التنقيب في البيانات الكبيره

هـــدى عـــمر مـــحمد بـانقيطه

المستخلص

نعيش اليوم في عصر البيانات الكبيرة (BD) . وفي ظل تصاعد جمع البيانات الكبيرة وتطبيقاتها , أصبحت قدرة أدوات ووسائل البرمجيات الحالية لادارة وتنفيذ هذه التطبيقات أقل . الحجم ليس هو السمة الوحيد التي تحدد البيانات الكبيرة، ولكن أيضا السرعة والتنوع والقيمة هي أيضا من مميزات البيانات الكبيره. يوجد العديد من الموارد التي تحتوي على بيانات كبيره و التي ينبغي معالجتها. فنجد مثلا البحوث الطبية الحيوية وهي واحده من بين العديد من المجالات الغنيه بالبيانات و تخفي الكثير من المعرفة. Medline عبارة عن قاعدة بيانات ضخمة تحتوي على العديد من الاوراق البحثية الطبية, والتي لا تزال مصدر غير مستغل إلى حد كبير للحصول على المعلومات الطبيه البيولوجية. اكتشاف المعرفة المفيدة من مثل هذا المورد الضخم يؤدي إلى حل مشاكل مختلفة تتعلق بنوع المعلومات المتعلقه ببعض المفاهيم والعلاقة الدلالية المرتبطة بها. في هذه البحث العلمي تم اقتراح نموذجا من مستويين لإستخراج العلاقة بين الكاينات الطبيه كالامراض وعلاجاتها من قاعدة بيانات MEDLINE , باستخدام قاعدة المعرفة للغات الطبية (UMLS) . ويستخدم هذا النموذج نهج الإشراف الذاتي لاستخراج العلاقات (RE)من خلال بناء نماذج التدريب المتطورة باستخدام معلومات من UMLS . وفي هذا لنموذج تم استخدام تكنولوجيا البياتات الكبيره Spark مع تقنيات متعدده لإستخراج البيانات, بالإضافة إلى نظام تعدد الوكلاء. يظهر النموذج نتيجة أفضل بالمقارنة مع الانظمة الحاليه في نفس المجال.

Big Data Knowledge Mining

Huda Umar Banuqitah

ABSTRACT

The era of Big Data (BD) has arrived. The rise of big data applications where data collection has grown beyond the capability of the current software tool to capture, manage and process within tolerable elapsed time. Volume is not the only the characteristic that defines big data, but also velocity, variety, and value. Many resources generate BD that should be processed. The biomedical research literature is one among many other domains that hides rich knowledge. MEDLINE is a huge database of biomedical research papers which remain a significantly underutilized source of biological information. Discovering the useful knowledge from such huge corpus leads to various problems related to the type of information such as the concepts related to the domain of texts and the semantic relationship associated with them. In this paper, we propose a Two-level model for Self-supervised relation extraction from MEDLINE using Unified Medical Language System (UMLS) Knowledgebase. The model uses a Self-supervised Approach for Relation Extraction (RE) by constructing enhanced training examples using information from UMLS and incorporates Spark BD technology with multiple Data Mining and machine learning technique with Multi Agent System (MAS). The system shows a better result in comparison with the current state of the art and naïve approach in terms of Accuracy, Precision, Recall and F-score.