PRÍRODOVEDECKÁ FAKULTA UPJŠ V KOŠICIACH

UNIFIKÁCIA PRODUKTOV INTERNETOVÝCH OBCHODOV

Mgr. Peter Šinaľ

Školiteľ[1]*:* RNDr. Peter Gurský, PhD.

[1]*Ústav Informatiky, Prírodovedecká fakulta UPJŠ, Jesenná 5, 040 01 Košice*

V súčasnej dobe používatelia internetu čoraz viackrát využívajú služby rôznych internetových obchodov pre zakúpenie ľubovoľných produktov. Každý internetový obchod je tvorený práve jedným produktovým katalógom. Produktový katalóg obsahuje informácie o jednej alebo viacerých doménach produktov. Hlavná požiadavka používateľov je mať maximálne množstvo informácií o produkte a porovnanie týchto informácií skrz viaceré internetové obchody.
Problém ktorý riešime je, ako zjednotiť dáta o produktoch danej domény z rôznych zdrojov. Ako riešenie sme navrhli algoritmus unifikácie/zjednocovania produktov z rôznych zdrojov podľa atribútov. Počiatočným krokom je získanie informácií o produktoch z viacerých internetových obchodov. Po získaní súborov, ktoré predstavujú produktové stránky (obsahuje informácie o produkte) konkrétnej domény, sa vykoná anotácia a následne extrakcia dát. Po získaní reprezentácie produktov prostredníctvom hodnôt svojich atribútov sa vykonáva unifikácia skrz jednotlivé produktové katalógy, teda identifikovanie rovnakých produktov v rôznych produktových katalógoch. Pri návrhu algoritmu sme sa čiastočne inšpirovali riešením problému unifikácie, ktoré bolo použité v rámci stohového systému[1], avšak prioritným zdrojom pre návrh nášho algoritmu unifikácie bol všeobecný model unifikácie[2]. S predpokladom získania presnejších výsledkov sme rozšírili všeobecný model unifikácie o vlastnosti atribútov - zdrojovú závislosť a relevantnosť atribútov. Vo výsledkoch algoritmu unifikácie chceme prezentovať množinu produktov bez prítomnosti duplicít - rovnakých produktov.

Litratúra:

1. Jiří Dokulil, Jakub Yaghob, Filip Zavoral: Evoluce replikačních algoritmů v stohově orientovaných systémech, report, 2006
2. Peter Christen: Data Matching, Concepts and Techniques for Record Linkage, Entity resolution, and Duplicate Detection, 2012

Prosíme, ohraničte dĺžku svojho abstraktu maximálne na jednu stranu!