چهارمین کنفرانس بین المللی
مطالعات نوین درعلوم کامپیوتر و فناوری اطلاعات
4th International Conference
On New studies Of Computer And IT
مشهد
۰۵ اسفند ماه ۱۳۹۶
FEBRUARY 24 ,2018

مروری بر مفاهیم و انواع روشهای داده کاوی در کلان داده

مهدی یوسف زاده اقدم[۱]، مریم فرشچیان یزدی[۲]، الهه کاشانی[۳]، سید رضا کامل طباخ[۴]

۱-دانشجوی دکتری مهندسی کامپیوتر، دانشگاه آزاد اسلامی ، مشهد ، ایران

۲-دانشجوی دکتری مهندسی کامپیوتر، دانشگاه آزاد اسلامی ، مشهد ، ایران

۳-کارشناسی ارشد فناوری اطلاعات،مدرس مرکز اموزش علمی کاربردی مالیاتی مشهد، ایران

۴-استادیار گروه کامپیوتر، دانشکده مهندسی، دانشگاه آزاد اسلامی ،مشهد ، ایران

چکیده

با گسترش علوم در دنیای امروزی، حجم انبوهی از داده‌ها به وجود آمده است و در هر لحظه تعداد زیادی داده تولید می‌شود. جهت استخراج و کشف دانش از این داده‌ها، باید بتوان آنها را ذخیره و پردازش کرد. داده کاوی یکی از روش‌هایی است که اطلاعات مفید و روابط مخفی بین داده‌ها را استخراج می‌کند ولی به علت حجم بالا و ساختارهای متنوع داده‌های حجیم امروزی، نمی‌توان از این روش‌ها جهت استخراج دانش استفاده کرد. همچنین ذخیره سازی و پردازش چنین حجمی از داده‌ها با روش‌های معمول و قدیمی از نظر زمان و هزینه مقرون به صرفه نیست. بنابراین یا باید ساختار الگوریتم‌های داده کاوی تغییر کند و یا با روش‌های جدیدی جایگزین شوند. داده‌های حجیم به دو صورت دسته‌ای و جریان‌های در حال حرکت وجود دارند که باید بتوان با استفاده از موازی سازی سخت افزاری و نرم افزاری و پردازش‌های جریانی، اطلاعات مفید را از آنها استخراج کرد. در حال حاضر مهم‌ترین مدل برای پردازش داده‌های حجیم، مدل نگاشت-کاهش است که توسط شرکت‌های زیادی برای پردازش داده‌هایشان استفاده می‌شود. نسخه متن‌باز نگاشت-کاهش توسط هدوپ ارائه شد. در این مقاله ابتدا سیر تکاملی انواع پردازش‌ها روی داده‌های حجیم مورد بررسی قرار گرفته و سپس روش‌های تجزیه و تحلیل این داده‌ها معرفی شده است و در نهایت الگوریتم k-means که یکی از مهمترین روش‌های خوشه‌بندی است در محیط هدوپ پیاده‌سازی شده است.

واژگان کلیدی: داده های حجیم،داده کاوی،خوشه بندی،الگوریتم K-means،بستر هدوپ