{tag} {/tag} International Journal of Computer Applications
© 2010 by IJCA Journal.

Number 3 - Article 15

Year of Publication: 2010

Authors:

Mayank Sharma

Navin Rajpal

B.V.R.Reddy

10.5120/77-172

{bibtex}pxc387172.bib{/bibtex}

Abstract

Performance of the data warehouse depends on physical design. Index selection and storage of multidimensional data bases are important activities of physical designing process. Conventional indexing techniques such as bitmaps, B-trees and hash based indexing systems need large storage space for storing indexes along with data itself. Spelling variants, misspellings and transliteration differences are source of uncertainty in data with in the databases. Misspelled and distorted key values are also hard to map in present indexing systems. In this paper neural network based physical design is suggested, a class of artificial neural network known as self-organizing net is used for indexing data warehouse at physical level. Indexes of active neurons will be used for generating indexes for the data values. In conventional indexing techniques every key value is mapped to a specific point in space, while in neural network based database indexing system, every key value is mapped to a region in space. This region is a class to which the key values of similar type belong. Indexes generated through this method used optimal space for storage, as only final weight matrices after training of neurons are stored. Self-organizing net based indexing is very robust as distorted key values get indexed to right classes. Accuracy of our self-organizing net based indexing system in

mapping key values with distorted keys is found to be high.

Reference

[1] Abraham Silberschatz, Henry F. Korth, and Sudarshan (2002). Database System Concepts (pp 445-489). 4th Edition, McGraw Hill.

[2] Agarwal Sameet, Agarwal Rakesh, Deshpande Prasad M. Gupta Ashish, Naughton Jeffrey F., Ramakrishnan Raghu, Sarawagi, Suntia (1996). On the Computation of multidimensional Aggregates, Proc. 22nd VLDB Conf. Mumbai, India,1996.

[3] Agrawal, S., Narasayya V., and Yang, B (2004). Integrating vertical and horizontal partitioning into automated physical database design. In Proc. ACM SIGMOD international Conference on Management of Data (Paris, France, June 13 - 18, 2004). SIGMOD '04. ACM, New York, NY, 359-370.

[4] Almasi S. George, Lawreence Douglas, and Rushmeier Edith (2001) Scalable Parallel algorithm for self-organizing maps with applications to sparse data-mining problems, United States Patent, Patent No. US 6,260,036 B1 July 10, 2001.

[5] Bennett, K. P., Fayyad, U., and Geiger, D (1999). Density-based indexing for approximate nearest-neighbor queries, In Proceedings of the Fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Diego, California, United States, August 15 - 18, 1999). KDD '99. ACM, New York, NY, 233-243.

[6] Dainel C. Zilio, Jun Rao San Lightstone, Guy Lohman, Adam Strom, Christian Garcia-Arellano, and Scott Fadden (2004), Recommending Materialized views and indexes with IBM's DB2 Design Advisor, International Conference on Autonomic Computing 2004.

[7] Daniel C. Zilio, Jun Rao, San Lightstone, Guy Lohman, Adam Strom, Christian Garcia-Arellano, and Scott Fadden (2004). DB2 Design Advisor :Integrated Automatic Physical Database Design, Proc. 30th VLDB Conf. Toronto, Canada, 2004, pp. 1087-1097.

[8] Davis, J. and Goadrich, M.(2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international Conference on Machine Learning, (Pittsburgh, Pennsylvania, June 25 - 29, 2006), ICML '06, vol. 148. ACM, New York, NY, 233-240. DOI=http://doi.acm.org/10.1145/1143844.1143874

[9] Elmasri Ramez, Somayajulu V. L. N. Durvansula, Navathe B. Shamkant, and Gupta K. Shyam (2007). Fundamentals of database systems(pp 297-356). 3rd edition: Pearson Education.

[10] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recogn. Lett. 27, 8 (Jun. 2006), 861-874. DOI= http://dx.doi.org/10.1016/j.patrec.2005.10.010

[11] Fawcett, T. (2003). ROC graphs: Notes and practical considerations for data mining researchers, Tech report HPL-2003-4. HP Laboratories, Palo Alto, CA, USA. Available:http://www.purl.org/net/tfawcett/papers/HPL-2003-4.pdf.

[12] Freeston, M(1995). A general solution of the n-dimensional B-tree problem, In Proceedings of the 1995 ACM SIGMOD international Conference on Management of Data (San Jose, California, United States, May 22 - 25, 1995). M. Carey and D. Schneider, Eds. SIGMOD '95. ACM, New York, NY, 80-91.

[13] French J.C., Powell, A. L., and Schulman, E.(1997). Applications of approximate word matching in information retieval. In Proceeding of the sixth international conference on information and knowledge Management, Las Vegas, Nevada, US, November 10-14,1997, CIKM'97. ACM. New York, NY, 9-15.

[14] Goil, S. and Choudhary, A. 1997. High Performance OLAP and Data Mining on Parallel

Computers. Data Min. Knowl. Discov. 1, 4 (Dec. 1997), 391-417. DOI= http://dx.doi.org/10.1023/A:1009777418785

[15] Goil Sanjay, Choudhary Alok(1996). Design and Implementation of a scalable parallel system for multidimensional analysis and OLAP, 13th Int'l symposium on parallel and distributed processing.

[16] Gray J., Reuter A., Layman A., and Pirahesh H.(1996). Data cube: A relational aggregation operator generalizing group-by, cross-tabs, and sub-totals. In Proc. of the 12th Int'l Conference on Data Engineering, pp 152-159.

[17] Harinarayan, V., Rajaraman, A., and Ullman, J. D. (1996). Implementing data cubes efficiently. In Proceedings of the 1996 ACM SIGMOD international Conference on Management of Data (Montreal, Quebec, Canada, June 04 - 06, 1996). SIGMOD'96. ACM, NewYork, NY, 205216. DOI=http://doi.acm.org/10.1145/233269.2333333

[18] Kesheng Wu, Ekow Otoo, and Arie Shoshani(2004). On the performace of bitmap indices for high cardinality attributes, Proceedings of the 30th VLDB Conf. Toronto, Canada, 2004 pp. 24-35.

[19] Kolovson, C. P. and Stonebraker, M. 1991. Segment indexes: dynamic indexing techniques for multi-dimensional interval data. SIGMOD Rec. 20, 2 (Apr. 1991), 138-147. DOI= http://doi.acm.org/10.1145/119995.115807

[20] Lanka, S. and Mays, E. 1991. Fully persistent B+-trees. SIGMOD Rec. 20, 2 (Apr. 1991), 426-435. DOI= http://doi.acm.org/10.1145/119995.115861

[21] Li Jianzhong, Srivastava Jaideep(2002), Efficient Aggregation Algorithms for Compressed Data Warehouses, IEEE Trans. Knowledge and data engineering, Vol. 14. No.3, pp 515-529.

[22] Malinowski, E. and Zimnyi, E.(2008). Advanced Data Warehouse Design: from Conventional to Spatial and Temporal Applications (Data-Centric Systems and Applications).1st ed.2008., pp 51-55,Springer Publishing Company, ISBN: 978-3-540-74404-7.

[23] Md. Mehedi Masud, Gopal Chandra Das, Md. Anisur Rahman, and Arunashis Ghose(2006). A Hasing Technique Using Separate Binary Tree", Data Science Journal, Volume 5, 19, October 2006, pp 143-161.

[24] Pao Y.H.(1989). Adaptive Pattern Recognition and Neural Networks,Addison-Wesley, Reading, MA, 1989.

[25] Pearson, P. K. 1990. Fast hashing of variable-length text strings. Commun. ACM 33, 6 (Jun. 1990), 677-680. DOI= http://doi.acm.org/10.1145/78973.78978

[26] Ramakrisha M.V., Justin, Zobel (1997).,"Performance in Practice of String Hashing Functions", Proceedings of the fifth International Conference on Database Systems for Advanced Applications, Melbourane, Australia, April 1-4, 1997.

[27] Sarawagi S.(1997). Indexing OLAP data, IEEE Data Engineering Bulletin, March.

[28] Seeger, B. and Larson, P. 1991. Multi-disk B-trees. SIGMOD Rec. 20, 2 (Apr. 1991), 436-445. DOI= http://doi.acm.org/10.1145/119995.115862

[29] Xin Dong, Han Jiawei, Li Xiaolei, Shao Zheng, and Wah. Benjamin W.(2007). Computing Iceberg Cubes by Top-Down and Bottom-Up Integration : The StarCubing Approach, IEEE Trans. Knowledge and data engineering, Vol. 19. No.1, pp 111-126.

[30] Zhao Y., Deshpande P., and Naughton J.(1997). An array-based algorithm for simultaneous multi-dimensional aggregates. In Proc. ACM-SIGMOD International Conferences on Management of Data, pp 159-170.

[31] Zhao Yihong, Tufte Kristin, Naughton F Jeffrey (1996), On the Performance of an Array-based ADT for OLAP workloads, Technical Report CS-TR-96-1313, University of

Wisconsin-Madison, CS Department, May, 1996.

Index Terms

Computer Science Databases

Key words

Self-organizing net

multidimensional databases

indexing