تطوير نموذج تنقيب البيانات المتدفقة لتطبيقات البيانات الكبيرة

ابتسام حامد عبد الرحيم المالكي

بحث مقدم لنيل درجة الماجستير في العلوم
(علوم الحاسبات)

د. منال عبد الله

كلية الحاسبات وتقنية المعلومات
جامعة الملك عبد العزيز
جدة – المملكة العربية السعودية
رجب ١٤٣٩هـ – أبريل 2018 م

المستخلص

أن التقدم الهائل في تكنولوجيا الأجهزة في السنوات الأخيرة سمح لنا تلقائيا بتسجيل التعاملات التجارية الإلكترونية وغيرها من المعلومات في الحياة اليومية بمعدل سريع. هذه العمليات تولد كميات هائلة من البيانات عبر الإنترنت والتي تنمو بمعدل غير محدود. كما أن بعض التطبيقات مثل المراقبة باستخدام شبكات الاستشعار اللاسلكية تعتبر من مصادر التي تسجل كميات هائلة من البيانات. حيث يشار إلى هذا النوع من البيانات على شبكة الإنترنت على أنها تيارات البيانات.

وحديثا ايضاً تمت مناقشة قضايا ادارة وتحليل تيارات البيانات بشكل واسع نظراً لأهميتها واتساع انتشارها ووجود العديد من التطبيقات الناشئة في هذا المجال. وهناك العديد من المشاكل الهامة مثل التجميع والتصنيف تم دراستها على نطاق واسع في مجال تنقيب البيانات الكبيرة. ومع ذلك، فإن غالبية هذه الأساليب قد لا تعمل بشكل فعال على تيارات البيانات. وعلاوة على ذلك، تيارات البيانات تتطلب التنقيب عن البيانات على الانترنت بشكل مستمر وسريع. لا شك ان هناك عدد متزايد من التطبيقات التي تولد تيارات هائلة من البيانات، هذه الكمية الهائلة من البيانات تحتاج الى معالجة ذكية وتحليل على شبكة الانترنت مثل نظم مراقبة الوقت الحقيقي، ونظم الاتصالات السلكية واللاسلكية، وشبكات الاستشعار وغيرها من البيئات الديناميكية. لذلك باتت الحاجة وشيكة لتحويل هذه البيانات لمعلومات ومعارف مفيدة وذلك من خلال تطوير أطر، أنظمة وخوارزميات تهتم بمعالجة التحديات القائمة في مجال تدفق البيانات. التخزين، الاستعلام والتنقيب عن البيانات المتدفقة يعتبر من المهام والتحديات عالية الصعوبة حسابيا.

في هذا البحث سوف يتم التركيز على التنقيب في البيانات المتدفقة. التنقيب في تدفق البيانات يُعنى باستخراج هياكل المعارف الممثلة بنماذج وانماط معينه من بيانات متدفقة بسرعة واستمرار. في هذا البحث سوف يتم تطوير وتمثيل نموذج جديد للتنقيب في تدفق البيانات، وسيتم اعتماد نوعين من التقنيات الرئيسية للتنقيب في تدفق البيانات وهي إما تقنيات تعتمد على البيانات مثل العينات، تقنيات تقليل تكتل البيانات أو الرسوم. اما التقنية الثانية فهي تعتمد على المهام مثل التقريب أو نافذة الانزلاق. والنموذج الذي سيتم تطويره في هذا البحث سوف يستفيد

من التقنيات المذكورة أعلاه لإنتاج نموذج جديد. اخيرا ستتم محاكاة هذا النموذج المطور واختباره على مجموعه من تدفق البيانات الكبيرة باستخدام احدى الأدوات المختصة بتدفق البيانات. اهداف هذا البحث تتحقق بخطوتين اساسيتين: الأولى من خلال تنفيذ نموذج لتصنيف البيانات المستخدمة باستخدام عنصرين لتكوين النموذج وهم نافذة الانزلاق والمصنف من خلال هذين الجزئيين يقوم النموذج بعملية التنقيب الأساسية. الخطوة الثانية تتمثل بإضافة عنصر جديد الى النموذج السابق وهذا العنصر وظيفته الكشف عن التغيير الطارئ في البيانات نتيجة لبعض المشكلات التي تطرأ على الشبكة او عملية مزامنة البيانات. اخيرا كلا النموذجين سوف يمثلان باستخدام لغات البرمجة الخاصة بالتنقيب في البيانات المتدفقة. بعد ذلك سوف تتم اختبار هذه النماذج وتحليل البيانات لكلا النموذجين ومقارنتهما ببعضهما، كما سيتم مقارنة هذه النتائج بأعمال اخرى في المجال ذاته.

Developing Data Stream Mining Model for Big Data Applications

Ebtesam Hamed Almalki

A Thesis Submitted in Partial Fulfillments of the Requirements for the Degree of Master in Computer Scince

Developing Data Stream Mining Model for Big Data Applications

Ebtesam Hamed Almalki

ABSTRACT

In recent decades, advances in hardware technology have enabled us to automatically record transactions and other crucial information of everyday life, at a rapid rate. Applications which monitor and scrutinize information using sensor networks are the major sources recording massive amount of data with high pace, speed and velocity. These processes usually generate huge amounts of online data which grow and multiply at an extra ordinary rate. Such type of online data is referred as "data stream". Various knowledge analysis techniques and extraction procedures are studied and investigated to solve the real-world problems. The data mining community use various approaches and methods such as clustering and classification to discover diverse solutions. But all elucidations and answers may not be effectively applied on data streams, since they require online mining which is a continuous process executed in a fashion. Thus, the storage, querying and mining of these data streams, with high speed and huge volume is extremely challenging task.

Stream mining is concerned with extracting knowledge structures represented in models and patterns in non-stopping streams of information. In this thesis, two main contributions are made. Firstly, a data stream mining algorithm named Sliding Window Random Decision Tree (SWRT) for classifying the data stream, is developed .The proposed model adopts sliding window method as data estimator. Secondly, a change detector method is added to improve SWRT performance, by removing drift caused by time change. (ASWRT) is developed, where ADWIN change detector has been used. Both models are verified against accuracy and time. The models are implemented using MOA (Massive Online Analysis) framework. The results of both model are analyzed and compared with each other, to evaluate the effect of employing change detector module. The results showed that, SWRT algorithm achieved 85.78% accuracy with big data of 1,800,000 and consumed total time of 15.41s. It is also noticed that the accuracy of ASWRT algorithm is better, as it attained 98. 88% of accuracy when compared with SWRT, and the total time consumed is 18.80s.