Forráskód hibák jellemzése GitHub-on végzett adatbányászat segítségével

Gyimesi Gábor
II. évf. programtervező informatikus MSc
Gyimesi Péter
II. évf. programtervező informatikus MSc

Témavezető: Tóth Zoltán Gábor és Dr. Ferenc Rudolf

SZTE TTIK Szoftverfejlesztés Tanszék

A szoftverekben elkerülhetetlenek a kódolási hibák a gyakori változások, a szűk határidők és a nem megfelelő specifikáció miatt. Ebből kifolyólag fontos, hogy minél több eszköz a segítségünkre legyen e hibák felkutatásában. Ennek egyik lehetséges módja, hogy megismerjük a korábbi hibák jellemzőit, és ezen jellemzők alapján próbáljuk meg beazonosítani azokat. A dolgozat célja az ilyen kódolási hibák jellemzése.

Manapság egyre nagyobb népszerűségnek örvendenek a szoftverek fejlesztését segítő forráskód hosting szolgáltatások, mint például a GitHub, SourceForge vagy a Google Code. Ezek számos szolgáltatást nyújtanak, melyek közül a számunkra legfontosabbak a verzió- és a hibakövető rendszerek. A szoftverek fejlesztéséhez elengedhetetlenek a verziókövető rendszerek, mert lehetővé teszik, hogy egyszerre több fejlesztő is dolgozhasson hatékonyan a projekten, valamint eltárolja minden verzióját a forráskódnak. A hibakövető rendszerek egységes felületet biztosítanak a hibák bejelentéséhez, illetve a hibához tartozó információkat ezen a felületen követhetjük. Bizonyos rendszerek lehetővé teszik ezen információk elérését egy jól definiált API-n keresztül. Az ilyen forráskód hosting szolgáltatásokon egyre több a nyílt forráskódú projekt, melyek kihasználják ezeket a lehetőségeket. Az így kinyerhető adatok felhasználhatóak elemzésére abból a célból, hogy a szoftverekben előfordult hibákat vizsgáljuk. A hibabejelentéseket felhasználva beazonosíthatjuk a hibás, majd pedig a már kijavított forráskódokat, és ily módon jellemezhetjük azokat akár a statikus forráskód metrikákkal, akár egyéb, a kigyűjtött adatokból előállítható metrikákkal.

Az adatok gyűjtéséhez a GitHub-ot választottuk és kijelöltünk 13 java projektet, melyek használják a hibakövetést, és bejelölik a hibával kapcsolatos forráskód módosításokat. A jellemzés céljából meghatároztuk a hibás, valamint a már kijavított kódrészek statikus forráskód metrikáit, illetve definiáltunk további 3 fájl szintű és egy forráskód verzió szintű metrikát. Ennek eredményeképp előállítottunk egy adatbázist, mely jellemzi a vizsgált projektekben előforduló hibákat, így felhasználható többek között a szoftverhibák automatikus felismerésének az elősegítésére.

Összességében kidolgoztunk egy rendszert, mely egy forráskód hosting szolgáltatásról, jelen esetünkben a GitHub-ról képes kigyűjteni adatokat, melyekből aztán felépít egy adatbázist a bejelentett szoftverhibák jellemzéséhez.