Webkeresés Monte Carlo módszerrel

Rácz Balázs

PhD értekezés összefoglalója

Témavezető:
Dr. Benczúr A. András

Adatbányászat és Webes Keresés Kutatócsoport
Informatikai Kutatólaboratórium
MTA Számítástechinkai és Automatizálási Kutatóintézet

valamint

Matematikai Intézet, Algebra Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem

Budapest
2009

1. Bevezetés

A XXI. századot gyakran nevezik az *információs társadalom* századának. Az információ megszerzésének egyik leggyorsabb, és így legfontosabb forrása az Internet. Ennek megfelelően a szoftvertechnológiai ipar legdinamikusabban növekedő területe az internetes szolgáltatásoké, melyek vezető képviselői a *keresőmotorok*, mint például a Google, Yahoo vagy Microsoft Live, amelyek napjainkra majdnem kizárólagos módját képezik az Interneten való tájékozódásnak.

A terület tudományos és technológiai kihívását a probléma mérete adja: A Web konzervatív becslések szerint is többszázmilliárd oldalt tartalmaz, és ez a szám már majdnem két évtizede töretlenül tartó exponenciális növekedésnek örvend. Ilyen méreteknél még a legegyszerűbb matematikai konstrukció is – mint például egy lineáris egyenletrendszer megoldása – megoldhatatlan feladat lehet.

A dolgozatban bemutatott eredmények ilyen skálázódási problémák megoldását adják bizonyos a keresőmotor magjában futó algoritmikus feladatokra. Maguk a problémák és az absztrakt megoldások a terület úttörőitől származnak, komoly érdeklődést kiváltva. Ennek ellenére az általunk kidolgozott algoritmusok voltak az elsők, amelyek a teljes Web méreteire ésszerű költséggel és komoly megszorítások nélkül alkalmazhatóak.

Egy különlegesen fontos tulajdonsága az általunk adott megoldásoknak, hogy nemcsak elméletileg alkalmazhatók a Webre, hanem rendkívül gyakorlatiasak is: közelről követik a keresőmotorok belső felépítését, erősen párhuzamosíthatóak, a kiszolgálórutinok pedig az ipari alkalmazhatóság olyan szigorú kritériumainak is megfelelnek, mint például a hibatűrés.

Ezen előnyök elérése érdekében a módszereink csupán *közelítő algoritmusok*. A dolgozat fontos eleme, hogy a közelítés pontosságát elemezzük, valamint formálisan bebizonyítjuk, hogy a tekintett problémákra csak közelítő megoldásokban reménykedhetünk, mivel a pontos válasz kiszámításához szükséges erőforrásokra általunk bebizonyított alsó becslések meghaladják a teljes Föld számítási kapacitását. Ezen kívül kísérletileg is ellenőriztük az algoritmusaink Webre való alkalmazhatóságát, mind a skálázódási, mind az eredmények minőségének szempontjából.

2. Áttekintés

A dolgozat három téziscsoportban mutatja be az eredményeket.

Az első téziscsoport a személyre szabott kereséssel, illetve *személyre szabott rangsorolás*sal foglalkozik. Az általános Web-keresési megoldások a találatokat egy globális rangsor szerint rendezve mutatják meg a felhasználóknak, amely függ a lekérdezéstől, de nem függ a felhasználótól. Számtalan példát lehet mutatni arra, hogy a weblapok relevanciája az egyes felhasználók számára jelentősen különbözik, például egy történelemtanár és egy számítógépes szakember számára más és más lapok lehetnek érdekesek és relevánsak még ugyanazon lekérdezés esetén is. A személyre szabott keresés lehetővé teszi az egyes felhasználóknak, hogy kifejezzék egyéni preferenciájukat, amely a rangsorolási eljárás paraméterévé válik. Mivel a PageRank az egyik legkiterjedtebben kutatott és alkalmazott rangsorolási módszer, ennek paraméterezhető változata, a *személyre szabott PageRank* különös figyelmet érdemel [22]. Minden korábbi eljárás amelyet személyre szabott PageRank pontszámok kiszámítására dolgoztak ki [10, 16, 17], igen jelentős korlátozásokat tartalmazott a megengedett paraméterezésre nézve [13]. Az általunk kidolgozott algoritmus az első, amely a teljes Web méretében alkalmazható és tetszőleges személyre szabást lehetővé tesz.

A második téziscsoport *hiperlink-alapú hasonlóságkeresés*sel foglalkozik. A hasonlóságkeresés egyik motivációja a fejlett adatbányászati algoritmusok széles családja, amely könnyen kiértékelhető hasonlósági függvényeket alkalmaz a magas szintű elemzések elkészítéséhez. Az átlagfelhasználó számára ennél fontosabb azonban, hogy a Weben való keresés egy alternatív paradigmáját jelenti a hasonlóságkeresés, amelyben nem kulcsszavakkal próbáljuk meg kifejezni a kívánt információt, hanem példa alapján. Ebben a modellben megadunk a keresőmotornak egy számunkra érdekes oldalt, mire a keresőmotor az azon oldalakhoz hasonló további oldalakat fog ajánlani, amelyek így vélhetően szintén érdeklődésünkre tarthatnak számot. A klasszikus link-alapú hasonlósági függvények,

mint például a ko-citáció a szociális hálózatok kutatásából származnak, és a hasonlóság mértékét kizárólag az egyes pontok szomszédainak vizsgálatával határozzák meg – a ko-citáció esetén például a közös hivatkozó cikkek száma. Mivel a Web nem csak méretében, hanem komplexitásában, átmérőjében is nagyobb, mint ezek a vizsgált szociális hálózatok, ezért ezek a függvények nem képesek megragadni a kívánt mélységet, és meglehetősen rossz minőségű eredményeket szolgáltatnak. A vizsgálatunk tárgya a Jeh és Widom által javasolt *SimRank* hasonlósági függvény [15], amely a PageRank-hoz hasonló rekurzív definíciót alkalmaz. A publikációinkban bemutatott algoritmus volt az első, amely ezt a hasonlósági függényt néhány százezer lapnál nagyobb adathalmazon ki tudta számítani.

A két fenti területen hasonló felépítésű a kutatásunk: Először Monte-Carlo módszeren alapuló approximációs algoritmust adunk az absztrakt problémára. Ezek után a konvergenciasebesség becslésére mondunk ki tételeket. Majd alsó becsléseket fogalmazunk meg a pontos eredményeket előállító algoritmusok számítási igényeire. Végül valódi Webről származó nagyméretű adathalmazokon kísérletileg igazoljuk a módszereink működőképességét, minőségét illetve a paraméterek megfelelő beállítását.

A harmadik téziscsoport a hasonlósági függvényekre vonatkozó lehetetlenségi eredményeket terjeszti ki. Ebben a részben az eldöntési problémával foglalkozunk: létezik-e két olyan oldal a hálózatban, amelyeknek egy meghatározott számnál több közös szomszédjuk van? Absztraktan megfogalmazva, ez ekvivalens azzal, hogy tartalmazza-e a gráf a $K_{2,c}$ teljes páros gráfot részgráfként. A feladat tárkomplexitását vizsgáljuk az adatfolyam-modellben: egy $\mathcal{A}$ algoritmus szekvenciálisan végigolvashatja a gráf éleit egy, vagy konstans sok alkalommal, majd meg kell adnia a választ. Alsó becslést adunk egy tetszőleges, véletlent is felhasználó, helyes algoritmus által felhasznált ideiglenes tárterületre.

Ezen probléma abban kapcsolódik a Webes kereséshez, hogy a keresőmotor segítségével szimulálhatjuk egy $\mathcal{A}$ algoritmus működését. Az előfeldolgozás során a keresőmotor letölti és feldolgozza az adathalmazt, egy indexadatbázist előállítva. Ezután a kiszolgálórutin képes válaszolni a kérdéseinkre kizárólag az indexadatbázis felhasználásával, mégpedig megfelelő keresőkérdések feltételével megkaphatjuk az eldöntési probléma helyes megoldását. Ennek megfelelően a bizonyított alsó becslések érvényesek az indexelő vagy a lekérdező algoritmus ideiglenes tárigényére, vagy az indexadatbázis méretére. Mivel az alsó becsléseink kizárják a gyakorlatban való megvalósítás lehetőségét (például négyzetes átmeneti tárterületet igényelve), így további bizonyítékot nyerünk arra nézve, hogy az egzakt probléma nem megoldható és közelítő algoritmusok felé érdemes a kutatásainkat összpontosítani.

3. Kutatási célok

Kutatásaink során a következő kérdésekre próbálunk válaszolni a személyre szabott PageRank rangsoroló, illetve a SimRank hasonlóságkereső függény tekintetében:

- Keressünk *jól skálázódó* közelítő algoritmust, amely segítségével ezeket a pontszámokat a keresőkérdés feldolgozása közben ki tudjuk számítani!
- Bizonyítsuk be, hogy nem létezik jól skálázódó egzakt választ adó algoritmus!
- Mennyire pontos közelítést ad az algoritmusunk?
- Milyen erőforrás-igényt támaszt az algoritmusunk? Demonstráljuk, hogy a skálázódási követelményeknek megfelel a megoldásunk kellően nagy méretű bemenetre való tényleges futtatással!
- Jól alkalmazhatóak-e a matematikai definíciók a gyakorlatban? Milyen minőségű találati listákat kapunk a szóbanforgó definíciók segítségével? Kísérletileg értékeljük ki a függvényeket egy valódi webes adathalmazon! Adjunk a kísérleti eredmények alapján bizonyítékot arra, hogy a fenti függvények lényegesen jobban alkalmazhatóak a webre, mint a klasszikus megoldások.

- Paraméterhangolás: Hogyan kell beállítani az általunk adott algoritmus paramétereit, hogy az ésszerű erőforrás-költséggel megfelelő minőségű eredményt adjon?
- Adjunk új definíciókat és vizsgáljunk meg őket a fenti szempontok szerint!

Jól látható, hogy a kutatási célok középpontjában a *jól skálázódó* fogalom áll. A web, mint adathalmaz méretéből adódó különlegessége, hogy a feldolgozására rendkívül specializált rendszerek születtek [4, 2, 8]. Egy algoritmust a következő követelmények teljesítése esetén tekintünk a webkeresők szempontjából jól skálázódónak [23, 19]:

- **Előfeldolgozás:** Az algoritmus előfeldolgozhatja a bemeneti adathalmazt az ún. *streaming + sorting* számítási modellben, azaz az előfeldolgozás során konstans számú külső táras rendezési eljárást és az adatsor konstans számú végigolvasását igénylő adatfolyam-algoritmust alkalmazhat. Az előfeldolgozás kimenete az *index*. Az előfeldolgozásnak legfeljebb egy nap alatt lefuttathatónak kell lennie.
- **Lekérdezés:** A lekérdezési algoritmus az indexadatbázis konstans számú rekordját kérheti le, majd a beolvasott adat mennyiségében lineáris algoritmust futtathat. A lekérdezésnek egy másodpercen belül lefuttathatónak kell lennie.
- **Memória:** Az algoritmusoknak külső tárasaknak kell lenniük, azaz csak konstans méretű memóriát alkalmazhatnak. Bizonyos esetekben a gráf csúcsainak számában lineáris memóriaigényű megoldás is szóba jöhet.
- **Párhuzamosíthatóság:** Az előfeldolgozó és lekérdező algoritmusnak képesnek kell lennie többezer kisebb teljesítményű gyors hálózatba kötött számítógép erőforrásait felhasználni.

4. Kutatási módszerek

Az általunk adott új algoritmusok elsősorban ujjlenyomatokon alapuló adatbányászati algoritmusok családjába sorolhatóak, melynek rendkívül látványos alappéldáját Broder fektette le [6]. Ezekben a közelítő módszerekben a keresett mennyiséget kifejezzük mint egy valószínűségi változó várható értékét, majd N független mintát (ujjlenyomatot) véve Monte Carlo módszerrel becsüljük.

A PageRank és hozzá hasonló problémák valószínűségszámítási átfogalmazásánál erősen támaszkodunk a PageRank véletlen sétákkal való kifejezésére: egyrészt mint a gráfon való véletlen bolyongás (Markov-lánc) stacionárius állapota [22], másrészt mint egy geometriai eloszlású hosszúságú véletlen séta végpontja [9].

Az egzakt megoldás gyakorlatban való kiszámíthatatlanságát a tárigényre vonatkozó alsó becslések bizonyításával nyerjük. Ehhez a gráfalgoritmusok adatfolyam-modellben való vizsgálatánál [14] alkalmazott módszereket használjuk, melyben kommunikációs komplexitási problémákra [20], főként a bit-vektor tesztelés feladatára vezetjük vissza a vizsgált kérdést.

A kísérleti kiértékelésben a személyre szabott PageRank esetében a közelítés pontosságát vizsgáljuk a [16] algoritmushoz hasonlítva. A rangsorok összehasonlításához a PageRank-kutatásban elterjedt módszereket alkalmazzuk [18, 11, 25].

A hasonlósági függvények minőségének kiértékelésénél Haveliwala [12] módszerét követjük, amelyben egy jó minőségű Internet-katalógus, az Open Directory Project (DMOZ) [21] téma szerinti besorolását tekintjük alapigazságnak, és a hasonlósági függvény minőségét úgy mérjük, hogy átlagosan mennyire sikerül ezt a besorolást visszaadnia.

5. Új eredmények

1. téziscsoport: Monte Carlo algoritmus a személyre szabott PageRank rangsoroló függvény kiszámítására

A Webes keresés egyik legfőbb problémája, hogy a felhasználónak egy rövid, kulcsszavakat tartalmazó keresőkérdés formájában kell megfogalmaznia az információigényét. Ez egy nagyon nehéz feladat, különösen az átlagember számára. Ha a keresőkérdés túl specifikus, túl sok szót tartalmaz, akkor nagy az esély arra, hogy egy jó minőségű potencális találati oldal valamelyik szót épp nem tartalmazza, mert másként fogalmaz – ez a *recall* probléma. Ha a keresőkérdés túl általános, akkor nagyon sok találat lesz, amelyből nehéz kiválasztani a felhasználót érdeklő jó minőségű oldalakat – ez a *precision* probléma.

A recall-probléma miatt a felhasználók azt a viselkedést sajátították el, hogy a keresőkérdést általában csak néhány szóval fogalmazzák meg, vállalva, hogy több tízmillió találatot ad ki a keresőprogram. A kiszolgálót vezérlő algoritmusoknak a legfontosabb feladatuk pedig megbírkózni a precision-problémával, azaz hogy a találatokat olyan sorrendben prezentálják a felhasználónak, hogy az első néhány találat között ott legyen a kérdésre releváns jó minőségű weboldal.

A rangsorolás problémájával nagyon sokat foglalkoztak, és az algoritmusokat több skálán lehet osztályozni. Egy *lokális* rangsoroló algoritmus csak egyetlen oldalt vizsgál egyszerre, míg egy *globális* rangsoroló algoritmus a teljes adathalmazt. A *statikus* rangsoroló algoritmusok egy rögzített rangsort számítanak ki az adathalmazból és azt alkalmazzák az összes keresőkérdésre, míg a *dinamikus* rangsoroló algoritmusok a keresőkérdés függvényében rangsorolnak. A gyakorlatban sok módszer kombinációját használják, például egy lokális statikus algoritmust a rosszindulatú weboldalak (például vírusok) kiszűrésére, egy lokális dinamikus algoritmust a találati oldal keresőkérdésre való illeszkedésének pontozására (például az oldal címében vagy nagy betűkkel szereplő keresőszavak felsúlyozásával), valamint egy globális statikus algoritmust a találati oldal Weben való népszerűségének (ami a minőség jelzője szokott lenni) pontozására.

Az utóbbi kategóriába tartozó globális rangsoroló eljárások közül a legtöbbet kutatott a *PageRank* [22], mivel sokan ezt sejtik a Google piacvezető keresőmotor pontossága (és ennek nyomán kialakult sikere) mögött. Ez a következő megfigyelésre épül:

> Egy $u \to v$ hiperlink a Weben az u oldal tanúbizonysága arról, hogy a v oldal jó minőségű tartalmat mutat.

Ezt a megfigyelést a PageRank deficíciójában rekurzívan alkalmazza, azaz egy u weboldal PageRank-pontszáma a rámutató weboldalak PageRank-pontszámából számítható ki.

A PageRank algoritmus egyik fontos hátránya, hogy statikus, azaz a Weboldalak relevanciáját egyetlen számként állapítja meg, és ugyanazt az értékelést alkalmazza akár egy amerikai számítógépguru, akár egy mongol történelemtanár kérdéseire is kell válaszolnia. Ezt a hátrányt képes kiküszöbölni a személyre szabás, amely esetén a Web egy része, mint kiindulópont értékelése alapján rangsoroljuk az oldalakat, és ezt a kiindulópontot az egyes felhasználóknál külön-külön állítjuk be.

A személyre szabott PageRank [5, 22] esetén a legfőbb nehézség, hogy a kiindulópont csak a keresőkérdés feltételekor áll rendelkezésünkre, így a szokásos PageRank számítási módszerek nem alkalmazhatóak, mivel azok tipikusan több órányi számítást igényelnek, és ennyit még a legtürelmesebb felhasználók sem hajlandóak várni a válaszra a személyre szabás előnyeiért cserébe. Skálázható számítási módszerek keresésével sokan foglalkoztak [10, 16, 17, 13], azonban ezen korábbi munkák mindegyike jelentős korlátozásokat tartalmaz arra vonatkozóan, hogy hogyan lehet kifejezni a személyre szabást. A téziscsoport jelentős eredménye, hogy ezeket a korlátozásokat feloldja.

> **1.1. tézis. [J4, C9]** Skálázható véletlent használó közelítő algoritmus személyre szabott PageRank pontszámok számítására, amely egy a weboldalak számában lineáris

méretű indexadatbázis ismeretében tetszőleges kiindulóoldalra konstans sok adatbázis-hozzáféréssel torzítatlan becslést ad. A becslés pontosságának növelése ugyanazon indexadatbázisból a kiindulóoldal szomszédaira vonatkozó rekordok figyelembevételével.

Ezen eredmény alapján, mivel a súlyozott kiindulópont mint weboldalakon értelmezett vektor szerint lineáris a személyre szabott PageRank [10], tetszőleges személyre szabás elérhető.

Természetesen a fenti eljárás alkalmazhatóságához szükséges még az, hogy az indexadatbázist ki tudjuk számítani skálázható módon. Erre két megoldást is adtam, amelyekből a rendelkezésre álló erőforrások függvényében lehet a megfelelőt kiválasztani.

1.2. tézis. [J4, C9] Külső táras indexelési eljárás, amely az 1.1. tézis algoritmusa által igényelt adatbázist V csúcsú, d átlagos fokszámú gráf esetén M méretű belső memória felhasználásával $\Theta(V(N \log_M NV + Ld))$ I/O művelet segítségével kiszámítja, ahol N a közelítés pontosságának, L a gráf keverési sebességének megfelelő konstans.

A konstansokba a szokásos, illetve a dolgozatban szereplő kísérletek eredményeként kapott értékeket behelyettesítve ($L = 20$, $d = 10$, $N = 100$, $V = 10^{10}$, $M = 1$GB) az I/O igényre 256 TB adódik, amely kb. 60 diszk felhasználásával egy nap alatt lefuttatható. A tényleges felhasznált tárterület mindössze 8 TB, és mivel az algoritmus csupán külső táras rendezést és összefésülést alkalmaz a hozzáféréshez, ezért nagyméretű, akár többszáz MB-os blokkok alkalmazásával a lemezhozzáférés csupán szekvenciális olvasással és írással történhet.

1.3. tézis. [J4, C9] Indexelési eljárás, amely K gyors helyi hálózatba kötött számítógép felhasználásával, melyek együttes memóriája elegendő a teljes Web-gráf tárolására, kiszámítja az 1.1. tézis algoritmusa által igényelt adatbázist $\Theta(NV)$ várható összes kommunikáció segítségével.

A Web-gráf memóriában tárolására rendkívül komplex tömörítési technikákat dolgoztak ki az elmúlt években [1, 3], amelyek élenként mindössze néhány bit felhasználásával kódolják a gráfot, azonban ezeknél egyszerűbb, a gyors feldolgozást jobban lehetővé tevő kódolások esetén is már 100 közönséges számítógép belső tárterülete is elegendő lehet a teljes gráf tárolására. A fent említett konstansok behelyettesítésével 48 TB hálózati kommunikáció adódik, amely alapján a mai hétköznapi hálózati technológiákkal 100 gép felhasználása esetén az indexelési algoritmus mindössze egy óra alatt lefuttatható.

2. téziscsoport: Monte Carlo algoritmus elemzése és javítása a SimRank hasonlósági függvény kiszámítására

A Webes keresés egyik legfőbb problémája, mint azt az 1. téziscsoport bevezetőjében említettük, a keresési kulcsszavak megfogalmazásának nehézsége (a felhasználó oldaláról), valamint a kulcsszavak megértésének nehézsége (a keresőmotor oldaláról). Ennek egyik lehetséges megoldása, hogy a felhasználótól több információt kérünk keresési kérdésként. Természetesen mivel a keresési munkafolyamat gördülékenységét nem kívánjuk bonyolult felhasználói interfésszel és egyértelműsítési igényekkel vagy visszakérdezésekkel elrontani, így nagy előnyt jelent az, ha a keresőkérdés valamilyen implicit módon rejt többletinformációt.

Ilyen implcit többletinformációt tartalmaz a *példa alapján* való keresés. Ebben az esetben a felhasználó megad egy általa már ismert weboldalt, mint keresési kérdést, és a keresőmotor ahhoz hasonló, jó minőségű oldalak listájával válaszol, melyek vélhetően szintén a felhasználó érdeklődésére tarthatnak számot. Ez a funkcionalitás már a kezdetek óta elérhető a nagy keresőmotorok találati oldalán „Related results" (magyarul „Hasonló oldalak") címszó alatt. Annak ellenére, hogy vélhetően ez a böngészők által legtöbbet megjelenített link (révén minden találati oldalon tucatjával szerepel), viszonylag kevesen kattintanak rá, mivel a jelenlegi algoritmusok által adott találatok minősége nem kielégítő.

Joggal hihetjük, hogy a fejlett link-elemző algoritmusok ugyanúgy forradalmasíthatják a példa alapján való keresést, mint ahogy a PageRank a kulcsszavas keresés rangsorolási problémáját. Ezért válik elsődleges vizsgálataink középpontjává a *SimRank* hasonlósági függvény [15], mely a PageRank-hez hasonló rekurzív definícióval fogalmazza meg két weblap (vagy egy tetszőleges gráfban két csúcs) hasonlóságát.

A SimRank függvény legnagyobb problémája, hogy míg a PageRank hatványiterációval való naív kiszámítási algoritmusa ésszerű erőforrásokkal lefuttatható, a SimRank hatványiterációval való számításához a gráf csúcsainak számában négyzetes tárterület és idő szükséges, ami önmagában kizárja a Webre való alkalmazhatóságát. Korábbi eredmények a SimRank kiszámítását komoly heurisztikus apparátust bevetve is csupán kétszázezer csúcsú gráfra voltak képesek elvégezni.

Szerzőtársam, Fogaras Dániel eredménye az első olyan SimRank algoritmus [J5, C10], amely a Web méreteire való alkalmazhatóságot jelentő skálázódási követelményeknek megfelel. Ez egy véletlent használó közelítő algoritmus, mely ujjlenyomatokat számít ki a gráf minden csúcsához, majd az ujjlenyomatok ismeretében torzítatlan becslést ad a SimRank hasonlósági függvény értékére. Monte Carlo módszerrel, N ujjlenyomat felhasználásával érhető el megfelelő pontosság:

> **2.1. tézis. [J5, C10]** A SimRank hasonlósági függvényt közelítő ujjlenyomatokon alapuló algoritmus pontosságának elemzése, és annak bizonyítása, hogy tetszőleges rögzített abszolút hiba mellett a hibavalószínűség az ujjlenyomatok N számában exponenciálisan tart 0-hoz a gráftól függetlenül a csúcsokon uniform módon, valamint hogy a hasonlósági toplista lekérdezésénél tetszőleges rögzített küszöbérték esetén és tetszőleges rögzített abszolút hibától eltekintve a felidézés/teljesség (recall) exponenciálisan tart 1-hez a gráftól függetlenül a kérdezett csúcsban uniform módon.

Ennek a tézisnek fontos következménye, hogy rögzített hiba esetén az ujjlenyomatok N száma tetszőleges lekérdezés esetén, és a gráf növekedése mellett is konstansnak tekinthető.

Annak ellenére, hogy viszonylag erős tételeink vannak a közelítés konvergenciájára, joggal merül fel a kérdés, hogy található-e egzakt eredményeket szolgáltató algoritmus, vagy közelítéssel kell-e megelégednünk? Erre a kérdésre válaszolnak az alsó becslésekre vonatkozó tételeink:

> **2.2. tézis. [J5, C10]** Alsó becslés az indexadatbázis méretére, amely szerint tetszőleges egzakt eredményt szolgáltató SimRank algoritmus bizonyos V csúcsú gráfokon $\Omega(V^2)$ méretű indexadatbázist igényel, míg tetszőleges közelítő eredményt szolgáltató algoritmus $\Omega(V)$ méretű indexadatbázist igényel.

Ennek következménye, hogy általános érvényű egzakt megoldásban Web méretű gráfok esetén nem reménykedhetünk, mivel a szükséges indexadatbázis mérete meghaladja a Földön elérhető jelenlegi tárolókapacitást. Az általunk adott ujjlenyomat-alapú algoritmus a következő reprezentáció segítségével egy logaritmikus faktor erejéig beállítja ezt az alsó becslést:

> **2.3. tézis. [J5, C10]** Csökkentett tárigényű reprezentáció, amely a [P]SimRank algoritmus [C10] által igényelt csatolt ujjlenyomat-utakat csúcsonként két cellában kódolja.

Ezen kompakt reprezentáció aszimptotikusan $O(V \log V)$ tárterületet foglal, a szokásos méreteket ($V = 10^{10}$, $N = 100$) figyelembe véve a teljes Webet tartalmazó hasonlósági indexadatbázis mindössze 8 TB adatot igényel.

Az általunk adott algoritmus ipari alkalmazhatóság szempontjából rendkívül fontos tulajdonságokat is felmutat.

> **2.4. tézis. [J5]** Az általunk adott Monte Carlo hasonlósági függvények ipari [2] alkalmazhatóságra való felkészítése: párhuzamosíthatóság, hibatűrés, folytonos terheléselosztás lehetősége, a terheléshez való dinamikus adaptáció lehetősége. Inkrementális indexelési eljárások a gráf változásainak létező indexbe való beillesztésére. Kísérleti igazolása, hogy a párhuzamosítás során a klaszter teljes kapacitása az igénybe vett számítógépek számában lineárisan nő.

3. téziscsoport: A közös szomszédságok eldöntési probléma bizonyítása és általánosítása

Ebben a téziscsoportban egy absztrakt problémával foglalkozunk, amely egyfajta bonyolultságelméleti megfogalmazása lehet a gráf-alapú hasonlóságkeresésnek. Buchsbaum, Giancarlo és Westbrook vizsgálta [7]-ban a következő kérdést az adatfolyam-modellben: Egy irányított gráfban létezik-e két olyan pont, melyek közös be- (vagy ki-) szomszédsága egy c küszöbértéket meghalad? Ez a kérdés ekvivalens a gráfban a $K_{2,c}$ mint irányított részgráf előfordulásával.

Az adatfolyam-modell esetén a bemeneti gráf az algoritmus számára csupán egy egyirányban végigolvasható szalag formájában áll rendelkezésre. Két különösen érdekes eset van: az egy-futamos adatfolyam algoritmus esetén a gráfot tartalmazó szalag csupán egyszer olvasható végig, majd az algoritmusnak választ kell adnia. Ez különösen nagy mennyiségű valós időben érkező adat feldolgozása esetén hasznos modell, mivel ezeket az adatfolyamokat általában a puszta adatmennyiség miatt nincs lehetőségünk eltárolni és utólag feldolgozni. Az általános eset pedig a konstans sok futamot megengedő modell, amelyben a bementi adatot tartalmazó szalagot $O(1)$ alkalommal visszatekerhetjük az elejére. Ez jól modellezi az adatot másodlagos vagy háttértárolón feldolgozó algoritmusokat, amelyben a háttértároló szekvenciális végigolvasásra jól alkalmazható, azonban a véletlen hozzáférés költsége nem felvállalható. Ez még a mai merevlemezes technológiákra is igaz.

Az érdekes kérdések az adatfolyam-modellben mindig az átmeneti tárigény alsó becslésére vonatkoznak, azaz mi az a minimális belső tárterület, amelyet bármely helyes algoritmus igényel?

Sajnos a [7] cikkben a kimondott egyik alapvető lemma bizonyítása hibás, méghozzá nem kézenfekvően javítható módon.

3.1. tézis. [J2] Helyes bizonyítás a [7] egy-futamos adatfolyam-eredményeire.

Az új bizonyítás segítségével erősebb becsléseket is adhatunk mind az 1, mind az $O(1)$ futamos adatfolyam modellben. Az új alsó becslések egy logaritmikus faktor erejéig élesek is, azaz algoritmust is adunk egy logaritmikus faktorral nagyobb tárterület felhasználásával.

3.2. tézis. [J2] Alsó becslés az egy-futamos adatfolyam-modellben a közös szomszédságproblémára, amely szerint $\Omega(\sqrt{c}n^{3/2})$ a belső tárigény n csúcsú gráfokon c szomszédsági küszöbérték keresése esetén. Algoritmus, amely $O(\sqrt{c}n^{3/2}\log n)$ tárterület felhasználásával megoldja a közös szomszédság problémát.

3.3. tézis. [J2] Alsó becslés az $O(1)$ futamos adatfolyam-modellben a közös szomszédságproblémára, amely szerint $\Omega(\sqrt{c}n^{3/2})$ a belső tárigény n csúcsú gráfokon c szomszédsági küszöbérték keresése esetén. Algoritmus, amely $O(\sqrt{c}n^{3/2}\log n)$ tárterület felhasználásával megoldja a közös szomszédság problémát.

Eredmények alkalmazása

A második téziscsoport algoritmusait az Analog pályázat keretében implementálta Fogaras Dániel, ezeket a hazai oldalakból 2004 decemberében készült gyűjtésen lehet kipróbálni a `www.hasonlo.hu` oldalon. Egy példa lekérdezést mutat az alábbi táblázat:

	Hasonlók(www.bkv.hu) lekérdezés PSimRank hasonlósági függvény alapján
1	www.bkv.hu/
2	www.malev.hu/
3	www.elvira.hu/
4	www.mahart.hu/
5	www.turizmusonline.hu/adatbazis/kutatas_fejlesztes.php
6	www.turizmusonline.hu/heti_turizmus/bemutatkozo.php
7	www.volan.hu/
8	www.idojaras.hu/
9	www.met.hu/
10	www.worldtimeserver.com/

Köszönetnyilvánítás

Mindenekelőtt szeretnék köszönetet mondani Fogaras Dániel szerzőtársamnak, aki felhívta a figyelmemet a hiperhivatkozásokon alapuló keresési algoritmusok vizsgálatában rejlő lehetőségekre, és akivel sok éven át tartó közös kutatás nélkül a dolgozatban bemutatott eredmények töredéke készült volna csak el.

Külön köszönetet szeretnék mondani az MTA SZTAKI Adatbányászat és Webes keresés kutatócsoportja vezető kutatóinak, témavezetőmnek, Benczúr Andrásnak és Lukács Andrásnak, hogy a SZTAKI-ban töltött idő alatt segítették pályámat, és olyan szerteágazó feladatokat biztosítottak – a számítástudományi kutatástól kezdve a nagyméretű infrastruktúrális szoftverfejlesztésen és pályázati munkákon át a közbeszerzés labirintusaiban való eligazodásig –, melyekből rengeteget tanultam, és az általuk megszerzett sokoldalúság nélkül biztosan nem tartanék itt, ahol ma. Köszönettel tartozom a SZTAKIban dolgozó kollégáimnak és szerzőtársaimnak, Csalogány Károlynak és Sarlós Tamásnak az algoritmusaink implementálásáért és kísérleti eredmények előállításáért, valamint rengeteg a tudományos munka során elengedhetetlen visszajelzésért, amelyekkel helyes pályán tartották a gondolatainkat. A kutatásaim során benyújtott kéziratok színvonalát nagyban javították a fent említetteken kívül Rónyai Lajos, Marx Dániel, Glen Jeh, Andrew Twigg, Adam L. Buchsbaum és Raffaele Giancarlo észrevételei. Köszönetet szeretnék mondani Bodon Ferencnek is, akivel a dolgozatban nem szereplő kutatási területeken dolgoztunk együtt hosszú időn át.

Nagy köszönettel tartozom Rónyai Lajos, Recski András és Tóth Bálint tanszékvezető uraknak a pályafutásom során nyújtott szakmai támogatásért, tanácsokért, kihívásokért és lehetőségekért.

Publikációk

A Habilitációs Bizottság és a Doktori Tanács által előírt pontozás szerinti publikációs teljesítmény 29,66 pont.
Az összes publikációk száma: 16
Az összes lektorált publikációk száma: 14
Az összes ismert hivatkozások száma: 124
Az összes ismert független hivatkozások száma: 109

Folyóiratcikkek

[J1] András Benczúr, István Bíró, Károly Csalogány, Balázs Rácz, Tamás Sarlós, and Máté Uher. PageRank és azon túl: Hiperhivatkozások szerepe a keresésben. *Magyar Tudomány*, 2006(11):1325, 2006. L 2 / 6 = 0,33 pont.

[J2] A. L. Buchsbaum, R. Giancarlo, and B. Rácz. New results for finding common neighborhoods in massive graphs in the data stream model. *Theoretical Computer Science*, 407(1-3):302–309, 2008. LR 6 / 3 = 2 pont.

[J3] Z. Dezső, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási. The dynamics of information access on the web. *Physical Review E*, 73(6), 2006. LR 6 / 6 = 1 pont.

[J4] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards fully personalizing PageRank: Algorithms, lower bounds and experiments. *Internet Mathematics*, 2(3), 2005. LR 6 * 33% = 2 pont.

[J5] Dániel Fogaras and Balázs Rácz. Practical algorithms and lower bounds for similarity search in massive graphs. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):585–598, 2007. L 6 / 2 = 3 pont.

[J6] Alexei Vazquez, Balázs Rácz, András Lukács, and Albert-László Barabási. Impact of non-poissonian activity patterns on spreading processes. *Physical Review Letters*, 98(15):158702, 2007. LR 6 / 4 = 1,5 pont.

Idegen nyelvű konferencia-kiadványok

[C7] A. A. Benczúr, K. Csalogány, K. Hum, A. Lukács, B. Rácz, Cs. I. Sidló, and M. Uher. Architecture for mining massive web logs with experiments. In *Proceedings of the HUBUSKA Open Workshop on Generic Issues of Knowledge Technologies*, 2005. 2 / 6 = 0,33 pont.

[C8] D. Fogaras and B. Rácz. A scalable randomized method to compute link-based similarity rank on the web graph. In *Proceedings of the Clustering Information over the Web workshop, Conference on Extending Database Technology*, 2004. L 4 / 2 = 2 pont.

[C9] D. Fogaras and B. Rácz. Towards fully personalizing PageRank. In *Proceedings of the 3rd Workshop on Algorithms and Models for the Web-Graph (WAW2004), in conjunction with FOCS 2004.*, 2004. L 4 / 2 = 2 pont.

[C10] D. Fogaras and B. Rácz. Scaling link-based similarity search. In *Proceedings of the 14th Int'l World Wide Web Conference*, 2005. L 6 / 2 = 3 pont.

[C11] Balázs Rácz. nonordfp: An FP-Growth variation without rebuilding the FP-tree. In *FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, 2004. L 4 pont.

[C12] B. Rácz, F. Bodon, and L. Schmidt-Thieme. On benchmarking frequent itemset mining algorithms. In *Proceedings of the 1st International Workshop on Open Source Data Mining, in conjunction with ACM SIGKDD*, 2005. L 4 * 50% = 2 pont.

[C13] B. Rácz, A. Lukács, and Cs. I. Sidló. Two-phase data warehouse optimized for data mining. In *Proceedings of the First International Workshop on Business Intelligence for the Real-Time Enterprise, in conjunction with VLDB 2006*, 2006. L 4 * 50% = 2 pont.

[C14] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th Int'l World Wide Web Conference*, 2006. L 6 / 4 = 1,5 pont.

Technical report

[T15] B. Rácz. Adatbányászati és többváltozós statisztikai modellek elektronikus újságok látogatottsági adatainak elemzésére., 2002. TDK dolgozat, 1 pont.

[T16] B. Rácz. Tömörítés és hosszútávú tárolás. Technical Report 4, Adatrosta könyvtár, 2003. 2 pont.

Nem publikációértékű munkák

[N17] B. Rácz and A. Lukács. High density compression of log files. In *Proceedings of the Data Compression Conference*, page 557, 2004. (poster), L 0 pont.

Hivatkozások a publikációkra

Scaling Link-Based Similarity Search [C10]

31 független, összesen 33 hivatkozás

[H1] Mohammed Al-Badawi, Dr. Siobhán North, and Dr. Barry Eaglestone. Indexing xml databases: Classifications, problems identification and a new approach. Technical report, The University of Sheffield, 2007.

[H2] Ilaria Bartolini and Paolo Ciaccia. Towards an effective semi-automatic technique for image annotation. In Michelangelo Ceci, Donato Malerba, and Letizia Tanca, editors, *SEBD*, pages 258–265, 2007.

[H3] Ilaria Bartolini and Paolo Ciaccia. Imagination: Exploiting link analysis for accurate image annotation. In *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics: 5th International Workshop, AMR 2007, Paris, France, July 5-6, 2007 Revised Selected Papers*, pages 32–44, Berlin, Heidelberg, 2008. Springer-Verlag.

[H4] Luca Becchetti, Paolo Boldi, Carlos Castillo, and Aristides Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–24, New York, NY, USA, 2008. ACM.

[H5] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. Link-based similarity search to fight web spam. In *AIRWeb 2006, Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web*, pages 9–16, 2006.

[H6] A. Broder. CS598E course material on Princeton University, 2005. `http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/Princeton.pdf`.

[H7] Aurel Cami and Narsingh Deo. Techniques for analyzing dynamic random graph models of web-like networks: An overview. *Networks*, 51(4):211–255, 2008.

[H8] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 7–14, New York, NY, USA, 2007. ACM.

[H9] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-supervised learning: A comparative study for web spam and telephone user churn. In *Proc. of Graph Labelling Workshop and Web Spam Challenge 2007 in conjunction with ECML/PKDD 2007*, 2007.

[H10] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards fully personalizing PageRank: Algorithms, lower bounds and experiments. *Internet Mathematics*, 2(3), 2005.

[H11] Monika Henzinger. Hyperlink analysis on the world wide web. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 1–3, New York, NY, USA, 2005. ACM.

[H12] Srivatsa Iyengar. Entity reconciliation in spin. M.Tech project report, Indian Institute of Technology Bombay, 2005.

[H13] Quanzhi Li and Yi-fang Brook Wu. People search: Searching people sharing similar interests from the web. *J. Am. Soc. Inf. Sci. Technol.*, 59(1):111–125, 2008.

[H14] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Trans. Web*, 2(1):1–35, 2008.

[H15] Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, and Denis Turdakov. Accuracy estimate and optimization techniques for simrank computation. *Proc. VLDB Endow.*, 1(1):422–433, 2008.

[H16] Siddhartha Reddy, K. Srinath, Srinivasa Mandar, and R. Mutalikdesai. Measures of ignorance on the web. In *Proceedings of the International Conference on Management of Data COMAD 2006*, pages 140–149, 2006.

[H17] Elisa Rondini. Semi-automatic techniques for the semantic annotation of multimedia databases. Master's thesis, University of Bologna, 2005. (in Italian).

[H18] Takehiko Sakamoto and Keishi Tajima. Improved methods of structure calculation using link-based similarity functions. In *Proc. of the 18th IEICE Data Engineering Workshop*, 2007. (in Japanese).

[H19] T. Sarlós, A. A. Benczúr, K. Csalogány, D. Fogaras, and B. Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th Int'l World Wide Web Conference*, 2006.

[H20] Allan M. Schiffman. Hierarchy in web page similarity link analysis. Technical Report 06-02, Carnegie Mellon University and CommerceNet Labs, May 2006.

[H21] Song, Ma, Lian-Li, and Zhang Zhijun. The study on the comprehensive computation of the documents similarity. *Computer Engineering and Applications*, 42(30), 2006. (in Chinese).

[H22] Jessica Staddon. Sponsored ad-based similarity: an approach to mining collective advertiser intelligence. In *ADKDD '08: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 50–56, New York, NY, USA, 2008. ACM.

[H23] K Venkatraman. Pagerank by distributed computing: An empirical analysis. Technical report, Quotient Inc., 2008.

[H24] Benjamin N. Waber, John J. Magee, and Margrit Betke. Web mediators for accessible browsing. In Constantine Stephanidis and Michael Pieper, editors, *Universal Access in Ambient Intelligence Environments*, volume 4397 of *Lecture Notes in Computer Science*, pages 447–466. Springer, 2006.

[H25] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Block-based similarity search on the web using manifold-ranking. In *Web Information Systems - WISE 2006*, volume 4255/2006 of *Lecture Notes in Computer Science (LNCS)*, pages 60–71. Springer-Verlag, 2006.

[H26] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Towards a unified approach to document similarity search using manifold-ranking of blocks. *Inf. Process. Manage.*, 44(3):1032–1048, 2008.

[H27] Haixuan Yang. *Machine learning models on random graphs.* PhD thesis, The Chinese University of Hong Kong (People's Republic of China), 2007. Adviser: King, Irwin and Lyu, Michael R.

[H28] Haixuan Yang, Irwin King, and Michael R. Lyu. Predictive random graph ranking on the web. In *In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI*, pages 3491–3498, 2006.

[H29] Yunming Ye, Yan Li, Xiaofei Xu, Joshua Huang, and Xiaojun Chen. MFCRank: A web ranking algorithm based on correlation of multiple features. In *Computational Linguistics and Intelligent Text Processing*, volume 3878/2006 of *Lecture Notes in Computer Science (LNCS)*, pages 378–388. Springer-Verlag, 2006.

[H30] X Yin. *Scalable Mining and Link Analysis Across Multiple Database Relations.* PhD thesis, University of Illinois at Urbana-Champaign, 2007.

[H31] Xiaoxin Yin and Jiawei Han. Exploring the power of heuristics and links in multi-relational data mining. In *LNAI 4994, Proceedings of the 17th Int'l Symposioum in Foundations of Intelligent Systems*, pages 17–27, 2008.

[H32] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Linkclus: efficient clustering via heterogeneous semantic links. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 427–438. VLDB Endowment, 2006.

[H33] Yangbo Zhu. Distributed pagerank computation in search engine confederation. Master's thesis, Carnegie Mellon University, 2006.

Towards scaling fully personalized pagerank: Algorithms, lower bounds, and experiments. [J4]

13 független hivatkozás

[H34] Marin Bertier, Davide Frey, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy. Personalized web search by gossiping with unknown social acquaintances. Technical Report 6878, Institut National de Recherche en Informatique et en Automatique, 2009.

[H35] Marin Bertier, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy. Toward personalized query expansion. In *Social Network Systems 2009*, 2009. to appear.

[H36] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: model and applications. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 609–618, New York, NY, USA, 2008. ACM.

[H37] Paolo Boldi, Roberto Posenato, Massimo Santini, and Sebastiano Vigna. Traps and pitfalls of topic-biased pagerank. In *Algorithms and Models for the Web-Graph: Fourth International Workshop, WAW 2006, Banff, Canada, November 30 - December 1, 2006. Revised Papers*, volume 4936 of *Lecture Notes in Computer Science*, pages 107–116, Berlin, Heidelberg, 2008. Springer.

[H38] Aurel Cami and Narsingh Deo. Techniques for analyzing dynamic random graph models of web-like networks: An overview. *Networks*, 51(4):211–255, 2008.

[H39] Soumen Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 571–580, New York, NY, USA, 2007. ACM.

[H40] Prasad Chebolu and Páll Melsted. Pagerank and the random surfer model. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1010–1018, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[H41] William W. Cohen. Graph walks and graphical models, 2007. Machine Learning Department, Carnegie Mellon University.

[H42] Einat Minkov. *Adaptive Graph Walk Based Similarity Measures in Entity-Relation Graphs.* PhD thesis, Carnegie Mellon University, 2008.

[H43] Purnamrita Sarkar, Andrew W. Moore, and Amit Prakash. Fast incremental proximity search in large graphs. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 896–903, New York, NY, USA, 2008. ACM.

[H44] Aixin Sun, Maggy Anastasia Suryanto, and Ying Liu. Blog classification using tags: An empirical study. In *10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007, Proceedings*, volume 4822 of *Lecture Notes in Computer Science*, pages 307–316. Springer, 2007.

[H45] Jonathan Traupman. Resisting sybils in peer-to-peer markets. In *Trust Management*, volume 238 of *IFIP International Federation for Information Processing*, pages 269–284. Springer, 2007.

[H46] Jonathan David Traupman. *Robust reputations for peer-to-peer markets.* PhD thesis, University of California at Berkeley, Berkeley, CA, USA, 2007. Adviser J.D. Tygar.

Towards scaling fully personalized pagerank. [C9]

17 független hivatkozás

[H47] Sinan Al-Saffar and Gregory Heileman. Experimental bounds on the usefulness of personalized and topic-sensitive pagerank. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 671–675, Washington, DC, USA, 2007. IEEE Computer Society.

[H48] Sinan al Saffar and Gregory L. Heileman. Semantic impact graphs for information valuation. In *DocEng '08: Proceeding of the eighth ACM symposium on Document engineering*, pages 209–212, New York, NY, USA, 2008. ACM.

[H49] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcraft, Vahab S. Mirrokni, and Shang hua Teng. Local computation of pagerank contributions. In *Proceedings of the Third Workshop on Algorithms and Models for the Web Graph*, pages 150–165, 2007.

[H50] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, New York, NY, USA, 2008. ACM.

[H51] Reid Andersen and Fan Chung. Detecting sharp drops in pagerank and a simplified local partitioning algorithm. In *Theory and Applications of Models of Computation, Proceedings of TAMC 2007*, pages 1–12, 2007.

[H52] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.

[H53] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM J. Numer. Anal.*, 45(2):890–904, 2007.

[H54] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. On the feasibility of low-rank approximation for personalized pagerank. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 972–973, New York, NY, USA, 2005. ACM.

[H55] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máté Uher. Spamrank - fully automatic link spam detection. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[H56] Maurice Coyle and Barry Smyth. Supporting intelligent web search. *ACM Trans. Interet Technol.*, 7(4):20, 2007.

[H57] David Gleich and Marzia Polito. Approximating personalized pagerank with minimal use of web graph data. *Internet Mathematics*, 3(3):257–294, 2006.

[H58] Rahul Sami and Andy Twigg. Lower bounds for distributed markov chain problems. *The Computing Research Repository*, abs/0810.5263, 2008.

[H59] Yang Sun, Huajing Li, Isaac G. Councill, Jian Huang, Wang-Chien Lee, and C. Lee Giles. Personalized ranking for digital libraries based on log analysis. In *WIDM '08: Proceeding of the 10th ACM workshop on Web information and data management*, pages 133–140, New York, NY, USA, 2008. ACM.

[H60] Hanghang Tong, Christos Faloutsos, Brian Gallagher, and Tina Eliassi-Rad. Fast best-effort pattern matching in large attributed graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–746, New York, NY, USA, 2007. ACM.

[H61] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 613–622, Washington, DC, USA, 2006. IEEE Computer Society.

[H62] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346, 2008.

[H63] LV Yuanhua. A study of personalized information retrieval based on implicit feedback. Master's thesis, Institute of Software, Chinese Academy of Sciences, 2007. (in Chinese).

To Randomize or Not To Randomize: Space Optimal Summaries for Hyperlink Analysis [C14]

9 független, összesen 14 hivatkozás

[H64] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcraft, Vahab S. Mirrokni, and Shang hua Teng. Local computation of pagerank contributions. In *Proceedings of the Third Workshop on Algorithms and Models for the Web Graph*, pages 150–165, 2007.

[H65] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, New York, NY, USA, 2008. ACM.

[H66] Reid Andersen, Fan Chung, and Kevin Lang. Local partitioning for directed graphs using pagerank. In *Algorithms and Models for the Web-Graph*, volume 4863 of *Lecture Notes in Computer Science*, pages 166–178. Springer, 2007.

[H67] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. Link-based similarity search to fight web spam. In *AIRWeb 2006, Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web*, pages 9–16, 2006.

[H68] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. Personalized query expansion for the web. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 7–14, New York, NY, USA, 2007. ACM.

[H69] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-supervised learning: A comparative study for web spam and telephone user churn. In *Proc. of Graph Labelling Workshop and Web Spam Challenge 2007 in conjunction with ECML/PKDD 2007*, 2007.

[H70] Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Estimating pagerank on graph streams. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 69–78, New York, NY, USA, 2008. ACM.

[H71] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Softw. Pract. Exper.*, 38(2):189–225, 2008.

[H72] Paolo Ferragina and Antonio Gulli. Snaket: A personalized search-result clustering engine. *CEPIS Upgrade: Next Generation Web Search*, 8(1):20–27, 2007.

[H73] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós. Towards fully personalizing PageRank: Algorithms, lower bounds and experiments. *Internet Mathematics*, 2(3), 2005.

[H74] David Gleich and Marzia Polito. Approximating personalized pagerank with minimal use of web graph data. *Internet Mathematics*, 3(3):257–294, 2006.

[H75] Miklos Kurucz, Andras Benczur, Karoly Csalogany, and Laszlo Lukacs. Spectral clustering in telephone call graphs. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 82–91, New York, NY, USA, 2007. ACM.

[H76] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, Washington, DC, USA, 2006. IEEE Computer Society.

[H77] Rebecca S. Wills and Ilse C. F. Ipsen. Ordinal ranking for google's pagerank. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1677–1696, 2009.

nonordfp: An FP-Growth variation without rebuilding the FP-tree [C11]

8 független, összesen 9 hivatkozás

[H78] Ferenc Bodon. A trie-based apriori implementation for mining frequent item sequences. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 56–65, New York, NY, USA, 2005. ACM.

[H79] Ferenc Bodon and Lars Schmidt-thieme. The relation of closed itemset mining, complete pruning strategies and item ordering in apriori-based fim algorithms. In *In Proc. PKDD*, pages 437–444, 2005.

[H80] Li Liu, Eric Li, Yimin Zhang, and Zhizhong Tang. Optimization of frequent itemset mining on multiple-core processor. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 1275–1285. VLDB Endowment, 2007.

[H81] Wim Pijls and Walter A. Kosters. Mining frequent itemsets: A perspective from operations research. Technical Report 24, Econometric Institute, Erasmus University Rotterdam, 2008.

[H82] Balázs Rácz, Ferenc Bodon, and Lars Schmidt-Thieme. On benchmarking frequent itemset mining algorithms: from measurement to analysis. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 36–45, New York, NY, USA, 2005. ACM.

[H83] Peter Schonhofen and Andras A. Benczur. Feature selection based on word-sentence relation. In *ICMLA '05: Proceedings of the Fourth International Conference on Machine Learning and Applications*, pages 37–42, Washington, DC, USA, 2005. IEEE Computer Society.

[H84] Takeaki Uno, Masashi Kiyomi, and Hiroki Arimura. Lcm ver.3: collaboration of array, bitmap and prefix tree for frequent itemset mining. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 77–86, New York, NY, USA, 2005. ACM.

[H85] Liqiang War and Liu Da-Xin. Study on fast algorithms for frequent itemset mining. *Journal of Harbin Engineering University*, 28(3), 2008. (in Chinese).

[H86] Zhong-ping Zhang, Li Yan, Zhi-jie Lin, and Ai-jie Wang. Frequent itemsets mining algorithm based on index arrays. *Compuer Application Research*, 26(1), 2009. (in Chinese).

Architecture for mining massive web logs with experiments [C7]

2 független, összesen 3 hivatkozás

[H87] E Khorram and S M Mirzababaei. Finding an optimized discriminate function for internet application recognition. *Proceedings of the World Academy of Science, Engineering and Technology*, 4:160–163, 2005.

[H88] M Rahmati and S M Mirzababaei. Data mining on the router logs for statistical application classification. In *Proceedings of the Fourth World Enformatika Conference*, 2005.

[H89] Cs. I. Sidló and A. Lukács. Shaping sql-based frequent pattern mining algorithms. In *Knowledge Discovery in Inductive Databases*, volume 3933 of *Lecture Notes in Computer Science*, pages 188–201. Springer, 2006.

On benchmarking frequent itemset mining algorithms: from measurement to analysis [C12]

1 független, összesen 2 hivatkozás

[H90] Ferenc Bodon. A trie-based apriori implementation for mining frequent item sequences. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 56–65, New York, NY, USA, 2005. ACM.

[H91] Wim Pijls and Walter A. Kosters. Mining frequent itemsets: A perspective from operations research. Technical Report 24, Econometric Institute, Erasmus University Rotterdam, 2008.

High-density compression of log files [N17]

3 független, összesen 4 hivatkozás

[H92] S Grabowski and S Deorowicz. Web log compression. *Automatyka / Akademia Gorniczo-Hutnicza im. Stanislawa Staszica w Krakowie*, 11(3):417–424, 2007. (in English).

[H93] Kimmo Hatonen. *Data mining for telecommunications network log analysis.* PhD thesis, Department of Computer Science, University of Helsinki, 2009.

[H94] B. Rácz, A. Lukács, and Cs. I. Sidló. Two-phase data warehouse optimized for data mining. In *Proceedings of the First International Workshop on Business Intelligence for the Real-Time Enterprise, in conjunction with VLDB 2006*, 2006.

[H95] Przemyslaw Skibinski and Jakub Swacha. Fast and efficient log file compression. In *CEUR Workshop Proceedings of 11th East-European Conference on Advances in Databases and Information Systems (ADBIS 2007*, 2007.

Impact of Non-Poissonian Activity Patterns on Spreading Processes [J6]

6 független, összesen 9 hivatkozás

[H96] Julian Candia, Marta C Gonzalez, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-Laszlo Barabasi. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015 (11pp), 2008.

[H97] A. Gautreau, A. Barrat, and M. Barthelemy. Microdynamics in stationary complex networks. *ArXiv e-prints*, November 2008.

[H98] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL*, 81(4):48002, feb 2008.

[H99] A. Grabowski, N. Kruszewska, and R. A. Kosiński. Properties of on-line social systems. *European Physical Journal B*, 66:107–113, November 2008.

[H100] J. Gu, W. Li, and X. Cai. The effect of the forget-remember mechanism on spreading. *European Physical Journal B*, 62:247–255, March 2008.

[H101] Wei Hong, Xiaopu Han, Tao Zhou, and Binghong Wang. Heavy-tailed statistics in short-message communication. *Chinese Physics Letters*, 26(2):028902 (3pp), 2009.

[H102] A. Grabowskiand R. Kosinski. The SIRS model of epidemic spreading in virtual society. *Acta Physica Polonica A*, 114(3):589–596, 2008.

[H103] J. G. Oliveira and A. Vazquez. Impact of interactions on human dynamics. *Physica A Statistical Mechanics and its Applications*, 388:187–192, January 2009.

[H104] T. Zhou, H. A. T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme. Role of activity in human dynamics. *EPL (Europhysics Letters)*, 82(2):28002 (5pp), 2008.

The Dynamics of Information Access on the Web [J3]

19 független, összesen 20 hivatkozás

[H105] Pierpaolo Andriani and Bill McKelvey. Beyond gaussian averages: redirecting international business and management research toward extreme events and power laws. *Journal of International Business Studies*, 38(7):1212–1230, December 2007.

[H106] Julian Candia, Marta C Gonzalez, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-Laszlo Barabasi. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015 (11pp), 2008.

[H107] Raul Caruso. Information and global security, a cautionary tale. *NewsNotes of the Economists for Peace and Security*, November 2006.

[H108] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Review of Modern Physics*. (to appear).

[H109] Peter Geczy, Noriaki Izumi, Shotaro Akaho, and Koiti Hasida. Knowledge worker intranet behaviour and usability. *Int. J. Bus. Intell. Data Min.*, 2(4):447–470, 2007.

[H110] Peter Géczy, Noriaki Izumi, Shotaro Akaho, and Kôiti Hasida. Human-centric design of percipient knowledge distribution service. In *WSKS '08: Proceedings of the 1st world summit on The Knowledge Society*, pages 31–40, Berlin, Heidelberg, 2008. Springer-Verlag.

[H111] K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL*, 81(4):48002, feb 2008.

[H112] X.-P. Han, T. Zhou, and B.-H. Wang. Modeling human dynamics with adaptive interest. *New Journal of Physics*, 10(7):073010–+, July 2008.

[H113] Wei Hong, Xiaopu Han, Tao Zhou, and Binghong Wang. Heavy-tailed statistics in short-message communication. *Chinese Physics Letters*, 26(2):028902 (3pp), 2009.

[H114] Yoram M. Kalman. *Silence in Text Based Computer Mediated Communication: The Invisible Component*. PhD thesis, University of Haifa, 2007.

[H115] Yoram M. Kalman, Gilad Ravid, Daphne R. Raban, and Sheizaf Rafaeli. Pauses and response latencies: A chronemic analysis of asynchronous cmc. *Journal of Computer-Mediated Communication*, 12(1):1–23, 2009.

[H116] Andreas Kaltenbrunner. *Dynamics of message interchange between stochastic units in the contexts of human communication behaviour and spiking neurons*. PhD thesis, Universitat Pompeu Fabra, 2007.

[H117] Andreas Kaltenbrunner, Vicenc Gomez, and Vicente Lopez. Description and prediction of slashdot activity. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, pages 57–66, Washington, DC, USA, 2007. IEEE Computer Society.

[H118] Andreas Kaltenbrunner, Vicenç Gómez, Ayman Moghnieh, Rodrigo Meza, Josep Blat, and Vicente López. Homogeneous temporal activity patterns in a large online communication space. In *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web*, volume 245 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[H119] J. G. Oliveira and A. Vazquez. Impact of interactions on human dynamics. *Physica A Statistical Mechanics and its Applications*, 388:187–192, January 2009.

[H120] Filippo Radicchi. Human activity in the web, 2009.

[H121] D. Ralt. No netting, health and stress - studying wellness from a net perspective. *Med. Hypotheses*, 70(1):85–91, 2008.

[H122] Xiaodan Song, Yun Chi, Koji Hino, and Belle L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 191–200, New York, NY, USA, 2007. ACM.

[H123] B. Ulicny, K. Baclawski, and A. Magnus. New metrics for blog mining. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6570 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, April 2007.

[H124] T. Zhou, H. A. T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme. Role of activity in human dynamics. *EPL (Europhysics Letters)*, 82(2):28002 (5pp), 2008.

Hivatkozások

[1] Micah Adler and Michael Mitzenmacher. Towards compressing web graphs. In *Data Compression Conference*, pages 203–212, 2001.

[2] Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. Web Search for a Planet: The Google Cluster Architecture. *IEEE Micro*, 23(2):22–28, 2003.

[3] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. Technical Report 293-03, Universita di Milano, Dipartimento di Scienze dell'Informazione, 2003.

[4] E. Brewer. Lessons from giant-scale services.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[6] Andrei Z. Broder. On the Resemblance and Containment of Documents. In *Proceedings of the Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29, 1997.

[7] A. L. Buchsbaum, R. Giancarlo, and J. R. Westbrook. On finding common neighborhoods in massive graphs. *Theoretical Computer Science*, 299(1-3):707–18, 2004.

[8] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th USENIX Symposium on Operating System Design and Implementation (OSDI)*, San Francisco, CA, USA, 2004. USENIX Association.

[9] Dániel Fogaras. Where to start browsing the web? In *Proceedings of the 3rd International Workshop on Innovative Internet Community Systems (I2CS)*, volume 2877/2003 of *Lecture Notes in Computer Science (LNCS)*, pages 65–79, Leipzig, Germany, June 2003. Springer-Verlag.

[10] Taher H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th World Wide Web Conference (WWW)*, pages 517–526, 2002.

[11] Taher H. Haveliwala. Efficient encodings for document ranking vectors. In *Proceedings of the 4th International Conference on Internet Computing (IC)*, pages 3–9, Las Vegas, Nevada, USA, 2003.

[12] Taher H. Haveliwala, Aristides Gionis, Dan Klein, and Piotr Indyk. Evaluating strategies for similarity search on the web. In *Proceedings of the 11th World Wide Web Conference (WWW)*, pages 432–442, 2002.

[13] Taher H. Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing PageRank. Technical Report 2003-35, Stanford University, 2003.

[14] Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. Computing on data streams. In *External Memory Algorithms, DIMACS Book Series vol. 50.*, pages 107–118. American Mathematical Society, 1999.

[15] Glen Jeh and Jennifer Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 538–543, 2002.

[16] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th World Wide Web Conference (WWW)*, pages 271–279. ACM Press, 2003.

[17] Sepandar Kamvar, Taher H. Haveliwala, Christopher Manning, and Gene Golub. Exploiting the block structure of the web for computing PageRank. Technical Report 2003-17, Stanford University, 2003.

[18] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating PageRank computations. In *Proceedings of the 12th World Wide Web Conference (WWW)*, pages 261–270. ACM Press, 2003.

[19] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[20] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[21] Open Directory Project (ODP). `http://www.dmoz.org`.

[22] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford University, 1998.

[23] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. ANF: A fast and scalable tool for data mining in massive graphs. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 81–90. ACM Press, 2002.

[24] Tamás Sarlós, András A. Benczúr, Károly Csalogány, Dániel Fogaras, and Balázs Rácz. To randomize or not to randomize: Space optimal summaries for hyperlink analysis. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 297–306, 2006. Full version available at http://www.ilab.sztaki.hu/websearch/Publications/.

[25] Pavan Kumar C. Singitham, Mahathi S. Mahabhashyam, and Prabhakar Raghavan. Efficiency-quality tradeoffs for vector score aggregation. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 624–635, 2004.