Enunciats dels projectes de PROP

Quadrimestre de primavera, curs 2014/15

Enunciat 1: Diputats dels EUA

Als Estats Units no existeix la disciplina de vot. És a dir, tots els diputats tenen llibertat per votar el que vulguin a qualsevol afer del Congrés, siguin del partit que siguin. Això fa que sovint no es pugui preveure el resultat d'una votació a priori, ja que al no haver-hi disciplina de vot, la proporció de diputats de cada partit no aporta tota la informació.

Tot i això, un periodista ha estat revisant el resultat de les últimes votacions i les reunions que han mantingut diferents membres del Congrés, i té la sospita que hi ha alguna mena de patró en els resultats de les votacions, és a dir, creu que hi ha grups de diputats que voten sempre el mateix tot i no tenir aparentment res en comú i tenir llibertat de vot.

El que vol aquest periodista és un sistema que li permeti detectar grups de diputats afins entre sí, independentment del partit al qual pertanyin, segons diferents criteris modificables com ara votar el mateix a les votacions parlamentaries, assistir al mateix tipus d'actes, celebrar reunions conjuntament, etc; per tal de confirmar la seva teoria. Podem suposar que el periodista té accés a la informació que calgui per detectar aquestes coincidències (amb una llista d'assistència a actes, registre de reunions conjuntes, etc.).

Enunciat 2: YouTube Mix

YouTube Mix és una funcionalitat de YouTube que genera de forma automàtica llistes de reproducció formades per cançons que els usuaris han pujat a Internet. Tot i que un criteri clar d'agrupació és l'autoria de la cançó o l'estil musical, es dóna molta importància també a l'activitat dels usuari, és a dir, per exemple, si dues cançons acostumen a ser escoltades per les mateixes persones amb poc temps de diferència, serà molt probable que les dues siguin agrupades dins de la mateixa llista de reproducció, encara que musicalment siguin cançons completament diferents.

Es demana que es creï un sistema d'emmagatzematge de cançons (tant sols la informació de les cançons, no la cançó en si) i que incorpori la funcionalitat d'auto-generar llistes de reproducció segons criteris modificables. El sistema també comptarà amb una llista d'usuaris i per cada usuari un registre amb les cançons que ha escoltat.

Enunciat 3: Wikipedia

La Wikipedia d'una llengua es pot considerar com un graf dirigit on les pàgines i categories son els nodes del graf i els enllaços entre aquests elements corresponen als arcs del graf. Els enllaços són de diferent tipus, en particular ens interessaran els enllaços entre diferents categories.

- <Pàgina, categoria>: Donada una pàgina ens dona les categories a les que pertany la pàgina.
- <Categoria, pàgina>: Donada una categoria ens dona les pàgines que conté.
- <Categoria, subcategoria> : Donada una categoria ens dona les seves subcategories.

- <Categoria, supercategoria> : Donada una categoria ens dona les seves supercategories.
- Enllaços de sortida: Donada una pàgina ens dona les pàgines a les quals aquesta pàgina apunta.

Apart d'aquests enllaços hi ha d'altres que no tenim en compte, com els interwikies que apunten a Wikipedies d'altres llengües o *external links* que apunten a direccions de fora de la Wikipedia.

El graf d'una Wikipedia es sol considerar dividit en dos subgrafs, el graf de pàgines i el graf de categories (WP^{pag} i WP^{cat}). Nosaltres ens limitarem al WP^{cat}.

Idealment el graf WP^{cat} hauria de ser un arbre o almenys una taxonomia, però en realitat no és així. Hi ha cicles, salts cap a endarrere, categories de servei i altres problemes. Per això ens plantegem buscar agrupaments de categories fortament relacionades entre sí dintre de WP^{cat}. Com a informació per obtenir les comunitats haureu de fer servir per cada categoria les seves subcategories, les seves supercategories i les seves pàgines associades. No cal fer servir el text de les categories ni el de les pàgines. Podeu fer servir els títols de pàgines i categories.

Comentaris comuns per a tots tres enunciats:

- Tots tres problemes són variants de la *detecció de comunitats en grafs*. Cada projecte ha de funcionar amb tres algorismes: Louvain, Newmann-Girvan, i *Clique percolation*. Això no vol dir necessàriament que *cada* grup de *cada* cluster hagi d'implementar els tres algorismes
- Les dades s'han de poder definir via el programa o importar des d'un fitxer de text.
- El sistema haurà de permetre la modificació posterior de la solució proposada.

Dates dels lliuraments:

- Primer: dilluns 16 de març
- Segon: dilluns 4 de maig (especificació de classes compartides: 13 d'abril; acceptació de classes compartides: 14 de maig)
- Tercer: dimecres 3 de juny (lliuraments interactius: a partir del 8 de juny).