

1. Raziskovalna organizacija:

Univerza v Ljubljani, *Fakulteta za računalništvo in informatiko*

2. Ime in priimek mentorja:

Janez Demšar

3. Področje znanosti iz šifranta ARRS:

2.07 Računalništvo in informatika

4. Kontaktni e-naslov mentorja:

janez.demsar@fri.uni-lj.si

5. Kratek opis programa usposabljanja:

Razvoj tehnik za implementacijo spletnih orodij za odkrivanje znanj iz podatkov

Orodja za odkrivanje znanje iz podatkov (data mining) na principu pretoka podatkov (data-flow) temeljijo na uporabi grafično predstavljenih gradnikov (widget). Vsak gradnik opravlja določeno preprosto nalogo, kot je branje podatkov, izbor podatkov ali značilk, vizualizacija in podobno. Uporabnik definira postopek obdelave podatkov tako, da izbira in povezuje gradnike v shemo. Shema se "izvaja" tako, da si gradniki med seboj pošiljajo podatke skladno z vzpostavljenimi povezavami. Takšen pristop k odkrivanju zakonitosti se je v preteklosti izkazal za zelo praktičnega, saj tudi laičnim uporabnikom omogoča preprosto programiranje analize podatkov. Na njem temelji več popularnih komercialnih in odprtokodnih orodij, kot sta recimo Knime, ki ga razvija Univerza v Koblenzu in Orange, ki ga razvijamo v naši skupini.

V zadnjem času se veliko programske opreme seli na splet, s čimer je poenostavljeno sodelovanje med uporabniki, prenašanje med napravami, uporabo na tabličnih računalnikih, uporabo kapacitet v oblakih ipd. Našteto bi bilo smiselno tudi na področju odkrivanja znanj iz podatkov, vendar zanj še ne obstajajo spletna orodja, ki bi temeljila na principu pretoka podatkov.

Razlogi za to niso zgolj praktični, temveč tudi teoretični. Orodja za odkrivanje znanj iz podatkov morajo delovati z velikimi količinami podatkov, kar zahteva veliko procesorskih zmogljivosti. Procesiranje zato ne more potekati na odjemalcu, v brskalniku, temveč na strežniku. Če želimo zagotoviti odzivnost aplikacije, se zato pojavi ozko grlo pri sprotnem prenosu podatkov (ki so predvsem vizualni) od strežnika do odjemalca. Obenem je potrebno zaradi varčevanja z računskimi kapacitetami poskrbeti, da se vsi izračuni opravijo šele ko in če je potrebno.

Kandidat se bo moral spopasti z več problemi.

- Določiti arhitekturo za leni izračun, ki je potrebna za implementacijo tovrstnih orodij. V skupini smo že opravili nekaj preliminarnih raziskav na tem področju, s katerimi pa smo odkrili predvsem omejitve trenutne tehnologije in nekaj možnih poti k rešitvi.
- Razviti tehnike za razdelitev dela med strežnikom in odjemalcem, ki bo minimizirala prenos prek internetnih povezav.
- Zasnovati vtičnike za podatkovne baze, ki bodo prenesli del obdelav v samo bazo in s tem omogočili strežnikom delo z večjimi zbirkami podatkov.
- Prilagoditi standardne metode vizualizacij za uporabo prek spleta in za delo z ogromnimi zbirkami podatkov. Mnoge trenutne metode so neprimerne, ker bi zahtevale bodisi preveč procesiranja na strani odjemalca bodisi hiter prenos slik oz. videa prek spleta.
-

Delo se bo smiselno vklapljal v trenutne projekte raziskovalne skupine, kot je projekt AXLE v okviru FP7 ter sodelovanje z akademskimi ustanovami (npr. Baylor College of Medicine, ZDA) in industrijskimi partnerji (npr. Astra Zeneca).