DoIt Hệ thống nâng cao chất lượng văn bản

Sản phẩm "DoIt – Hệ thống nâng cao chất lượng văn bản" do nhóm nghiên cứu Khoa Công nghệ thông tin, Trường Đại học Công nghệ (Đại học Quốc gia Hà Nội) thực hiện đã đạt giải Nhì Giải thưởng Nhân tài Đất Việt 2017. Sản phẩm lần này đã góp phần nâng thành tích đạt giải Nhân tài Đất Việt của Nhà trường lên con số 5.

■ TUYẾT NGA

SẢN PHẨM VIỆT DÀNH CHO NGƯỜI VIỆT

Sự phát triển của công nghệ thông tin đã mang lại nhiều đột phá trong cuộc sống của con người. Trong lĩnh vực giáo dục, máy tính và Internet đã giúp cho người dạy và người học tiếp cận được nhiều nguồn thông tin, nhiều công cụ phục vụ cho việc dạy và học. Hiện nay, nhiều trường đại học trên thế giới đang sử dụng một số các công cụ như Turnitin, CheckforPlagiarism hay PlagScanđể hỗ trợ cho việc kiểm tra và đánh giá văn bản được tạo ra bởi người học (bao gồm các bài tập lớn cho đến các đồ án, khóa luận, luận văn,…). Tuy nhiên các công cụ này chủ yếu chỉ phù hợp với các tài liệu viết bằng tiếng Anh và có thu phí sử dụng khá cao. Trong nước đã có khá nhiều nghiên cứu về kiểm tra lỗi chính tả, ngữ pháp của văn bản tiếng Việt, và một vài nghiên cứu về phát hiện trùng lặp văn bản nhưng việc triển khai xây dựng một công cụ để sử dụng thì chưa có.

Trước thực tế như vậy, nhóm các nhà khoa học của Trường Đại học Công nghệ, gồm các thành viên chính TS. Võ Đình Hiếu, PGS.TS Phạm Bảo Sơn, PGS.TS Lê Anh Cường, PGS. TS. Nguyễn Việt Hà là các giảng viên của Khoa Công nghệ thông tin (CNTT), cùng một số nghiên cứu sinh, học viên và sinh viên của Khoa đã xây dựng hệ thống DoIT (Document

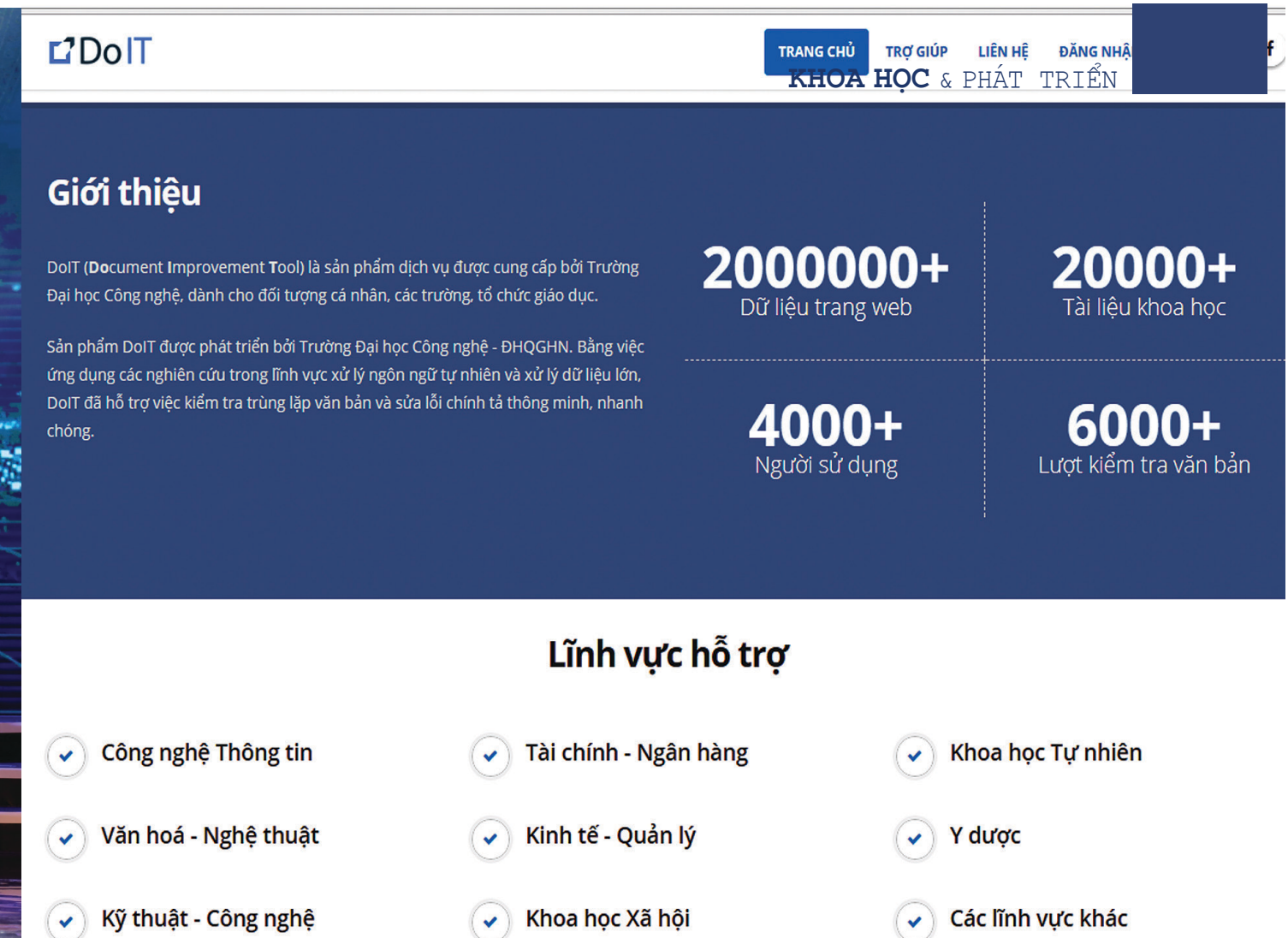
Improvement Tools, http://doit.uet.vnu.edu.vn/) với hai chức năng chính gồm kiểm tra lỗi chính tả và phát hiện trùng lặp. Hệ thống hứa hẹn sẽ góp phần nâng cao chất lượng của các đồ án, khóa luận, luận văn, luận án của người học nói riêng và chất lượng giáo dục và đào tạo nói chung.

TS. Võ Đình Hiếu cho biết, DoIT là kết quả của đề tài nghiên cứu cấp ĐHQGHN được triển khai từ năm 2014. Mục tiêu ban đầu đặt ra là nghiên cứu các phương pháp kiểm lỗi chính tả; các phương pháp so sánh văn bản để phát hiện sao chép; xây dựng hệ thống hướng dịch vụ hỗ trợ kiểm lỗi chính tả và phát hiện sao chép văn bản ứng dụng cho quản lý khóa luận tốt nghiệp, luận văn thạc sĩ và luận án tiến sĩ trong lĩnh vực Công nghệ thông tin tại ĐHQGHN. Đề tài cũng hướng đến việc phát triển chuyên môn cho giảng viên, đào tạo và phát triển kỹ năng nghiên cứu khoa học cho nghiên cứu sinh, học viên, sinh viên trong các lĩnh vực liên quan đến đề tài như xử lý ngôn ngữ tự nhiên, học máy, kiến trúc hướng dịch vụ.

Hệ thống DoIT ngoài việc chỉ ra các từ bị lỗi chính tả còn đề xuất từ đúng thay thế. Chức năng phát hiện trùng lặp sẽ chỉ ra câu/đoạn trong văn bản được kiểm tra trùng lặp với câu/đoạn của tài liệu có trong cơ sở dữ liệu (CSDL) của hệ thống. Hiện nay, nguồn cơ sở dữ liệu ngày càng được mở rộng, kết quả kiểm tra sao chép ngày càng chính xác. Hệ thống hỗ trợ hầu hết các định dạng văn bản phổ biến như pdf, doc, docx, ppt, txt, odt, v.v. Hệ thống được xây dựng trên nền Web và được thiết kế theo các mô đun. Thiết kế này tạo môi trường để giảng viên của Khoa Công nghệ thông tin thử nghiệm và đưa vào sử dụng các kết quả nghiên cứu mới trong các lĩnh vực liên quan. Bạn đọc có thể trải nghiệm sản phẩm tại http://doit.uet.vnu.edu.vn hoặc tại http://doit.lic.vnu.edu.vn.

ĐƯỢC VINH DANH VÀ CON ĐƯỜNG PHÍA TRƯỚC

Với những lý do và mục đích như vậy, sản phẩm DoIT – Hệ thống hỗ trợ nâng cao chất lượng văn bản được trao giải Nhì Nhân tài Đất Việt 2017, thuộc hệ thống sản phẩm Công nghệ thông tin Tiềm năng. TS. Võ Đình Hiếu cho biết, nhóm tác giả đã rất vui khi sản phẩm vào vòng chung khảo



và đạt giải nhì. Điều này phần nào thể hiện sự quan tâm của xã hội đến việc nâng chất lượng của tài liệu, đặc biệt là các khóa luận, luận văn, luận án trong các trường đại học. Giải thưởng là sự khích lệ lớn đối với các thành viên của nhóm để tiếp tục nghiên cứu và hoàn thiện hệ thống.

Trong quá trình nghiên cứu, xây dựng hệ thống, nhóm đã gặp không ít khó khăn. Ban đầu, nhóm đã thử nghiệm phương án phát hiện trùng lặp bằng cách sử dụng thông tin từ các máy tìm kiếm trên Internet (như Google). Dù kết quả chạy thử rất tốt nhưng không thể triển khai trên diện rộng vì chi phí sử dụng các dịch vụ của máy tìm kiếm quá cao. Nhóm lại phải quay về với phương án dùng máy tìm kiếm nội bộ. Với phương án này, nhóm lại đối mặt với thách thức về nguồn dữ liệu và cách quản lý dữ liệu hiệu quả. TS. Võ Đình Hiếu chia sẻ, may mắn là Khoa CNTT có nhiều giảng viên trẻ, đầy nhiệt huyết nên chúng tôi rất dễ chia sẻ, hợp tác để cùng nghiên cứu, giải quyết vấn đề. Nhóm thực hiện đề tài cũng đã nhận được sự tạo điều kiện của lãnh đạo ĐHQGHN và của Trường Đại học Công nghệ. Trong thời gian sắp tới, nhóm rất mong muốn nhận được nhiều sự ủng hộ hơn nữa để có thể triển khai hệ thống trên diện rộng, phục vụ cho nhiều đối tượng người dùng thuộc nhiều lĩnh vực khác nhau.