Human/Machine Co-Authoring: Advanced Knowledge Synthesis through Mining Massive Publication Network

Jack Hooper
Advisor: Dr. Javed Khan

Introduction
Automatic report generation is an emerging technology that mechanically generates documents in the form of a report consisting of text, tables, and figures about a specific topic. The scholarly database utilized in this project is a snapshot of the Microsoft Academic Graph sourced from AMiner. The solution requires handling of a massive data set at a high throughput computing. In this project we are putting the previously prototyped system onto Ohio Supercomputer Center as a robust data lake.

Data Lake
A data lake is essentially a massive database with hundreds of thousands of entries. It can be used as a foundation for users and programs, as well as a base for many different kinds of research. Normally a data lake is sustained for a long period of time which allows an ecosystem of many different services and researches to thrive. The data lake I am using for this project is the Microsoft Academic Graph (MAG). The MAG consists of more than 250 million scientific publications and is larger than 500 gbs. We are creating an application for policy makers in various countries to study impact of scientific collaboration, for understanding their investment strategies in scientific research and education.

Process
The program takes in the ISO code for the selected country, a starting and ending study year, and the name of the database being used. The program then takes that information and executes a series of queries in order to acquire the necessary information for the report. The program and data are being hosted by the Ohio Supercomputer Facility.

After all of the variables needed have been acquired the program then creates a new text file that contains the advanced knowledge synthesized report for the country. The report is formed by using a static template (DNA) with the variables filling in the blanks. The template was extracted from a high value past paper written by a team of experts for a specific country. Going forward I will research implementing an AI text generator like ChatGPT to generate the report instead of using a static framework. Our process uses AI to extract the analysis of this paper and reproduces analogous analysis for hundreds of other countries using the MAG data lake.

Implementation
This is an ongoing, multi-step, project. It involves,

1. Sourcing and moving the data
2. Creating SQL queries
3. Creating a User Interface
4. Implementing AI Text generating technology

We are currently in stage 2.

Resources Used
“Automatic Generation of a Metric Report: A Case Study of Scientometric Analytics” Amal Babour, Javed I. Khan
Microsoft Academic Graph Ohio Supercomputer Center

QR Code for Man/Machine synthesized paper

QR Code for Original Paper which contributed to the DNA