Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Aljaž Srša 63120233 in Gregor Sušnik 63120102

Implementacija spletnega pajka

Poročilo pri predmetu Iskanje in ekstrakcija podatkov s spleta

Asistent: asist. prof. dr. Slavko Žitnik

**Povzetek**

Spletni pajek [1] imenovan tudi spletni robot je program, ki na strukturiran način avtomatično preiskuje spletišče z namenom indeksiranja vsebine. V seminarski nalogi poskusiva implementirati preprostega spletnega pajka. Pri implementaciji ne uporabljava knjižnic, ki delno že implementirajo delovanje spletnega pajka. Posledično sva naletela na številne težave, ki sva jih morala reševati sistematično in premišljeno. Na koncu preizkusiva izdelanega spletnega pajka na v naprej določenih spletnih mestih in pridobljene rezultate vizualno predstaviva in pokomentirava.

1. **Uvod**

Iskanje vsebine po spletišču je za uporabnika nadvse preprosto. V priljubljen iskalnik vnese iskano geslo in že v nekaj trenutkih dobi ogromno spletnih strani, ki se navezujejo na iskano geslo. V ozadju delovanja iskalnika pa se neprestano dogaja veliko več. Spletni pajek obiskuje in preiskuje ogromno količino spletnih strani in hkrati obdeluje še večjo količino podatkov z namenom, da poskuša kategorizirati (indeksirati) spletno stran. Sama arhitektura spletnega pajka je odvisna od njegovega namena in je v večini primerov kompleksna. Spletni pajek mora za dobro delovanje obiskati in obdelati čim večje število spletnih strani, to pa pomeni izkoriščanje vseh možnih strojnih virov, ki so na voljo.

V prvi seminarski nalogi smo se lotili problema implementacije spletnega pajka. Pri implementaciji smo bili omejeni pri uporabi knjižnic. Uporabili smo lahko le tiste, s katerimi si lahko nekoliko olajšamo programiranje, ne pa tistih, ki v neki meri že implementirajo spletnega pajka. Ideja seminarske naloge je v poljubnem programskem jeziku izdelati bolj preprost primerek spletnega pajka in poizkusiti njegovo delovanje na določenih spletnih mestih.

1. **Metode pridobivanja podatkov**
   1. **Uporaba regularnih izrazov**
   2. **Uporaba XPath**
   3. **Uporaba RoadRunner algoritma**

1. **Rezultati**

**Literatura**