تحليل البيانات الضخمة باستخدام محسن صقور هاريس و اطار تقليل الخريطة

إعداد

محمد قاسم البشابشه

إشراف

الدكتور ليث ابوعليقه

الملخص

يعد تجميع البيانات الكبيرة والتي هي أحد المشكلات الصعبة الحديثة والشائعة التي يتم استخدامها في تطبيقات مختلفة مثل الشبكات الاجتماعية والمعلوماتية الحيوية والعديد من التطبيقات الأخرى. تحتاج خوارزميات التجميع التقليدية إلى التحسين للتعامل مع أحجام البيانات المتزايدة. في هذا البحث تم اقتراح مُحسِّن صقور هاريس هوكس باستخدام إطار تقليل الخريطة للتعامل مع البيانات الضخمة، ويُطلق عليه محسن صقور هاريس كوكس الهجين. يستفيد محسن صقور هاريس هوكس الهجين المقترح من قدرة خوارزمية التجميع التقليدية على حل مشاكل التجميع، حيث تقوم خوارزمية التجميع التقليدية بتهيئة الحلول الأولى لمحسن صقور هاريس هوكس ومحسن صقور هاريس هوكس لتحسين الحلول المرشحة لإيجاد الحل الأمثل. إطار تقليل الخريطة هو نموذج برمجة متوازي يستخدم للمعالجة الموزعة ويقوم بإنشاء مجموعات بيانات ببرنامج موازٍ على مجموعة. يستخدم محسن هاريس هوكس الهجين إطار عمل تقليل الخريطة للتوازي لأنه يوفر التسامح مع الأخطاء وموازنة التحميل ومكان البيانات للتغلب على أوجه القصور هذه في خوارزمية التجميع التقليدية والتركيز على إيجاد المجموعات المثلى. واظهرت النتائج الدراسة ان الطريقة المقترحة جيده جداً في تحسين وإيجاد المجموعات المثلى والمتقاربة من البيانات مما يعكس فعالية الاطروحة. وتم الحكم على تقييم النتائج التي تم الحصول عليها من حيث مقاييس الدقة ومعدل الخطأ وتقييم النتائج المقترحة بناءً على بيئة محاكاة باستخدام بايثون. وتم استخدام مجموعة من مجموعات البيانات المعيارية القياسية للتحقق من أداء الطريقة المقترحة.

الكلمات المفتاحية: البيانات الضخمة، تقليل الخريطة، خوارزمية التجميع التقليدية، محسن هاريس هوكس.

Big Data Analysis using Hybrid Harris Hawks Optimizer and MapReduce Framework.

Prepared by:

Mohammad Qassem Bashabsheh

Supervised by:

Dr. Laith Abualigah

Abstract

Clustering large data is a recent and popular challenge that is used in various applications, including social networking, bioinformatics, and many others. In order to manage the rapidly growing data sizes, traditional clustering algorithms must be improved. In this research, a hybrid Harris Hawks Optimizer (HHHO) with K-mean clustering and MapReduce framework is proposed to solve the various data clustering problem. The proposed scheme uses the K-means' ability to solve the various clustering problems. More specifically, the K-means are utilized as initial solutions to the traditional Harris Hawks Optimizer (HHO). In

general, HHO tries to refine the candidate solutions to find the best one. MapReduce is a distributed processing computing paradigm that produces datasets using a parallel program on a cluster. In particular, it is adopted in the developed HHHO for parallelization since it offers fault tolerance, load balancing, and data locality. The performance of the presented methodology has been evaluated by means of numerical comparisons which proved the efficiency of the proposed HHHO, where it produces better results than other existing computation methods. Moreover, it has a very good ability in improving and finding optimal and converging sets of data. In addition, the accuracy and error rate of the obtained results are assessed. The proposed method is implemented and evaluated using PAITHON simulation settings.

Keywords: Big Data Analysis, Hybrid Harris Hawks Optimizer, MapReduce Framework.