بررسی تطبیقی الگوریتم‌های ساخت هیستوگرام‌های درخت R و مقایسه آن با مجموعه درخت‌های R^+ و R^*

محبوبه شمسی ، مرضیه دوستی
عضو هیئت علمی دانشکده برق و کامپیوتر دانشگاه صنعتی قم

A comparative study of algorithms to build the tree histograms comparison with the R and R + and R * of trees

Mahboobeh shamsi , marzie dousti
Faculty members of electrical and computer engineering university of qom
Email: shamsi@qut.ac.ir
Email: Dousti ٥@gmail.com

چکیده

امروزه طبقه‌بندی اطلاعات برای استفاده مناسب از آنها به هنگام نیاز امری بسیار مهم و حیاتی است ، در این راستا روش‌های مختلفی برای طبقه‌بندی معرفی شده است. که هریک به نوبه خود بهبودی را در راستای بهینه‌سازی اکتشاف و طبقه‌بندی داده بوجود آورده‌اند ، یکی از روش‌های مهم در این مقوله ، استفاده از شاخص‌های فضایی است. نظریه اصلی شاخص فضایی که مهمترین نظریه اساسی پرس و جو می‌باشد ، محاسبه نظریه‌ی نزدیکی است. معروف‌ترین ساختار شاخص ، درخت R می‌باشد. استراتژی اصلی درخت R جمع آوری نودهای فضایی چند بعدی با مستطیل حداقل محدوده (MBR) می‌باشد که کوچکترین نود فضایی داخلی مستطیل باشد. بعد از شاخص بندی ، بهینه سازی بازیابی اطلاعات در پایگاه داده‌ی فضایی مهم است[۲]. از اینرو جهت بررسی بیشتر موضوع ، در این مقاله ، شاخص فضایی جدید که متعلق به خانواده R^+tree می‌باشد و مفهوم همپوشانی گره ها را حفظ می‌کند و R^{++}tree نام دارد ، بررسی شده است. نتایج این مقاله نشان می‌دهد که R^{++}tree در دامنه ی پرس و جو ، پرس و جو KNN و پرس و جو Top-k بسیار کارآمدتر از R^*tree می‌باشد.

کلمات کلیدی:
شاخص فضایی، هیستوگرام های ساخت درخت های **R** ، درخت های **R**

۱. مقدمه

هیستوگرام ها ، ساختارهای مهمی‌هستند که در ابتدا در سیستم‌های پایگاه داده به منظور تخمین انتخابی پرس و جوها مورد استفاده قرار گرفتند. آنها برای کسب پاسخ تقریبی سریع برای پرس و جوهای متراکم مهم استفاده می‌شوند. بیشتر آنها متدهای مبتنی بر شبکه ای هستند که تنها برای دیتا پوینت دو بعدی قابل اجرا هستند. مشکل طراحی بهینه هیستوگرام‌های چندبعدی به NP-hard مشهور هستند[۳].

وضعیت فعلی بیانگر این است که حین اجرای یک پرس و جو در نظام مدیریت یک پایگاه داده ی فضایی (SDBMS) ، بهینه ساز پرس و جو تمام طرح‌های موجود در ارزیابی پرس و جو را به وجود می‌آورد. تمام این طرح ها در بازده نهایی مشابه اند اما در هزینه اجرا و زمان اجرا با یکدیگر فرق می‌کنند[۲].

فلسفه تخصیص باکت‌های هیستوگرام اختصاص آنها به زیرفضاهایی که بدرستی خوشه‌های اشیا را تشخیص داده اند می‌باشد. بنابراین ، ابتدا یک روش برای پیدا کردن مرکز خوشه‌های اشیائ پیشنهاد می‌شود. سپس، یک الگوریتم برای ساخت باکت‌های هیستوگرام از این مراکز پیشنهاد می گردد. این باکت ها از مراکز خوشه ها ، مقداردهی اولیه می‌شوند ، سپس برای پوشش خوشه ها گسترش می‌یابند. بهترین طرح توسعه براساس مفهوم افزایش چولگی انتخاب شده است[۶].

۲. تاریخچه

یکی از اولین متدها که برای دیتای چندبعدی پیشنهاد شد htree می‌باشد. این متد پارتیشن بندی ناهمپوشی فضای چند بعدی را بر اساس فرکانس به عنوان پارامتر منشا ایجاد می‌کند[۲] نویسندگان دو استراتژی را پیشنهاد کرده اند:

۱) متغیر بنیادی بصورت زیر عمل می‌کند: در فاز اول، الگوریتم شبکه معینی را محاسبه می‌کند و تعداد اشیای فضایی زیر گروه را برای هرخانه تعریف می‌کند. براساس شبکه محاسبه شده تقسیم بندی ساخت دوتایی بازگشتی (BSP) برای محاسبه هیستوگرام استفاده می‌شود.

۲) برای کاهش تاثیر استراتژی دوم ساختاری یعنی Minskew-Progressive-Refinement شبکه هایی با تفکیک‌های مختلف مورد استفاده قرار می‌گیرند. هر تفکیک شبکه ای برای ساخت بخش‌های یکسان منحنی‌های پیوند هیستوگرام استفاده می‌شود[۲].

MinSkew یک روش شناخته شده ساخت هیستوگرام برای داده‌های فضایی است.درابتدا ، MinSkew مجموعه داده‌های اصلی را با استفاده از یک شبکه یکنواخت تخمین میزند. سپس، با یک باکت تک متشکل از تمام اشیاء داده شروع می‌شود ، برای هر باکت ، انحراف مکانی آن و نقطه تقسیم در طول ابعادش که حداکثر کاهش در انحراف مکانی را تولید خواهد کرد ، محاسبه می کند و بعد ، MinSkew باکت هایی که تقسیم آنها منجر به بزرگترین کاهش در چوله مکانی می‌شود ، را بر می‌دارد. این باکت را به دو باکت فرزند تقسیم می‌کند ، و داده را ازباکت قدیمی به باکت جدید اختصاص می‌دهد. پس از اتمام MinSkew ، هیستوگرام ساخته شده مجموعه ای از باکت غیر متداخل است [۶].

بسادگی می‌توان از الگوریتم ساخت باکت Bichromatic برای دریافت یک نسخه جدید از هر باکت ساخته شده استفاده کرد. مجموعه باکت‌های Bichromatic سپس به عنوان هیستوگرام نهایی گزارش می‌شود. این استراتژی ، به عنوان مثال ، تبدیل هر باکت به Bichromatic پس از اتمام فرآیند ساخت هیستوگرام ، می‌تواند برای هر روشی که در آن باکت ساخته شده با یکدیگر همپوشانی ندارند ، بکار گرفته شود[۶].

STHist یک روش ساخت هیستوگرام برای داده‌های جغرافیایی گرافیکی دو یا سه بعدی است. با توجه به مجموعه داده همراه با فضای داده ها ، STHist ابتدا تمام فضای داده ها را به تعدادی از بخش‌های داده پارتیشن بندی می‌کند. سپس ، برای هر بخش ، STHist به صورت بازگشتی نقاط متراکم که به باکت هیستوگرام تبدیل شده است را تشخیص می‌دهد. در اینجا ، نقاط متراکم یک منطقه داده است که شرایط خاصی در فرکانس جسم ، شکل و اندازه را برآورده می‌کند. همه باکت‌ها دریک بخش داده که یک درخت باکت سازمان یافته است شناسایی شدند. با توجه به روش ساخت باکت Bichromatic ، ما می‌توانیم از الگوریتم ساخت باکت Bichromatic برای تبدیل هرباکت ساخته شده توسط STHist در طول فرآیند ساخت هیستوگرام به نسخه Bichromatic استفاده کنیم. به عبارت دیگر ، برای هر بخش داده ها ، پس از اینکه باکت ریشه ایجاد می‌شود ، الگوریتم ساخت باکت Bichromatic برای این باکت ریشه استفاده می‌شود. سپس ، STHist نقاط داخل باکت ریشه بهبود یافته را تشخیص می‌دهد و این نقاط را به باکت فرزند تبدیل می‌کند. برای هر یک از این باکت‌های فرزند ، ما الگوریتم ساخت باکت Bichromatic را برای دریافت نسخه بهبود یافته اعمال می‌کنیم. این روند تا زمانی که هیچ باکت جدیدی ساخته نشده باشد ادامه می‌یابد[۶].

STHist-c روش بهبود یافته STHist می‌باشد که همان چهارچوب STHist را بکار می‌برد. روش STHist شرایط سفت و سخت را بر روی شکل و اندازه برای تشخیص مناطق تراکم بکار می‌برد. این دلیل عمدتا برای یک راه مرتبط آسان برای یافتن مناطق تراکم می‌باشد. بهرحال بخاطر این شرایط سفت و سخت ، روش STHist گاهی اوقات موفق به تشخیص

خوشه‌های اشیائ درمجموعه داده ها نمی‌شود. برای ساخت نمودار هیستوگرام دقیق ، STHist-c باکت ها را در مکان خوشه‌های داده اختصاص می‌دهد. ابتدا یک الگوریتم خوشه بندی و دو روش آماری ، برای پیدا کردن تعداد و مکان خوشه‌های اشیا با هم ترکیب می‌شوند. سپس یک الگوریتم جدید برای ساخت باکت ها از مرکز این خوشه ها ارائه می‌شود. مناطق باکت ها به تدریج از مرکز خوشه به تمام جهات گسترش می‌یابند. در زمان گسترش یک باکت ، از میان بسیاری از گزینه‌های احتمالی گسترش ، بهترین روش بر اساس مفهوم جدیدی از افزایش چولگی انتخاب می‌شود[۶].

۳. آزمایشات

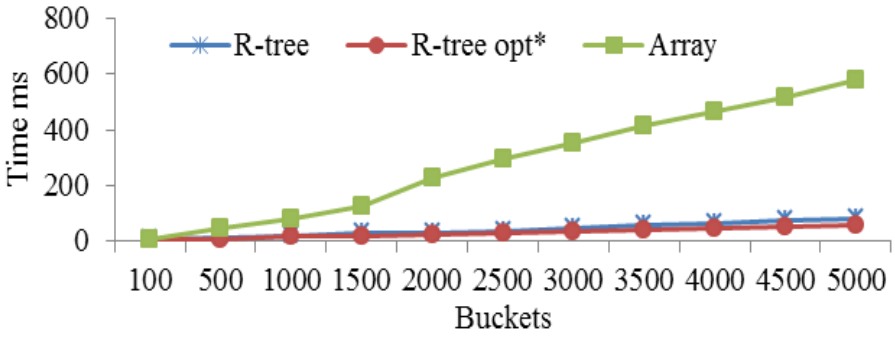
در این مقاله ، اجرای هیستوگرام‌های مختلف را مورد بررسی و مقایسه قرار گرفته شده است. از MinSkew بعنوان متد مرجع استفاده شده که بسیار خوب عمل می‌کند ، دومی را با عنوان MinSkewProg معرفی شده است. برای d=۲ شبکه ای با خانه‌های $۲^۴$ مورد استفاده قرار داده شد. برای d=۳ خانه های $۲^{۱۵}$ برای MinSkew و چهار شبکه با خانه های $۲^۶$ ، $۲^۹$ ، $۲^{۱۲}$ ، $۲^{۱۵}$ برای MinSkewProg مورد استفاده قرار داده شده است. دیگر متد ها که مورد آزمایش قرار گرفتند در جدول شماره ۱ لیست شده اند[۲].

جدول ۱: (بررسی تابع هزینه متدهای مطالعه شده)[۲]

تابع هزینه	توضیحات
C_V	حجم MBR
C_{QP}	گسترش C_V با میانگین ، طولهای طرف پرس و جو
C_{RK}	معیار k-Uniformity
C_{SK}	Spatial skew of MBR
هیستوگرام	**توضیحات**
MinSkew	MinSkew ، شبکه ثابت
MinSkewProg	MinSkew,Prog ، پالایش
rkHist	rkHist withα=۰. ۱
Rtree	پارتیشن بندی سایز ثابت ، منحنی Hilbert
R-V	C_V منحنی Hilbert
R-VQP	C_{QP} منحنی Hilbert
R-RK	C_{RK} منحنی Hilbert
R-SK	C_{SK} منحنی Hilbert
FST	جنگل STHist

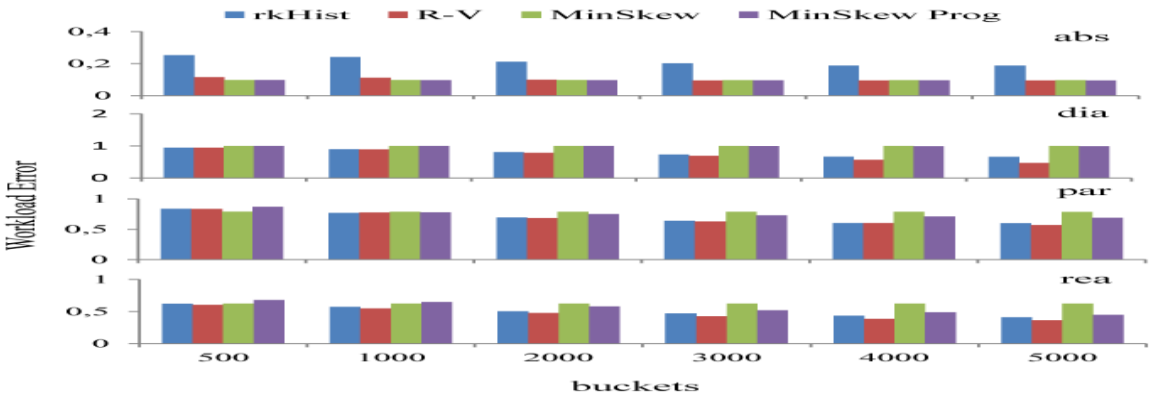
۴. زمان ساخت و تخمین

برای کاهش زمان تقریبی هیستوگرام می‌تواند بعنوان حافظه‌ی اصلی R-tree ارائه شود. شکل ۴ عملکرد اندازه‌ی منحنی‌های پیوندی و زمان مجموع حجم کار برای ارائه هیستوگرام را نشان می‌دهد ، دوتای اول R-tree هستند. سومی آرایه‌ای برای منحنی‌های پیوند است. برای R-tree تنظیمات حافظه‌ی اصلی مورد استفاده قرار داده شده و گنجایش خروجی به ۱۲ مدخل برای هر گروه تنظیم شده است (دوباره ذکر شده بهترین مجموعه در آزمایشات بود). علاوه بر آن یک R-tree با استفاده از منحنی‌های پیوند هیستوگرام با یک متد پارتیشن بندی otp و C_V بعنوان تابع هزینه‌ای درست شده است[۲].



شکل۱، [۲]

شکل۲ نتایج متدهای R-tree را در مقایسه با استراتژی پارتیشن بندی اندازه ثابت برای دیتاپونیت d=۲ را نشان می‌دهد. مشاهده شد که متدهایی که از چهارچوب پارتیشن بندی بهینه بر اساس نوع استفاده می‌کنند دقت بهتری را نسبت R-tree از خود نشان می‌دهند.



شکل۲، [۲]

یک راه حل ممکن پارتیشن بندی توزیع دیتایی بر اساس اندازه و شکل هدف و ساخت هیستوگرام یا شاخص‌گذاری بطور مستقل برای هر طبقه بندی می‌باشد[۲].

۵. تفاوت‌های درخت‌های R و R^+ و R^*

درخت R^+ یک روش برای دنبال کردن داده با استفاده از یک محل اغلب مختصات (x,y) ، و اغلب برای مکان‌های بر روی سطح زمین است. جستجو در یک عدد یک مشکل حل شده است. جستجو در دو یا بیشتر، و درخواست برای مکان هایی که در نزدیکی هر دو مختصات x و y ، نیاز به الگوریتم‌های craftier دارد. اساسا ، درخت R^+ ساختمان داده ی یک نوع درخت R ، مورد استفاده برای اطلاعات مکانی شاخص گذاری است[۵].

۵.۱. تفاوت Rtree و R^+tree :

R^+tree یک مصالحه بین Rtree و kdtree : آن‌ها از تداخل گره‌های داخلی با واردکردن یک شیئ به برگ‌های متعدد در صورت لزوم اجتناب می‌کنند. همپوشانی کل منطقه است که در دو یا چند گره موجود است. همپوشانی حداقل مجموعه ای از مسیرهای جستجو به برگ (حتی بیشتر برای زمان دسترسی از حداقل پوشش بحرانی) را کاهش می‌دهد. جستجوی کارآمد نیاز به حداقل پوشش و همپوشانی دارد[۵].

درخت R^+ و درخت R در موارد زیر با هم تفاوت دارند:

گره ها حداقل نیمی پر تضمین نمی‌شود. ورودی‌های هر یک از گره‌های داخلی همپوشانی ندارند. یک شناسه شیئ می‌تواند در بیش از یک گره برگ ذخیره شود[۵].

مزایا:

از آنجا که گره ها با یکدیگر همپوشانی ندارند. مزایای کارایی پرس و جوی نقطه از مناطق فضایی توسط حداکثر یک گره پوشیده شده است. یک مسیر واحد دنبال می‌شود و گره‌های کمتری نسبت به Rtree مشاهده می‌شوند[۵].

معایب:

از آنجا که مستطیل ها تکرار می‌شوند ، یک درخت R^+tree می‌تواند بزرگتر از درخت Rtree بر روی مجموعه داده یکسان ساخته شود. ساخت و نگهداری R^+tree پیچیده تر از ساخت و ساز و تعمیر و نگهداری دیگر انواع درخت R است[۵].

درخت *R یک نوع از درختان R هستند که برای شاخص گذاری اطلاعات مکانی استفاده می‌شوند. هزینه ساخت و ساز R^*tree کمی بالاتر از استاندارد درختان R است. از آنجا که داده ها ممکن است نیاز به جاگذاری دوباره داشته باشند ، اما درخت نتیجه معمولا عملکرد پرس و جو بهتر خواهد داشت. مانند درخت استاندارد R ، می‌تواند در هر دو نقطه و داده‌های مکانی ذخیره شود[۵].

۲.۵. تفاوت بین Rtree و R^*tree :

به حداقل رساندن همپوشانی و پوشش هر دو به کارایی درخت R بستگی دارد. همپوشانی بدان معنی است که ، در پرس و جوی داده ها و یا درج ، بیش از یک شاخه از درخت نیاز به گسترش دارد. پوشش حداقلی باعث بهبود عملکرد هرس می‌شود ، در اغلب موارد به مستثنی کردن کل صفحات از جستجو ، به ویژه برای پرس و جوهای دامنه منفی ، اجازه می‌دهد.

درخت *R تلاش می‌کند برای کاهش هر دو; با استفاده از ترکیبی از یک الگوریتم تقسیم گره تجدید نظر و مفهوم درج مجدد اجباری از سرریز گره جلوگیری کند. این براساس مشاهدات ساختار درخت R که بسیار حساس به ترتیبی که در ورودی ها درج شده است ، می‌باشد. بنابراین ساختار ساخت درج به احتمال زیاد زیر بهینه است. حذف و درج مجدد ورودی ها به پیداکردن مکانی در درخت‌ها که مناسبتر از مکان اصلی خودشان می‌باشد ، اجازه می‌دهد[۵]. در جدول ۲ ، مقایسه بین درخت R و *R و R^+ نشان داده شده است.

جدول ۲، مقایسه بین درخت R و درخت‌های*R و R^+

مقایسه	R
R^+	در R^+ گره‌های کمتری نسبت به R مشاهده می‌شود.
R^*	ساختار ساخت درج به احتمال زیاد زیر بهینه است.

۶. نتایج

آزمایشات انجام شده نشان می‌دهد که STHist نسبت به متدهای پیشنهادی دیگر برتر است. با این وجود STHist پیچیدگی زمانی $O(n^2)$ برای داده ی ۲بعدی و $O(n^3)$ برای داده ی ۳ بعدی را دارد. روش دیگر برای دستیابی به هیستوگرام فضایی تولید آن است که از ساختار شاخص گذاری فضایی مثل Rtree استفاده می‌کند.

دراین مقاله R^{++}tree با R^+tree و R^*tree بعنوان شاخص متعارف از خانواده Rtree ها مورد مقایسه قرار داده شد. R^{++}tree بعنوان نسخه اصلاحی R^+tree ارائه گردید. نتیجه حاصل نشان داد که R^{++}tree در دامنه ی پرس و جو ، پرس و جو KNN و پرس و جو Top-k بسیار کارآمدتر از R^*tree می‌باشد. کارآمدی R^{++}tree با رشد ابعاد به آرامی‌کاهش می‌یابد ، چون مقادیر تکرار هم افزایش می‌یابد. در انتها نشان داده شد که راندمان زمان جستجوی R^{++}tree بسیار بهتر از R^+tree است ، آن هم زمانی که دیتاپوینت تکراری مد نظر قرار می‌گیرد می‌باشد.

۷. منابع

[۱] Combining R-Tree and B-Tree to Enhance Spatial Queries Processing Marwa A. M. Abd Elwahab ۱, Khaled M. Mahar ۲, HatemAbdelkader ۳, HatemAwad Khater ٤

[۲] New Database Architecture for Smart Query Handler of Spatial Database ParthasarathiBoyala, RituparnaChaki b aWest Bengal University of Technology, BF ٤۲ SaltLake City, Kolkata-۷۰۰۰٦٤, India bWest Bengal University of Technology, BF ٤۲ SaltLake City, Kolkata-۷۰۰۰٦٤, India

[۳] A class of R-tree histograms for spatial databases Technical Report DaniarAchakeev Department of Mathematics and Computer Science Philipps-Universität Marburg, Germany achakeye@mathematik. uni-marburg. de Bernhard Seeger Department of Mathematics and Computer Science Philipps-Universität Marburg, Germany seeger@mathematik. uni-marburg. de

[٤] International Conference on Computational Science, ICCS ۲۰۱۲ Workshop on using Emerging Parallel Architectures (WEPA ۲۰۱۲) Speeding up spatial database query execution using GPUs Bogdan Simion ۱,∗, Suprio Ray∗, Angela Demke Brown∗Department of Computer Science, University of Toronto

[٥]R++-tree: an efficient spatial access method for highly redundant point data Martin Šumák, Peter Gurský P. J. Šafárik University in Košice, Jesenná ٥, ۰٤۰۰۱ Košice, Slovakia martin. sumak@student. upjs. sk, peter. gursky@upjs. sk

[٦] Improving the Accuracy of Histograms for Geographic Data Objects ,HaiThanh Mai, Jaeho Kim, and Myoung Ho Kim
Department of Computer Science, KAIST
۲۹۱ Daehak - r o, Yu seong- Gu, D aeje on ۳۰٥- ۷۰۱, R epublic of Korea

[۷] {mhthanh,jaeho,mhkim }@dbserver.kaist. ac. k r [۷]Y. Lifang; L. Rui; H. Xianglin; L. Yueping, "Performance of Rtree with slim-down & Reinsertion Algorithm," in ProcInternational Conference on Signal Acquisition and Processing, pages: ۲۹۱-۲۹٤, ۲۰۱۰.

[۸] Z. Shaohui, C. Zhanwei , "The research of Hilbert R-tree spatial index algorithm based on Hybrid clustering," in Proc International Conference on Electronic & Mechanical Engineering and Information Technology, pages: ۳٤۹٥-۳٤۹۷, ۲۰۱۱.

[۹] Kao, B. Lee, S. Lee, F. Cheung, D. " Clustering Uncertain Data Using Voronoi Diagrams and R-Tree Index,"IEEE Transactions on Knowledge and data engineering, Vol. ۲۲ No. ۹, ۲۰۱۰.

[۱۰] Shengnan, K. " Integrating R-tree and Levels of Detail,"Eighth International Conference on Fuzzy Systems and KnowledgeDiscovery (FSKD), ۲۰۱۱.

[۱۱] Yu,B. Kim,H. Choi,W. Kwon,D. , " Parallel Range Query Processing on R-Tree with Graphics Processing Unit," Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, ۲۰۱۱