Ministry of Higher Education and Scientific Research
University of Mosul
College of Computer Science and Mathematics
Department of Computer Science

A Cloud Big Data Parallel Computing Approach for Academic Capital Dynamics of the Scientific Research in Iraq

A Thesis Submitted to the Council of the College of Computer Science and Mathematics University of Mosul as a Partial Fulfillment of Requirements for the Degree of Doctor of Philosophy in Computer Science

By

Nagham Ajeel Sultan Al-Ajeely

Supervised by

Prof. Dr.Dhuha Basheer Abdullah Albazaz

2024 A.D. **1445 A.H.**

ABSTRACT

The massive usage of modern technologies and the Internet with smart devices yields large-scale data called Big Data. The development of cloud computing technologies enables handling big data in terms of processing and storage; therefore, several techniques have been developed by companies and institutions aiming to come up with efficient approaches. Previous works optimizing these approaches can be performed by involving advanced technologies such as parallel processing, containerization, and virtualization.

Moreover, in sociology, the concept of social capital represents an individual's collective behavior, and norms in a community. In the same context, academic capital means the collective performance of an author or an institution. This concept, has not been studied in previous literature. On the other hand, Google Scholar is considered the most well-known web search engine for accessing papers around the world. Its data can be used to assess the academic performance of a particular author, university, or institution with citations, h-index, publishers, etc. However, processing, analyzing , and visualizing these data need to adopt dedicated dynamic methods. Therefore, developing an dynamic efficient approach to convert big data to knowledge for evaluation by authors or universities is critical.

This thesis introduces a dynamic system by proposing a novel rank called Academic Capital Distributed Network (ACDN). It dynamically collects extensive data from Google Scholar using several Cloud Computing technologies. The collected data was then processed, visualized, and analyzed to assess the dynamic academic performance of Iraqi authors and universities.

Web scrapers were designed to gather dynamic data from Google Scholar of Iraqi universities and authors based on their official universities domains. These scrapers are based on concepts inspired by parallel processing techniques using dedicated virtual machines. The collected data is seamlessly stored using a Replica Cluster in MongoDB Atlas to facilitate efficient data management and provide a robust and scalable dynamic infrastructure for storing and managing the collected data. This method offers high availability, fault tolerance, read scalability, and data integrity. Moreover, a Cloud-based parallel computing environment is adopted to facilitate the efficient organization and processing of the collected data, leveraging big data frameworks, with a particular focus on Apache Spark.

Two novel datasets have been created: one for authors and one for universities. By using these datasets, two network models were generated; the first is for Iraqi universities; termed as Iraqi Universities Network (IUN), and the latter is for Iraqi authors; termed as Iraqi Scholars Network (ISN). Acadmic capital of authors and universities was calculated using data from the two network models IUN and ISN, and Google Scholar indicators. Furthermore, the performance of the Iraqi authors and universities was evaluated also, ranks were assigned based on their respective metrics of academic capital which change dynamiclly in response to the indicators and search criteria employed by the author or university.

The findings show the efficiency of the proposed approach in terms of data quality and reliability in dynamic collecting big data from Google Scholar Cloud. The results also offer interesting facts about the Iraqi universities and authors. In addition, collaboration patterns among authors and universities were also extracted using the two network models. The proposed rank was also benchmarked with other international ranks. The main results show that the average degree of the ISN is 1.8, which means that each Iraqi scholar has approximately coauthored two papers on average. This results is considered acceptable compared to ACM international network. On the other hand, the average shortest path length of IUN is 3.2 meaning that the scientific collaboration among the Iraqi universities was indicates a moderate distance between nodes on average. However, the clustering coefficient of IUN is 0.249, which reflects that the Iraqi universities are not strongly connected and the tendancy to cluster together is low (far from 1). Furthermore, The results of Academic Capital of authors show that a Computer Science author from University of Babylon has the highest Academic Capital among all the authors in the Iraqi universities. The Academic Capital of the Iraqi universities show that the University of Baghdad has the highest value followed by Al-Mustansiriyah University, University of Basrah, University of Mosul, University of Babylon, and so on. This result is relatively in agreement of the other international ranks such as QS, Times, Webometrics.This work is considered the first kind of analysis that rates all the Iraqi universities and authors as the main focus, which is of interest to the Ministry of Higher Education and Scientific Research officials in Iraq.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

طريقة حوسبة البيانات الضخمة السحابية المتوازية لديناميكيات رأس المال الأكاديمي للبحث العلمي في العراق

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

نغم عجيل سلطان العجيلي

بإشراف
أ.د. ضحى بشير عبدالله البزاز

١٤٤٥ هـ
٢٠٢٤ م

الـخــلاصـــة

إن الاستخدام المكثف للتقنيات الحديثة والإنترنت مع الأجهزة الذكية ينتج عنه بيانات واسعة النطاق تسمى البيانات الضخمة. إن تطور تقنيات الحوسبة السحابية يتيح التعامل مع البيانات الضخمة من حيث المعالجة والتخزين؛ لذلك، تم تطوير العديد من التقنيات من قبل الشركات والمؤسسات بهدف التوصل إلى أساليب فعالة. يمكن تنفيذ الأعمال السابقة التي تعمل على تحسين هذه الأساليب من خلال إشراك التقنيات المتقدمة مثل المعالجة المتوازية، والحاويات، والمحاكاة الافتراضية.

علاوة على ذلك، في علم الاجتماع، يمثل مفهوم رأس المال الاجتماعي السلوك الجماعي للفرد، والأعراف في المجتمع. وفي نفس السياق، يعني رأس المال الأكاديمي الأداء الجماعي للمؤلف أو المؤسسة. هذا المفهوم، لم تتم دراسته في الأدبيات السابقة. من ناحية أخرى، يعتبر Google Scholar محرك بحث الويب الأكثر شهرة للوصول إلى الأبحاث حول العالم ويمكن استخدام بياناته لتقييم الأداء الأكاديمي لمؤلف معين، أو جامعة، أو مؤسسة معينة من خلال الاستشهادات، ومؤشر h-index ، وما إلى ذلك. ومع ذلك، فإن معالجة هذه البيانات وتحليلها وتصورها تحتاج إلى اعتماد أساليب ديناميكية مخصصة. ولذلك، فإن تطوير نهج ديناميكي فعال لتحويل البيانات الضخمة إلى معرفة للتقييم من قبل المؤلفين أو الجامعات أمر بالغ الأهمية

تقدم هذه الأطروحة نظامًا ديناميكيًا من خلال اقتراح رانك جديد يسمى Academic Capital Distributed Network (ACDN). يقوم بجمع بيانات واسعة النطاق ديناميكيًا من Google Scholar باستخدام العديد من تقنيات الحوسبة السحابية. ومن ثم تمت معالجة البيانات التي تم جمعها وتصورها وتحليلها لتقييم الأداء الأكاديمي الديناميكي للمؤلفين والجامعات العراقية.

تم تصميم أدوات استخراج البيانات من الويب لجمع البيانات الديناميكية من Google Scholar للجامعات العراقية والمؤلفين بناءً على نطاقات جامعاتهم الرسمية (domain). تعتمد أدوات الكشط هذه على مفاهيم مستوحاة من تقنيات المعالجة المتوازية باستخدام أجهزة افتراضية مخصصة بشكل ديناميكي. يتم تخزين البيانات المجمعة بسلاسة باستخدام مجموعة النسخ المتماثلة في MongoDB Atlas لتسهيل إدارة البيانات بكفاءة وتوفير بنية تحتية ديناميكية قوية وقابلة للتطوير لتخزين وإدارة البيانات المجمعة. توفر هذه الطريقة توفرًا عاليًا وتسامحًا مع الأخطاء وقابلية التوسع للقراءة وتكامل البيانات. علاوة على ذلك، تم اعتماد بيئة حوسبة متوازية قائمة على السحابة لتسهيل التنظيم والمعالجة الفعالة للبيانات المجمعة، والاستفادة من أطر البيانات الضخمة، مع التركيز بشكل خاص على Apache Spark.

تم إنشاء مجموعتي بيانات جديدتين: واحدة للمؤلفين والأخرى للجامعات. باستخدام مجموعات البيانات هذه، تم إنشاء نموذجين للشبكة؛ الأول خاص بالجامعات العراقية؛ والمعروفة باسم شبكة الجامعات العراقية (IUN)، والأخيرة مخصصة للمؤلفين العراقيين؛ تُعرف باسم شبكة العلماء العراقيين (ISN). تم حساب رأس المال الأكاديمي للمؤلفين والجامعات باستخدام بيانات من نموذجي الشبكة IUN وISN ومؤشرات Google Scholar. علاوة على ذلك، تم تقييم أداء المؤلفين والجامعات العراقية أيضًا، وتم تحديد الرتب بناءً على مقاييس رأس المال الأكاديمي الخاصة بهم والتي تتغير ديناميكيًا استجابة للمؤشرات ومعايير البحث التي يستخدمها المؤلف أو الجامعة.

أظهرت النتائج الرئيسية أن Average Degree هو ١,٨، مما يعني أن كل باحث عراقي شارك في تأليف ورقتين كمعدل. تعتبر هذه النتائج مقبولة مقارنة بشبكة ACM الدولية. من ناحية أخرى، فإن متوسط أقصر طول مسار لـ IUN هو ٣,٢ مما يعني أن التعاون العلمي بين الجامعات العراقية كان يشير إلى مسافة متوسطة بين العقد في المتوسط. ومع ذلك، فإن معامل التجمع لـ IUN هو ٠,٢٤٩، مما يعكس أن الجامعات العراقية ليست مرتبطة بقوة وأن الميل إلى التجمع معًا منخفض (بعيدًا عن ١). علاوة على ذلك، أظهرت نتائج رأس المال الأكاديمي للمؤلفين أن مؤلف من كلية علوم الحاسوب / جامعة بابل لديه أعلى رأس مال أكاديمي بين جميع المؤلفين في الجامعات العراقية. وكذلك رأس المال الأكاديمي الخاص بالجامعات العراقية اظهر أن جامعة بغداد حصلت على أعلى قيمة تليها الجامعة المستنصرية، جامعة البصرة ، جامعة الموصل، جامعة بابل، وهكذا. وتتفق هذه النتيجة نسبياً مع التصنيفات العالمية الأخرى مثل QS، Times، Webometrics. ويعتبر هذا العمل أول نوع من التحليل الذي يصنف جميع الجامعات العراقية والمؤلفين كمحور رئيسي، وهو ما يحظى باهتمام مسؤولي وزارة التعليم العالي والبحث العلمي في العراق.