**LỜI MỞ ĐẦU**

Trong khóa luận này, chúng tôi phát triển một hệ thống dùng Web Crawler kết hợp với việc sử dụng cơ sở dữ liệu chỉ mục có sẵn để xây dựng và cập nhật kho dữ liệu chỉ mục bài báo khoa học trong lĩnh vực khoa học máy tính.

Bằng cách sử dụng Web Crawler để thu thập thông tin chỉ mục các bài báo trực tiếp trên các thư viện số, hệ thống đảm bảo thông tin thu thập được là chính xác và cập nhật. Cùng với đó, hệ thống sẽ sử dụng các cơ sở dữ liệu chỉ mục có sẵn để dữ liệu xây dựng được là đầy đủ.

Việc kết hợp nguồn dữ liệu chỉ mục các bài báo được rút trích trên các thư viện số và các cơ sở dữ liệu chỉ mục có sẵn giúp cho dữ liệu mà hệ thống xây dựng được có tính cập nhật, chính xác và đầy đủ từ đó có thể phát triển các công cụ tìm kiếm thông tin bài báo khoa học đảm bảo nhu cầu tìm kiếm của người dùng.

Hệ thống trong khóa luận thực hiện rút trích thông tin chỉ mục trên các thư viện số ACM, IEEEXplore, CiteSeer và sử dụng cơ sở dữ liệu chỉ mục có sẵn DBLP. Cùng với chức năng tự động cập nhật các bài báo mới từ các thư viện số thì dữ liệu thông tin chỉ mục các bài báo khoa học thu thập được sẽ chính xác, đầy đủ và cập nhật.