystä poistettiin automaattisesti lisätyt termit (nykyinen kansio yms). He havaitsivat, että noin neljännes hauista rajoittaa tuloksia sijainnin perusteella kun taas kielen perusteella hakuja rajattiin vain noin viidessä prosentissa hauista. He havaitsivat paljon tilanteita, jossa tehtiin kaksi peräkkäistä hakua siten, että hakujen välillä ei ollut muute interaktiota koodihaun kanssa. Tarkemmin tutkittuaan he havaitsivat että suuri osa näistä oli joko kyselyn muokkaamista tai hakualalueen rajaamista tiettyyn kansioon. He myös havaitsivat että keskimäärin näiden kahden kyselyn

välissä oli vain kahdeksan sekunnin ero; siinä ajassa ohjelmoija ehti tarkastelemaan alkuperäiset hakutulokset ja tarkentamaan hakuaan.

4. **Tyypillisen hakusession sisältää** Logianalyysistä he havaitsivat, että hakusessio kestää keskimäärin 3 minuuttua ja 30 sekuntia ja sisältää 2 selaimen välilehteä. Ohjelmoijat tekivät keskimäärin 12 hakua päivän aikana (tutkimus kesti 15days, siirrä ylemmäs). He tunnistivat myös useita erilaisia patterneja TOOD: miten kivasti tähän.

5. **Kontekstin vaikutus hakupatterneihin** Kyselyistä ja logianalyysistä saamillaan tuloksia yhdistämälle he pystyivät tunnistamaan tarkempia patterneja. Mikäli ohjelmoija halusi oppia koodin attributesista TODO he harvoin klikkasivat yhtään tiedostoa, tämä luultavasti johtus joko siitä että vastaus näkyi hakutlosten esikatselussa, tai koska haku tuotti vain yhden tuloksen. He myös vahvistivat oletuksensa, että ohjelmoija joka ei tunne hakemaansa koodia kunnolla navigoi ja hakee siihen liittyen enemmän kuin koodin tunteva henkilö.

3 discussion

Tutkimuksessa yhdistelty data (sekä log että kyselyt mahdollistivat tarkemman insights miten ohjelmoijat etsivät koodia. He myös havaitsivat, että ohjelmoijat etsivät koodia useasti, he hakevat paljon esimerkkejä ja he etsivät pääosin jo itselleen tuttua koodia. He myös havaitsivat, että ohjelmoija useasti tarkentaa hakutermejään iteratiivisesti.

Ajattele käyttäjän tarpeita, tarjoa esimerkkejä Tulosten pohjalta heillä on koodinhakutyökalujen luojille muutamia ehdoituksia:

- Tarjoa tuloksien esikatselukka.
- Tarjoa mahdollisuus filtteröidä tiedoston sijainnin ja/tai tekijöiden perusteella.
- Tarjoa tulokset huomioiden monipuolisempi konteksti, mahdollisesti jopa seuraamalla käyttäjän toimia muualla.
- Harkitse työkalun integraatiota kehytysympäristöön.

Sadowski, Stolee ja Elbaum mainitsevat vielä lopuksi, että tulokset eivät

Lähteet

[1] Sadowski, Caitlin, Stolee, Kathryn T. ja Elbaum, Sebastian: *How Developers Search for Code: A Case Study.* Teoksessa *Joint Meeting of the European Software Engineering Conference and the Symposium on the*

Foundations of Software Engineering (ESEC/FSE), 1600 Amphitheatre Parkway, 2015.