

بررسی تکنیک‌های خوشه بندی در داده‌های حجیم

اکرم گل محمدی^۱، سید دانیال عزیزاده جواهری^۲ و رضا قائمی^۳

^۱ دانشجوی دکتری، گروه کامپیوتر، واحد نیشابور، دانشگاه آزاد اسلامی، نیشابور، ایران،

golmohammadi.akram@gmail.com

^۲ دانشجوی دکتری، گروه کامپیوتر، واحد نیشابور، دانشگاه آزاد اسلامی، نیشابور، ایران، itcjavaheerii@gmail.com

^۳ استادیار گروه کامپیوتر، واحد قوچان، دانشگاه آزاد اسلامی، قوچان، ایران، r.ghaemi@iau.ac.ir

چکیده-داده‌های حجیم پتانسیل بسیار خوبی در بهینه سازی داده‌ها، تصمیم گیری، آشکار سازی روندهای تجاری در زمینه‌های مختلفی از قبیل تولید، امور مالی و فناوری اطلاعات را نشان داده است. برنامه‌های کاربردی داده‌های حجیم می‌توانند از میان حجم بزرگ داده‌ها تراکنش‌ها را با مقیاس پذیر بالا، سرعت بالا و قابلیت اطمینان بالا انجام دهند. منظور از داده‌های حجیم داده‌هایی است که بسیار بزرگ، پویا و پیچیده با درجه خاصی از دقت هستند. به همین دلیل، تجزیه و تحلیل داده‌ها با استفاده از تکنیک‌های سنتی به خاطر پیچیدگی بالا و هزینه محاسباتی دشوار می‌شود. در این مقاله، یک مرور کلی از تکنیک‌های مختلف خوشه بندی استفاده شده برای تجزیه و تحلیل داده‌ها ارائه می‌شود. بررسی‌ها نشان می‌دهد روش نگاشت و کاهش به طور بالقوه برای خوشه بندی داده‌های حجیم بسیار مفید است و می‌تواند یک بستر بسیار مناسب برای پیاده سازی الگوریتم‌های موازی ارائه دهد.

واژگان کلیدی: داده‌های حجیم- تکنیک‌های خوشه بندی- نگاشت و کاهش- چند ماشین- تحلیل داده

۱- مقدمه

که از آن می‌توان حجم عظیمی از اطلاعات ناشناخته، ناشناس، بلا استفاده استفاده کرد، از الگوهای و دانش آنها استفاده کرد [10-11]. حتی اگر این داده‌های گسترده به سؤال از الگوها و اطلاعات ناشناس در کنار جنبه‌های تحلیلی و آماری کمک می‌کنند، ذخیره. پردازش و بازیابی این داده‌های عظیم هنوز دشوار و وقت گیر است. فن آوری‌های جدید برای ذخیره و استخراج اطلاعات مفید از این حجم داده‌های حجیم مورد نیاز است زیرا کشف و استخراج اطلاعات و دانش مفید از این حجم داده دشوار است، از این رو، بانک‌های اطلاعاتی سنتی نمی‌توانند نیازهای کاربران را برآورده سازند [48].

داده‌های حجیم معمولاً به مجموعه‌ای از داده‌ها گفته می‌شود که از ابزار استخراج، پالایش، مدیریت و پردازش با ابزارهای استاندارد مدیریتی و بانک‌های اطلاعاتی فراتر می‌رود. به عبارت دیگر، اصطلاح داده‌های حجیم به داده‌هایی گفته می‌شود که از نظر حجم و تنوع پیچیده هستند. با این وجود، نمی‌توان آنها را با ابزارهای سنتی

جمع آوری مجموعه داده‌های گسترده، به عنوان داده‌های حجیم^۱ نیز شناخته می‌شود و کانون چشم انداز فناوری فعلی تبدیل شده است. تجزیه و تحلیل داده‌های حجیم به دست آوردن اطلاعات با ارزش از مجموعه داده‌هایی است که بسیار بزرگ، سازمان یافته و پردازش آنها با استفاده از تکنیک‌های سنتی داده کاوی کاری دشوار است و دارای طیف گسترده‌ای از برنامه‌ها از جمله، استنباط، پیش بینی، ارتباطات و... است [۱].

رشد چشمگیر و قدرتمند اینترنت و نفوذ فناوری در کلیه زمینه‌های زندگی بشر، تولید داده‌های عظیم را آغاز کرده است.

تعداد زیادی از داده‌ها دقیقه به دقیقه از منابع مختلف و خدمات مورد استفاده انسان و سایر ماشین آلات و اینترنت اشیاء تولید می‌شوند. این حجم عظیمی از داده‌ها می‌توانند به عنوان یک اطلاعات بالقوه در نظر گرفته شوند

^۱ Big Data