Összetett rendszer vállalkozások címeinek Webről történő automatikus összegyűjtésére

Nagy István
V. évf. közgazdasági programozó matematikus

Témavezetők: Farkas Richárd tudományos segédmunkatárs, Csirik János egyetemi tanár
MTA-SZTE Mesterséges Intelligencia Tanszéki Kutatócsoport

Egy vállalkozás számára rengeteg információ található a Weben a potenciális partnerekről, vevőkről vagy szállítókról. Célom egy olyan on-line kereséseken alapuló rendszer megvalósítása, amely automatikusan képes ezen információk összegyűjtésére. Ennek megvalósíthatóságát illusztrálandó, dolgozatomban egy olyan rendszert mutatok be, amely az egyes tevékenységi körökhöz tartozó magyarországi vállalkozások neveit és címeit automatikusan gyűjti össze. A problémára a megoldást egy olyan összetett rendszer szolgáltatja, amely számos részprobléma azonosítását és megoldását igényli (ez a dolgozat egyik legfontosabb eredménye). Az egyes részproblémák megoldása során egyaránt alkalmaztam gépi tanuló algoritmusokat, szabályalapú módszereket és egyéb heurisztikákat. A főbb komponensek és az egész rendszer empirikus kiértékelésére létrehozott keretrendszerrel a dolgozatban számszerű eredményeket közlök.

A három főbb komponens:

- Weblapok osztályozása: Az egyes vállalatok internetes oldalainak az azonosítása érdekében a kereséshez használt on-line keresők eredményeit automatikusan „céges” és „nem céges” csoportokba kell sorolni. A probléma megoldásához a pozitív és jelöletlen példákból való tanulás megközelítését alkalmaztam, ahol a standard módszer egy módosított változatát is ismertetem.
- Címek és cégnevek azonosítása: a letöltött weboldalakon az egyes cégek neveinek és címeinek automatikus jelölésére volt szükség (tulajdonnév felismerési feladat). Ez a részprobléma mind szabály alapú megközelítéssel, mind gépi tanulással megvalósításra került, ezáltal lehetőség nyílt a két módszer összehasonlítására is.
- Címek és cégnevek automatikus normalizálása és összerendelése.