Applied Algorithms for Large-Scale Problems

János Demetrovics
Informatics Research Laboratory
Computer and Automation Research Institute
Hungarian Academy of Sciences
demetrovics@sztaki.hu

1. Bevezetés

A rangsorolás, hasonlóságkeresés és klasszifikáció a Weben megjelenő tartalom vizsgálatának nemtriviális feladata. Elért eredményeink egyrészt a nagy adatmennyiségek kezelésével, és az adatbányászat lineáris algebrán alapuló módszereivel kapcsolatos matematikai kutatás, másrészt valós adathalmazokon történő kísérleti jellegűek. Kutatási eredményeinket az elmélettől a gyarorlati alkalmazásokig haladó sorrendben mutatjuk be.

2. Algebra, statisztika, kombinatorika

Legfontosabb eredményünk a véges testek feletti polinomok felbontására szolgáló determinisztikus algoritmusokhoz tartozó kombinatorikai struktúrákkal kapcsolatos, amelyek az asszociációs sémák magasabb dimenziós általánosításaiként foghatók fel. Ezek vizsgálatával közel 20 év után sikerült áttörést elérnünk abban a kérdésben, hogy lehet-e determinisztikus polinom időben felbontani véges test feletti polinomokat. Több pozitív részeredményt értünk el. Egy általános módszert dolgoztunk ki, ami lényegében az algoritmikus Galois-elmélet kiterjesztése testek helyett véges féligyegyszerű algebrákra.

Véges testek feletti polinomok felbontására gyors véletlent használó algoritmusok ismertek, a determinisztikus polinomidejű módszer létezese máig nyitott kérdés. Szubexponenciális futási idejű determinisztikus módszerek csak bizonyos számelméleti hipotézis - az úgynevezett általanosított Riemann-sejtés (GRH) feltetélezese mellett ismertek. (Vonatozik ez már a másodfokú esetre, azaz a négyzetgyökvonásra is.) Az ilyen eljárások jelentős részéből sikerült kiküszöbölni ezt a feltevést azon az áron, hogy az algoritmus a polinom felbontása helyett a kapcsolódó faktorgyűrű egy automorfizmusát konstruálja meg. GRH feltételezése mellett az automorizmus használható a polinom felbontására, így eredményeink úgy is értelmezhetők, hogy a GRH felhasználását az algoritmus végére helyezik át. Eredményeinket sikerrel alkalmaztuk nemkommutatív algebrák nullosztóinak GRH-t nem feltételező hatékony keresésére. Módszereinkben kitüntett szerepet játszik a véges testek Galois-elméletenek egyfajta algoritmikus kiterjesztése nullosztókat is tartalmazó véges gyűrűkre.

Kapcsolatot találtunk véges testek feletti polinomokat faktorizáló determinisztikus algoritmusok és bizonyos kombinatorikus struktúrák - az asszociációs sémák általánosításai - között. Módszerünkkel többek között fel tudunk bontani polinom időben olyan polinomokat, amelyek foka speciális alakú prímszám. Kidolgoztunk egy polinom idejű kvantum-algoritmus a rejtett részcsoport problémájára 2 osztályú nilpotens csoportokban. Egy több rendezvényből álló tematikus félévhez kapcsolódó egyik kötetben összefoglalót írtunk arról (felkérésre), hogy a GB-technikák hogyan alkalmazhatók extremális kombinatorikai kérdések vizsgálatára. Véges félcsoportok testek feletti (lineáris) reprezentációival

kapcsolatban azzal a központi kérdéssel foglalkoztunk, hogy milyen félcsoportok esetén lesznek a reguláris reprezentáció mátrixai lineárisan függetlenek az alaptest felett. Ez automatikusan teljesül, ha a félcsoportunk valójában csoport. A korábbi, ún, l-széles családokkal kapcsolatos eredményeinket jelentős mértékben sikerült kiterjeszteni a moduláris esetre. Egy sor érdekes állítás adódik a ponthalmaz Gröbner-bázisára, illetve Gröbner-normálisára. A dolgozat kombinatorikus következményeket is tárgyal.

Felkérésre összefoglaló cikket írtunk a 4th Conference on Algebraic Informatics kötetébe, ahol Rónyai Lajos a nyáron meghívott előadó lesz. A korábbi eredmények áttekintésén túl a dolgozat tartalmaz eddig nem publikált eredeti eredményeket is az S-extremális halamzrendszerekről. Ezeknek jellemzését adtuk Gröbner-bázisok segítségével, és egy a korábbninál gyorsabb módszert javasoltunk a felismerésükre.

A Noga Alontól származó kombinatorikai Nullahelytételt arra az esetre terjesztjük ki, amikor halmazok helyett multihalmazokon vett polinomfüggvényekkel dolgozunk. Az általánosabb tételt a véges differenciák módszerével, illetve gyűrűelméleti úton is igazoljuk. Néhány alkalmazást is tárgyalunk, ezek multiplicitásokat tartalmazó kiterjesztései ismert tételeknek. Test helyett egységelemes kommutatív gyűrűkre is kiterjesztettük Alon tételét.

Vizsgáltuk a többváltozós algebrai egyenletrendszerek strukturális tulajdonságait. Az egyenletrendszer és a gyökök olyan jellemzőivel foglalkoztunk, amelyek globális módszerekkel, a gyökök meghatározása nélkül is számíthatók. Különösen akkor fontosak ezek a módszerek, amikor olyan egyenletrendszerünk adódik, amelyik többszörös megoldásokat tartalmazó rendszerhez van közel. Gyakran ilyenek a mérések útján nyert (kissé) hibás adatokból felállított közelítő egyenletrendszerek.

Vizsgáltuk geometriai struktúrák egy pont híján való fedéseit, és sikerült bizonyítanunk ezek méretére az affin síkokra érvényes Brouwer-Schrijver tételhez hasonló jelenséget. Foglalkoztunk másrészt halmazrendszerekre vonatkozó eredmények vektorterekre vonatkozó analogonjainak vizsgálatával. Ebben az irányban több társszerzővel sikerült igazolnunk a Hilton-Milner tétel q-analogonját.

3. Nagyméretű algoritmikus problémák

Extrém méretű mátrixokon történő szinguláris felbontással kapcsolatos kutatásokat végeztünk, amelynek új eredménye, hogy a klaszterezés nehézségét a sok kis számú sűrű közösség és az ezeket összekötő nyúlványok jelenlétére vezettük vissza és a híres Barabási, Watts-Strogatz és Kleinberg hálózatmodelleket kiterjesztő olyan modellt adtunk, amely megmagyarázza a sűrű közösségek és nyúlványok létrejöttét.

Egy további kutatásunkban a kommunikációs mérnöki munkában, optikai hálózatok hibalelemzése kapcsán felmerülő algoritmikus és tervezési kérdésekkel foglalkozunk. Matematikai szempontból a kombinatorikus csoport-tesztelés (CGT) strukturált változatáról van szó. Itt a tesztelendő elemek egy gráf (optikai hálózat) éleit jelentik. A teszthalmazok nem lehetnek akármilyenek, általában összefüggőknek kell lenniük, de más, ennél szigorúbb modellek is értelmesek. Ha egy él meghibásodik, akkor az őt tartalamzó teszthalmazok (és csak azok) hibát jeleznek. E hibajelzésekből kell tudnunk azonosítani a hibás élet. A dolgozatban azzal az esettel foglalkozunk, amikor legfeljebb 1 él lehet hibás.

Egyebek között majdnem optimális erdményt adunk az ún. m-trail (amikor minden teszthalmaznak van nyitott vagy zárt Euler-bejárása) modellben teljes gráfokra. Egy új heurisztikus algoritmust (RCS) javasolunk, amely elég jó eredményeket (kevés halmazból álló tesztrendszereket) ad gyakorlati méretű és szerkezetű gráfokra.

Az előző kutatási témának azzal az esetével is foglalkoztunk, amikor a gráf egy négyzetrács, a teszthalmazoknak pedig összefüggő élhalmazoknak kell lenniük (m-tree modell). Lényegében opti-

mális, körülbelül $\log_2 |E|$ méretű tesztrendszert sikerült megadni, ahol E a gráf élhalmaza. A konstrukció érdekessége, hogy algebrai eszközöket, pontosabban véges testeket használ.

Polinomidejű kvantum-algoritmust adtunk 2 osztályú nilpotens csoportokban a rejtett részcsoport problémájának megoldására. A módszer két részből áll: Az első rész konstans nilpotencia-osztályú csoportok rejtett részcsoportjainak megkeresését vezeti vissza arra az esetre, amikor a csoport exponense egy p prímszám és a rejtett részcsoport rendje 1 vagy p. A második rész az ilyen speciális esetre ad kvantum-algoritmust bizonyos véges testek feletti kvadratikus egyenletrendszerek hatékony megoldása segítségével. Az egyenletrendszer hatékony megoldhatóságáról szóló eredmény a Chevalley–Warning-tétel speciális esetének algoritmikus változataként értelmezhető.

Az operációkutatás egyik fontos ága többtényezős döntések elmélete. Ebben jelentős szerepet kapnak a páros összehasonlítás-mátrixok. Ezek olyan pozitív valós elemű négyzetes mátrixok, amelyekben az átlóra szimmetrikus elempárok szorzata 1. Az ideális mátrixok itt éppen az 1 rangúak. A gyakorlatban több ok miatt is nehéz ezt elérni. Az optimális mátrixok két modell szerinti közelítésével (Saaty, illetve LS) foglalkoztunk abban az esetben, amikor az input mátrix hiányosan adott. Mindkét modellre vonatkozóan tisztázni tudtuk, mikor lesz egyértelmű az input adatokkal összhangban levő, ezen belül pedig az ideálisat legjobban közelítő mátrix. A feltétel mindkét modell esetén egy az input mátrixhoz rendelt gráf összefüggősége. A tételek alapján érdekes algoritmusok adódnak.

A páronként összehasonlítási mátrixok fontos eszköznek számítanak a döntéstámogatás területén. Ebben a munkában azt vizsgáltuk, hogy miként alkalmazhatók az elmélet eredményei akkor, ha a szakértő a mátrixot csak részlegesen tölti ki. A strukturális eredmények mellett foglalkozunk a problémák algoritmikus vonatkozásaival is.

4. Az adatbányászat matematikai alapjai

A Látens Dirichlet Allokáció (LDA) módszert 2003-ban Blei és társai publikálták. Az LDA egy generatív valószínűségi modell, melynek célja egy dokumentumhalmaz reprezentálása rögzített számú téma keverékével. A témákat a korpusz szókészlete (szótár) felett vett multinomiális valószínűségeloszlásokkal reprezentálja, egy dokumentum pedig ezen eloszlások keveréke. A fenti konstrukcióhoz azt a feltételezést teszi, hogy a dokumentum szavainak sorrendje lényegtelen (ezzel gyakran élnek más szövegbányászati alkalmazások is), a dokumentumot ún. szózsákkal (olyan halmaz melyben ugyanaz az elem többször is előfordulhat) reprezentálja. Új eredményeink között megtalálható az LDA skálázhatóságának javítása, a Web dokumentumok klasszifikációjában történő felhasználása, a modell átalakítása úgy, hogy a hivatkozások generálására is alkalmas legyen, illetve utóbbi feladat esetében a Gibbs mintavételezés megvalósítása a módosított modell esetében.

Az MTA SZTAKI nemzetközi hírnevét öregbíti a KDD Cup 2009 versenyén nagyon szoros versenyben elért 6. helyezés. A 2009. évi KDD Cup versenyét az idén Párizsban tartott KDD (ACM SIGKDD Conference on Knowledge Discovery and Data Mining) konferenciához kapcsolódóan rendezték. A versenyen az egyik nagy nemzetközi mobilszolgáltató, az Orange anonimizált ügyféladatain kellett három, üzletileg rendkívül fontos tulajdonságot előre jelezni. Az első helyezett IBM Research csapatának pontossága 0.852, az MTA SZTAKI-é 0.846 volt, miközben a módszer szórása 0.03 körüli - a folyamatosan a teszt adatok 10%-án értékelt eredmények szerint például a verseny zárásakor az MTA SZTAKI az első helyen állt! Annak ellenére, hogy 50,000 tanító és 50,000 teszt ügyfél közel 50,000 jellemzőjét, tehát óriási adatmennyiséget tettek közzé a verseny céljára, hatalmas érdeklődést keltett a feladat: a végeredmény táblázatban 79, a módszeréről és a csapattagokról valamennyi információt közzétevő csapat szerepel. Néhány prominens résztvevő: IBM Research, Melbourne University, Taiwan University, Neo Metrics, LatentView Analytics India, Inductis, University of Waikato, HP Labs.

5. Keresés és Web Spam szűrés

A Web Spam szűrésének területén a világ egyik legerősebb kutató helyei közé kerültünk. A keresők találati oldalain elfoglalt előkelő (első) helyezés nagy forgalmat és így üzleti lehetőséget biztosít az adott weboldal üzemeltetőjének. Emiatt egyes weboldal üzemeltetők olyan technikákat (spamdexing) alkalmaznak, amelyek a felhasználók számára semmiféle többletszolgáltatást nem nyújtanak, egyetlen céljuk, hogy a céloldal helyezését a kereső rangsorokban manipulálják. Ezek szűrése elengedhetetlen mind a kereskedelmi célú keresőrendszerek, mind az Internet archiválásával foglalkozó intézetek, mind például egy elemzést, kutatást Web letöltések elemzésével végző intézmény részére.

A Web Spam kutatás nemzetközi elismertségét jelzi, hogy az adatbányászat és gépi tanulás legjelentősebb európai rendezvényéhez, a 2010. évi ECML/PKDD (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases) konferenciához kapcsolódó Discovery Challenge 2010 feladatának az Internet archívumok számára történő tartalom minőség előrejelzést választották és az Európai Internet Archívummal együttműködésben az MTA SZTAKI szervezte a versenyt. A labor a rendezvény szervezésével jelentős feladatot vállalt magára: nagy mennyiségű adat feldolgozását, terjesztését, a tanító és teszt adatok címkézését, a résztvevők tájékoztatását, eredményeik kiértékelését és a legsikeresebb csapatok eredményeinek publikálását. Az adatok előkészítéséről és a verseny tapasztalatairól a World Wide Web konferencia kapcsolódó workshopjain is beszámoltunk.

Az információkereső, rangsoroló eljárások minőségének tekintetében jelentősen továbbfejlesztettük magyar képi információkereső eljárásainkat. A módszer integrálja mind a képi, mind a természetes szöveg alapú módszereket és többek között tartalmazza a képi szegmentáció és feature szelekció hatékony implementációját. Az eredmények a CLEF (Cross-Lingual Evaluation Forum), TREC (Text Retrieval Conference) és TRECVID (NIST Text Retrieval Conference Video Retrieval Evaluation) évi rendszeres konferenciáin kerültek bemutatásra, ahol az MTA SZTAKI mára már a kutatói közösség elfogadott részévé vált.

6. Összefoglalás

Az egyes területeken elért legfontosabb eredményeink a következők:

Algebra: Hosszú idő, mintegy 20 év után el?relépést értünk el az aritmetikai/algebrai algoritmusok témakörének egyik alapkérdését illetően: nevezetesen, hogy lehet-e determinisztikus polinom időben véges test feletti polinomokat felbontani.

Algoritmusok: Algoritmust adtunk nagyon nagy méretű hálózatok partícionálására, illetve kiterjesztettük a valós hálózatok ismert modelljeit a klaszterezés nehézségének megmagyarázására.

Adatbányászat: Helyezést értünk el a 2009. év legjelentősebb adatbányászati versenyén, a KDD Cup 2009-en.

Információ-visszakeresés: A Web Spam szűrés területén a legnagyobb ipari kutatóhelyek támogatásával fő szervezői voltunk a 2010. évi Web minőség automatikus értékélésével kapcsolatos nemzetközi versenynek.

A projekt eredményeinek hasznosítása ipari partnereinken keresztül folyamatosan történik, amelyet a projekt záró beszámoló Hasznosítás fejezete ismertet.