Eötvös Loránd Tudományegyetem
Informatika Kar
Informatika Doktori Iskola
Információs Rendszerek Doktori Program

Tézisfüzet

Dr. Fogarassyné Vathy Ágnes

Új gráf alapú klaszterező és vizualizációs adatbányászati algoritmusok

Témavezető:

Dr. Kiss Attila

egyetemi docens, Ph.D.

Budapest, 2007.

1. Bevezetés

Az információs társadalom fejlődése során napról-napra egyre nagyobb mennyiségű adat halmozódik fel a különféle információs rendszerekben. Ezen adathalmazok nem csak a meglévő ismereteket, hanem új tudást is magukban rejtenek. Az adatbányászat célja ezen új ismerek feltárása, s interpretálása oly módon, hogy az a továbbiakban hasznosítható legyen. A klaszterezés az adatbányászat egy speciális területe, melynek célja az elemzendő adathalmazon belül jól elkülöníthető alcsoportok feltárása.

Az elmúlt évtizedekben számos klaszterező módszer és algoritmus látott napvilágot, azonban a problémák és az adatok sokszínűségéből fakadóan, nem létezik olyan módszer vagy algoritmus, amely minden esetben alkalmazható lenne. A gráf alapú klaszterező algoritmusok egy speciális szeletét képezik ennek a kutatási területnek. A gráfok nem csupán az adatokra, vagy az őket tömörebb formában definiáló reprezentatív elemekre vonatkozóan tartalmaznak információkat, hanem egyúttal feltárják az adatok, illetve a reprezentatív elemek között lévő kapcsolatokat is. Ez a tömör információtárolási forma megfelelő alapot biztosít a klaszterező algoritmusok hatékony működéséhez.

Mindemellett, mivel az emberi szem alacsony (2 vagy 3) dimenzióban hatékonyan képes felismerni az egyes alcsoportokat, az adatok klaszterezésében hatékony segítséget nyújthat az elemzendő adatok alacsony dimenzióban történő megjelenítése. Mivel a gráfok az adatok kapcsolati rendszerét is magukban foglalják, ezért a gráf alapú vizualizációs technikák az adatok elhelyezkedésén túlmenően a kapcsolatokra vonatkozó információkkal is elősegítik a vizuális adatelemzés folyamatát.

2. A dolgozat célkitűzése

A dolgozat célja új, gráf alapú klaszterező és vizualizációs módszerek fejlesztése. A kutatás során számos korábbi kapcsolódó eredményt használtam fel, s célom az volt, hogy ezen módszerek hiányosságait kiküszöbölve olyan új módszereket, algoritmusokat hozzak létre, melyek hatékonyabb csoportosítást és adatmegjelenítést tesznek lehetővé. A dolgozatban alkalmazott módszerek a kapcsolódó témakörök területeit ölelik fel. Ezen témakörök a következőek: adatbányászat, klaszterező módszerek és algoritmusok, gráfelmélet, neurális hálózatok, dimenziócsökkentési technikák, fuzzy módszerek, alakzatok topológiája.

3. Új tudományos eredmények

1. tézis – A Jarvis-Patrick csoportosítási módszer továbbfejlesztése

A dolgozat első tézisében a kémia tudományterületén gyakran alkalmazott Jarvis-Patrick klaszterező algoritmus továbbfejlesztését és kiterjesztését adtam meg. A Jarvis-Patrick algoritmus az objektumok közös szomszédjainak száma alapján csoportosítja az adatokat. A módszernek a következő két fő hátránya van: (i) a döntési kritérium, amely meghatározza, hogy két objektum azonos csoportba kerüljön-e, nagyon merev; valamint (ii) ezen csoportosítási feltétel kiszámításához az algoritmus az objektumoknak csak a közvetlen szomszédjait veszi figyelembe. Ezen problémák megoldására egy olyan új hasonlósági mérték bevezetését javasoltam, amely szintén az objektumok közös szomszédjain alapul, de nem korlátozódik csupán a közvetlen szomszédokra. Továbbá, a javasolt hasonlósági mérték fuzzifikálja a Jarvis-Patrick algoritmus merev döntési kritériumát. A javasolt hasonlósági mérték hierarchikus klaszterező algoritmusokban alkalmazva hatékony eszközt nyújt az adathalmazokban fellelhető alcsoportok feltárására. A javasolt módszer hatékonyságát számos alkalmazási példa támasztja alá.

A témával foglalkozó publikációm: [4].

2. tézis – A minimális feszítőfákon és a Gath-Geva algoritmuson alapuló új csoportosítási algoritmus

A dolgozat második tézise egy új élvágási feltételen alapuló klaszterező algoritmust definiál. A gráf alapú klaszterező algoritmusok nagy része azon az elven alapul, hogy először felépítenek egy, az elemzendő adatokat reprezentáló gráfot, majd ezen gráfban keresik azokat az éleket, amelyek elhagyásával kialakult részgráfok az egyes alcsoportoknak feleltethetők meg. Ezeknek az inkonzisztens éleknek a felismerése nem egyszerű feladat, hiszen a keresendő adatcsoportok különböző sűrűségűek és méretűek lehetnek. A dolgozatban javaslatot tettem egy új élvágási módszer alkalmazására, amely a lehetséges csoportok teljes hipertérfogata alapján határozza meg az elhagyandó éleket. Ezen az új élvágási feltételen alapulva szintén javaslatot tettem egy új, gráf alapú klaszterező algoritmus alkalmazására, amely hatékony kombinációja a minimális feszítőfa alapú csoportosítási módszereknek és a Gath-Geva klaszterező algoritmusnak. A javasolt algoritmus képes feloldani a gráf alapú

klaszterező algoritmusok tipikus problémáját, a láncolási effektust, illetve megoldást nyújt a Gath-Geva algoritmus inicializációs problémájára is. Az alkalmazott módszernek köszönhetően, az eredményképpen létrejött csoportok matematikai leírása könnyen megadható. A dolgozat több példán keresztül mutatja be a javasolt módszer hatékonyságát.

A témával foglalkozó publikációim: [4,6].

3. tézis – Új topológia alapú vizualizációs módszer

A dolgozat harmadik tézise egy új gráf alapú vizualizációs módszert definiál. A gyakorlati adatbányászati feladatok során általában magas dimenzionalitású adatokat kell elemezni, ezért rendkívül informatív segítséget nyújt az elvégzendő feladatban az elemzendő adathalmaz alacsony dimenziójú, az emberi szem által is értelmezhető megjelenítése. Az elemzendő adatok azonban gyakran tartalmazhatnak olyan adatsokaságot, illetve adatsokaságokat, melynek belső dimenziója kisebb, mint az adatokat jellemző tulajdonságok számossága. Ezen adathalmazok megjelenítése főként akkor jelent problémát, ha a kisebb dimenzionalitású adatsokaság nemlineárisan van beágyazva a magasabb dimenzionalitású térbe. A dolgozat ezen tézisében egy olyan új gráf alapú vizualizációs módszert javasoltam, amely alkalmas az ilyen nemlineárisan beágyazott adatsokaságok alacsony dimenzióban történő megjelenítésére. A javasolt módszer a topológiát reprezentáló hálózatok és a többdimenziós skálázás fő erényeit ötvözi oly módon, hogy eredményképpen egy hatékony megjelenítési módszer jön létre. A javasolt új módszer részletes elemzése biztató eredményeket mutat. Összevetve a javasolt módszert más hasonló céllal fejlesztett gráf alapú módszerekkel azt tapasztaltam, hogy az új módszer jobb minőségi mutatókkal rendelkezik mind az adatok távolságának, mind az adatok szomszédosságának megőrzése terén.

A témával foglalkozó publikációim: [1,2,3]

4. Eredmények hasznosítása

A dolgozatban javasolt módszerek a tudományos kutatások számos területén alkalmazhatóak. Például az orvostudomány területén a gráf alapú módszerek alkalmazhatóak a betegutak elemzése során, vagy a biokémiai kutatásban a fehérjék domén struktúrájának meghatározása céljából. Természetesen az informatika témakörén belül is számos alkalmazási területet találunk. A világháló kapcsolati rendszerének elemzése napjainkban egy rendkívül nagy érdeklődésre számot tartó kutatási téma. Mivel a weboldalak megadhatók gráf formájában is, ahol az egyes csomópontok az egyes oldalakat, az őket összekötő élek pedig a közöttünk lévő hiperlink kapcsolatokat jelentik, a dolgozatban definiált módszerek és algoritmusok alkalmazhatóak a weboldalakkal kapcsolatos kutatások során is. Egy másik példát említve, a világhálón számos olyan weboldal található (pl. iWiW), amely lehetővé teszi a felhasználóknak, hogy kiépítsék saját ismerősi, kapcsolati rendszerüket. Az emberek szociális kapcsolati rendszerének elemzése napjainkban szintén egy időszerű téma. Mivel ezen kapcsolati rendszerek szintén kitűnően megadhatók gráf formájában, ezért a javasolt módszerek hatékony segítséget nyújthatnak az ilyen témájú kutatások során is. Továbbá, a javasolt új módszerek és algoritmusok tetszőleges adatbányászati programcsomagba történő beépítése tetszőleges alkalmazási lehetőségek előtt nyitja meg az utat.

5. További kutatási lehetőségek

A dolgozatban felvázolt eredmények számos új érdekes kutatási témát vetnek fel. Az adathalmazok lokális, belső dimenzionalitása azon jellemző tulajdonságokat foglalja magában, melyek nélkülönözhetetlenek az adatpontok jellemzéséhez. Miután láttuk, hogy a leképezések távolságőrzését minősítő mérőszámok kapcsolatban állnak az adathalmazok belső dimenzionalitásával, ezért érdemes lenne megvizsgálni, hogy ezen mérőszámok alapján megadható-e olyan módszer, amely az adathalmazok belső dimenzióját becsüli meg. Ezen kérdéskörhöz kapcsolódóan érdemes lenne megvizsgálni a fraktálok dimenziójának meghatározására vonatkozó lehetőségeket is. Továbbá, érdekes kutatási témának tűnik egy olyan leképezés létrehozása is, amely a fraktál dimenzióját őrzi meg.

Publikációs lista

A dolgozat témáihoz kapcsolódó referált publikációk

[1] **A. Vathy-Fogarassy**, J. Abonyi: Local and Global Mappings of Topology Representing Networks. *Information Sciences*, Elsevier (benyújtva)

[2] **A. Vathy-Fogarassy**, A. Werner-Stark, B. Gal, J. Abonyi: Visualization of Topology Representing Networks. In *Proceedings of IDEAL 2007, Lecture Notes in Computer Science series*, Springer Verlag (elfogadva, megjelenés alatt)

[3] **A. Vathy-Fogarassy**, A. Kiss, J. Abonyi: Topology Representing Network Map – A new Tool for Visualization of High-Dimensional Data. *LNCS Transactions on Computational Science*, Springer-Verlag, 2007. (elfogadva, megjelenés alatt)

[4] **A. Vathy-Fogarassy**, A. Kiss, J. Abonyi: Improvement of Jarvis-Patrick Clustering Based on Fuzzy Similarity. *Lecture Notes in Computer Science: Applications of Fuzzy Sets Theory*, Volume 4578/2007, ISSN: 0302-9743, pp. 195–202, 2007.

[5] **A. Vathy-Fogarassy**, A. Kiss, J. Abonyi: Hybrid Minimal Spanning Tree and Mixture of Gaussians based Clustering Algorithm. *Lecture Notes in Computer Science: Foundations of Information and Knowledge Systems*, Volume 3861/2006, ISSN: 0302-9743, pp. 313–330, 2006.

[6] **Á. Vathy-Fogarassy**, B. Feil, J. Abonyi: Minimal Spanning Tree based Fuzzy Clustering. *Transactions on Enformatika, Systems Sciences and Engineering*, Volume 8, ISSN: 1305-5313, pp. 7–12, 2005.

Egyéb publikációk

[7] **Dr. Fogarassyné Vathy Ágnes**: Csoportosítás (klaszterezés) – 4. fejezet (131–184. o.) az *Adatbányászat a hatékonyság eszköze – Gyakorlati útmutató kezdőknek és haladóknak* című könyvben, Szerk: Dr. Abonyi János, ISBN: 9636183422, ComputerBooks Kiadói Kft, 2006.

[8] Starkné W. Á., **Fogarassyné Vathy Á.**, Csoma Á.: Szakértő szoftverágens a diszlexia lehetőségének megállapítására. *Acta Agraria Kaposváriensis*, Vol. 10, No 1, 65–82. o., ISSN 1418-1789, 2006.

[9] Vassányi I., Rovnyai J., **Fogarassyné Vathy Á.**, Tobak T.: Adatbányászati alkalmazások az egészségügyben. *Informatika és Menedzsment az Egészségügyben*, 2006/5 szám, 49–53. o., 2006.

[10] **Á. Vathy-Fogarassy**, G. Balázs, T. Tobak, I. Vassányi: Intelligent Data Analysis Center: A Client/Server Mining Model over the Internet. In *Proceedings of 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'2005)*, pp. 57–65, 2005.

[11] **Dr. Fogarassyné Vathy Á.**, Dr. Vassányi I.: Adatbányászati technológiák az egészségügyben. *Informatika és Menedzsment az Egészségügyben*, 2005/3 szám, 46–50. o., 2005.

[12] **Fogarassy-Vathy Ágnes**: Application of Data Mining Methods in Health Care Databases. In *Proceedings of the 6th International Conference on Applied Informatics*, Vol. I., pp. 261–266, 2004.

[13] **Dr. Fogarassyné Vathy Á.**, Dr. Patericza A.: Intelligens Adatelemző Központ létrehozása IKTA 142/2002. *XXIII. Centenáriumi Neumann Kollokvium elektronikus kiadványa*, 2004.

[14] **Dr. Fogarassyné Vathy Á.**, Dr. Fogarassy Gy.: Egészségügyi adatok előkészítése elemzések céljából. *Informatika és Menedzsment az Egészségügyben*, 2003/8 szám, 36–41. o., 2003.

[15] **Vathy Ágnes**: Adatbázisok biztonságának tervezési módszere és megvalósítási kérdései. *Informatika a felsőoktatásban 2002, elektronikus konferencia kiadvány*, Debreceni Egyetem, 2002.

[16] **Vathy Ágnes**, Kiss Attila: Database Security – Access Rights from Design to Implementation. In *Proceedings of the 5th International Conference on Applied Informatics*, Vol. I., pp. 85–94, 2001.

[17] **Vathy Ágnes**, Timár Lajos: Az EER modell koncepcióinak érvényesülése egy relációs adatbázis-rendszerben. *Informatika a felsőoktatásban '99 Konferencia kiadvány*, I. kötet, 92–97. o., Debreceni Egyetemi Szövetség, 1999.

[18] Timár Lajos, **Vathy Ágnes**: A jó minőségű EER modell helye az adatbázis-tervezésben. *Informatika a felsőoktatásban '99 Konferencia kiadvány*, I. kötet, 86–91. o., Debreceni Egyetemi Szövetség, 1999.

[19] Timár L., **Vathy Á.**, Vígh K. Telekesi É., Tátrai J., Szigeti J. Kocsis T., Vass I.: *Építsünk könnyen és lassan adatmodellt!*, Veszprémi Egyetem és Műszertechnika Kft., 1996.

[20] **Vathy Ágnes**, Németh Krisztián: *Adatmodellezési feladatok I.*, Veszprémi Egyetemi Kiadó, 1996.