بررسی استفاده از خوشه‌بندی جهت کاهش زمان پرس و جوهای تجمیع رستری داخل پایگاه داده مکانی

مطالعه موردی: رسترهای بارش

جواد سدیدی [۱] سعیده صاحبی وایقان [۲]

هانی رضائیان [۳]

تاریخ دریافت مقاله: ۹۵/۱۱/۲۰ تاریخ پذیرش مقاله: ۹۶/۰۳/۲۳

****

چکیده

در سال‌های اخیر با پیشرفت فن‌آوری‌های جمع‌آوری و مدیریت‌داده، پایگاه‌داده‌های بسیار بزرگ پدیدار شده‌اند. بسیاری از پرس‌وجوهای تجزیه‌وتحلیل بر اساس ماهیتشان به تجمیع و خلاصه‌سازی بخش‌های بزرگی از داده‌های در حال تجزیه‌وتحلیل نیاز دارند. مسئله اصلی در حیطه‌ی پایگاه داده پردازش کارآمد پرس‌وجو مخصوصاً در سیستم‌های لحظه‌ای [۴] است که نیازمند رسیدن به جواب آنی می‌باشد تا اینکه کاربر زمان زیادی را برای دریافت پاسخ صرف نکند. AQP (Approximate Query Processing) به‌عنوان روشی جایگزین برای پردازش پرس‌وجو در محیط‌هایی که ارائه یک پاسخ دقیق زمان‌بر است، با هدف ارائه پاسخ تخمینی، کاهش زمان پاسخ را با حذف یا کاهش تعداد دسترسی‌ها به داده‌ی پایه میسر می‌سازد. پردازش In-Database [۵] عملکرد شبکه‌های کامپیوتری را بهبود بخشیده و به طراحی مناسب پرس‌وجوها با نتایج نسبتاً سریع و دقیق کمک می‌کند. در این پژوهش عملیات تجمیع (Sum) در پایگاه داده PostgreSQL روی داده‌های رستری بارش به دو روش معمولی و بهینه پیشنهاد شده، انجام شده است. بررسی نتایج نشان می‌دهد که سرعت اجرای تابع Sum با خوشه‌بندی، ۲۷/۲ برابر اجرای این تابع بدون خوشه‌بندی است و میانگین اختلاف عددی پیکسل‌های حاصل از اجرای تابع Sum بهینه با اجرای تابع معمولی آن ۰/۰۲۸ می‌باشد. میانگین زمان اجرای پرس‌وجوهای معمولی و بهینه برای تابع Sum به ترتیب ۲۱۱ و ۷/۷۵٤ ثانیه می‌باشد که نشانگر کارآمد بودن روش پیشنهاد شده در این تحقیق می‌باشد. نتایج تحقیق حاضر که در حقیقت کاهش معنی دار زمان پاسخ آنالیزهای داخل پایگاه داده‌ای در داده‌های رستری می‌باشد، می‌تواند در ارائه سرویس‌های رئال تایم تحت وب مانند هواشناسی، ترافیک و ... که نیازمند تحلیل‌های آنی و جواب لحظه‌ای می‌باشند مورد استفاده قرار گیرد.

واژه‌های کلیدی: بهینه‌سازی تجمیع، پردازش تقریبی پرس و جو، پردازش In-Database، آنالیز رستری، Sum.

****

۱- استادیار گروه سنجش از دور و سیستم اطلاعات جغرافیایی، دانشکده علوم جغرافیایی، دانشگاه خوارزمی، تهران ، ایران (نویسنده مسئول). jsadidi@gmail.com

۲- کارشناس ارشد سنجش از دور و سیستم اطلاعات جغرافیایی، دانشگاه خوارزمی، تهران، ایران. saiedehsahebi@yahoo.com

۳- استادیار گروه سنجش از دور و سیستم اطلاعات جغرافیایی، دانشکده علوم جغرافیایی، دانشگاه خوارزمی، تهران ، ایران. hani.rezayan@gmail.com

4- Real time

۵- درون پایگاه‌داده

۱- مقدمه

رسترها در مقایسه با داده‌های مکانی معمولی مانند اعداد و رشته‌ها دارای پیچیدگی و ویژگی‌های خاصی هستند و ماهیتاً می‌توان آنها را «Big Data» نامید. پردازش داده‌های رستری و تصاویر مکانی فرآیندی به طور ویژه سنگین و زمان بر است.

سیستم مدیریت پایگاه داده‌های مکانی (SDBMS)[1] زمینه‌ی فناوری پایگاه داده را برای سیستم‌های اطلاعات جغرافیایی (GIS) و دیگر برنامه‌ها فراهم می‌کند و یک پایگاه داده معمولی تصویر مکانی دارای ده‌ها و یا صدها ترابایت یا حتی پتابایت داده هستند. با توجه به اینکه پردازش‌ها و تجزیه و تحلیل داده حجیم و زمان بر است، محل داده فاکتور مهمی در پردازش‌ها می‌باشد.

پردازش‌های In-Database ، پردازشگر را به داده نزدیک می‌کند به جای اینکه داده به پردازشگر نزدیک شود. این امر باعث بهبود عملکرد با غلبه بر محدودیت شبکه‌های کامپیوتری می‌شود(*Xie,Q., et al., 2013:2*).

با وجود پیشرفت‌های اخیر فنآوری‌های جمع‌آوری و مدیریت داده، پایگاه داده‌های بسیار بزرگ پدیدار شده‌اند. درحالی‌که جمع‌آوری و ذخیره‌سازی مجموعه عظیم داده‌ها نسبتاً ساده شده است، رسیدن به آنالیزهای کارآمد بسیار مشکل می‌باشد. یک دلیل بر این امر این است که بسیاری از پرس‌وجوهای تجزیه وتحلیل بر اساس ماهیتشان به تجمیع[2]و خلاصه‌سازی بخش‌های بزرگی از داده‌های در حال تجزیه‌وتحلیل نیاز دارند (*Babcock, B., et al, 2003:539*). عوامل زیادی وجود دارند که انگیزه برای مطالعه در حیطه پردازش تقریبی پرس‌وجو و توسعه آن را تقویت می‌کنند. یکی از این عوامل افزایش فزاینده تعداد اپلیکیشن‌ها، فن‌آوری‌ها، کاربران و در نتیجه حجم داده‌ها است. افزایش حجم مجموعه داده‌ها و پیچیدگی آن‌ها منجر به کند شدن و پرهزینه شدن نسبی پرس‌وجوها گشته است. عامل مهم دیگر تحلیل داده در پایگاه داده‌های بزرگ با استفاده از داده‌کاوی و سیستم‌های پشتیبانی تصمیم‌گیری است. این روش‌ها اساساً از پرس‌وجوهای تجمیع برای تجزیه‌وتحلیل داده‌ها استفاده می‌کنند. همچنین آن‌ها از فرایندهای تکراری استفاده می‌کنند و بسیاری از پرس‌وجوها را اجرا می‌کنند تا تمرکز تنها روی داده‌های دلخواه صورت گیرد. هزینه این پرس‌وجوها بسیار بالاست و به منابع گسترده‌ای نیاز دارند. پردازش تقریبی پرس‌وجو روشی بسیار کارآمد برای تشخیص داده‌های دلخواه و به حداقل رساندن پرس‌وجوهای تجمیع است(*Babcock, B., et al, 2003:539*).

مسئله اصلی در حیطه‌ی پایگاه داده پردازش کارآمد پرس‌وجو است تا اینکه کاربر زمان زیادی را برای دریافت پاسخ صرف نکند. با این‌حال در بسیاری از موارد رفع این نیاز به‌سادگی صورت نمی‌پذیرد.

پاسخ پرس‌وجو به‌صورت سنتی روی ارائه پاسخ‌های دقیق به پرس‌وجو تمرکز دارد. در بسیاری از موارد جواب تقریبی سریع به‌جای پاسخ دقیق زمان‌بر، کافی است. به‌عنوان مثال مجموع، میانگین یا درصد برای زمانی که تنها چند رقم اول از دقت موردنظر است(*Acharya et al., 1999:275*).

AQP[3] به‌عنوان روشی جایگزین برای پردازش پرس‌وجو در محیط‌هایی که ارائه‌ی یک پاسخ دقیق زمان‌بر است، پدید آمده است(*Azevedo et al.,2007:70*).

محیط‌هایی که یک پاسخ دقیق ارائه می‌کنند، زمان پاسخ نامطلوب طلب می‌کنند و همین امر انگیزه برای مطالعه‌ی تکنیک‌هایی جهت ارائه‌ی پاسخ تقریبی برای پرس‌وجو را ایجاد کرده است.

هدف از ارائه پاسخ تخمینی کاهش زیاد زمان نسبت به محاسبه پاسخ دقیق با حذف یا کاهش تعداد دسترسی‌ها به داده‌ی پایه است(*Acharya et al., 1999:70*). یکی از مهم‌ترین کاربردهای این تکنیک‌ها بهینه‌سازی پرس‌وجو است (*Ioannidis and Poosala, 1995:234*) (*Papadias et al., 2001:443*). گاهی اوقات پردازش تقریبی پرس‌وجو می‌تواند برای داشتن پیش‌نمایشی از پرس‌وجو برای گرفتن اطلاعات

1- Someplace Data Base Management System
2- Aggregation
3- Approximate Query Processing

بیشتر درباره داده قبل از اجرای پرس‌وجو مفید باشد. این امر به طراحی پرس‌وجوی مناسب با ارائه نتایج نسبتاً دقیق و سریع کمک می‌کند. در نهایت استفاده از پردازش تقریبی پرس‌وجو روی سیناپس‌های داده‌های ذخیره‌شده می‌تواند راه‌حلی عالی برای مقابله با محدودیت‌های شبکه‌ای و ساختاری باشد(Liu, 2009:113) (Garofalakis, 2001:725).

سیستم اطلاعات جغرافیایی با دو دسته از داده‌های مکانی و غیر مکانی سر و کار دارد. در یک پایگاه داده مکانی تعداد اشیا به آسانی به میلیون می‌رسد. ازآنجایی‌که بهره‌گیری از روش AQP جهت پردازش پرس‌وجوهای مکانی که دارای داده‌های پیچیده با حجم بالایی هستند، باعث افزایش سرعت پرس‌وجو و کاهش هزینه‌ها می‌شود، حائز اهمیت است. AQP پاسخ‌های تقریبی مفیدی را برای پرس‌وجوهای پایگاه داده فراهم می‌کند.

از دهه هشتاد قرن گذشته استفاده از روش تقریب در پرس‌وجوهای پشتیبانی تصمیم‌گیری آغاز شده است و تحقیقاتی در ده سال اخیر در زمینه مشکلات AQP مورد توجه بوده است(Chakrabarti et al, 2001:199). از آن جمله می‌توان به موارد زیر اشار نمود:

Perera و همکاران (۲۰۱۵) روشی جدید برای پرسشگری روی داده‌های سری زمانی با استفاده از تبدیلات فوریه، کسینوسی و تجمیع و قطعه مبنا[1] ارائه دادند. رویکرد آن‌ها بار ارتباطات را با تبادل مدل به‌جای داده‌های اصلی کاهش می‌دهد. نتایج حاصل از این پژوهش نشان داد که پرسشگری روی مدل‌های ذخیره‌شده به‌جای داده‌های اصلی اجرا شده و زمان پاسخ به پرس‌وجو نسبت به اجرای آن روی داده‌های اصلی تا ۸۰ درصد کاهش می‌یابد.

Xie و همکاران (۲۰۱۳) به توصیف استراتژی پیاده‌سازی موتور Oracle Spatial GeoRaster برای پردازش تصویر در درون پایگاه داده و مزایای عملکرد آن پرداختند و به این نتیجه رسیدند که در ابتدا نه تنها پایگاه داده را با قابلیت‌های پرس‌وجوهای پیشرفته مانند پرس‌وجوهای تجزیه وتحلیل و پرس‌وجوهای تجمیع تصویر یا موزاییک کردن آن‌ها بهبود می‌بخشد، بلکه پردازش حجم عظیمی از تصاویر را در داخل پایگاه‌داده میسر می‌سازد و به طور قابل توجهی مدیریت و دستکاری داده GeoRaster را افزایش می‌دهد. همچنین سه قابلیت اصلی برای عملکرد بالاتر توصیف کردند. قابلیت اول این است که به‌جای اینکه تصویر جهت پردازش به خارج از پایگاه داده انتقال یابد، پردازش تصویر را به تصویر نزدیک می‌سازد ویژگی دوم، پردازش موازی است که باعث بهبود عملکرد می‌شود.

ویژگی سوم، پردازش هم زمان است. نتایج قابلیت و عملکرد نشان داد که این موتور پردازش تصویر داخل پایگاه داده نه تنها به طور چشمگیری پرس‌وجوهای مکانی و توانایی پردازش پایگاه داده‌های تصاویر بزرگ مقیاس را بهبود می‌بخشد، بلکه به طور مؤثر برخی از چالش‌های عملکرد را حل می‌کند.

Azevedo و همکاران (۲۰۱۰) به بررسی پردازش تقریبی پرس‌وجو به‌عنوان یک رویکرد جایگزین برای ارائه‌ی پاسخ به کاربر در زمان کوتاه‌تر با هدف فراهم نمودن یک نتیجه‌ی تقریبی در زمان کوتاه‌تر نسبت به پاسخ دقیق پرداختند. همچنین الگوریتم جدیدی برای مجموعه عملیات مکانی پیشنهاد نمودند که می‌تواند پردازش تقریبی با استفاده از 4CRS[2] انجام دهد. Thomas Brinkhoff (۲۰۰۸) با توجه به پیچیدگی زیاد اشیاء و پرس‌وجوها و همچنین نظر به حجم بسیار بزرگ داده‌ها، به بررسی ذخیره‌سازی و معماری دسترسی در سیستم‌های پایگاه داده مکانی پرداخت. نتایج حاصل از عملکرد تجربی این تحقیق گویای بهبود قابل‌توجهی از عملکرد پرس‌وجوها برای منطقه بزرگ شده است. Bernardino و همکاران (۲۰۰۲) به بررسی مشکل اصلی روش‌های DWS[3] که عبارت است از اینکه به مجموعه‌ای بسیار گران‌قیمت از کامپیوترهای با قابلیت تحمل خطا برای پیشگیری از نقص در یک کامپیوتر برای متوقف کردن کل سیستم، نیاز است، پرداختند. روش پیشنهاد

1- Piecewise

2- Four-Color Raster Signature

3- Data Warehouse Striping

شده در این تحقیق مورد تکنیک‌های بر پایه پرس‌وجوهای تقریبی است که ارائه پاسخ تقریبی به کاربران را حتی زمانی که یک یا چند کامپیوتر در مجموعه در دسترس نباشد، امکان‌پذیر می‌سازد. نتایج حاصل از ارزیابی ارائه‌شده در هر دو روش تحلیلی و تجربی در این مقاله نشان داد که نتایج به‌دست‌آمده از روش تقریبی دارای خطای بسیار کوچک است که می‌تواند در بسیاری از موارد قابل‌اغماض باشد.

Chakrabarti و همکاران (۲۰۰۱) با استفاده از تکنیک موجک‌های چندبعدی پردازش تقریبی پرس‌وجو در اپلیکیشن‌های چندبعدی مدرن را پیشنهاد دادند که به‌طور مستقیم روی سیناپس‌های ضرایب موجک از جداول رابطه‌ای عمل می‌کنند و پردازش دلخواه پرس‌وجوهای پیچیده را به‌طور کامل در دامنه‌ی ضرایب موجک میسر می‌سازد و این امر باعث کاهش زمان موردنیاز برای پاسخ به پرس‌وجو می‌شود. همچنین یک الگوریتم جدید برای تجزیه‌ی موجک پیشنهاد دادند که می‌تواند این سیناپس‌ها را در یک شیوه‌ی کارآمد ورودی – خروجی[1] (I-O) بسازد.

یکی از روش‌های تجمیع داده‌ها تجمیع Sum می‌باشد. تجمیع رسترها به روش Sum مستلزم صرف زمان زیادی است. از اینرو اعمال تقریب در انجام تجمیع Sum به کاهش زمان این عملیات می‌انجامد. در این پژوهش، یک روش جدید با خوشه‌بندی رسترها با استفاده از تابع مشابهت برای بهینه‌سازی عملیات تجمیع گریدهای رستری درون پایگاه داده‌ها ارائه‌شده است. در این رویکرد امکان تنظیم دقت و صحت نتایج با قابلیت تغییر پارامترهای دخیل در تابع مشابهت ممکن می‌باشد. هدف از انجام این پژوهش کاهش زمان پردازش پرس‌وجو در پایگاه داده‌های مکانی و بهبود پردازش لحظه‌ای In-Database است.

۲- دیدگاه ها و مبانی نظری

AQP بهترین روش برای سناریوهای تجزیه‌وتحلیل داده است که در مواقعی که دقت، عامل بسیار مهمی محسوب نمی‌شود راه‌حلی با هزینه کارآمد می‌باشد(*Yassin et al, 2015:11*).

AQP یک روش با هزینه کارآمد برای مدیریت حجم عظیمی از داده‌ها توسط پرس‌وجوهای پاسخ توسط سیناپس‌های داده است. ایده اصلی پشت AQP کاهش دقت به‌منظور بهبود زمان پاسخ است(*Liu, 2009:113*).

تجمیع به مفهوم جمع‌کردن رکوردهای مختلف در یک رکورد توسط گروه‌بندی (به‌وسیله مثلاً ارزش‌ها) و یا استفاده از یک تابعی که مجموعه‌ای از ارزش‌ها را می‌گیرد و به یک ارزش بازمی‌گرداند، گفته می‌شود. به توابعی که مجموعه‌ای از ارزش‌ها را می‌گیرد و به یک ارزش برمی‌گرداند توابع تجمیع گفته می‌شود. در یک پایگاه داده رابطه‌ای استاندارد بسیاری از توابع تجمیع رایج مورداستفاده عبارتنداز:
Avg, COUNT, Sum, Min, Max (Obe, R. O., & Hsu, L. S., 2011:227).

توابع تجمیع تنها یک مقدار را بر اساس داده‌های یک ستون برمی‌گردانند.

۲-۱- تجمیع در پایگاه داده

در پایگاه‌داده‌ها که روی تعامل با کاربر تمرکز دارد، ارائه نتیجه پرس و جو در زمان مناسب بسیار ضروری است. در حال حاضر، عملکرد پایگاه‌داده‌ها برای تجمیع داده‌های گرید مکانی مناسب نیست(*Kang, et al, 2014:2*). نتیجه حاصل از تجمیع p گرید رستر با اندازه m*n توسط تابع Sum یک گرید m*n است. همه رسترها دقیقاً همان ناحیه مکانی را پوشش می‌دهند و دارای اندازه یکسان می‌باشند. روش تجمیع ساده زیر می‌تواند جهت محاسبه Sum برای هر سلول (x, y) استفاده شود(*Kang, et al, 2014:2*):

$$result\_grid(x,y) = \sum_{i=1}^{p} grid_i(x,y)$$ رابطه ۱

این روش مجموع ماتریس‌ها است (هر گرید یک ماتریس است).

نگاره ۱ تجمیع سه گرید کوچک را که با تابع Sum تجمیع شده‌اند، نشان می‌دهد.

1- Input-Output

نگاره ۱: تجمیع رستری

Sum(| ۱ | ۲ | ۳ | / | ۳ | ۴ | ۸ | / | ۷ | ۹ | ۱ |) ; | ۳ | ۲ | ۱ | / | ۵ | ۱ | ۶ | / | ۰ | ۵ | ۱ | ; | ۲ | ۵ | ۲ | / | ۴ | ۳ | ۴ | / | ۸ | ۷ | ۱ |) = | ۶ | ۹ | ۶ | / | ۱۲ | ۹ | ۱۸ | / | ۱۵ | ۲۱ | ۳ |

پایگاه داده‌های معمول برای تجمیع‌های اشیا ساده مناسب می‌باشند. در رابطه با تجمیع گریدهای رستری به علت پیچیدگی و حجیم بودن داده ها، روش‌های سنتی ساده برای تجمیع داده‌های رستری مناسب نمی‌باشند. با فرض اینکه ماهانه یک گرید کوچک ۵۰۰*۵۰۰ از داده‌های بارش تولید شود، اگر بخواهیم مجموعه‌ای از گریدهای تولید شده به مدت ۵ سال را با تابع Sum تجمیع کنیم (برای مثال ۶۰ گرید به اندازه ۵۰۰*۵۰۰)، بایستی ۱۵ میلیون ارزش با یکدیگر جمع شوند (۵۰۰*۵۰۰*۱۲*۵). این حجم از داده مشکلاتی در رابطه با عملکرد را در پی خواهد داشت.

تجمیع سلول به سلول[1] (با استفاده از توابع جمع و میانگین) p گرید رستری با اندازه m*n حداقل به m*n*p عملیات ریاضی (جمع (+)) نیاز دارد. در این حالت تمامی سلول‌ها از تمامی رسترها در عملیات تجمیع مورد محاسبه قرار می‌گیرند. هدف از انجام تقریبی تجمیع این است که تنها برخی از گریدها در محاسبات مورد استفاده قرار گیرند و در این حالت تعداد رسترهای دخیل در محاسبه تجمیع کاهش می‌یابد و به‌موجب آن تعداد عملیات ریاضی جمع نیز کاهش می‌یابد که این امر باعث تقریب تجمیع و کاهش زمان محاسبه آن می‌شود.

۳ - بحث

۳-۱- مواد تحقیق

در پژوهش حاضر از یک سیستم با مشخصات سخت افزاری Processor: Intel(R) dual-core,2.60 GHz, RAM: 6 GB, Windows 7(64 bit) جهت اجرای تحقیق و از PostgreSQL9.5 برای تشکیل پایگاه داده و انجام آنالیزها استفاده شده است. PostgreSQL دارای یک افزونه به نام PostGIS می‌باشد که حاوی یک کتابخانه‌ی متن‌باز[2] است که قابلیت مکانی در سیستم مدیریت پایگاه داده-های شی-رابطه‌ای[3] PostgreSQL را ایجاد می‌کند. PostGIS یک توسعه‌دهنده‌ی پایگاه‌داده مکانی برای سیستم مدیریت پایگاه‌داده PostgreSQL است *(Obe et al,. 2015)*. برای ورود داده ها به داخل پایگاه داده از raster2pgSQL استفاده شده است. Raster2pgSQL قابلیت ورود رسترها به داخل PostGIS فراهم می‌کند و فرمت‌های رستری مورد پشتیبانی GDAL را به صورت فرمت مناسب جهت ورود به جداول رستری PostGIS بارگذاری می‌کند.

داده‌های رستری بارش ایستگاه‌های هواشناسی محدوده استان لرستان به مدت ۵ سال آماری از ۲۰۱۰ تا ۲۰۱٤ استفاده شده و در مرحله بعد با استفاده از روش درون‌یابی کریجینگ برای هر ماه بارش، درونیابی انجام و رسترهای مربوطه تولید شده است. به‌این‌ترتیب ٦۰ تصویر رستری (۵ سال) به‌عنوان ورودی در پایگاه‌داده ایجاد گردیده است.

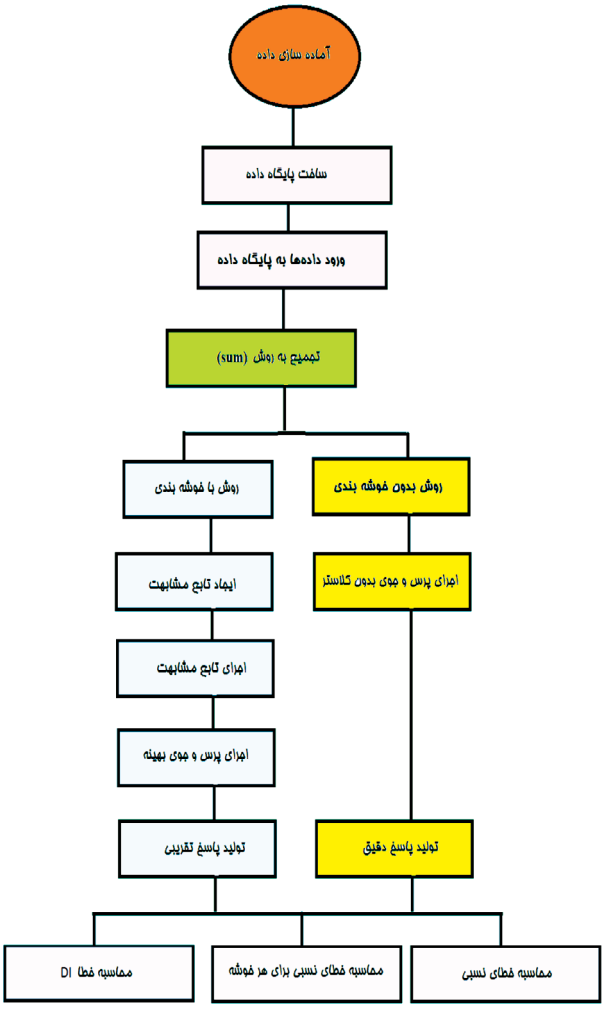
۳-۲- روش انجام تحقیق

ابتدا داده‌های مورد استفاده در فرمت‌های مناسب، جهت ورود به پایگاه داده آماده شده و پس از نصب PostgreSQL9.5 و PostGIS و کتابخانه GDAL، پایگاه داده ایجاد شده است. در ابتدا جداول و ارتباطات بین جداول در درون پایگاه داده تشکیل شده، سپس با بهره گیری از raster2pgSQL رسترها به پایگاه داده وارد شده اند. در ادامه اطلاعات پیکسلی رسترها در جدولی مجزا استخراج و سپس پرس‌وجوهای تجمیع Sum در دو نوع بدون اعمال خوشه‌بندی و اعمال خوشه‌بندی بهینه آماده شده است و عملیات تجمیع Sum به دو روش خوشه‌بندی شده و بدون

1- cell-by-cell

2- FOSS
3- ORDBMS

خوشه‌بندی بر روی داده‌ها انجام گرفته است. مراحل بدون خوشه‌بندی شامل اجرای پرس‌وجو بدون کلاستر و تولید پاسخ دقیق می‌باشد. برای انجام خوشه‌بندی از تابع مشابهت استفاده شده است. توابع مورد استفاده به زبان PL/pgSQL در محیط PostGIS ساخته شده‌اند. مراحل با خوشه‌بندی به ترتیب شامل ایجاد تابع مشابهت، اجرای تابع مشابهت، اجرای پرس‌وجوی بهینه و تولید پاسخ تقریبی است. در پایان جهت ارزیابی صحت هر یک از این متدها با محاسبه خطای نسبی میانگین، خطای DI[1] و خطای نسبی مجزا برای هر خوشه، ارزیابی شده است. نگاره۲ فلوچارت انجام پژوهش حاضر را نشان می‌دهد.



نگاره ۲: فلوچارت انجام پژوهش

1- Difference - Indicator

در ادامه مراحل مختلف این رویکرد شرح داده‌شده است:

۳-۲-۱- خوشه‌بندی[2] گریدها به دسته‌ها

گام اول: خوشه‌بندی

در گام اول اجرای این تحقیق، گریدهای مختلف ذخیره شده در پایگاه‌داده‌ها بر طبق شباهت آن‌ها با یکدیگر خوشه‌بندی گردیدند. برای خوشه‌بندی گریدها از تابع مشابهت[3] استفاده شده است. گریدهای مشابه در خوشه‌های یکسان قرار می‌گیرند. هر گرید تنها و تنها به یک خوشه اختصاص می‌یابد.

تابع مشابهت به نوع و ماهیت داده‌های رستر بستگی دارد. این انتخاب بایستی توسط فرد متخصص در حوزه مربوطه صورت گیرد. نتایج خوشه‌بندی کاملاً به تابع شباهتی که برای خوشه‌بندی استفاده می‌شود، بستگی خواهد داشت. برای مثال پارامترهای خوشه‌بندی می‌تواند تعداد خوشه‌ها و یا درصد شباهت بین گریدها در داخل خوشه‌ها باشد *(Kang, et al, 2014:4)*.

درگام اول، تابع مشابهت برای ساخت خوشه‌های رستر با استفاده از درصد فراوانی سلول‌های مشابه توسعه داده شده است. این توابع مشابهت در نظر گرفتند که زمانی می‌توانند مجموعه‌ای از گریدها {...gridi, …, gridj ,…} در خوشه یکسان قرار گیرند که شرط زیر برای مجموعه‌ی SC از سلول‌ها برقرار باشد:

$$SC = \left\{ \left((x,y) \middle| |grid_i(x,y) - grid_j(x,y)| \leq c \right) \right\}$$

رابطه ۲

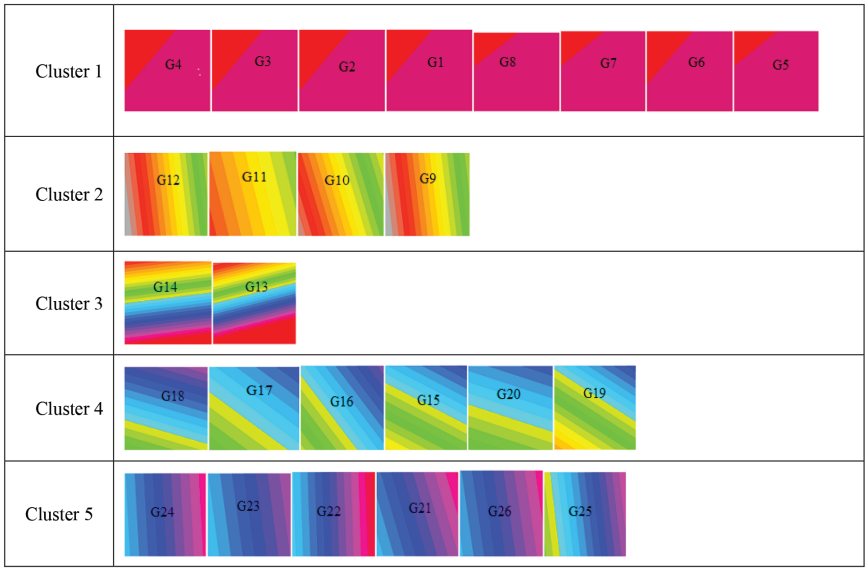
$$\text{Cardinality (SC)} \geq m \times n \times t\%$$

در اینجا c و t مقادیر ثابت از پارامترهای تابع مشابهت هستند که C مقدار ثابت و T درصد شباهت است *(Kang, et al, 2014:12)*.

سناریوهای مختلفی برای تعیین c و t اجرا شده است و در نهایت مقادیر بهینه انتخاب شده است. اندازه رسترهای بارش است. n×m

2- Clustering

3- Similarity function



نگاره ۳: نمونه‌ای رسترهای خوشه‌بندی‌شده

برای درک بهتر تفاوت بین پرس‌وجوی A و B، خوشه‌های پایگاه داده نمایش داده شده در نگاره ۳ آمده است. ۲۶ گرید در پنج خوشه (براساس شباهتشان) در پایگاه داده گروه‌بندی شده‌اند: خوشه ۱ دارای ۸ رستر، خوشه ۲ دارای ٤ رستر، خوشه ۳ دارای ۲ رستر، خوشه ٤ و ۵ نیز دارای ٦ رستر می‌باشد. هر رستر با یک شناسه خاص برای انجام محاسبات همراه است.

برای مثال، پرس‌وجوی A که از خوشه‌بندی استفاده نمی‌کند، ۸ رستر خوشه ۱ را جمع خواهد کرد. پرس‌وجوی B تنها از رستر September 2011 استفاده می‌کند و ارزش سلول‌های آن به ۸ ضرب می‌شود تا رستر نتیجه ایجاد گردد (زیرا ۸ رستر در خوشه ۱ وجود دارد)، از آنجایی که همه گریدهای یک خوشه مشابه هستند و می‌توان یک رستر را انتخاب کرد و ارزش‌های سلول‌های آن را در تعداد رسترهای موجود در آن خوشه ضرب کرد (به‌جای اینکه همه رسترها وارد عملیات جمع گردند) تا تخمینی از نتیجه دقیق به دست آید، با این رویکرد تعداد عملیات ریاضی استفاده شده کاهش می‌یابد.

در ادامه به‌طور تقریبی تعداد عملیات ریاضی اعمال‌شده توسط دو پرس‌وجو مقایسه شده است. مقایسه این شاخص بسیار حائز اهمیت است، زیرا در مورد مجموعه داده‌های بزرگ عملیات اعمال شده برای محاسبه تجمیع ، مجموعه عملیات وقت‌گیر هستند. پرس‌وجوی A محاسبات نشان داده در زیر را اعمال می‌کند.

Sum of grids for the cluster 1:
G1 + G2 + G3 + G4 + G5 + G6 + G7 + G8;
7× m × n additions of cells.
Sum of grids for the cluster 2:
G9 + G10 + G11 + G12;
3 × m × n additions of cells.
Sum of grids for the cluster 3:
G13 + G14;
1 ×m × n additions of cells.
Sum of grids for the cluster 4:
G15 + G16 + G17 + G18+ G19 + G20;
5 × m × n additions of cells.
Sum of grids for the cluster 5:
G21+ G22 + G23+ G24 + G25 +g26;
5 ×m × n additions of cells.

بنابراین درمجموع m*n*25 عملیات ریاضی موردنیاز است. m*n برابر اندازه رستر است.

پرس‌وجوی B محاسبات زیر را اعمال خواهد کرد؛ همان‌طور که در محاسبات پرسوجوی B نشان داده شده است ارزش‌های سلولی فقط به یک مقدار ثابت ضرب

شده‌اند. محاسبات اعمال شده در پرسوجوی B به‌صورت زیر است:

Sum of grids for the cluster 1:
G8 × 7;
1 × m × n multiplications.
Sum of grids for the cluster 2:
G12 × 4;
1 × m × n multiplications of cells.
Sum of grids for the cluster 3:
G14 × 2;
1 ×m × n multiplications of cells.
Sum of grids for the cluster 4:
G20 × 6;
1 ×m × n multiplications of cells.
Sum of grids for the cluster 5:
g26 × 6;
1 ×m × n multiplications of cells.

جدول ۱: مقایسه تعداد عملیات در دو پرس و جوA و B

کلاستر	تعداد عملیات ریاضیاتی پرس‌وجوی A	تعداد عملیات ریاضیاتی پرس‌وجوی B
۱	7 × m × n	1 × m × n
۲	3 × m × n	1 × m × n
۳	1 × m × n	1 × m × n
٤	5 × m × n	1 × m × n
٥	5 × m × n	1 × m × n
عملیات مجموع	4 × m × n	4 × m × n
کل عملیات	25 × m × n	9 × m × n

بنابراین درمجموع 9*m*n عملیات ریاضی برای پرس‌وجوی B اعمال شده است. در این مثال پرس‌وجوی B تعداد عملیات ریاضی را 64% کاهش می‌دهد(جدول ۱). همان‌طوری که در مثال ساده بالا نشان داده شده است ، نتیجه ساده بدست آمده در مثال بالا در عمل جهت آنالیز داده‌های رستری حجیم و زیاد می‌تواند بسیار کارآمد باشد. کاربر به آسانی می‌تواند گریدهای تقریبی را با اعمال پرس‌وجوی B استخراج کند. کاربر پس از دریافت نتیجه تقریبی، می‌تواند درصورتی‌که دقت قابل‌قبول نباشد، تصمیم بگیرد و در صورت لزوم نتیجه دقیق ارائه شود. پایگاه-داده‌ها همچنین می‌توانند به‌آسانی نتایج دقیق را با استفاده از اجرای پرس‌وجوی A به‌جای B آماده سازند.

۳-۲-۲- گام دوم: پرسشگری از داده

در پایگاه داده‌های رابطه‌ای، پرس‌وجوها می‌توانند به زبان SQL بیان شوند. همان‌طوری که در پرس‌وجو A زیر نشان داده شده است؛ در این پرس‌وجو رسترهای پایگاه‌داده به روش Sum تجمیع شده است.

Query A:
```
SELECT id_cell, Sum(value)
FROM cells, grids
WHERE
grids.id_grid = cells.id_grid
GROUP BY id_cell
```

پرس‌وجوی A مجموع رسترها را به‌صورت سلول به سلول محاسبه می‌کند. رویکرد دیگر برای تجمیع، استفاده از مدل B برای پرس‌وجو است که برآوردی از نتیجه نهایی است.

این مدل پرس‌وجو از خوشه معرفی‌شده در مرحله خوشه‌بندی به‌منظور کاهش زمان محاسبه استفاده می‌کند. پرس‌وجوی B برآوردی از نتیجه دقیق گرید محاسبه شده در پرس‌وجوی A تولید خواهند کرد. در زیر به‌طور نمونه پرس‌وجوی B آورده شده است:

Query B:
```
SELECT id_cell, Sum (value × number_of_grids)
FROM cells, grids,
SELECT max (id_grid) id_grid_rep, id_cluster, count (*)
number_of_grids
FROM grids
GROUP BY id_cluster) as TempCluster
WHERE grids.id_grid = cells.id_grid and grids.id_grid =
TempCluster.id_grid_rep
GROUP BYid_cell
```

به بیانی واضح‌تر، برای محاسبه رستر تجمیع یافته با استفاده از خوشه‌بندی، یک گرید از هر خوشه را (به‌جای همه گریدها) به‌عنوان نماینده آن خوشه انتخاب و ارزش‌های سلول‌های آن، به تعداد گریدهای موجود در خوشه ضرب می‌گردند (در گام اول: خوشه‌بندی شرح داده شده است).

جدول۲: مقایسه عملکرد پرس‌وجوی A و B و خطاهای تابع مشابهت (T,15)

Similarity (c,t)	Query A (بدون خوشه بندی)(ثانیه)				Query B (با خوشه بندی)(ثانیه)				DI error	Relative (%) error
	تکرار ۱	تکرار ۲	تکرار ۳	میانگین	تکرار ۱	تکرار ۲	تکرار ۳	میانگین		
(۱۵,۳۰)	۲۱۸	۲۲۲	۲۱٤	۲۱۸	٦/۷	٦/۸	٦/۹	٦/۸	۷۵/۹۷	۰/۰۳۸٤
(۱۵,٤۰)	۲۱۲	۱۹۸	۲۰۵	۲۰۵	۷/۲	۷/۱	۷/٦	۷/۳	۷۵/۳۸	۰/۰۳۸۲
(۱۵,۵۰)	۲۲۱	۲۱۱	۲۱۹	۲۱۷	۷/۳	۷/۲	۷/۹	۷/٤۷	۹۷/۷٤	۰/۰٤۹
(۱۵,٦۰)	۲۱٦	۲۰۹	۲۱۱	۲۱۲	۷/۹	۸/٤	۹/۲	۸/۵	۱۰۵/۵۵	۰/۰۵۳۵
(۱۵,۷۰)	۲۰۷	۱۹۹	۲۰۳	۲۰۳	۸/۳	۸/۱	۹/٦	۸/۷	۱۷۲/۲۹	۰/۰۸۷
میانگین				۲۱۱				۷/۷۵٤	۱۰۷/۳	۰/۰۲۸۷٤

٤- جمع بندی و نتیجه‌گیری

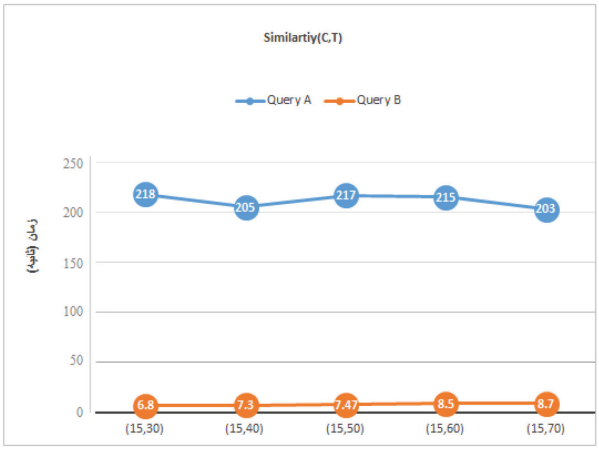
قبل از اجرای هر پرس‌وجو، کامپیوتر برای جلوگیری از 'cash effect' ایجاد شده توسط سیستم مدیریت پایگاه داده یا سیستم عملیاتی به‌طور مجدد راه‌اندازی[1] شده است. دامنه ارزش‌های سلولی در مجموعه داده‌های مورد استفاده (۱۲۰-۰) است.

با توجه به زمان اجرای تابع مشابهت(رابطه ۲) و خطای حاصل از پرس و جوی تقریبی بهینه، برای اعمال تابع مشابهت C=15 انتخاب گردیده است. ۵ سناریو برای انتخاب بهترین T اجرا شده است. در این پنج سناریو ۷۰ و٦۰و ۵۰و٤۰ و ۳۰ = T در نظر گرفته شده است. تعداد خوشه‌های تولیدشده در هر سناریو و تعداد گریدها در هر خوشه به همراه متوسط تعداد گریدها در هر خوشه در سناریوهای مختلف در جدول ۱ ذکر شده است.

جدول ۱: مشخصات کلاسترها (T,15)

Similarity (C,T)	تعداد خوشه	متوسط تعداد گرید در هر خوشه
(۱۵,۳۰)	۵	۱۲
(۱۵,٤۰)	۵	۱۲
(۱۵,۵۰)	٦	۱۰
(۱۵,٦۰)	٦	۱۰
(۱۵,۷۰)	۷	۹

پس از انجام خوشه‌بندی در پنج سناریو، پرس‌وجوهای A و B در سه تکرار روی داده ها اجرا شده است. در جدول ۲ زمان اجرای پرس‌وجوهای A و B در هر سناریو ارائه شده است. در ادامه دقت پرس‌وجوی B در سناریوهای اجرا شده محاسبه گردید. نتایج حاصل شده در جدول شماره ۲ آمده است. در نگاره ٤ زمان اجرا برای دو پرس و جوی A و B نشان داده شده است.



نگاره ٤: مقایسه زمان اجرای تابع مشابهت Sum با خوشه‌بندی (B) و بدون خوشه‌بندی (A)

یک جدول 'Grid' شناسه گریدها (ID_grid) و شناسه خوشه‌های (ID_cluster) مربوطه را ذخیره می‌نماید. سلول‌ها، مختصات (x,y) و ارزش اندازه‌گیری شده آن‌ها در جدول

1- Restart

'cell' ذخیره شده است.جداولی مانند زمان نیز برای درج اطلاعات زمانی رسترها ایجاد شده است.

٤-١- ارزیابی دقت پرس‌وجوی بهینه

برای تعیین دقت روش رسترهای تخمینی، اختلاف بین رسترهای نتیجه واقعی و رسترهای تخمینی در تمامی سناریوهای اجرا شده محاسبه شده است. روش‌های مورد استفاده برای محاسبه با تجزیه تحلیل داده‌های آزمایش‌ها در زیر ارائه‌شده است.

در ابتدا تفاوت بین سلول‌های نظیر به نظیر رسترهای تخمینی و سلول‌های رسترهای واقعی با استفاده از رابطه ۳ محاسبه شد *(Kang, et al, 2014:13)*.

رابطه ۳:

$$DI = \frac{\left(\sum_{i=1}^{u}\sum_{j=1}^{v}|estimated\_gride_i(cell_k) - real\_gride_i(cell_k)|\right)}{(u \times v)}$$

U تعداد گریدهای بازگشتی توسط پرس‌وجو است

V اندازه گریدهاست(تعداد سطر در ستون).

$estimated\_grid_i(cellk)$ مقدار k امین سلول در i امین گرید است.

$real\_grid_i(cellk)$ مقدار k امین سلول در i امین گرید است.

$estimated\_grid_i$ گرید تخمینی از $real\_grid_i$ است.

در ادامه به توضیح محاسبه خطای نسبی بیان پرداخته شده است. محور x تخمینی از ارزش صحیح y می‌باشد. اختلاف بین x و y را می‌توان به‌وسیله خطای قدر مطلق $|x-y|$ یا در صورتی که $y \neq 0$ باشد خطای نسبی $|x-y|/|y|$ محاسبه کرد. اگر خطای نسبی x برابر با 10^{-k} باشد می‌توان گفت که x برای k رقم اعشار دقیق است. یک رستر می‌تواند به‌صورت یک وکتور مشاهده شود. این وکتور norm اندازه‌ای برای فاصله بین یک وکتور دلخواه و وکتور صفر فراهم می‌کند؛ بنابراین فاصله بین دو وکتور می‌تواند به‌عنوان norm اختلاف وکتورها تعریف گردد. همچنین استفاده از normها برای محاسبه اختلاف مطلق $\|x-y\|$ و تفاوت نسبی $\|x-y\|/\|y\|$ بین وکتور x و y استفاده شود. L^1-norms نشان‌دهنده یک مقدار خطای میانگین است *(Burden and Faires 1993; Golub and Van Loan 1993)*.

اگر $x = (x1, x2, \ldots, xn)$ و $y = (y1, y2, \ldots, yn)$ وکتورهای Rn باشند برخی normهای نسبی به‌صورت زیر تعریف می‌شوند *(Kang, et al, 2014:15)*:

رابطه ٤:

$$L^1\ relative\ error: \|x_i - y_i\|_1 / \|y_i\|_1 = \Sigma|x_i - y_i| / \Sigma|y_i|$$

نتایج حاصل از اجرای پرس‌وجوهای محاسبه خطا به دو روش (DI) و خطای نسبی به دست آمده است. از مشاهده نتایج چنین برمی‌آید که خطای پرس‌وجوی B در سناریوهای مختلف انجام‌شده بین ۰/۰۰٦ تا ۰/۰۹۵ است.

٤-٢- بررسی کیفیت روش

آزمایش عملی اجرای پرس‌وجوی بهینه روی کارایی این روش را ثابت می‌کند. در این بخش برای بررسی کیفیت تئوری از روشی که توسط Kang و همکاران در سال ۲۰۱٤ بر روی نتایج تولیدشده معرفی شد، استفاده گردید. این روش محاسبه خطا، خطا را برای هر خوشه به‌صورت مجزا محاسبه می‌کند؛ بنابراین برخی شرایط ایجادکننده نتایج ضعیف را می‌توان برجسته نمود. در نظر بگیرید که یک فرایند خوشه‌بندی تعداد NC خوشه را که هرکدام دارای P رستر هستند، ایجاد کرده است. گریدها همگی دارای تعداد سلول M*N می‌باشند. در هر خوشه دقیقاً t% از سلول‌ها، $|grid_i(x, y) - grid_j(x, y)| \leq c$ را در نظر گرفته شده در هر خوشه K تعداد زیرمجموعه آخر سلول‌ها و L مکمل K می‌باشد؛ Cardinality(K) = t/100 × m × n و Cardinality(L) = (1 – t/100) × m × n می‌باشد. ارزش‌های m و n و t و c و p برای همه خوشه‌ها ثابت هستند. مجموعه سلول‌ها شامل t% از یک خوشه به خوشه دیگر مشابه است. بیشترین خطا حالت زمانی رخ می‌دهد که:

۱ – t % سلول‌های مشابه از یک خوشه به خوشه دیگر متفاوت باشد.

۲– در هر خوشه $\Sigma_{i=1}^{p}\Sigma_{k=1}^{w}|representative\_grid(cell_k) - grid(cell_k)|$

با L ûcellk و w = cardinality(L) بالا باشد (*Kang, L., et al, 2015:16*).

در این پژوهش، یک روش جدید با خوشه بندی رسترها با استفاده از تابع مشابهت برای بهینه‌سازی عملیات تجمیع گریدهای رستری درون پایگاه‌داده‌ها ارائه‌شده است.

در این رویکرد امکان تنظیم دقت و صحت نتایج با قابلیت تغییر پارامترهای دخیل در تابع مشابهت ممکن می‌باشد.

در این پژوهش عملیات تجمیع Sum به دو روش خوشه‌بندی شده و بدون خوشه‌بندی بر روی داده‌ها انجام گرفته است(نگاره ۱).

بررسی نتایج نشان می‌دهد که سرعت اجرای تابع Sum با خوشه‌بندی ۲۷/۲ برابر اجرای این تابع بدون خوشه‌بندی است. همچنین میانگین اختلاف عددی پیکسل‌های حاصل از اجرای تابع Sum بهینه و اجرای معمولی آن۰/۰۲۸ محاسبه شد و میانگین زمان اجرای پرس‌وجوهای A و B به ترتیب ۲۱۱ و ۷/۷۵ می‌باشد (جداول ۱و ۲).

نتایج حاصل از تابع مشابهت (T,۱۵) که در این تابع C ثابت و متغیر T است نشان می‌دهد که با افزایش T زمان اجرای تابع مشابهت افزایش می‌یابد و همچنین تعداد خوشه‌ها نیز افزایش می‌یابد.

بنابراین متوسط تعداد گرید در هر خوشه کاهش می‌یابد. البته همیشه کاهش زمان اجرا و تعداد خوشه‌ها مطلوب نیست چراکه با کاهش تعداد خوشه‌ها، خطای انجام پرس‌وجوها افزایش می‌یابد. نتایج حاصل از تحقیق حاضر می‌تواند در کاهش زمان آنالیزهای رستری داخل پایگاه داده که با انبوهی از داده‌ها مخصوصاً در سیستم‌های لحظه-ای مانند هواشناسی و ترافیک و مانورهای نظامی جهت مکان‌گزینی لحظه‌ای و ... مواجه هستند، جهت ارائه جواب در کمترین زمان ممکن و در لحظه استفاده شود. چرا که گاهی اوقات ارائه جواب دقیق ولی دیرتر از زمان مطلوب فایده‌ای در پی نخواهد داشت و در مقابل ارائه جوابی با دقتی کمی پایین‌تر از دقیق ولی در زمان خواسته شده ارزش بیشتری از جواب دقیق خواهد داشت.

۵- پیشنهادات

از پردازش تقریبی پرس و جو می‌توان برای سایر توابع تجمیع نیز استفاده نمود و همچنین این الگو را روی سایر رسترها مانند تصاویر ماهواره‌ای با قدرت تفکیک‌های مکانی و طیفی متفاوت و رسترهایی دیگر با سایر ماهیت ها اعمال نمود. این الگو را می‌توان با سایر روش‌های خوشه‌بندی و با توابع مشابهت دیگری انجام داد.

٦- منابع و مآخذ

1. Acharya, S., Gibbons, P. B., Poosala, V., & Ramaswamy, S. (1999, June). Join synopses for approximate query answering. In ACM SIGMOD Record (Vol. 28, No. 2, pp. 275-286). ACM.

2. Azevedo, L. G., Zimbrão, G., & De Souza, J. M. (2007). Approximate query processing in spatial databases using raster signatures. In Advances in Geoinformatics (pp. 69-86). Springer Berlin Heidelberg.

3. Babcock, B.; Chaudhuri, S. and Das, G. (2003), Dynamic Sample Selection for Approximate Query Processing., in Alon Y. Halevy; Zachary G. Ives & AnHai Doan, ed., 'SIGMOD Conference', ACM., pp. 539-550.

4. Bernardino, J. R., Furtado, P. S., & Madeira, H. C. (2002). Approximate query answering using data warehouse striping. Journal of Intelligent Information Systems, 19(2), 145-167.

5. Brinkhoff, T., Horn, H., Kriegel, H. P., & Schneider, R. (1993, June). A storage and access architecture for efficient query processing in spatial database systems. In International Symposium on Spatial Databases (pp. 357-376). Springer Berlin Heidelberg.

6. Burden, R., and D. Faires. 1993. Numerical Analysis. 5th ed. Boston, MA: PWS.

7. Burden, R., and D. Faires. 1993. Numerical Analysis. 5th ed. Boston, MA: PWS.

8. Chakrabarti, K., Garofalakis, M., Rastogi, R. and Shim, K. (2001). Approximate query processing using wavelets. The VLDB Journal, 10(2-3), pp.199-223.

9. Garofalakis, M. and Gibbon, P. (2001). Approximate Query Processing: Taming the TeraBytes. In: VLDB

'01 Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., p.725.

10. Gibbons, P. B., Matias, Y., Poosala, V., 1997, Aqua project white paper, Technical report, Bell Laboratories (1997).

11. Golub, G., and C. Van Loan. 1993. MATRIX Computations. 4th ed. Baltimore, MD: Johns Hopkins University Press.

12. Golub, G., and C. Van Loan. 1993. MATRIX Computations. 4th ed. Baltimore, MD: Johns Hopkins University Press.

13. Güting, R. H., & Schneider, M. (1993, June). Realms: A foundation for spatial data types in database systems. In International Symposium on Spatial Databases (pp. 14-35). Springer Berlin Heidelberg.

14. Ioannidis, Y. E., & Poosala, V. (1995, June). Balancing histogram optimality and practicality for query result size estimation. In ACM SIGMOD Record Vol. 24, No. 2, pp. 233-244.

15. Kang, M. A., Zaamoune, M., Pinet, F., Bimonte, S., & Beaune, P. (2015). Performance optimization of grid aggregation in spatial data warehouses. International Journal of Digital Earth, 8(12),1-19.

16. Liu, Q. (2009). Approximate Query Processing. Encyclopedia of Database Systems, pp.113-119.

17. Mehanna, Y. S., Mahmuddin, M., & Abdelaziz, H. S. Approximate Query Processing Concepts and Techniques,pp.11-19.

18. Obe, R. O., & Hsu, L. S. (2015). PostGIS in action. Manning Publications Co.

19. Papadias, D., Kalnis, P., Zhang, J., & Tao, Y. (2001, July). Efficient OLAP operations in spatial data warehouses. In International Symposium on Spatial and Temporal Databases (pp. 443-459). Springer Berlin Heidelberg

20. Perera, K. S., Hahmann, M., Lehner, W., Pedersen, T. B., & Thomsen, C. (2015, April). Modeling Large Time Series for Efficient Approximate Query Processing. In International Conference on Database Systems for Advanced Applications (pp. 190-204). Springer International Publishing.

21. Wang, J. (2009). Encyclopedia of data warehousing and mining. Hershey: Information Science Reference.

22. Xie, Q., Chen, F., Zhang, Z., & Lucena, I. (2013). In-database image processing in Oracle Spatial GeoRaster. In ASPRS 2013 Annual Conference Baltimore, Maryland.