Adatbányászat

Tankönyv:

Angol nyelven elérhető:
Mining of Massive Datasets
Jure Leskovec. Anand Rajaraman, Jeffrey D. Ullman,
3rd Edition, Stanford University http://i.stanford.edu/~ullman/mmds/book0n.pdf
2nd Edition, Stanford University http://infolab.stanford.edu/~ullman/mmds/bookL.pdf

Magyar nyelven elérhető:
Adatbányászat
Bodon Ferenc, Buza Krisztián (2014)
http://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011-0064_55_adatbanyaszat/index.html
Bevezetés az adatbányászatba
Pang-Ning Tan, Michael Steinbach, Vipin Kumar
Panem Kft.
http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/adatok.html

Az angol nyelvű tankönyvnek megfelelő kurzus a Stanford University képzésében:
Mining of Massive Datasets, CS246, Jure Leskovec, Anand Rajaraman, Jeff Ullman
http://www.mmds.org/

Témák:

1. **Mi az adatbányászat?**

Modellezés, Statisztikai modellezés, Gép tanulás, A modellezés kiszámítási kérdései, Leíró jellemzők számának csökkentése, Az adatbányászat statisztikai határa, Teljes információ, Bonferroni-elv, Példa a Bonferroni-elvre, Dokumentumok szavainak fontossági mértékei, Hash-függvények, Indexek, Másodlagos tárolók.

References

1. L. Breiman, “Statistical modeling: the two cultures,” Statistical Science 16:3, pp. 199–215, 2001.
2. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner, “Graph structure in the web,” Computer Networks 33:1–6, pp. 309–320, 2000.
3. M.M. Gaber, Scientific Data Mining and Knowledge Discovery — Principles and Foundations, Springer, New York, 2010.
4. H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: The Complete Book Second Edition, Prentice-Hall, Upper Saddle River, NJ, 2009.
5. D.E. Knuth, The Art of Computer Programming Vol. 3 (Sorting and Searching), Second Edition, Addison-Wesley, Upper Saddle River, NJ, 1998.
6. C.P. Manning, P. Raghavan, and H. Sch¨utze, Introduction to Information Retrieval, Cambridge Univ. Press, 2008.
7. R.K. Merton, “The Matthew effect in science,” Science 159:3810, pp. 56–63, Jan. 5, 1968.
8. P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison-Wesley, Upper Saddle River, NJ, 2005.

2. **Osztott számítási környezetek és a MapReduce**

Osztott fájlrendszerek, A számolást végző csomópontok fizikai szervezése, Nagyskálájú fájlrendszerek szervezése, A MapReduce, Map taszkok, Kulcs szerinti csoportosítás, Reduce taszkok, Combiner

taszkok, A MapReduce végrehajtásának részletei, Megoldás a csomópontok kieséséne problémájára, Algoritmusok MapReduce elvű kiszámítása, Mátrix-vektor szorzás kiszámítása MapReduce módszerrel, Mi a teendő, ha a vektor nem fér el a memóriában, A relációs algebra műveletei, A szelekció kiszámítása MapReduce módszerrel, A vetítés kiszámítása MapReduce módszerrel, Az egyesítés, metszet, különbség kiszámítása MapReduce módszerrel, A természetes összekapcsolás kiszámítása MapReduce módszerrel, A csoportosítás, aggregálás kiszámítása MapReduce módszerrel, Mátrixszorzás, Mátrixszorzás egy MapReduce lépéssel, A MapReduce kiterjesztése, A Workflow rendszerek, Spark, A Spark implementálása, TensorFlow, A MapReduce rekurzív kiterjesztései, Bulk-Synchronous rendszerek, A kommunikációs költség modellje, Kommunikációs költség Taszk hálózatokban, Wall-Clock idő, Többszörös összekapcsolások, A MapReduce elméleti bonyolultsága, A Reducer mérete és a Replication Rate, Hasonlósági összekapcsolások, A MapReduce problémák gráfmodellje, Mapping sémák, Mi a teendő, ha nincs meg az össze input adat egyszerre, Alsó korlátok a Replication Rate becslésére, Teljes esettanulmány: A mátrixszorzás.

References

1. F.N. Afrati, V. Borkar, M. Carey, A. Polyzotis, and J.D. Ullman, “Cluster computing, recursion, and Datalog,” to appear in Proc. Datalog 2.0 Workshop, Elsevier, 2011.
2. F.N. Afrati, A. Das Sarma, S. Salihoglu, and J.D. Ullman, “Upper and lower bounds on the cost of a MapReduce computation.” to appear in Proc. Intl. Conf. on Very Large Databases, 2013. Also available as CoRR, abs/1206.4377.
3. F.N. Afrati and J.D. Ullman, “Optimizing joins in a MapReduce environment,” Proc. Thirteenth Intl. Conf. on Extending Database Technology, 2010.
4. F.N. Afrati and J.D. Ullman, “Matching bounds for the all-pairs MapReduce problem,” IDEAS 2013, pp. 3–4.
5. A. Alexandrov, R. Bergmann, S. Ewen, J.-C. Freytag, F. Hueske, A. Heise O. Kao, M. Leich, U. Leser, V. Markl, F. Naumann, M. Peters, A. Rheinlander, M.J. Sax, S. Schelter, M. Hoger, K. Tzoumas, and D. Warneke, “The Stratosphere platform for big data analytics,” VLDB J. 23:6, pp. 939–964, 2014.
6. V.R. Borkar, M.J. Carey, R, Grover, N. Onose, and R. Vernica, “Hyracks: A flexible and extensible foundation for data-intensive computing,” Intl. Conf. on Data Engineering, pp. 1151–1162, 2011.
7. Y. Bu, B. Howe, M. Balazinska, and M. Ernst, “HaLoop: efficient iterative data processing on large clusters,” Proc. Intl. Conf. on Very Large Databases, 2010.
8. F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber, “Bigtable: a distributed storage system for structured data,” ACM Transactions on Computer Systems 26:2, pp. 1–26, 2008.
9. B.F. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H.-A. Jacobsen, N. Puz, D. Weaver, and R. Yerneni, “Pnuts: Yahoo!’s hosted data serving platform,” PVLDB 1:2, pp. 1277–1288, 2008.
10. J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Comm. ACM 51:1, pp. 107–113, 2008.
11. D.J. DeWitt, E. Paulson, E. Robinson, J.F. Naughton, J. Royalty, S. Shankar, and A. Krioukov, “Clustera: an integrated computation and data management system,” PVLDB 1:1, pp. 28–41, 2008.
12. flink.apache.org, Apache Foundation.
13. S. Ghemawat, H. Gobioff, and S.-T. Leung, “The Google file system,” 19th ACM Symposium on Operating Systems Principles, 2003.
14. giraph.apache.org, Apache Foundation.
15. hadoop.apache.org, Apache Foundation.
16. hadoop.apache.org/hive, Apache Foundation.
17. M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. “Dryad: distributed data-parallel programs from sequential building blocks,” Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems, pp. 59–72, ACM, 2007.
18. Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J.M. Hellerstein, “Distributed GraphLab: a framework for machine learning and data mining in the cloud,” —em Proc. VLDB Endowment 5:8, pp. 716–727, 2012.
19. G. Malewicz, M.N. Austern, A.J.C. Sik, J.C. Denhert, H. Horn, N. Leiser, and G. Czajkowski, “Pregel: a system for large-scale graph processing,” Proc. ACM SIGMOD Conference, 2010.

20. C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, “Pig latin: a not-so-foreign language for data processing,” Proc. ACM SIGMOD Conference, pp. 1099–1110, 2008.
21. spark.apache.org, Apache Foundation.
22. spark.apache.org/graphx, Apache Foundation.
23. spark.apache.org/sql, Apache Foundation.
24. www.tensorflow.org.
25. J.D. Ullman and J. Widom, A First Course in Database Systems, Third Edition, Prentice-Hall, Upper Saddle River, NJ, 2008.
26. Y. Yu, M. Isard, D. Fetterly, M. Budiu, I. Erlingsson, P.K. Gunda, and J. Currey, “DryadLINQ: a system for general-purpose distributed dataparallel computing using a high-level language,” OSDI, pp. 1–14, USENIX Association, 2008.
27. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, and I. Stoica, “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing,” Proc. 9th USENIX conference on Networked Systems Design and Implementation, USENIX Association, 2012.

3. Hasonló elemek keresése

Halmozok hasonlóságán alapuló alkalmazások. Halmazok Jaccard-hasonlósága, Dokumentumok hasonlósága, Collaborative Filtering megfogalmazása hasonló halmazok keresési problémájaként, Dokumentumok egységekre bontása: a Shingling, A k-Shingle, A Shingle méretének megválasztása, A Shingle Hash-elése, Szavakból épített Shingle, Halmazok hasonlónságőrző kivonatai, összesítései, Halmazok mátrixreprezentációja, A Minhashing, A Minhashing és a Jaccard-hasonlóság, A Minhash aláírás, A Minhash aláírás kiszámítása a gyakorlatban, A Minhashing gyorsítása, Gyorsítás Hash függvényekkel, Dokumentumok Locality-Sensitive Hash-elése, LSH a Minhash aláírás esetére, A Banding technika elemzése, A technikák kombinálása, Távolságmértékek, A távolságmérték definíciója, Euklideszi-távolság, Jaccard-távolság, Koszinusz-távolság, Szerkesztési távolság, Hamming-távolság, A Locality-Sensitive függvények elmélete, A Locality-Sensitive függvények, Locality-Sensitive családok Jaccard-távolsára, Locality-Sensitive családok erősítése, LSH családok másfajta távolságokra, LSH családok a Hamming-távolságra, Véletle hipersíkok és a koszinusz-távolság, LSH családok az Euklideszi-távolságra, További LSH családok az Euklideszi-terekben, A Locality-Sensitive Hash-elés alkalmazásai, Entity Resolution, Példák az Entity-Resolution alkalmazására, A rekordok megegyezésének validálása, Megegyezés ujjlenyomatok alapján, Egy LSH család az ujjlenyomat alapján történő illesztésre, Hasonló újságcikkek, Magasabb fokú hasonlósági módszerek, Azonos elemek keresése, Halmazok sztring típusú reprezentálása, Hosszúságalapú szűrés, Prefix indexelés, A pozícióinformációk kihasználása, A pozíció és hossz használata az indexekben.

References

1. A. Andoni and P. Indyk, “Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions,” Comm. ACM 51:1, pp. 117–122, 2008.
2. A.Z. Broder, “On the resemblance and containment of documents,” Proc. Compression and Complexity of Sequences, pp. 21–29, Positano Italy, 1997.
3. A.Z. Broder, M. Charikar, A.M. Frieze, and M. Mitzenmacher, “Min-wise independent permutations,” ACM Symposium on Theory of Computing, pp. 327–336, 1998.
4. M.S. Charikar, “Similarity estimation techniques from rounding algorithms,” ACM Symposium on Theory of Computing, pp. 380–388, 2002.
5. S. Chaudhuri, V. Ganti, and R. Kaushik, “A primitive operator for similarity joins in data cleaning,” Proc. Intl. Conf. on Data Engineering, 2006.
6. M. Datar, N. Immorlica, P. Indyk, and V.S. Mirrokni, “Locality-sensitive hashing scheme based on p-stable distributions,” Symposium on Computational Geometry pp. 253–262, 2004.
7. A. Gionis, P. Indyk, and R. Motwani, “Similarity search in high dimensions via hashing,” Proc. Intl. Conf. on Very Large Databases, pp. 518–529, 1999.
8. M. Henzinger, “Finding near-duplicate web pages: a large-scale evaluation of algorithms,” Proc. 29th SIGIR Conf., pp. 284–291, 2006.
9. P. Indyk and R. Motwani. “Approximate nearest neighbor: towards removing the curse of dimensionality,” ACM Symposium on Theory of Computing,

pp. 604–613, 1998.
10. P. Li, A.B. Owen, and C.H. Zhang. “One permutation hashing,” Conf. on Neural Information Processing Systems 2012, pp. 3122–3130.
11. U. Manber, “Finding similar files in a large file system,” Proc. USENIX Conference, pp. 1–10, 1994.
12. M. Theobald, J. Siddharth, and A. Paepcke, “SpotSigs: robust and efficient near duplicate detection in large web collections,” 31st Annual ACM SIGIR Conference, July, 2008, Singapore.
13. C. Xiao, W. Wang, X. Lin, and J.X. Yu, “Efficient similarity joins for near duplicate detection,” Proc. WWW Conference, pp. 131-140, 2008.

4. Adatfolyamok adatbányászata

Az adatfolyam adatmodellje, Egy adatfolyam-kezelő rendszer, Példák adatfolyamokra, Adatfolyam lekérdezések, Adatfolyamok feldolgozásának elvei, Minta kiválasztása az adatfolyamból, Egy kiindulási példa, Reprezentatív minták kinyerése, Általános mintavételezési probléma, A mintaméret változtatása, Adatfolyamok szűrése, Példák a szűrésre, A Bloom-szűrő, A Bloom-szűrés elemzése, Az adatfolyamban előforduló különböző elemek megszámolása, A Count-Distinct probléma, A Flajolet-Martin algoritmus, A becslések kombinálása, Az algoritmusok helyigénye, A momentumok becslése, A momentumok definíciója, A második momentumokra vonatkozó Alon-Matias-Szegedy algoritmus, Miért működik az Alon-Matias-Szegedy algoritmus, Magasabb rendű momentumok, Végtelen adatfolyamok, Az 1-esek megszámolása egy ablakban, A pontos megszámlálás öltsége, A Datar-Gionis-Indyk-Motwani algoritmus, A DGIM algoritmus tárhelyigénye, Lekérdezések megválaszolása a DGIM algoritmus esetén, A DGIM feltételek fenntartása, A hiba csökkentése, Az 1-esek összeszámolási problémájának kiterjesztései, Lecsengő ablakok, A legtöbb közös elem problémája, A lecsengő ablakok definíciója, A legnépszerűbb elemek megtalálása.

References

1. N. Alon, Y. Matias, and M. Szegedy, “The space complexity of approximating frequency moments,” 28th ACM Symposium on Theory of Computing, pp. 20–29, 1996.
2. B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, “Models and issues in data stream systems,” Symposium on Principles of Database Systems, pp. 1–16, 2002.
3. B.H. Bloom, “Space/time trade-offs in hash coding with allowable errors,” Comm. ACM 13:7, pp. 422–426, 1970.
4. M. Datar, A. Gionis, P. Indyk, and R. Motwani, “Maintaining stream statistics over sliding windows,” SIAM J. Computing 31, pp. 1794–1813, 2002.
5. P. Flajolet and G.N. Martin, “Probabilistic counting for database applications,” 24th Symposium on Foundations of Computer Science, pp. 76–82, 1983.
6. M. Garofalakis, J. Gehrke, and R. Rastogi (editors), Data Stream Management, Springer, 2009.
7. P.B. Gibbons, “Distinct sampling for highly-accurate answers to distinct values queries and event reports,” Intl. Conf. on Very Large Databases, pp. 541–550, 2001.
8. H.V. Jagadish, I.S. Mumick, and A. Silberschatz, “View maintenance issues for the chronicle data model,” Proc. ACM Symp. on Principles of Database Systems, pp. 113–124, 1995.
9. W.H. Kautz and R.C. Singleton, “Nonadaptive binary superimposed codes,” IEEE Transactions on Information Theory 10, pp. 363–377, 1964.
10. J. Vitter, “Random sampling with a reservoir,” ACM Transactions on Mathematical Software 11:1, pp. 37–57, 1985.

5. Link analízis

A PageRank, Korai keresőmotorok és a spamok fogalma, A PageRank formális definíciója, A Web szerkezete, A zsákutcák elkerülése, A pókhálócsapdák és a büntető súlyozása, A PageRank használata a keresőmotorokban, A PageRank hatékony kiszámítása, Az átmenetmátrix reprezentálása, A PageRank iterációja MapReduce módszerrel, Combiner-ek használata az eredményvektor kiszámítására, Az átmenetmátrix blokkjainak reprezentálása, További hatékony módszereke a PageRank iterációjának kiszámítására, Topic-Sensitive PageRank, A Topic-Sensitive Page Rank motivációs példája, Torzított véletlen bolyongás, A Topic-Sensitive PageRank használata, Témák

meghatározása a szavak alapján, A Link Spam, A Spam farmok felépítése, A Spam farmok elemzése, A Link Spam elkerülése, A TrustRank, A Spam Mass, A Hub-ok and Authority-k, A HITS mögötti intuíció, A Hub és Authority formális leírása.

References

1. S. Brin and L. Page, “Anatomy of a large-scale hypertextual web search engine,” Proc. 7th Intl. World-Wide-Web Conference, pp. 107–117, 1998.
2. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Weiner, “Graph structure in the web,” Computer Networks 33:1–6, pp. 309–320, 2000.
3. Z. Gy¨ongi, P. Berkhin, H. Garcia-Molina, and J. Pedersen, “Link spam detection based on mass estimation,” Proc. 32nd Intl. Conf. on Very Large Databases, pp. 439–450, 2006.
4. Z. Gy¨ongi, H. Garcia-Molina, and J. Pedersen, “Combating link spam with trustrank,” Proc. 30th Intl. Conf. on Very Large Databases, pp. 576–587, 2004.
5. T.H. Haveliwala, “Efficient computation of PageRank,” Stanford Univ. Dept. of Computer Science technical report, Sept., 1999. Available as http://infolab.stanford.edu/~taherh/papers/efficient-pr.pdf
6. T.H. Haveliwala, “Topic-sensitive PageRank,” Proc. 11th Intl. WorldWide-Web Conference, pp. 517–526, 2002
7. J.M. Kleinberg, “Authoritative sources in a hyperlinked environment,” J. ACM 46:5, pp. 604–632, 1999.

6. Gyakori elemhalmazok

A vásárlói kosár modellezése, A gyakori elemhalmazok definíciója, A gyakori elemhalmazok alkalmazása, Asszociációs szabályok, Magas konfidenciájú asszociációs szabályok keresése, a vásárlói kosarak és az A-Priori algoritmus, A vásárlói kosarak adatreprezentációja, Az elemszám kiszámításának memóriaigénye, Az elemhalmazok monotonitása, A párok megszámolásának nehézsége, Az A-Priori algoritmus, A-Priori algoritmus az összes gyakori elemhalmaz kiszámítására, Nagy adathalmazok kezelése a memóriában, Park, Chen és Yu algoritmusa, A Multistage algoritmus, A Multihash algoritmus, A Limited-Pass algoritmus, Véletlen mintavételezésen alapuló algoritmusok, Hibacsökkentés a mintavételezési algoritmusokban, A Savasere, Omiecinski és Navathe algoritmus, A SON algoritmus és a MapReduce, A Toivonen-algoritmus, Miért működik a Toivonen-algoritmus, Gyakori elemek megszámlálása adatfolyamokban, Mintavételezési módszerek adatfolyamok esetében, Gyakori elemhalmazok a lecsengő ablakokban, Hibrid módszerek.

References

1. R. Agrawal, T. Imielinski, and A. Swami, “Mining associations between sets of items in massive databases,” Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 207–216, 1993.
2. R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Intl. Conf. on Very Large Databases, pp. 487–499, 1994.
3. M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J.D. Ullman, “Computing iceberg queries efficiently,” Intl. Conf. on Very Large Databases, pp. 299-310, 1998.
4. J.S. Park, M.-S. Chen, and P.S. Yu, “An effective hash-based algorithm for mining association rules,” Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 175–186, 1995.
5. A. Savasere, E. Omiecinski, and S.B. Navathe, “An efficient algorithm for mining association rules in large databases,” Intl. Conf. on Very Large Databases, pp. 432–444, 1995.
6. H. Toivonen, “Sampling large databases for association rules,” Intl. Conf. on Very Large Databases, pp. 134–145, 1996.

7. Klaszterezés

Bevezetés a klaszterezési technikákba, pontok, terek, távolságok, klaszterezési stratégiák, A dimenziók átka, A hierarchikus klaszterezés, Hierarchikus klaszterezés Euklideszi-térben, A hierarchikus klaszterezés hatékonysága, Alternatív szabályok a kontrolált hierarchikus klaszterezésre, Hierarchikus klaszterezés nem-Euklideszi terekben, A K-means algoritmus, A K-Means algoritmus elemzése, A klaszterek inicializálása a K-Means algoritmusban, A k érték helyes megválasztása, Bradley, Fayyad és Reina algoritmusa, Adatfeldolgozás a BFR algoritmusban, A CURE algoritmus, A CURE inicializálása, A

CURE algoritmus végrehajtása, Klaszterezés nem-Euklideszi terekben, Klaszterek reprezentálása a GRGPF algoritmusban, A klaszterfa inicializálása, Pontok hozzáadása a GRGPF algoritmusban, Klaszterek kettévágása és összevonása, Folyamok klaszterezése és párhuzamosítása, A Stream-Computing modell, Egy példa folyamklaszterező algoritmusra, A kosarak inicializálása, a kosarak összevonása, Lekérdezések megválaszolása, klaszterezés párhuzamos környezetben.

References

1. B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan, "Maintaining variance and k-medians over data stream windows," Proc. ACM Symp. on Principles of Database Systems, pp. 234–243, 2003.
2. P.S. Bradley, U.M. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," Proc. Knowledge Discovery and Data Mining, pp. 9–15, 1998.
3. V. Ganti, R. Ramakrishnan, J. Gehrke, A.L. Powell, and J.C. French:, "Clustering large datasets in arbitrary metric spaces," Proc. Intl. Conf. on Data Engineering, pp. 502–511, 1999.
4. H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: The Complete Book Second Edition, Prentice-Hall, Upper Saddle River, NJ, 2009.
5. S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 73–84, 1998.
6. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 103–114, 1996.

8. Hirdetés a Weben

Az On-line hirdetések lényege, Hirdetési lehetőségek, Hirdetések direkt elhelyezése, A hirdetések megjelenítésének elvei, On-line algoritmusok, On-line és Off-line algoritmusok összehasonlítása, Greedy algoritmusok, A Competitive Ratio fogalma, A párosítási probléma, Párosítás és tökéletes párosítás, Greedy algoritmus a maximális párosítás kiszámítására, A Greedy párosításra vonatkozó Competitive Ratio, Az Adwords probléma, A keresési hirdetések története, Az Adwords probléma formális definíciója, Greedy megközelítés az Adwords problémára, A Balance algoritmus, Alsó korlát a Balance algoritmus Competitive Ratio értékének becslésére, A Balance algoritmus több licitáló esetére, Az általánosított Balance algoritmus, További észrevételek az Adwords problémával kapcsolatban, Az Adwords megvalósítása, Licitek párosítása és kereső lekérdezések, További összetett párosítási problémák, Egy párosítási algoritmus a dokumentumok és licitek esetére.

References

1. N. Craswell, O. Zoeter, M. Taylor, and W. Ramsey, "An experimental comparison of click-position bias models," Proc. Intl. Conf. on Web Search and Web Data Mining pp. 87–94, 2008.
2. B. Kalyanasundaram and K.R. Pruhs, "An optimal deterministic algorithm for b-matching," Theoretical Computer Science 233:1–2, pp. 319–325, 2000.
3. A Mehta, A. Saberi, U. Vazirani, and V. Vazirani, "Adwords and generalized on-line matching," IEEE Symp. on Foundations of Computer Science, pp. 264–273, 2005.

9. Ajánló rendszerek

Ajánló rendszerek modelljei, A Utility mátrix, A hosszú farkú eloszlások, Ajánló rendszerek alkalmazhatósága, A Utility mátrix kitöltése, A Content-Based ajánlás, Az Item Profile, A dokumentumok jellemzőinek megkeresése, Item Feature-k meghatározása leírók Tag-ek alapján, Az Item Profile reprezentálása, A User Profile, Elemek ajánlása a tartalom a alapján, Osztályozó algoritmusok, A Collaborative Filtering, A hasonlóság mértékei, A hasonlóság dualitása, Felhasználó és elemek klaszterezése, Dimenziócsökkentés, UV-dekompozíció, A Root-Mean-Square Error, Az UV-dekompozíció növekményes kiszámítása, Optimalizálás, Egy teljes UV-dekompozíciós algoritmus, A Netflix Challenge.

References

1. G. Adomavicius and A. Tuzhilin, “Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” IEEE Trans. on Data and Knowledge Engineering 17:6, pp. 734–749, 2005.
2. C. Anderson, http://www.wired.com/wired/archive/12.10/tail.html 2004.
3. C. Anderson, The Long Tail: Why the Future of Business is Selling Less of More, Hyperion Books, New York, 2006.
4. Y. Koren, “The BellKor solution to the Netflix grand prize,” www.netflixprize.com/assets/GrandPrize2009 BPC BellKor.pdf 2009.
5. G. Linden, B. Smith, and J. York, “Amazon.com recommendations: itemto-item collaborative filtering,” Internet Computing 7:1, pp. 76–80, 2003.
6. M. Piotte and M. Chabbert, ”The Pragmatic Theory solution to the Netflix grand prize,” www.netflixprize.com/assets/ GrandPrize2009 BPC PragmaticTheory.pdf 2009.
7. A. Toscher, M. Jahrer, and R. Bell, “The BigChaos solution to the Netflix grand prize,” www.netflixprize.com/assets/GrandPrize2009 BPC BigChaos.pdf 2009.
8. L. von Ahn, “Games with a purpose,” IEEE Computer Magazine, pp. 96–98, June 2006.

10. Közösségi hálók adatbányászata

A közösségi hálók gráfja, Mik egy közösségi háló jellemzői?, Példák közösségi hálókra, Eltérő adattípusú csúcsokat tartalmazó gráfok, Közösségi hálók klaszterezése, Távolságmértékek a közösségi hálók gráfjában, Alapvető klaszterező eljárások alkalmazása, A Betweenness, A Girvan-Newman algoritmus, Közössége keresése a Betweenness segítségével, Közösségek közvetlen keresése, Klikkek keresése, Teljes páros gráfok, Teljes páros részgráf keresése, Miért létezik teljes páros gráf, Gráfok particionálása, Mitől jó egy partíció?, Normalizált vágások, Gráfok megadása mátrixokkal, A Laplace-mátrix sajátértékei, További particionáló módszerek, Átfedő közösségek keresése, A közösségek természete, Maximum-Likelihood becslések, Az Affiliation-Graph modell, Közösségek keresésének fiszkrét optimalizációs módszerei, A Simrank, Véletlen bolyongás a közösségi gráfon, Véeletlen bolyongás újrakezdéssel, Háromszögek számlálása, Miért pont a háromszögeket számoljuk?, Algoritmus a háromszögek megtalálására, A háromszögkereső algoritmus optimalitása, Háromszögek keresése MapReduce módszerrel, Csökkentsük a Reduce Taszkokat, Gráfok szomszédsági tulajdonságai, Irányított gráfok és a szomszédság, Gráfok átmérője, Elérhetőség és tranzitív lezárás, Elérhetőség kiszámítása MapReduce módszerrel, Szeminaiv kiértékelés, Lineáris tranzitív lezárás, Tranzitív lezárás rekurzív duplázással, Javított Smart tranzitív lezárás, A módszerek összehasonlítása,Tranzitív lezárás gráfredukcióval, A szomszédság méretének közelítő kiszámítása.

References

1. F. N. Afrati, D. Fotakis, and J. D. Ullman, “Enumerating subgraph instances by map-reduce,” http://ilpubs.stanford.edu:8090/1020
2. F.N. Afrati and J.D. Ullman, “Transitive closure and recursive Datalog implemented on clusters,” in Proc. EDBT (2012).
3. L. Backstrom and J. Leskovec, “Supervised random walks: predicting and recommending links in social networks,” Proc. Fourth ACM Intl. Conf. on Web Search and Data Mining (2011), pp. 635–644.
4. P. Boldi, M. Rosa, and S. Vigna, “HyperANF: approximating the neighbourhood function of very large graphs on a budget,” Proc. WWW Conference (2011), pp. 625–634.
5. S. Fortunato, “Community detection in graphs,” Physics Reports 486:3–5 (2010), pp. 75–174.
6. M. Girvan and M.E.J. Newman, “Community structure in social and biological networks,” Proc. Natl. Acad. Sci. 99 (2002), pp. 7821–7826.
7. Y.E. Ioannidis, “On the computation of the transitive closure of relational operators,” Proc. 12th Intl. Conf. on Very Large Data Bases, pp. 403–411.

8. G. Jeh and J. Widom, “Simrank: a measure of structural-context similarity,” Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002), pp. 538–543.
9. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, “Trawling the Web for emerging cyber-communities, Computer Networks 31:11–16 (May, 1999), pp. 1481–1493.
10. J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney, “Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters,” http://arxiv.org/abs/0810.1355.
11. S. Melnik, H. Garcia-Molina, and E. Rahm, “Similarity flooding: a versatile graph matching algorithm and its application to schema matching, Proc. Intl. Conf. on Data Engineering (2002), pp. 117–128.
12. C.R. Palmer, P.B. Gibbons, and C. Faloutsos, “ANF: a fast and scalable tool for data mining in massive graphs,” Proc. Eighth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (2002), pp. 81–90.
13. J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Trans. on Pattern Analysis and Machine Intelligence,” 22:8 (2000), pp. 888–905.
14. Stanford Network Analysis Platform, http://snap.stanford.edu.
15. S. Suri and S. Vassilivitskii, “Counting triangles and the curse of the last reducer,” Proc. WWW Conference (2011).
16. H. Tong, C. Faloutsos, and J.-Y. Pan, “Fast random walk with restart and its applications,” ICDM 2006, pp. 613–622.
17. C.E. Tsourakakis, U. Kang, G.L. Miller, and C. Faloutsos, “DOULION: counting triangles in massive graphs with a coin,” Proc. Fifteenth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (2009).
18. P. Valduriez and H. Boral, “Evaluation of recursive queries using join indices,” Expert Database Conf. (1986), pp. 271–293.
19. U. von Luxburg, “A tutorial on spectral clustering,” Statistics and Computing bf17:4 (2007), 2007, pp. 395–416.
20. J. Yang and J. Leskovec, “Overlapping community detection at scale: a nonnegative matrix factorization approach,” ACM International Conference on Web Search and Data Mining, 2013.
21. J. Yang, J. McAuley, J. Leskovec, “Detecting cohesive and 2-mode communities in directed and undirected networks,” ACM International Conference on Web Search and Data Mining, 2014.
22. J. Yang, J. McAuley, J. Leskovec, “Community detection in networks with node attributes,” IEEE International Conference On Data Mining, 2013.

11. Dimenziócsökkentés

Szimmetrikus mátrixok sajátértékei és sajátvektorai, A sajátértékek, sajátvektorok definíciói, kiszámítása, A sajátvektorok, sajátértékek kiszámítása hatványiterációval, A mátrixok sajátvektorainak jellemzői, Főkomponens analízis, Példa főkomponens analízisre, Sajátvektorokkal történő dimenziócsökkentés, Távolságmátrix, Az SVD definíciója, Az SVD szemléletes jelentése, Dimenziócsökkentés SVD kiszámításával, Miért máködik, hogy az alacsony szinguláris értéket kinullázzuk?, Lekérdezések fogalmak használatával, Az SVD kiszámítása, A CUR dekompozíció, A CUR definíciója, Oszlopok és sorok megfelelő választása, A Middle mátrix konstruálása, A teljes CUR dekompozíció, Dupla sorok és oszlopok törlése.

References

1. S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” J. American Society for Information Science 41:6 (1990).
2. P. Drineas, R. Kannan, and M.W. Mahoney, “Fast Monte-Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition,” SIAM J. Computing 36:1 (2006), pp. 184–206.
3. G.H. Golub and W. Kahan, “Calculating the singular values and pseudoinverse of a matrix,” J. SIAM Series B 2:2 (1965), pp. 205–224.
4. G.H.Golub and C.F. Van Loan, Matrix Computations, JHU Press, 1996.
5. M.W. Mahoney, M. Maggioni, and P. Drineas, Tensor-CUR decompositions For tensor-based data, SIGKDD, pp. 327–336, 2006.
6. K. Pearson, “On lines and planes of closest fit to systems of points in space,” Philosophical Magazine 2:11 (1901), pp. 559–572.
7. J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, “Less is more: compact matrix decomposition for large sparse graphs,” Proc. SIAM Intl. Conf. on Data Mining, 2007.
8. M.E. Wall, A. Reichtsteiner and L.M. Rocha, “Singular value decomposition

and principal component analysis," in A Practical Approach to Microarray Data Analysis (D.P. Berrar, W. Dubitzky, and M. Granzow eds.), pp. 91–109, Kluwer, Norwell, MA, 2003.

12. Nagyskálájú gépi tanulás

A gép tanulás modelje, Tanító halmazok, Példák gépi tanulásra, A gépi tanulás különböző megközelítései, Gépi tanulás architektúrái, A Perceptronok, Egy Perceptron tanítása zéró köszöbbel, Perceptronok konvergenciája, A Winnow-algoritmus, Különböző változó küszöbök használata, Többosztályos Perceptronok, A tanító halmaz transzformáció,A Perceptronok használatának problémája, Perceptronok párhuzamos implementációja, A Support-Vector Machines, Az SVM működésének lényege, Hipersíkok normalizálása, Optimális közelítő szeparátor keresése, SVM módszer Gradient Descent használatával, A sztrochasztikus Gradient Descent, Az SVM pűárhuzamos megvalósításai, Legközelebbi szomszéd szerinti tanulás, A legközelebbi szomszéd szerinti tanulás kiszámítása eljárásai, Egyetlen legközelebbi szomszéd használata, Egydimenziós függvények tanulása, Kernel regresszió, Magas dimenziójú Euklideszi terek kezelése, Nem-Euklideszi távolságok kezelése, Döntési fák és használatuk, Tisztasági mértékek, Döntési fák csúcspontjainak kiválasztása, Numerikus értékek esetére megfelelő tesztek, kategorikus értékek esetére megfelelő tesztek, Döntési fák építésének párhuzamosítása, csúcsok levágása, Általánosítások: Decision Forest, A tanulási módszerek összehasonlítása.

References

1. H. Blockeel and L. De Raedt, "Top-down induction of first-order logical decision trees," Artificial intelligence 101:1–2 (1998), pp. 285–297.
2. A. Blum, "Empirical support for winnow and weighted-majority algorithms: results on a calendar scheduling domain," Machine Learning 26 (1997), pp. 5–23.
3. L. Bottou, "Large-scale machine learning with stochastic gradient descent," Proc. 19th Intl. Conf. on Computational Statistics (2010), pp. 177–187, Springer.
4. L. Bottou, "Stochastic gradient tricks, neural networks," in Tricks of the Trade, Reloaded, pp. 430–445, Edited by G. Montavon, G.B. Orr and K.-R. Mueller, Lecture Notes in Computer Science (LNCS 7700), Springer, 2012.
5. C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery 2 (1998), pp. 121–167.
6. N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000.
7. C. Cortes and V.N. Vapnik, "Support-vector networks," Machine Learning 20 (1995), pp. 273–297.
8. Y. Freund and R.E. Schapire, "Large margin classification using the perceptron algorithm," Machine Learning 37 (1999), pp. 277–296.
9. H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: the Complete Book, Prentice Hall, Upper Saddle River NJ, 2009.
10. T. Joachims, "Training linear SVMs in linear time." Proc. 12th ACM SIGKDD (2006), pp. 217–226.
11. N. Littlestone, "Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm," Machine Learning 2 (1988), pp. 285–318.
12. M. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry (2nd edition), MIT Press, Cambridge MA, 1972.
13. J. R. Quinlan, "Induction of decision trees," Machine Learning 1 (1986), pp. 81–106.
14. F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," Psychological Review 65:6 (1958), pp. 386–408.