# Tên đề tài

Xây dựng dữ liệu chỉ mục các bài báo khoa học (Science Article Bibliography Database - DBSA).

# Đặt vấn đề:

Cùng với sự phát triển của internet, số lượng các bài báo khoa học, số tạp chí điện tử, cũng như số các tác giả viết báo khoa học ngày càng tăng, điều này gây ra một số khó khăn khi mà người người dùng muốn tìm kiếm các bài báo về một vấn đề mà mình nghiên cứu - cũng như gây ra một thách thức lớn với các hệ thống đánh dấu,lưu trữ chỉ mục trong việc cập nhật thông tin các bài báo mới một cách đẩy đủ và nhanh nhất. Trong bài báo này chúng tôi đề cập đến phát triển một hệ thống dùng để xây dựng và cập nhật thông tin dữ liệu chỉ mục các bài báo khoa học từ các thư viện số và dữ liệu chỉ mục có sẵn. Hệ thống có chức năng tìm kiếm các bài báo khoa học được công bố trên các thư viện số (ACM, IEEE, CiteSeer) sau đó rút các thông tin về chỉ mục của bài báo, bao gồm các thông tin tên tác giả, tựa đề, tóm tắt, năm công bố, nơi công bố và địa chỉ đường dẫn của bài báo. Từ những thông tin chỉ mục này, hệ thống sẽ kết hợp với dữ liệu chỉ mục có sẵn trong DBLP[1] để xây dựng lên dữ liệu chỉ mục các bài báo khoa học. Như vậy hệ thống đảm bảo được tính cập nhật với các bài báo mới được công bố trên các thư viện số và hệ thống sử dụng được dữ liệu thông tin chỉ mục rất lớn có sẵn từ DBLP. Việc xây dựng dữ liệu chỉ mục các bài báo khoa học là rất cần thiết, thông qua dữ liệu xây dựng được, ta có thể phát triển các công cụ tìm kiếm bài báo khoa học cung cấp cho người dùng thông tin những bài báo mới đầy đủ và cập nhật nhất.

# Mục tiêu và phạm vi của đề tài:

## Mục tiêu của đề tài:

* Xây dựng một cơ sở dữ liệu chỉ mục của các bài báo khoa học từ nhiều nguồn khác nhau: từ các cơ sở dữ liệu chỉ mục có sẵn như DBLP và các thư viện số ACM, Citeseer, IEEExplore , ScienceDriect.
* Phân loại các bài báo sau khi hệ thống thu thập về theo từng chuyên mục, chuyên đề mà bài báo đề cập đến.

## Phạm vị của đề tài:

* Thông tin về các bài báo được thu thập từ cơ sở dữ liệu chỉ mục DBLP và các thư viện số ACM, IEEExplore, Citeseer, Scient Driect.
* Khung phân loại của bài báo khoa học được sử dụng trong đề tài thuộc lĩnh vực khoa học máy tính được tham khảo trên Wiki [[1]](#footnote-1). Khung phân loại này được sử dụng vì đây là nguồn tham khảo đáng tin cậy được cộng đồng công nhận và sử dụng. Đối với lĩnh vực khoa học máy tính khung phân loại này chi tiết hơn chứ không khái quát như những khung phân loại của thư viện số cũng như ACM Computing Classification.

1. http://en.wikipedia.org/wiki/Computer\_science [↑](#footnote-ref-1)