

SỬ DỤNG AI TỰ ĐỘNG PHÂN TÍCH BÀI BÁO, CUNG CẤP BẢNG XẾP HẠNG NGHIÊN CỨU

MINH ANH

SẢN PHẨM "HỆ THỐNG CHỈ SỐ VÀ PHÂN TÍCH TỔ CHỨC ĐƯỢC TĂNG CƯỜNG BẰNG CÔNG NGHỆ MÁY - TRÍ TUỆ NHÂN TẠO VỚI TÍNH MINH BẠCH CAO" (METRIXA) CỦA ĐỘI STARFOX - ĐỘI TUYỂN TIN HỌC TRƯỜNG THPT CHUYÊN KHOA HỌC TỰ NHIÊN ĐẠT GIẢI "TRIỂN VỌNG" CUỘC THI DATA FOR LIFE 2025.

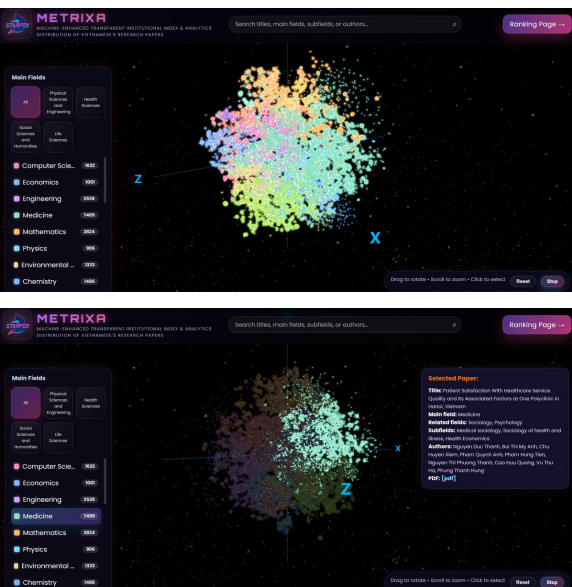
Nhóm 8 học sinh Trường THPT chuyên Khoa học Tự nhiên (thuộc Trường ĐH Khoa học Tự nhiên, ĐHQGHN) đã phát triển METRIXA, hệ thống AI đầu tiên ở Việt Nam có thể tự động thu thập, phân tích và trực quan hóa dữ liệu hàng chục nghìn bài báo khoa học, tạo nên bản đồ nghiên cứu 3D giúp đánh giá năng lực khoa học của các tổ chức một cách minh bạch và chính xác hơn.

Xuất phát từ thực tế hiện nay, các bảng xếp hạng ở Việt Nam chủ yếu tập trung vào cơ sở vật chất và chất lượng giảng dạy của các trường đại học, tổ chức, đội Starfox lựa chọn hướng tiếp cận khác. Đó là dựa trên hoạt động nghiên cứu khoa học. Dữ liệu được sử dụng là các bài báo khoa học từ các tổ chức, trường đại học đã được công bố quốc tế. Dữ liệu được cập nhật và chuẩn hóa tự động từ các kho học thuật lớn như Semantic Scholar và S2ORC, hai nguồn được giới nghiên cứu quốc tế đánh giá cao về độ tin cậy.

Về đối tượng mà METRIXA hướng tới, nhóm chia sẻ: đối với nhà nước, hệ thống cung cấp dữ liệu để theo dõi tiến độ nghiên cứu khoa học một cách nhất quán và có cơ sở; đối với nhà nghiên cứu, nền tảng cho phép tìm cộng sự, phát hiện các nhóm nghiên cứu liên ngành và nhận diện xu hướng khoa học mới còn với học sinh, sinh viên, biểu đồ 3D giúp họ quan sát sự phân bố của các lĩnh vực tại Việt Nam, từ đó định hướng ngành nghề phù hợp hơn.

Hệ thống có khả năng thu thập, phân tích và trực quan hóa dữ liệu bài báo khoa học của Việt Nam, từ đó tạo ra bảng xếp hạng nghiên cứu chuyên sâu cho các trường đại học và viện nghiên cứu. Điểm nổi bật của METRIXA là hệ thống đi sâu vào nội dung nghiên cứu thay vì chỉ xét đến cơ sở vật chất hoặc các tiêu chí đào tạo như các bảng xếp hạng thông thường.

Hệ thống gồm hai thành phần chính. Trong đó, với Biểu đồ 3D đánh giá cao về độ tin cậy.





cứ, METRIXA phân tích trực tiếp các bài báo, xác định lĩnh vực nghiên cứu và thể hiện chúng qua một biểu đồ 3D trực quan. Mỗi điểm trên biểu đồ là một bài báo, màu sắc tương ứng với từng ngành, khoảng cách phản ánh mức độ liên quan giữa các lĩnh vực. Biểu đồ này cho phép người dùng nhận thấy ngay những cụm ngành có sự tương tác mạnh, ví dụ như khoa học máy tính luôn gần toán học vì tính chất nền tảng của toán trong thuật toán và mô hình tính toán.

Ở phần Bảng xếp hạng tổ chức nghiên cứu tự động, AI phân tích đóng góp của từng trường, từng viện thông qua số lượng và chất lượng bài báo được công bố quốc tế. Dữ liệu được cập nhật và chuẩn hóa tự động từ các kho học thuật lớn như Semantic Scholar và S2ORC, hai nguồn được giới nghiên cứu quốc tế đánh giá cao về độ tin cậy.

METRIXA hướng tới ba nhóm đối tượng. Đối với nhà nước, hệ thống cung cấp dữ liệu để theo dõi tiến độ nghiên cứu khoa học một cách nhất quán và có cơ sở. Đối với nhà nghiên cứu, nền tảng cho phép tìm cộng sự, phát hiện các nhóm nghiên cứu liên ngành và nhận diện xu hướng khoa học mới. Với học sinh, sinh viên, biểu đồ 3D giúp họ quan sát sự phân bố của các lĩnh vực tại Việt Nam, từ đó định hướng ngành nghề phù hợp hơn.

“Điểm khiến METRIXA nổi bật là hệ thống bảng xếp hạng đi rất sâu vào từng lĩnh vực và chuyên ngành nghiên cứu. Cùng với đó là biểu đồ trực quan giúp theo dõi sự phân bố của các bài báo theo lĩnh vực, tạo ra một bức tranh khoa học mà ai cũng có thể nhìn thấy và hiểu ngay”, Trần Bảo Khánh cho hay.

Để xây dựng METRIXA, nhóm Starfox phải tự học và kết hợp nhiều mô hình AI tiên tiến. Việc đảm bảo dữ liệu đầu vào chính xác và minh bạch là yêu cầu bắt buộc của cuộc thi Data for Life. Nhóm sử dụng hai nguồn dữ liệu uy tín toàn cầu gồm Semantic Scholar và S2ORC, vốn được cộng đồng khoa học đánh giá cao về tính minh bạch.

Phần khó nhất là thiết kế hệ thống tự động thu thập, phân tích, trực quan hóa dữ liệu sao cho ổn định và chính xác. Trong đó, một thách thức lớn là việc tách riêng bài báo của Việt Nam khỏi dữ liệu quốc tế. Tên tác giả thường được viết bằng tiếng Anh và không dấu, khiến việc phân biệt tác giả Việt Nam với tác giả từ các quốc gia châu Á khác trở nên cực kỳ phức tạp.

Trong module phân tích, nhóm sử dụng mô hình ngôn ngữ lớn (LLM) có khả năng tìm kiếm website, xác minh thông tin bài báo và tự động điền các thông tin còn thiếu. Một lớp AI khác được sử dụng để đọc, phân tích

nội dung khoa học và trích xuất đặc trưng của bài báo dưới dạng vector 768 chiều.

Để chuyển những vector nhiều chiều này thành hình ảnh trực quan, nhóm áp dụng thuật toán t-SNE (T-distributed Stochastic Neighbor Embedding), một kỹ thuật giảm chiều dữ liệu trong khoa học máy tính. Thuật toán giúp đưa dữ liệu từ 768 chiều về 3 chiều mà vẫn giữ được “cấu trúc lân cận, tức các bài báo liên quan sẽ nằm gần nhau.

“t-SNE cho phép dựng biểu đồ 3D chính xác và dễ hiểu. Người xem có thể xoay, chạm, nhìn thấy cả hệ sinh thái nghiên cứu”, Khánh giải thích.

Thời gian tới, nhóm sẽ tiếp tục nghiên cứu để hoàn thiện hơn giải pháp này, trong đó hướng đi cụ thể bao gồm sẽ cải thiện công thức tính toán để bảng xếp hạng trở nên chính xác và toàn diện hơn. Nhóm cũng đặt mục tiêu so sánh nghiên cứu giữa Việt Nam và thế giới đồng thời mong muốn giảm thiểu chi phí hoạt động ở mức tối đa để đảm bảo tính bền vững lâu dài của sản phẩm.