

Projecte final de carrera

Grup de recerca ITAKA

Directors: Montserrat Batet i David Sánchez

Extracció d'informació a partir de recursos web semi-estructurats

La Web és la font d'informació més gran que existeix. Actualment, co-existeixen tot tipus de recursos digitals de qualsevol tipus de domini. Tot i això, el profit que se'n treu és baix degut a la manca d'estructura (és a dir, són documents escrits en llenguatge natural sense seguir cap patró fix). Això dificulta molt el seu processament doncs, des del punt de vista informàtic, és molt complicat "interpretar" la informació que contenen.

Tot i això, a la Web també existeixen recursos semi-estructurats en forma de documents que, tot i estar escrits en llenguatge natural, sí segueixen unes certes pautes en la seva forma. Analitzant aquesta estructura bàsica, és possible extreure una sèrie de característiques estructurades que serien més fàcils de processar. Aquest tipus de característiques poden ser tant numèriques, com textuales.

L'objectiu d'aquest projecte és el d'estudiar diversos d'aquest repositoris (tipus Wikipèdia o e-bay) per tal de dissenyar un sistema capaç d'extreure'n un conjunt de característiques estructurades (ex: donat un anunci d'un producte a e-bay, extreure'n la categoria en que s'ha inclòs, el preu, el nombre de visites, la condició, etc., així com un conjunt de termes rellevants extrets de la descripció textual de l'anunci). Aquests resultats es representaran en forma de taula de manera que permeti aplicar-hi eines d'anàlisi de dades (data-mining).

Les principals tasques d'aquest projecte són:

- Buscar i estudiar diversos repositoris on-line de documents semi-estructurats (anuncis, notícies, articles, etc.).
- Estudiar i escollir les característiques que es poden extreure per a cada tipus de repositori.
- Estudiar les eines disponibles per extreure informació a partir de documents semi-estructurats a partir d'un conjunt de regles (wrappers).
- Dissenyar i implementar un sistema, el més genèric possible, d'extracció de característiques d'un repositori determinat. S'extrauran dos tipus de característiques:
 - Característiques dependents del format (ex: el preu d'un anunci a l'e-bay) que es farà a partir de la seva representació al document.
 - Característiques de text lliure, que s'extrauran a partir de l'anàlisi del text en llenguatge natural (en anglès) contingut al document. Aquest és un punt especialment important doncs, per una banda, caldrà estudiar i utilitzar diverses eines per tal d'analitzar sintàcticament el text i, per altra, caldrà detectar i escollir aquells termes que són més importants de cara a descriure semànticament el recurs actual.
- A partir de les característiques extretes, compondre una sèrie de fitxers de resultats en diversos formats estàndards (ex: excel) per tal que facilitar-ne el seu processament.