مجله محاسبات نرم، جلد ۹، شماره ۱ (بهار و تابستان ۱۳۹۹)، صفحه ۲۶ـ۳۹
https://dx.doi.org/10.22052/scj.2021.111450

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

دانشگاه کاشان
University of Kashan

استفاده از تکنولوژی داده‌های عظیم در نظرکاوی✧

فاطمه پورغلامعلی[۱]*، استادیار، محسن کاهانی[۲]، استاد، احسان عسگریان[۲]، دکتری

[۱] دانشکده مهندسی، دانشگاه ولیعصر (عج) رفسنجان، رفسنجان، ایران.

[۲] دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران.

اطلاعات مقاله

تاریخچه مقاله:

دریافت ۱۹ اسفند ماه ۱۳۹۸

پذیرش ۲۹ دی ماه ۱۳۹۹

کلمات کلیدی:

نظرکاوی

داده‌های عظیم

هدوپ

مدل برنامه‌نویسی نگاشت_کاهش

پردازش زبان طبیعی

بازیابی اطلاعات

چکیده

نظرات، نقش مهم و تعیین کننده‌ای در فرآیند تصمیم‌گیری چه برای مشتریان و چه برای سازمان‌های تجاری ایفا می‌کنند. از این رو، وجود سیستم‌های خودکار نظرکاوی برای داده‌های نظری موجود در وب، ضروری به نظر می‌رسد. از طرفی، با حجم بالا و رشد روزافزون داده‌های نظری روی وب، فرآیند نظرکاوی می‌تواند با چالشی بزرگ روبه‌رو شود؛ چرا که پردازش و تحلیل این حجم عظیم از داده‌ها با تکنولوژی‌های متداول، ممکن است عملی نباشد. در صورتی که سیستم‌های کاوش نظرات به تکنولوژی‌های داده‌های عظیم مجهز شوند، بدون نگرانی از مدیریت، ذخیره‌سازی و مدیریت حجم روزافزون داده‌های نظری، می‌توانند به کار خود ادامه دهند. با آنکه در سال‌های اخیر تحقیقات زیادی در حوزه تحلیل حسی نظرات انجام شده است، اما تلاش‌های کمی در حوزه کاوش داده‌های نظری در حیطه زبان فارسی در مقیاس بالا انجام گرفته است. از این رو، در این تحقیق، دو روش نظرکاوی برای داده‌های زبان فارسی با استفاده از یک لغت‌نامه حسی زبان فارسی در بستر تکنولوژی داده‌های عظیم ارائه شده است. برای ذخیره‌سازی و پردازش داده های نظری از چارچوب متداول و کارای هدوپ و مدل برنامه‌نویسی نگاشت_کاهش (MapReduce) استفاده شده است. نتایج به دست آمده از آزمایش‌ها حاکی از آن است که چارچوب پیشنهادی برای نظرکاوی به شکلی کارا عمل کرده و نه تنها برای حجم‌های بالا بلکه در حجم‌های حدود ۲۰ مگابایت شاهد افزایش ۱۰۰ برابری در کارامدی هستیم. این ضریب، در حجم‌های بالاتر به شکل محسوس‌تری افزایش می‌یابد.

۱. مقدمه

نظرات کاربران در مورد محصولات، منبعی ارزشمند برای تصمیم‌گیری می‌باشند. استفاده از نظرات مشتریان هم برای سایر مشتریان و هم برای صاحبان کسب و کار بسیار سودمند است. نظرکاوی شاخه‌ای فعال در حوزه پردازش زبان طبیعی و متن کاوی می‌باشد که مورد توجه محققین زیادی قرار گرفته است [۱]. از طرف دیگر امروزه که ابزارهای فراوانی برای تولید محتوا در اختیار کاربران قرار می‌گیرد، مساله مدیریت حجم زیاد داده، به عنوان یکی از مسایل اساسی و در عین حال جذاب و چالش برانگیز ظهور و بروز پیدا کرده است [۲].

✧ نوع مقاله: پژوهشی

* نویسنده مسئول

پست(های) الکترونیک: f.pourgholamali@vru.ac.ir (پورغلامعلی)

kahani@um.ac.ir (کاهانی)

ehasna.asgarian@gamil.com (عسگریان)

نحوه ارجاع به مقاله: پورغلامعلی، فاطمه، کاهانی، محسن، عسگریان، احسان، «استفاده از تکنولوژی داده‌های عظیم در نظرکاوی»، مجله محاسبات نرم، جلد ۹، شماره ۱، ص۲۶ـ۳۹، بهار و تابستان ۱۳۹۹.

شرکت‌های تجاری برای ارائه بهتر محصولات و فراهم کردن بستری برای فروش برخط محصولات خود، وبسایت‌هایی ایجاد نموده‌اند که از طریق آن بسیاری از امکانات و فعالیت‌های شرکت تسریع و تسهیل می‌گردد. از جمله این امکانات، نظردهی وبسایت‌ها می‌باشد. با توجه به استفاده وسیع کاربران از امکانات برخط فروشگاه‌ها و با در نظر گرفتن تعداد زیاد کاربران در این فروشگاه‌ها، مدتی پس از شروع به کار وبسایت‌ها، شاهد حجم عظیم داده‌های نظری در مورد محصولات خواهیم بود. با این تفاصیل، بکارگیری تکنولوژی-های مربوط به مدیریت و پردازش داده‌های عظیم در حوزه نظرکاوی ضروری به نظر می‌رسد. ثابت شده است که توانایی مدیریت حجم زیاد داده، مزایای اقتصادی زیادی به همراه خواهد داشت [۳].

با وجود آنکه پیشرفت‌های بسیاری در حوزه کاوش نظرات انجام شده است [۱و۴] تلاش‌های کمی برای پردازش کارای حجم عظیم نظرات به خصوص در حیطه زبان فارسی انجام گرفته است. سازمان‌ها و شرکت‌های تجاری تا زمانی‌که با حجم‌های غیرکلان داده‌های نظری سر و کار داشته باشند، با روش‌های معمول می‌توانند به کار خود ادامه دهند؛ اما زمانی‌که با گذشت زمان، توسعه‌ی سازمان، افزایش کاربران و محصولات و محبوبیت فعالیت‌های برخط رو به رو شوند، حجم‌های معمول داده تبدیل به حجم‌های کلان شده و روش‌های پیشین دیگر پاسخگو نخواهد بود.

با توجه مسایل ذکر شده و همچنین به نیاز انجام پژوهش‌های مهم و کاربردی در حوزه زبان فارسی [۵] در این پژوهش چارچوبی برای تحلیل نظرات فارسی کاربران در بستر تکنولوژی داده‌های عظیم ارائه می‌گردد. این راه حل با بهره‌گیری از امکانات گسترده چارچوب هدوپ[1] و مدل قوی برنامه‌نویسی نگاشت-کاهش[2]، کمک شایانی به توسعه‌پذیری و مقیاس‌پذیری سیستم‌های نظرکاوی می‌کند.

نوآوری این پژوهش را می‌توان در ارائه دو چارچوب برای پردازش حجم عظیم داده‌های نظری فارسی در بستر هدوپ و برنامه نویسی نگاشت-کاهش برشمرد. همچنین در این پژوهش آزمایش‌های متعدد جهت آزمودن جنبه‌های مختلف تکنولوژی داده‌های عظیم و بررسی نقاط قوت و ضعف آن انجام گرفته است.

در بخش بعد مفاهیم اصلی مورد نیاز این پژوهش با جزئیات بیشتری توضیح داده می‌شوند. سپس مروری بر پیشینه کارهای انجام شده در این حوزه خواهیم داشت. در بخش سوم مجموعه داده مورد استفاده و جزئیات چارچوب پیشنهادی مطرح خواهد شد. بخش چهارم حاوی نتایج پژوهش می‌باشد و بخش آخر نتیجه‌گیری و جمع بندی ارائه می‌گردد.

۲. پیشینه

در این بخش ابتدا مروری بر مفاهیم اصلی این حوزه خواهیم داشت. پس از آن، روش‌های کلی نظرکاوی و سپس تحقیقات انجام شده برای بکارگیری تکنولوژی داده‌های عظیم برای نظرکاوی را بیان خواهیم کرد.

۲,۱. نظرکاوی

نظرکاوی به رشته مطالعاتی اطلاق می‌گردد که در آن به تجزیه و تحلیل احساسات، ارزیابی‌ها و گرایش‌های افراد در مورد موجودیت‌هایی مثل محصولات و سرویس‌ها و خصیصه‌های آن‌ها پرداخته می‌شود. تحقیق و پژوهش در مورد نظرات و احساسات مردم تقریبا از سال ۲۰۰۰ شروع شده و از آن زمان، به زمینه‌ای جذاب و رو به رشد تبدیل شده است. نظرکاوی، بسیاری از زیرمساله‌های جدیدی را مطرح می‌نماید و غیر از NLP در زمینه‌های داده‌کاوی، وب‌کاوی، بازیابی اطلاعات و ... بطور وسیع مورد مطالعه قرار گرفته‌است [۶]. به شکل رسمی، یک نظر یک چهارگانه (g, s, h, t) است که در آن g هدف (آنچه که "نظر" در مورد آن بیان شده است)، s حس بیان شده، h صاحب نظر و t زمان بیان نظر می‌باشد [۴]. در این تعریف، دو جزء اول اجزاء مهمتری نسبت به دوجزء دیگر هستند و در اغلب کاربردها تنها همان دو جزء اول به عنوان

[1] Hadoop
[2] MapReduce

اجزاء "نظر" در نظرگرفته می‌شوند. به جای واژه هدف می‌توان واژه "موجودیت" را نیز بکار برد.

- **لغت‌نامه حسی:** مهم‌ترین نشانه‌ای که ما را به حس نظرات نزدیک می‌سازد واژگان حسی می‌باشد. این واژه‌ها اغلب برای بیان نظرات مثبت و منفی به کار برده می‌شوند. به عنوان مثال، خوب، عالی و شگفت انگیز، واژه‌های حسی مثبت و بد، ضعیف، وحشتناک واژه‌های حسی منفی هستند. علاوه بر خود کلمات، تعدادی عبارت و اصطلاح هم در هر زبان وجود دارد که می‌توانند حسی باشند. یک لیست از واژه‌ها و عبارات حسی، لغت‌نامه حسی نام دارد. روش‌های مختلفی برای جمع آوری لغت‌نامه حسی ارائه شده است [۶]. اگرچه داشتن لغت‌نامه حسی بسیار ضروری است، اما کافی نمی‌باشد؛ چرا که یک واژه حسی مثبت یا منفی در کاربردهای مختلف، جهات متضادی را می‌تواند به خود بگیرد. یعنی در کاربردی مثبت و درکاربردی دیگر منفی باشد. علاوه بر این، ممکن است یک جمله، کلمه حسی نداشته باشد اما بار حسی داشته باشد.
- **لایه‌های تحلیل در نظرکاوی:** روش‌های موجود برای نظرکاوی در سطوح مختلفی از تحلیل از نظر ریزدانگی قرار دارند. این لایه‌ها عبارتند از الف) سطح سند، ب) سطح جمله، ج) سطح ویژگی. در بالاترین سطح و درشت‌ترین حالت **سطح سند** قرار دارد. در این سطح، با داشتن یک سند نظری به دنبال آن هستیم که بدانیم کل این سند، نظر مثبت یا منفی را بیان می‌کند. در اصل این کار با دسته‌بندی سند به دسته مثبت یا منفی انجام می‌پذیرد [۷و۸]. به عنوان مثال، با داشتن یک مقاله از یک محصول، سیستم نظرکاوی مشخص می‌نماید که مقاله نظر کلی مثبت یا منفی در مورد محصول دارد. در سطح جمله با نگاه جزئی‌تر به سراغ جملات می‌رویم. جملات به دسته‌های مثبت، منفی، و خنثی دسته‌بندی می‌شوند. دو سطح بیان شده، به شکل دقیق مشخص نمی‌کنند که کاربران چه چیزی دوست دارند و چه چیزی دوست ندارند. در سطح **ویژگی** به جای پرداختن به ساختارهای زبان مثل جمله، پارگراف، عبارت و ... مستقیما به سراغ خود **نظر** می‌رویم. از آنجا که اغلب، موجودیت و حس مربوط به آن (دو جزء اصلی نظر) همراه با هم می‌آیند، در این سطح، به دنبال کشف حس روی موجودیت‌ها و جنبه‌ها (ویژگی‌ها)ی مختلف آن‌ها هستیم.

۲،۲. داده‌های عظیم

داده‌های عظیم به مجموعه داده‌هایی اطلاق می‌شود که مدیریت، کنترل و پردازش آن‌ها فراتر از توانایی ابزارهای نرم‌افزاری در یک زمان قابل تحمل و مورد انتظار است. نمونه‌هایی از داده‌های عظیم، سامانه‌های بازشناسی با امواج رادیویی، شبکه‌های حسگر، شبکه‌های اجتماعی، متون و اسناد اینترنتی، نمایه‌های جستجوهای اینترنتی، آرشیو عکس و ویدیو هستند. در سال ۲۰۰۱، تحلیلگر صنعت، داگ لنی، داده‌های عظیم را با سه خصوصیت بارز (3V) تعریف کرد: حجم (Volume)، سرعت (Velocity) و تنوع (Variety). به این معنی که داده‌های عظیم از نظر حجم، سرعت ایجاد و تنوع ساختاری مورد توجه می‌باشند. در سال‌های بعد خصوصیات بیشتری از قبیل صحت (Veracity)، نوسان (Volatility) و مقدار (Value) به تعریف داده‌های عظیم اضافه شد [۹].

- **هدوپ**: یکی از بسترهای مناسب برای ذخیره‌سازی و پردازش داده‌های عظیم، چارچوب متن باز هدوپ می‌باشد. این تکنولوژی با ترکیب و توزیع داده‌های ساختمند و غیرساختمند به ذخیره‌سازی آن می‌پردازد و به زبان جاوا پیاده‌سازی شده است. سیستم فایل توزیع شده هدوپ یا HDFS[1]، با توزیع و تکثیر داده‌های فایل روی گره‌های داده (data node) و مدیریت آنها از طریق گره هماهنگ‌کننده به نام، گره نام (name node) امکان ذخیره و بازیابی کارا برای داده‌های عظیم را فراهم می‌آورد.
- **مدل برنامه‌نویسی نگاشت-کاهش**: در سال ۲۰۰۴ مدل

[1] Hadoop Distributed File System

برنامه نویسی نگاشت-کاهش برای پردازش مقادیر عظیمی از داده‌های بدون ساختار در سیستم‌های توزیع شده و موازی ارائه شد [۱۰]. مدل برنامه‌نویسی نگاشت-کاهش دارای دو مزیت مهم است: ۱) برنامه‌نویس درگیر جزئیات توزیع داده‌ها، تکثیر، متعادل‌سازی بار، ارتباطات و همگام‌سازی پردازنده‌ها و غیره نمی‌شود و ۲) دارای ساختاری سرراست می‌باشد و تنها نیاز به ارائه دو تابع نگاشت و کاهش است. نکته اصلی این چارچوب فهم مفهوم *جفت‌های کلید مقدار* است. برنامه‌نویس برای نوشتن توابع *نگاشت*[1] و *کاهش*[2] باید پردازش‌ها و داده-های خود را در قالب جفت‌های کلید-مقدار مدل کند. در واقع قسمت پیچیده و زمان‌بر کار همین جاست. برنامه‌نویس باید درک عمیقی نسبت به مفهوم نگاشت-کاهش داشته باشد و با تسلطی که بر روی الگوریتم اصلی دارد آن را در قالب توابع نگاشت و کاهش بازنویسی کند.

در یک فاز نگاشت-کاهش ابتدا داده‌های ورودی در قالب جفت‌های کلید-مقدار روی نگاشت‌دهنده‌ها توزیع می-شوند. هر نگاشت‌دهنده، تابع نگاشت را بر روی داده‌های خود اجرا کرده و خروجی تمام نگاشت‌دهنده‌ها در قالب جفت‌های کلید-مقدار مرتب می‌شود. هر کلید با مجموعه‌ای از مقادیر حاصل شده به کاهش‌دهنده‌ها داده می‌شود تا تابع کاهش را اجرا کنند. هدوپ، پیاده‌سازی متن بازی از نگاشت-کاهش را ارائه کرده است.

۲,۳. روش‌های کلی نظرکاوی

به طور کلی نظرکاوی یک کار دسته‌بندی است و در حالت کلی متون نظری سه دسته مثبت و منفی و خنثی جای خواهند گرفت. یکی از روش‌های انجام این کار، استفاده از روش‌های نظارتی است. گام‌های اصلی در این زمینه عبارتند از: ایجاد بانک نظرات برچسب خورده، ایجاد لغت‌نامه حسی، انتخاب ویژگی‌های مناسب (مانند کلمات مرتبط با موضوع، نرخ رخداد عبارت، برچسب‌های اجزای سخن، عبارات حسی، تغییردهنده معنا، وابستگی نحوی و غیره،) و در نهایت استفاده از الگوریتم‌های دسته‌بندی مانند ماشین پشتیبان بردار (SVM)، بیزین ساده و ... و محاسبه مجموع امتیاز سند [۱و۱۱].

به دلیل وجود مشکلات روش‌های نظارتی از جمله هزینه بالای تولید پیکره‌ی آموزش، روش‌های غیرنظارتی ارائه گردیدند. در روش ارائه شده در [۱۲] الگوهای متداول نحوی که معمولا برای بیان احساس بکار می‌روند، استخراج می-گردند. این الگوها معمولا از روی برچسب‌های اجزای سخن[3] ساخته می‌شوند. عبارات دو واژه‌ای که با این الگوها تطابق داشته باشند استخراج می‌گردند و گرایش حسی عبارت با استفاده از وابستگی آماری دو واژه با دو واژه حسی مرجع مثبت و منفی محاسبه می‌گردد.

۲,۴. نظرکاوی در مقیاس بالا

همزمان با بزرگ شدن حجم داده‌ها و اهمیت یافتن کشف دانش از این منبع نامتناهی، یکی از کاربردهای آن یعنی نظرکاوی نقش خود را نشان می‌دهد. نظرکاوی‌های بزرگ مقیاس در سال‌های اخیر مورد توجه محققان قرار گرفته است که در ادامه به برخی از آنها اشاره می‌کنیم.

روشی برای تحلیل حسی روی داده‌های عظیم در [۱۳] ارائه شده است و اشاره شده است که فاکتورهای انسانی روی حس استخراج شده از داده‌های عظیم تاثیر می‌گذارد. تحقیق مذکور بر روی پیکره Yahoo! Answers، پیکره عظیمی از سوال و جواب‌های کاربران می‌باشد و این امکان برای کاربران فراهم شده تا به سوالات رای بدهند. پیکره شامل ۳۴ میلیون سوال و ۱۳۲ میلیون جواب است. جنبه‌های مختلفی از سوال برای پیش‌بینی حس مربوط به سوالات مورد بررسی قرار گرفت؛ از جمله جنبه‌های ظاهری مثل طول سوال، جنبه‌های نقطه‌گذاری مثل تعداد علامات سوال و کلمات شروع، جنبه‌های آماری مثل سن، تحصیلات و جنسیت.

[1] Map
[2] Reduce
[3] Part Of Speech - POS

نویسنده مقاله در [۱۴] از داده‌های شبکه اجتماعی توییتر استفاده کرده است و عنوان نموده است که برای این داده‌ها ابزارهای عادی پردازش متن کارا نیستند. چرا که عملیات و کارکترهای خاص این داده‌ها، صورتک ها و کلماتی مثل veeerrryyyyy goooodddd در لغت‌نامه های معمول وجود ندارد و برای مدیریت آنها بایستی راهکارهای جدیدی اتخاذ نمود. سپس با استفاده از عبارات دوتایی در چارچوب نگاشت_کاهش یک گراف هم‌رخدادی ایجاد کردند؛ به این ترتیب که برای مشابهت بین کلمات از معیار فاصله کسینوسی بین کلمات استفاده می‌شود. محاسبه فاصله کسینوسی و ساخت گراف هم‌رخدادی در چارچوب نگاشت_کاهش انجام می‌پذیرد. تعدادی لغت حسی به عنوان بذر اولیه در نظر گرفته شده و درجه حسی آنها از طریق یال‌های گراف به سایر لغات منتشر میشود. این عمل نیز در چارچوب نگاشت_کاهش انجام می‌پذیرد.

تحقیق دیگری در [۸] با استفاده از ابزار متن کاوی KNIME و با استفاده از تکنولوژی داده‌های عظیم به استخراج حس از مجموعه نظرات وب‌سایت‌های هتل‌ها و توییتر پرداخته است. پس از آنکه خزنده وب متن نظرات را از وب جمع‌آوری می‌کند، پیش‌پردازش های لازم جهت نظرکاوی صورت می‌پذیرد و داده‌های پیش‌پردازش شده در پایگاه داده غیررابطه‌ای HBase که یکی دیگر از تکنولوژی‌های داده‌های عظیم محسوب می‌شود، جمع آوری می‌شود. سپس KINME از طریق نود HBase Reader رکوردها را در بسته های ۳۰۰۰۰ تایی خوانده و با استفاده از لغت‌نامه‌ای که از قبل ایجاد شده-است به استخراج حس ویژگی‌ها می‌پردازد. ویژگی‌ها در این کار محدود به ۴ ویژگی خاص هتل‌ها است که به صورت تجربی و دستی استخراج شده اند.

در کارهای جدیدتر، چارچوب هدوپ برای پردازش داده‌های نظری و تحلیل حسی مورد استفاده قرار گرفته است. در [۱۵] مجموعه ای از روش‌ها مبتنی بر هدوپ مورد تحلیل و بررسی قرار گرفته است. در [۷] پیش‌بینی حس توییت‌ها در بستر هدوپ انجام گرفته است. در [۱۶] پیشنهاد شده است که برای بهره‌مندی از مکانیزم‌های ذخیره و بازیابی کارا، داده‌های نظری در سیستم فایل توزیع شده هدوپ (HDFS) ذخیره‌سازی شوند. در این تحقیق داده‌های حاصل از تعاملات دانشجویان از توییتر و فیسبوک استخراج شده و پس از برچسب‌گذاری داده‌ها و ذخیره‌سازی آن‌ها در سیستم فایل توزیع شده هدوپ، روش‌های مختلف یادگیری ماشین برای پیش‌بینی بار حسی نظرات دانشجو مورد استفاده قرار می‌گیرد.

اسپارک هم یکی از چارچوب‌های متن‌باز متداول می‌باشد که پردازش داده‌های عظیم را در یک معماری توزیع شده با قابلیت تحمل خطا ممکن می‌کند. با استفاده از این چارچوب [۱۷] یک روش برای موازی سازی مدل JST [۱۸] پیشنهاد می‌شود که در آن مدلسازی عنوان و حس به شکل همزمان از متن نظرات انجام می‌شود.

در [۱۹] اشاره شده است که به‌کارگیری تکنولوژی داده‌های عظیم مثل پایگاه داده‌های غیر رابطه‌ای HBase و MongoDB ، هدوپ و تکنیک نگاشت_کاهش، در کنار همه مزایایی که دارد، نیازمند تجهیز سازمان به این منابع می‌باشد و هزینه‌های زیادی را تحمیل می‌نماید. استفاده از محاسبات ابری می‌تواند صرفه‌جویی قابل ملاحظه‌ای را در بر داشته باشد. چرا که، ملزومات خرید زیرساخت‌های شبکه‌ای و نرم افزارهای مربوطه برداشته شده و تنها سرویس‌های مورد نیاز با هزینه‌ای کمتر، قابل بکارگیری از طریق ابر محاسباتی خواهند بود. البته در مورد این ادعا باید این نکته را مد نظر قرار داد که سازمانی که خود را به تکنولوژی داده‌های عظیم مجهز می‌کند، گرچه در ابتدا ممکن است هزینه بالایی متحمل شود، اما می‌تواند از این امکانات خود برای همیشه استفاده نماید، در حالیکه استفاده از محاسبات ابری مستلزم این است که اولا سرویس مورد نظر به شکل درستی شناسایی شود که این کار دشواری است [۲۰]. ثانیا هر بار که نیاز به کار محاسباتی جدیدی باشد، بایستی هزینه‌ای پرداخت شود. نکته دیگر آن است که بسیاری از سازمان‌ها تمایلی برای به اشتراک‌گذاری داده‌های خصوصی خود در ابر محاسباتی ندارند. اینکه سازمان‌ها مجهز به زیرساخت‌های لازم برای مدیریت و پردازش داده‌های عظیم

باشند یا از امکانات محاسبات ابری استفاده نمایند بستگی به مشخصات و نیازمندی‌های سازمان دارد.

با مرور کارهای انجام شده در این زمینه می‌توان به این نتیجه رسید که کارهای مرتبط به چند دسته تقسیم بندی می‌شوند، یک دسته آن‌هایی هستند با هدف تولید لغت‌نامه حسی از تکنولوژی داده‌های عظیم بهره برده‌اند و دسته دیگر هدف تحلیل حسی نظرات را دنبال کرده‌اند. در این پژوهش به دلیل آنکه لغات نامه حسی در فازهای قبلی پژوهش تولید شده بود، تمرکز اصلی بر فازهای بعدی گذاشته شد. هدف اصلی این پژوهش استفاده از روش‌های غیرنظارتی نظرکاوی در بستر تکنولوژی داده‌های عظیم بوده است.

از یک دیدگاه دیگر، کارهای مرتبط از نظر گستره‌ی استفاده از امکانات تکنولوژی داده‌های عظیم دسته‌بندی می‌شوند. به شکل دقیق‌تر، برخی کارها، تنها از مکانیزم‌های ذخیره‌سازی داده‌های عظیم استفاده کرده‌اند و از امکانات پردازشی کارای آن (احتمالا به دلیل دشوار بودن پیاده‌سازی آن) دوری جسته‌اند. در این پژوهش هدف آن است که هم از امکانات ذخیره‌سازی و هم پردازشی موجود برای مدیریت داده‌های نظری استفاده شود.

در جدول (۱) مهمترین کارهای مرتبط با نظرکاوی در مقیاس‌های بزرگ خلاصه‌سازی شده است.

جدول (۱) مقایسه کارهای مرتبط

روش	مجموعه داده	ساخت لغت نامه	تحلیل حسی	تکنولوژی مورد استفاده	ذخیره سازی عظیم	پردازش عظیم
[۱۳]	توییتر	*	*	MapReduce HBase Mahout	*	*
[۱۶]	نظرات هتل‌ها		*	HBase KNIME	*	
[۱۱]	داده‌های آموزشی		*	HDFS	*	
[۱۲]	Answer Yahoo!		*	Traditional		
[۲۳]	آمازون		*	SPARK		*
روش پیشنهادی	پیکره فارسی دیجیکالا		*	MapReduce HDFS	*	*

۳. مواد و روش انجام تحقیق

در این تحقیق، سعی بر آن است تا با استفاده از امکاناتی که تکنولوژی داده‌های عظیم فراهم آورده است چارچوبی معرفی نماییم که قابلیت انجام نظرکاوی بر روی داده‌های عظیم نظری را داشته باشد. در این راستا داده‌های مورد استفاده، الگوریتم‌های مورد نظر و تکنولوژی‌های مناسب را معرفی خواهیم نمود.

۳,۱. مجموعه داده و منابع

مجموعه داده مورد استفاده پیکره Digikala2013 [1] می‌باشد که نظرات موجود در زیل محصولات موجود در سایت دیجی کالا[2] را جمع‌آوری نموده است [۲۱]. این پیکره شامل ۳۱,۷۳۰ نظر به زبان فارسی درباره ۱۱ گروه محصول تجاری مختلف می‌باشد. در کنار این پیکره، از لغت‌نامه‌ای[3] استفاده خواهد شد که حاوی حدود ۶۴,۰۰۰ لغت حسی همراه با بار حسی می‌باشد و به شکل دستی توسط چندین کاربر خبره ایجاد شده است و از کارایی مناسبی برخوردار است [۲۱و۲۲].

لازم به ذکر است همانطور که پیشتر ذکر شد، مجموعه داده‌ای که مدیریت و پردازش آن با ابزارها و روش‌های معمول غیرممکن و یا کاملا ناکارامد باشد را می‌توان داده عظیم در نظر گرفت. از این رو مجموعه داده مذکور گرچه حجمی در مقیاس‌های ترا و پتا ندارد، اما از آنجا که اجرای الگوریتم‌های نظرکاوی به شکل معمول برای آن، روزها و شاید هفته‌ها زمان لازم داشته باشد، مطرح کردن مساله داده‌های عظیم و استفاده از تکنولوژی داده‌های عظیم برای آن به جا و در خور توجه می‌باشد. علاوه بر این، بردن مساله در فضای داده‌های عظیم پتانسیل گسترش ابعاد مساله به مقیاس‌های کلان را به ارمغان خواهد آورد.

از آنجا که قصد داریم چارچوبی پیشنهاد دهیم که قابلیت تحلیل حسی نظرات را در ابعاد بالا داشته باشد به سراغ

[1] https://github.com/Text-Mining/Persian-Sentiment-Resources
[2] www.digikala.com
[3] https://github.com/Text-Mining/Persian-Sentiment-Resources/blob/master/PersianSWN.csv

جنبه‌های خاص تکنولوژی داده‌های عظیم می‌رویم. به این منظور، از چارچوب هدوپ و سیستم فایل توزیع شده آن (HDFS) و مدل برنامه نویسی نگاشت_کاهش استفاده خواهیم نمود که پیشتر به معرفی آن‌ها پرداختیم. سیستم مورد استفاده برای نصب هدوپ و اجرای برنامه‌های نگاشت_کاهش، کلاستری شامل ۷ گره با ۱۶ گیگابایت حافظه اصلی و پردازنده‌های core i7 3.2 Ghz 4 core می‌باشد. نسخه هدوپ مورد استفاده، ۲,۵ و پهنای باند شبکه ۱۰۰ مگابیت بر ثانیه است. سیستم عامل مورد استفاده اوبونتو نسخه ۱۴,۱۰ می‌باشد.

۳,۲. روش پیشنهادی

همانگونه که قبلا ذکر شد در نظرکاوی، سطوح تحلیل مختلفی وجود دارد. در این تحقیق، دو روش مختلف نظرکاوی در بستر داده‌های عظیم مورد استفاده قرار خواهند گرفت :

- **الگوریتم ۱- نظرکاوی سطح سند مبتنی بر لغت‌نامه**

در این مرحله با استفاده از لغت‌نامه حسی که در دسترس داریم به تحلیل حسی متن مورد نظر می‌پردازیم. لغت‌نامه، حاوی دو گروه واژه می‌باشد: گروه واژه‌های مثبت و گروه واژه‌های منفی. در هر گروه، میزان حس بیان شده نیز ذکر گردیده است. در صورتیکه متن نظری، حاوی یکی از لغات موجود در لغت‌نامه حسی باشد، بار حسی آن، متناظر با مثبت یا منفی بودن واژه، لحاظ می‌شود. بار حسی مثبت کل متن بر اساس بار حسی لغات مثبت آن محاسبه می‌شود. همچنین بار حسی منفی کل متن بر اساس بار حسی لغات منفی آن محاسبه می‌گردد. سپس بار حسی کل متن حاصل تفاضل این دو مقدار (رابطه (۱)) و یا حاصل تقسیم آن‌ها (رابطه (۲)) خواهد بود. در نتیجه خواهیم داشت:

$$SO_{text} = \sum_{w \in pw \cap text} SO_w - \sum_{w \in nw \cap text} SO_w \quad (۱)$$

$$SO_{text} = \frac{\sum_{w \in pw \cap text} SO_w}{\sum_{w \in nw \cap text} SO_w} \quad (۲)$$

در این دو معادله SO_{text} به معنی بار حسی کلی متن مورد نظر و SO_w به معنی بار حسی کلمه می‌باشد. pw و nw به ترتیب اشاره بر مجموعه لغات مثبت و منفی لغت‌نامه دارد. در این پژوهش، از تفاضل بار حسی مثبت و منفی، معادله (۱)، استفاده شده است.

- **الگوریتم ۲- نظرکاوی سطح سند مبتنی بر برچسب‌های اجزای سخن**

از آنجا که ممکن است تمامی اجزا یک جمله حاوی بار حسی نباشند بسیاری از محققین پیشنهاد داده اند تنها از بخشی از اجزاء جمله که پتانسیل داشتن بار حسی را دارند به منظور تحلیل حسی استفاده شود. از این رو یکی از روش‌های متداول نظرکاوی بهره‌گیری از برچسب‌های اجزاء سخن مانند اسم، فعل، صفت و غیره می‌باشد. به این معنا که از الگوهای نحوی که احتمالا حاوی نظر هستند استفاده می‌شود. به این ترتیب که چند الگوی کلیدی نحوی را در نظر گرفته و در صورتیکه سند نظری حاوی این الگوها باشد، بار حسی الگو محاسبه شده و با بار حسی کل جمع می‌گردد [۱۲]. به منظور برچسب‌زنی مجموعه داده مورد استفاده، از ابزار برچسب‌زن اجزای سخن تهیه شده در آزمایشگاه فن‌آوری وب دانشگاه فردوسی بهره گرفته شده است[1]. در [۱۲] برای محاسبه بار حسی الگو از وابستگی آماری لغات استفاده می‌شود، اما از آنجا که برای ما این معیار در دسترس نبود، از لغت‌نامه استفاده کردیم. به این ترتیب که اگر واژه‌های موجود در الگوی پیدا شده در لغت‌نامه حاوی بار حسی باشند، بار حسی آن واژه لحاظ می‌گردد. الگوهای مورد استفاده با الهام از [۱۲] استخراج شده و در جدول (۲) نمایش داده شده‌اند.

جدول (۲): الگوهای نحوی مورد استفاده برای نظرکاوی

کلمه اول	کلمه دوم	کلمه سوم
اسم/گروه اسمی	صفت	-
قید	صفت	-
صفت	حرف ربط	صفت
صفت	فعل	-

[1] http://wtlab.um.ac.ir/index.php?option=com_content&view=article&id=326&Itemid=224&lang=en

به این ترتیب، بار حسی متن مورد نظر بر اساس بار حسی لغات موجود در الگوهای نحوی موجود در جمله به دو روش بیان شده در رابطه (۳) و (۴) می‌تواند سنجیده شود.

$$SO_{text} = \sum_{w \in pw \cap spt} SO_w - \sum_{w \in nw \cap spt} SO_w \qquad (۳)$$

$$SO_{text} = \frac{\sum_{w \in pw \cap spt} SO_w}{\sum_{w \in nw \cap spt} SO_w} \qquad (۴)$$

در این معادلات منظور از spt (syntax patterns of text) کلماتی از جمله است که با الگوهای نحوی مد نظر تطابق دارند. در این پژوهش از معادله (۳) استفاده شده است.

- **الگوریتم ۳- تبدیل روش‌های نظرکاوی به مدل برنامه نویسی نگاشت-کاهش به منظور پردازش موازی**

همانطور که قبلا ذکر شد، به منظور استفاده از امکانات پردازش موازی در هدوپ بایستی الگوریتم مورد نظر را در قالب توابع نگاشت و کاهش مدلسازی کرد. در نظرگرفتن این نکته ضروری است که برای این مدلسازی، ورودی‌های مساله باید به فرم جفت‌های کلید-مقدار تبدیل شوند. این جفت‌ها در فاز نگاشت بین گره‌های پردازشی هدوپ توزیع می‌شوند. سپس این جفت‌ها درهم‌آمیزی و بر اساس کلیدها مرتب‌سازی می‌شوند و در فاز کاهش روی جفت‌هایی که کلید مشترک دارند یک عملیات محاسباتی انجام می‌شود. خروجی این فاز از نگاشت-کاهش نیز مجددا یک مجموعه از جفت‌های کلید-مقدار است که بنا به شرایط مساله ممکن است به فاز بعدی نگاشت-کاهش داده شود.

نکته جالب توجه آن است که اغلب الگوریتم‌ها را می‌توان در قالب یک یا چند فاز نگاشت-کاهش مدلسازی کرد. البته برنامه‌نویس بایستی با هوشمندی و تسلط بر مفهوم نگاشت-کاهش، محاسبات مورد نظر خود را در فازهای مختلف بگنجاند و ورودی و خروجی‌های هر فاز را تنظیم نماید.

در ادامه، شبه کد این مدلسازی برای دو روش نظرکاوی در شکل های (۱) و (۲)، آورده شده است.

```
mapper1:
foreach value word w in opinion and lexicon
    if w is in opinion emit <w, r>
    if w is in lexicon emit <w, SO>
reducer1:
foreach key word w
    find the SO of the w in lexicon
    SO = SO * w's frequency in the opinion
    emit <w, SO>
```

```
mapper2:
foreach pair <key, value> <w, SO> from reducer1
    is SO < 0
        emit <Negative, SO>
    else
        emit <Positive, SO>
reducer2:
    foreach key k
    emit <k,sum(values)>
```

شکل (۱): نظرکاوی سطح سند مبتنی بر لغت نامه با مدل برنامه نویسی نگاشت-کاهش

```
mapper1:
foreach value word w in POS taged opinion and lexicon
    if w is in lexicon emit <w, SO>
    else
        find matches for the given patterns
        foreach match emit <w, r>
reducer1:
foreach key word w
    find the SO of the w in lexicon
    SO = SO * w's frequency in the opinion
    emit <w, SO>
```

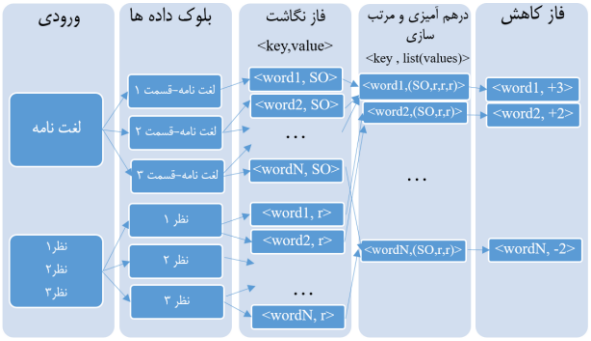
```
mapper2:
foreach pair <key, value> <w, SO> from reducer1
    is SO < 0
        emit <Negative, SO>
    else
        emit <Positive, SO>
reducer2:
    foreach key k
    emit <k,sum(values)>
```

شکل (۲): نظرکاوی سطح سند مبتنی بر برچسب‌های اجزاء سخن با مدل برنامه‌نویسی نگاشت-کاهش

همانگونه که مشاهده می‌شود، هر کدام از روش‌های نظرکاوی مورد استفاده در مدل نگاشت_کاهش دارای دو فاز نگاشت_کاهش می‌باشد. در ادامه چگونگی این مدلسازی شرح داده شده است.

برای روش اول پیشنهادی، یعنی نظرکاوی سطح سند مبتنی بر لغت‌نامه، هدف آن است که در اسناد نظری موجود، کلمات حسی موجود در لغت‌نامه جستجو شود. سپس با توجه به بار حسی کلمات لغت‌نامه که در سند نظری حضور داشته اند، بار حسی متن با توجه به فرمول (۱) یا (۲) محاسبه گردد. برای این کار دو فاز نگاشت-کاهش طراحی می‌گردد. در فاز اول کلمات حسی متن با استفاده از لغت‌نامه استخراج می‌شوند و مشخص می‌شود هر کلمه حسی با توجه به فراوانی‌اش در

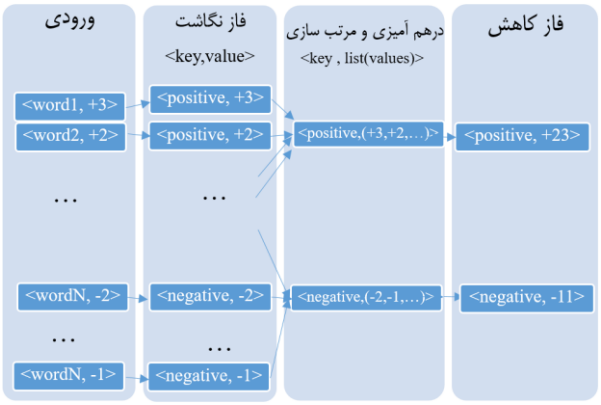
متن، چه تاثیری در بار حسی کل متن خواهد داشت. در فاز دوم، بار حسی متن با توجه به بار حسی لغات حسی در متن، محاسبه می‌گردد. در شکل (۳) معماری فاز اول نگاشت_کاهش معرفی شده در شکل (۱) نشان داده شده‌است. تمامی کلمات لغت‌نامه به همراه کلمات متون نظری به جفت-های کلید-مقدار نگاشته می‌شوند. برای کلید، خود کلمه در نظر گرفته خواهد شد و مقدار، برای کلمات لغت‌نامه، همان درجه حسی آن لغات و برای کلمات متن نظر مقدار ثابتی مانند 'r' در نظرگرفته می‌شود. این کار در واقع یک نکته طراحی کلیدی برای این فاز از نگاشت-کاهش می‌باشد. در مرحله‌ی مرتب‌سازی برای هر کلمه لیستی از مقادیر متناظر تهیه می‌شود. و در فاز کاهش، عملیات لازم روی آن لیست انجام می‌گیرد. به عنوان مثال اگر کلمه‌ای که بار حسی ۱+ داشته باشد و در متن نظری ۳ بار ظاهر شود، در فاز مرتب‌سازی دارای لیستی با ۴ عضو (یکی مربوط به لغت نامه و ۳ تا مربوط به متن نظری) می‌باشد. و در فاز کاهش بار حسی لغت مربوط به لغت‌نامه در فرکانس کلمات نظری ضرب می‌شود. برای مثال ذکر شده، مقدار ۳+ محاسبه می-شود.



شکل (۳): معماری پیشنهادی برای فاز اول نگاشت_کاهش، مربوط به شکل (۱)

خروجی فاز اول نگاشت-کاهش به ورودی نگاشت در فاز دوم می‌رود. شکل (۴) معماری این فاز را نشان می‌دهد. در این مرحله، مجموعه‌ای از کلمات نظری و بار حسی آنها را داریم و هدف آن است که بار حسی کلی متن استخراج شود. از آنجا که به دنبال محاسبه بار حسی مثبت و منفی هستیم، در این فاز، وظیفه نگاشت، جداسازی جفت‌های مثبت و منفی است. در نتیجه دو کلید 'positive' و 'negative' اتخاذ می-گردد. وظیفه کاهش در این فاز، محاسبه مجموع بار مثبت و منفی برای این دو کلید می‌باشد. این کار به این شکل انجام می‌شود که هر ورودی یک کلید دارد که نشان دهنده هر کلمه از متن نظری است، به همراه یک مقدار که بار حسی آن کلید را در متن نظری نشان می‌دهد. در فاز نگاشت اگر آن کلمه یک کلمه مثبت باشد (مقدار، مثبت باشد) به جفتی نگاشته می‌شود که کلید آن، کلمه positive و مقدار آن همان مقدار حسی خواهد بود. به همین ترتیب اگر کلمه مربوطه بار منفی داشته باشد به کلید negative نگاشته می‌شود. پس از در هم‌آمیزی و مرتب‌سازی، دو لیست خواهیم داشت، یک لیست از جفت‌هایی با کلید positive و یک لیست از جفت‌هایی با کلید negative. در مرحله کاهش، بار مثبت کلی و بار منفی کلی محاسبه می‌شود.

معماری نگاشت-کاهش برای روش دیگر پیشنهادی (یعنی نظرکاوی مبتنی بر برچسب‌های اجزاء سخن) هم مشابه روش دیگر دارای دو فاز نگاشت کاهش می‌باشد. در فاز اول در مرحله نگاشت، تمامی کلمات موجود در لغت‌نامه به همراه کلماتی از متون نظری که در الگوهای نحوی صدق می‌کنند، با مکانیزمی شبیه روش قبل به جفت‌های کلید-مقدار مدل می-شوند. بقیه مراحل تا انتهای فاز دوم نگاشت-کاهش شبیه روش قبل است؛ لذا از ذکر مجدد آن‌ها خودداری می‌کنیم.



شکل (۴): معماری پیشنهادی برای فاز دوم نگاشت_کاهش، مربوط به شکل (۱)

۴. نتایج اجرا

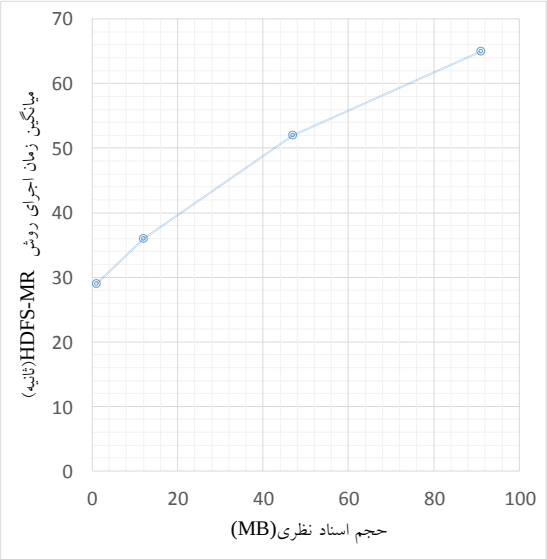
به منظور بررسی دقیق‌تر استفاده از تکنولوژی داده‌های عظیم، برای هر کدام از روش‌های مورد استفاده برای نظرکاوی، اندازه‌های متفاوتی برای فایل ورودی مورد پردازش در نظر گرفته شد. همانطور که قبلا ذکر شد برخی از کارهای مرتبط به منظور استفاده از سیستم کارای ذخیره و بازیابی سیستم فایل توزیع شده هدوپ پیشنهاد دادند که ذخیره‌سازی داده‌ها بر روی HDFS انجام گیرد [۱۶]. این روش به عنوان روش پایه برای آزمایش‌های این پژوهش در نظر گرفته شد (HDFS-based). در حالت دیگر علاوه بر استفاده از HDFS به عنوان سیستم ذخیره‌سازی داده، استفاده از روش پردازشی کارای نگاشت-کاهش پیشنهاد می‌شود که روش پیشنهادی این تحقیق می‌باشد (HDFS-MR).

به منظور ارزیابی روش نظرکاوی مربوطه، حس پیش‌بینی شده توسط سیستم نظرکاوی (مثبت و منفی) با حس واقعی که به صورت برچسب داده در مجموعه داده مشخص است مقایسه می‌گردد. برای مجموعه داده مذکور، دقت (accuracy) یعنی نسبت تعداد برچسب درست پیش بینی شده به تعداد کل، برای الگوریتم مبتنی بر لغت‌نامه ۶۸٪ و دقت الگوریتم مبتنی بر برچسب‌های اجزا سخن ۷۱٪ می‌باشد. لازم به ذکر است الگوریتم‌های ذکر شده برای نظرکاوی در بخش ۳ الگوریتم‌های مورد قبولی هستند و با توجه به کیفیت لغت‌نامه مورد استفاده دقت مناسبی را حاصل می‌نمایند. واضح است این دقت ارتباطی به نحوه پیاده‌سازی ندارد، و به همین دلیل، دو مدل پیاده‌سازی در روش‌های HDFS-based و HDFS-MR دارای دقت‌های یکسان می‌باشند.

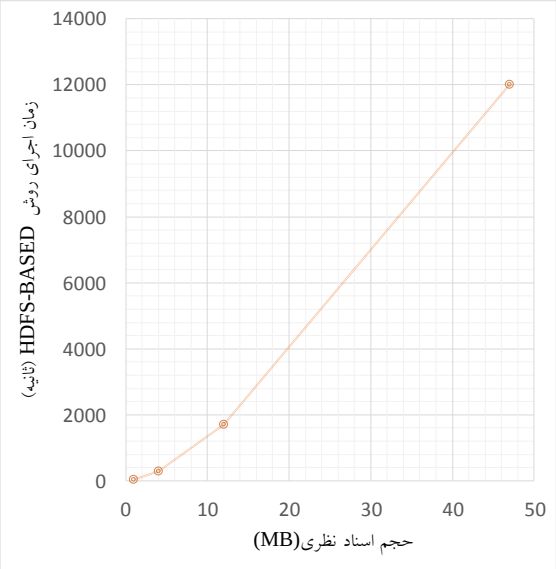
هدف اصلی از این پژوهش ارزیابی کارایی استفاده از تکنولوژی داده‌های عظیم در نظرکاوی می‌باشد. به این دلیل در ادامه، اجرای دو پیاده‌سازی مذکور را در حالات مختلف مورد مقایسه و بحث قرار می‌دهیم.

۴,۱. نظرکاوی سطح سند مبتنی بر لغت‌نامه

در شکل های (۵) و (۶) میانگین زمان اجرای برنامه بر روی اندازه‌های مختلف فایل ورودی حاوی سند نظری در دو روش مذکور نشان داده شده است. از آنجایی که بازه زمانی مشاهده شده در دو روش HDFS-based و HDFS-MR مقیاس‌های کاملا متفاوتی دارند، به منظور نمایش بهتر، نمایش آن‌ها را در نمودارهای مجزا انجام دادیم.



شکل (۵): میانگین زمان اجرای الگوریتم ۱ به روش HDFS-MR بر روی اندازه‌های مختلف ورودی



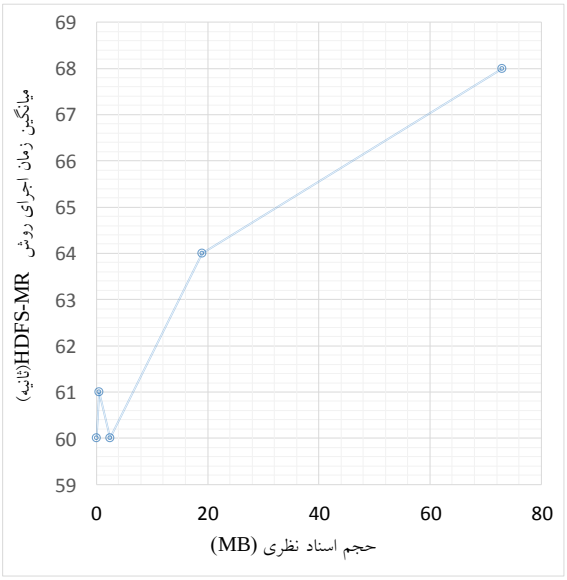
شکل (۶): زمان اجرای الگوریتم ۱ به روش HDFS-BASED بر روی اندازه‌های مختلف ورودی

همانگونه که مشاهده می‌شود استفاده تنها از سیستم فایل توزیع شده هدوپ بدون استفاده از امکانات پردازش و اجرای موازی از طریق نگاشت-کاهش گر چه برای حجم‌های کم داده به شکل کاراتری عمل کرده است، اما برای حجم بالا زمان‌های خیلی بالاتری را گزارش می‌کند. برای اندازه‌های بزرگ فایل زمان اجرای HDFS-MR به شکلی چشم‌گیر و با شیبی نسبتا تند افزایش می‌یابد. دلیل این امر آن است که گر چه سیستم فایل هدوپ به شکل توزیع شده داده‌ها را در گره‌ها پخش و مدیریت می‌کند و این موضوع برای مسائلی که با نگهداری و بازیابی داده و تغییر فرمت آن به عنوان مساله اصلی برخورد می‌کنند اهمیت دارند [۲۳]، اما کارایی این ذخیره و بازیابی کارا برای مسایلی که بار پردازشی بالا دارند چندان مشهود نیست و اگر پردازش و اجرای الگوریتم به شکل موازی انجام نشود، مثلا از روش نگاشت-کاهش استفاده نشود، کارایی آن، شبیه اجرای سری خواهد بود.

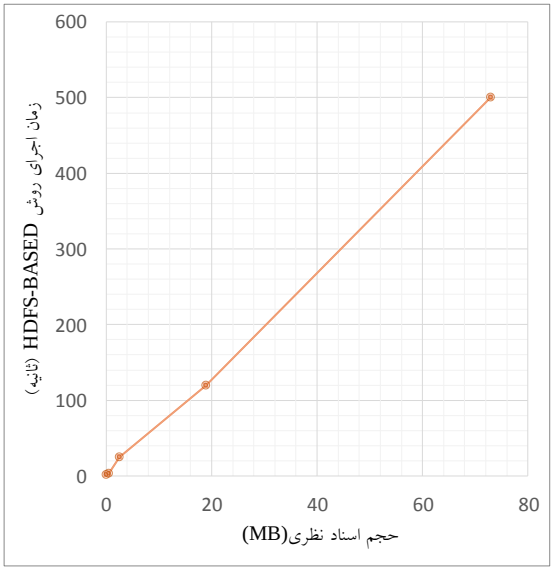
نکته جالب توجه این است که در زمان اجرای HFDS-MR برای اجرای برنامه نظرکاوی سطح سند مبتنی بر لغت‌نامه، برای حجم کوچک داده نوسانات زیادی مشاهده می‌شود. به شکلی که به عنوان مثال برای یک متن نظری در یک فایل ۲۰ کیلو بایتی اجراهای متوالی برنامه زمان اجراهایی بین ۱۸ تا ۴۰ ثانیه را گزارش می‌دهد. این زمان، علاوه بر آنکه دارای نوسان است، برای این حجم داده نسبتا زیاد می‌باشد. از این امر می‌توان اینگونه برداشت کرد که به دلیل وجود سربار آماده‌سازی اولیه گره‌ها و توزیع و تبادلات داده‌ها در فازهای نگاشت و کاهش، در حجم‌های کم داده، زمان سربار اضافه، خود را بیشتر نشان می‌دهد. علاوه بر این چون یک مرحله تصادفی در هر فاز پردازشی نگاشت-کاهش داریم (مرحله در هم‌آمیزی) یک میزان عدم قطعیت در هر اجرا وجود دارد و این، در حجم‌های کم بیشتر خود را نشان می‌دهد.

۴،۲. نظرکاوی سطح سند مبتنی بر برچسب‌های اجرای سخن

روش دوم ارائه شده برای نظرکاوی هم با مدل برنامه نویسی نگاشت_کاهش پیاده‌سازی و بر روی هدوپ اجرا شد. زمان اجرای این الگوریتم به دو روش HDFS-based و HDFS-MR بر روی اندازه‌های مختلف ورودی در شکل های (۷) و (۸) مشاهده می‌شود.



شکل (۷): میانگین زمان اجرای الگوریتم ۲ به روش HDFS-MR بر روی اندازه‌های مختلف ورودی

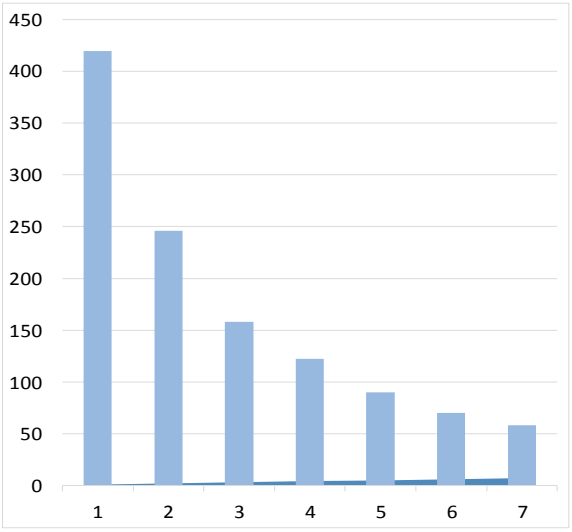


شکل (۸): میانگین زمان اجرای الگوریتم ۲ به روش HDFS-BASED بر روی اندازه‌های مختلف ورودی

۴،۳. تاثیر تعداد گره‌ها بر سرعت اجرا

در این بخش هدف آن است تا تاثیر تعداد گره‌ها در کارایی روش پیشنهادی مورد مطالعه قرار گیرد. همان طور که قبلا

اشاره شد در هدوپ گره نام بایستی داده‌ها را بین گره‌های داده توزیع نماید. سپس محاسبات به شکل همزمان در گره‌های داده انجام شده و نتایج هر گره باید تجمیع گردد. گرچه انجام محاسبات به شکل همزمان سرعت را بالا می‌برد اما سربار ناشی از هماهنگی‌ها و ارسال و دریافت پیام‌ها مساله حائز اهمیتی است. به این شکل که نباید انتظار داشت با افزایش تعداد گره ها الزاما زمان اجرا همچنان کاهش چشمگیری داشته باشد. چرا که هر چقدر که تعداد گره‌ها افزایش یابد، سربار هماهنگ‌سازی و ارسال و دریافت پیام‌ها هم بیشتر خواهد بود. در شکل (۹) زمان اجرای الگوریتم ۲ بر روی کلاسترهایی با تعداد گره‌های مختلف (از ۱ تا ۷) برای اندازه ورودی ۶۰ مگابایت نشان داده شده است. همانطور که مشاهده می‌شود با افزایش تعداد گره‌ها، گرچه زمان اجرا کاهش می‌یابد، اما به دلیل سربارهای مذکور شیب کمتری در نمودار مشاهده می‌شود.



شکل (۹): مقایسه زمان اجرای الگوریتم ۲ با تعداد مختلف گره‌های کلاستر

۵. نتیجه‌گیری و کارهای آتی

در این پژوهش به بررسی اثر استفاده از تکنولوژی داده‌های عظیم در نظرکاوی بر روی داده‌های زبان فارسی پرداختیم. هدوپ به عنوان یکی از بسترهای مناسب برای ذخیره‌سازی و پردازش حجم عظیمی از داده‌ها مورد استفاده قرار گرفت. به منظور بهره‌گیری از توان پردازشی هدوپ دو الگوریتم کارامد برای نظرکاوی مبتنی بر مدل برنامه نویسی نگاشت-کاهش ارائه شد. دو روش پیشنهادی هر کدام تنها با دو فاز نگاشت-کاهش طراحی و پیاده سازی شدند.

مشاهده شد در حالیکه استفاده تنها از مکانیزم ذخیره‌سازی هدوپ استفاده شود و از مکانیزم پردازش موازی استفاده نشود مجبور به انجام پردازش‌ها به صورت سری هستیم. از آنجا که برای هدف ما، محاسبات نقش مهمی دارند، برای حجم بالای داده، استفاده از پردازش به شکل موازی به کمک برنامه نویسی نگاشت-کاهش کاراتر می‌باشد.

استفاده از روش نگاشت-کاهش به حدی کاراست که برای حجم‌های بزرگ داده با افزایش حجم داده، منجر به افزایش چشمگیری در زمان اجرا نمی‌شود و همان‌طور که مشاهده شد نمودار مربوطه شیب بسیار ملایمی دارد.

همچنین مشاهده شد برای حجم‌های کوچک داده گاهی روش نگاشت-کاهش به خوبی عمل نمی‌کند. علاوه بر این، نوساناتی در طی اجراهای متوالی روش نگاشت-کاهش برای یک برنامه شاهد هستیم. دلیل این امر به ساختارهای درونی هدوپ و معماری مدل نگاشت_کاهش برمی‌گردد که پیشتر به آن اشاره شد. دامنه این نوسانات برای داده‌های کوچک خیلی بیشتر می‌باشد. به این دو دلیل، استفاده از روش نگاشت-کاهش برای حجم‌های کم داده توصیه نمی‌شود.

با مقایسه اجراهای دو الگوریتم نظرکاوی ارائه شده (مبتنی بر لغت‌نامه و برچسب‌های اجزای سخن) مشاهده می‌شود با اینکه تغییر الگوریتم تاثیر زیادی در زمان اجرای سری دارد، (روش مبتنی بر برچسب‌های اجزای سخن بسیار سریع‌تر از روش مبتنی بر لغت‌نامه است)؛ اما این تغییر در روش نگاشت-کاهش چندان مشهود نیست. می‌توان اینگونه نتیجه گرفت، در صورتی‌که تکنولوژی داده‌های عظیم در دسترس باشد، نیازی به صرف هزینه برای طراحی الگوریتم‌های کارا نیست؛ چرا که این تکنولوژی آنقدر در کارایی موثر هست که تاثیر طراحی الگوریتم را ناچیز می‌کند. این در حالیست که طراحی الگوریتم در حالت سری می‌تواند تاثیر زیادی بر

کارایی برنامه داشته باشد.

یکی از مزایای بزرگ مدلسازی مساله در چارچوب هدوپ و نگاشت-کاهش، کاربرد آن در تشخیص خطا در داده است. همانگونه که ذکر شد، یکی از چالش‌های داده‌های عظیم مقادیر آنها است [۹]. در اغلب کاربردهای کاوش اطلاعات ابتدا بایستی پیش‌پردازشی بر روی داده‌ها به منظور پاکسازی داده‌ها و تضمین کیفیت داده و جلوگیری از بروز خطا در مراحل بعدی صورت پذیرد. همین مساله با اینکه جزو پردازش اصلی نیست زمانبر می‌باشد و برای داده‌های عظیم چالشی جدی محسوب می‌شود. این مساله درحین آزمایشات این پژوهش به وضوح نشان داده شد به شکلی که مشاهده شد در حالیکه برطرف کردن باگ‌های داده به شکل معمول برای حجم بالای داده چندین روز به طول می انجامد، با استفاده از روش نگاشت_کاهش در عرض چند ثانیه خطاهای داده شناسایی و مجموعه داده پاکسازی شد.

به طور خلاصه مشاهدات صورت گرفته با در نظر گرفتن جنبه های نظری تکنولوژی داده‌های عظیم توجیه‌پذیر می‌باشند. با فراهم شدن بستری برای پردازش توزیع شده در هدوپ و تقسیم کارها بین واحدهای پردازشی، زمان پردازش کاهش چشمگیری خواهد داشت. از سوی دیگر ارتباطات و تعاملات بین واحدهای پردازشگر و مدیریت آنها مستلزم صرف زمان سربار خواهد بود. هر چقدر زمان پردازش هر یک از واحدهای پردازشی بیشتر باشد، این زمان سربار، تاثیر کمتری در زمان کل نشان خواهد داد. در نتیجه برای زمانیکه با حجم بالای پردازش مواجه باشیم استفاده از ساختارهای پردازش توزیع شده کارایی بیشتری نشان خواهد داد. در عوض برای حجم کم پردازشی، استفاده از ساختارهای توزیع شده با توجه به زمان سربار مقرون به صرفه نخواهد بود.

مورد مشاهده شده دیگر این بود که هدوپ نسبت به تعداد فایل ورودی حساس بوده و در صورت تعدّد فایل‌های ورودی، افت کارایی خواهد داشت. در نتیجه در صورتیکه بخواهیم الگوریتم‌های نظرکاوی را برای تعداد سند زیاد به طور جداگانه انجام دهیم (مثلا برای انجام کارهای آماری روی نظرات افراد مختلف) استفاده از HDFS مناسب نخواهد بود. یک راه حل آن است که نظرات را در پایگاه داده HBase قرار داده و ورودی مرحله نگاشت، داده‌های جدول HBase باشد. این مدل کاوش، در کارهای آتی این تحقیق قرار دارد.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[۵] فرهمندپور ز.، نیک مهر ه.، منصوری زاده م.، طیب‌زاده قمصری ا.، «یک سیستم نوین هوشمند تشخیص هویت نویسنده فارسی زبان بر اساس سبک نوشتاری»، مجله محاسبات نرم، جلد ۱، شماره ۲، ص ۳۵-۲۶، ۱۳۹۱.

[6] Park D., Lee J., Han I., "The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement", International Journal of Electronic Commerce, 11(4):125-148, 2007.

[7] Sehgal D., Agarwal A.K., "Real-time Sentiment Analysis of Big Data Applications Using Twitter Data with Hadoop Framework", Soft Computing: Theories and Applications, Springer, Singapore, pp. 765-772, 2018.

[1] Yadav A., Vishwakarma D. K., "Sentiment analysis using deep learning architectures: a review". Artificial Intelligence Review, 53(6): 4335-4385. 2020.

[2] Shayaa S., Jaafar N.I., Bahri S., Sulaiman A., Wai P.S., Chung Y.W., Piprani A.Z., Al-Garadi M.A., "Sentiment analysis of big data: Methods, applications, and open challenges". IEEE Access, 6:37807-37827, 2018.

[3] Hasan M.M., Popp J., Oláh J., "Current landscape and influence of big data on finance". Journal of Big Data, 7(1):1-17, 2020.

[4] Liu B., Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge university press, 2020.

[8] Mihanović A., Gabelica H., Krstić Ž., "Big data and sentiment analysis using KNIME_Online reviews vs. social media", In Information and Communication Technology, Electronics and Microelectronics, pp. 1464-1468, 2014.

[9] Cui Y., Kara S., Chan K. C., "Manufacturing big data ecosystem: A systematic literature review". Robotics and computer-integrated Manufacturing, 62:101861, 2020.

[10] Dean J., Ghemawat S., "Mapreduce: Simplified data processing on large clusters", Communications of the ACM 51(1):107-113, 2004.

[11] Pang B., Lee L., Vaithyanathan S., "Thumbs up? sentiment classification using machine learning techniques". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86, 2002.

[12] Turney P., "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews". In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424, 2002.

[13] Kucuktunc O., Cambazoglu B., Weber I., Ferhatosmanoglu H., "A large-scale sentiment analysis for Yahoo! Answers", Proceedings of the fifth ACM international conference on Web search and data mining, ACM. pp. 633-642, 2012.

[14] Khuc V., Shivade C., Ramnath R., Ramanathan J., "Towards building large scale distributed systems for twitter sentiment analysis", In Proceedings of the 27th annual ACM symposium on applied computing, pp. 459-464, 2012.

[15] Dipty S., "Study of Sentiment Analysis Using Hadoop", Big Data Analytics. Springer, Singapore, pp. 363-376, 2018.

[16] Jena R.K., "Sentiment mining in a collaborative learning environment: capitalising on big data". Behaviour & Information Technology, 38(9): 986-1001, 2019.

[17] Zahedi E., Baniasadi Z., Saraee M., "A distributed joint sentiment and topic modeling using Spark for big opinion mining". In Electrical Engineering (ICEE), Iranian Conference on. IEEE, pp. 1475-1480, 2017.

[18] Lin C., He Y., Everson R., Ruger S., "Weakly supervised joint sentiment-topic detection from text". IEEE Transactions on Knowledge and Data engineering, 24(6):1134-1145, 2012.

[19] Benedetto F., Tedeschi A., "Big Data Sentiment Analysis for Brand Monitoring in Social Media Streams by Cloud Computing", Sentiment Analysis and Ontology Engineering, Springer International Publishing, pp. 341-377, 2016.

[۲۰] هراتیان اول ن.، صفائی ع.، «کشف سرویس‌های ابری در زبان فارسی از طریق تکامل هستان شناسی»، مجله محاسبات نرم، جلد ۴، شماره ۲، ص ۹۳-۸۴، ۱۳۹۴.

[۲۱] عسگریان ا.، کاهانی م.، شریفی ش.، «حس‌نگار: شبکه واژگان حسی فارسی»، پردازش علائم و داده‌ها دوره ۱۵، شماره ۱-پ ۱۵، ص ۸۶-۷۱، ۱۳۹۷.

[22] Asgarian E., Kahani M., Sharifi S., "The Impact of Sentiment Features on the Sentiment Polarity Classification in Persian Reviews", Cognitive Computation, pp. 1-19, 2017.

[۲۳] نجفی ح.، دانش پور ن.، «بهبود فرآیند استخراج، تبدیل و بارگذاری در پایگاه داده تحلیلی با کمک پردازش موازی»، مجله محاسبات نرم، جلد ۴ شماره ۲، ص ۳۱-۱۸، ۱۳۹۴.