

Een evaluatieframework voor content-gebaseerde algoritmen

Probleemstelling:

Hoewel de enorme hoeveelheid van content op het internet tegemoetkomt aan de noden van de hedendaagse gebruiker, zorgt dit overaanbod ervoor dat gebruikers de meest relevante informatie niet meer kunnen terugvinden. Online shops, video-sharing website en nieuwssites hebben vaak duizenden items beschikbaar voor klant. De traditionele filtersystemen zoals sleutelwoord-gebaseerde zoekmechanisme voldoen vaak niet meer aan de vereisten van gebruikers die een gepersonaliseerd aanbod van relevant informatie verlangen.

Aanbevelingssystemen kunnen een oplossing bieden voor dit probleem door het filteren van het contentaanbod op basis van de persoonlijke interesses van de klant. Deze aanbevelingssystemen hebben echter nood aan profiel informatie en (in het geval van content-gebaseerde suggesties) ook aan relevante metadata die de items beschrijft. Vele websites beschikken echter vaak niet over deze relevante metadata, waardoor deze een beroep moeten doen op een algemene beschrijving van de content of op de content zelf (bijvoorbeeld nieuwsartikel).

Binnen WiCa is er een aanbevelings service ontwikkeld die de ontwikkelaar toelaat om aanbevelingen te generen op website die de beschikbare webservice gebruiken. Enkele belangrijke uitbreidingen zijn nodig om deze service ruim inzetbaar te maken:

- Kernwoord extractie: de service laat geen extractie van kernwoorden uit teksten toe. Het extraheren van keywoorden laat echter toe om de aanbevelingsalgoritmen efficiënter te laten werken.
- Een evaluatieframework voor content-gebaseerde algoritmen: Gangbare metrieken voor de evaluatie van de prestatie van aanbevelingsalgoritmen kunnen gebruikt worden voor het testen van de beschikbare extractietechnieken. Voorbeelden zijn precision en recall. Het framework zal moeten zoeken naar de beste combinatie van aanbevelingsalgoritme en extractietechniek voor een dataset van een grote sociale netwerksite.

Doelstellingen:

Deze thesis bestaat uit twee aansluitende delen en bevat volgende taken:

- Studie naar technieken voor keyword extractie uit teksten.
- Implementatie van een extractie mechanisme (eventueel aan de hand van bestaande libraries)
- Architectuur studie naar een raamwerk om snel verschillende algoritmen te evalueren aan de hand van reeds verzamelde data.
- Implementatie van een evaluatie systeem

Technologieën: Java

1 student

Onderzoek naar content-gebaseerde aanbevelingssystemen en sleutelwoord-extractie algoritmen: 15%

Architectuur onderzoek naar een test framework voor evaluatie van algoritmen: 15%

Implementatiewerk: 70%

Locatie: Universiteit Gent campus Zuiderpoort, thuis

Inlichtingen:

Promotor, begeleiders: prof. dr. ir. Luc Martens, ir. Toon De Pessemier, ing. Kris Vanhecke

Voor vragen kan je contact opnemen met Toon.DePessemier@intec.ugent.be of Kris.Vanhecke@intec.ugent.be

Meer informatie over de Wireless & Cable onderzoeksgroep is te vinden op <http://www.wica.intec.ugent.be>.