Dokumentumkollekciók vizualizálása kulcsszavak segítségével

Berend Gábor, Erdős Zoltán, Farkas Richárd

Szegedi Tudományegyetem,
TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.,
e-mail:berendg@inf.u-szeged.hu, erdos.zoltan91@gmail.com, rfarkas@inf.u-szeged.hu

Kivonat A dokumentumkollekciókban történő eligazodás kapcsán hasznos segítséget képesek nyújtani a különféle intelligens adatvizualizációs eljárások. Egy lehetőség, ha a dokumentumkollekció alkotóelemeire mint irányítatlan, súlyozott gráf csúcsaira tekintünk, a köztük lévő kapcsolatok erősségeit pedig a dokumentumpárokat jellemző hasonlóságértékek adják, majd az előzőek szerint definiált gráfot jelenítjük meg valamilyen gráfrajzoló eljárás segítségével.

Az efféle megközelítések alkalmazása során azonban – különösen nagy méretű adatbázisok esetében – gyakorlati nehézségekbe ütközhetünk. Nagy számú csúccsal és éllel rendelkező gráfok esetében nehézkessé válhat azok áttekinthetősége, valamint a csúcsok koordinátáinak meghatározásáért felelős optimalizációs számítások konvergenciája is lassú lehet.

Az általunk megvalósított alkalmazás – korábbi munkáinkra [1,2] is építkezve – dokumentumkollekciók vizualizációját hajtja végre azok kulcsszavaira támaszkodva. Előnye, hogy a vizualizációs szempontból nehézséget jelentő méretű korpuszok megjelenítését is lehetővé teszi azáltal, hogy a dokumentumok hierarchikus klaszterekbe történő besorolásának elvégzése után bizonyos csomópontokat összevonva ábrázol. A tematikus dokumentumegyüttesek aprólékos felépítésének megismerésére pedig felhasználói interakció útján nyílik lehetőség.

A teljes dokumentumgráf megjelenítésével kapcsolatos nehézségek olyan módon kerültek tehát áthidalásra, hogy az alkalmazás inicializálása során a dokumentumok kulcsszavaik alapján történő klaszterezéseként előálló főbb témák – melyek száma jellemzően jóval elmarad a dokumentumok számától – kirajzolása történik meg. A klaszterezés végrehajtása egy különösen jól skálázódó algoritmus segítségével [3] történik, amivel akár százezres nagyságrendű dokumentumkollekciók klaszterezése is megoldható az alkalmazás inicializálása során. A vizualizálandó korpusz klasztereinek feldolgozását megkönnyítendő, a dokumentumklasztereket összegző, azokat a többi klasztertől megkülönböztető kulcsszavak kiválasztása és megjelenítése történik meg információelméleti megfontolások mentén.

Demóalkalmazásunk a Magyar Nemzeti Szövegtárban található újságcikkek vizualizációját hajtja végre. Amiatt ugyanakkor, hogy az alkalmazás bemenetéül egy egyszerű – a megjelenítendő dokumentumok kulcsszavait tartalmazó – szöveges állomány szolgál, így adaptálása más jellegű

szövegekre könnyen végrehajtható. Természetesen a bemeneti állomány a kulcsszavak mellett tartalmazhat egyéb adatokat is (pl. dokumentumközi hivatkozással kapcsolatos információkat), így ezek beépítése sem okozna nehézséget a vizualizációs eljárásba.

Kulcsszavak: automatikus kulcsszókinyerés, dokumentumvizualizáció

Köszönetnyilvánítás

Berend Gábor publikációt megalapozó kutatása a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosítószámú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése országos program című kiemelt projekt keretében zajlott. A projekt az Európai Unió és az Európai Szociális Alap társfinanszírozásával valósult meg. A további szerzőket a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta.

Hivatkozások

1. Berend, G., Vincze, V., Farkas, R., Zsibrita, J., Jelasity, M.: Kulcsszókinyerés alapú dokumentumklaszterezés. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 251–262
2. Farkas, R., Berend, G., Hegedűs, I., Kárpáti, A., Krich, B.: Automatic free-text-tagging of online news archives. In: Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, Amsterdam, The Netherlands, The Netherlands, IOS Press (2010) 529–534
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008+