# TÌM KIẾM TÀI LIỆU HỌC TẬP ĐA NGÔN NGỮ VỚI KỸ THUẬT CHỈ MỤC NGỮ NGHĨA TIỀM ẨN (LATENT SEMANTIC INDEXING)

**Trần Cao Đệ**
*Khoa Công nghệ Thông tin và Truyền thông– Đại học Cần Thơ*
*Email: tcde@cit.ctu.edu.vn*

***Tóm tắt:*** *Việc tìm kiếm tài liệu học tập trên Internet đang trở thành một nhu cầu khách quan, tất yếu và thiết thực của mọi người. Các công cụ tìm kiếm có thể hỗ trợ việc tìm kiếm đa ngôn ngữ (ví dụ Google). Tuy nhiên, phần lớn các công cụ chưa hỗ trợ tìm kiếm đa ngôn ngữ với tiếng việt.*

*Bài viết này giới thiệu một kỹ thuật tìm kiếm đa ngôn ngữ với sự hỗ trợ của tự điển và kỹ thuật chỉ mục ngữ nghĩa tiềm ẩn (Latent semantic indexing). Một cách khái quát, các từ khóa tìm kiếm thuộc một ngôn ngữ nào đó, ví dụ tiếng việt. Chúng được dịch sang một ngôn ngữ mong muốn (ví dụ: tiếng anh, tiếng pháp…) nhờ tự điển, sau đó các từ dịch được chuyển qua các công cụ tìm kiếm thực hiện tìm kiếm các từ khóa trong các ngôn ngữ mong muốn. Như vậy kết quả trả về là đa ngôn ngữ.*

*Vấn đề đặt ra là làm sao cho việc dịch các từ khóa tìm kiếm một cách chính xác vì một từ khóa trong ngôn ngữ gốc có thể có nhiều từ tương ứng trong ngôn ngữ dịch. Vấn đề này được giải quyết bằng kỹ thuật chỉ mục ngữ nghĩa tìm ẩn (latent semantic indexing).*

***Từ khóa: tìm kiếm theo ngữ nghĩa, lập chỉ mục theo ngữ nghĩa, ngữ nghĩa tiềm ẩn.***

## 1. GIỚI THIỆU

Tìm kiếm thông tin đa ngôn ngữ (TKTTĐNN) cho phép người dùng có thể thực hiện câu truy vấn (query) để tìm kiếm thông tin bằng một ngôn ngữ nào đó (ví dụ tiếng việt) và thu hồi (retrieve) lại kết quả trả về trong nhiều ngôn ngữ khác nhau (ví dụ: tiếng anh, tiếng pháp, tiếng nga, tiếng hoa, tiếng việt…). Các công cụ tìm kiếm tài liệu trên Internet như Google, Yahoo đã hỗ trợ việc TKTTĐNN cho một số ngôn ngữ phổ biến như tiếng anh, tiếng pháp… Tuy nhiên chưa có công cụ nào hỗ trợ việc tìm kiếm đa ngôn ngữ với tiếng việt.

Hiện nay, các kỹ thuật TKTTĐNN tập trung vào việc dịch câu truy vấn từ ngôn ngữ gốc sang các ngôn ngữ muốn tìm kiếm rồi thực hiện truy vấn. Các cách tiếp cận chính là:

- Dịch dựa vào tự điển
- Dịch dựa vào mạng ngữ nghĩa và các kỹ thuật máy học.

Kỹ thuật dịch dựa vào tự điển là khá đơn giản:

- Dùng một tự điển có định dạng có thể đọc và phân tích được các mục giải nghĩa cho một từ, chẳng hạn các tự điển dạng tài liệu XML. Các từ điển dạng này có thể tải miễn phí từ Internet.
- Câu truy vấn tìm kiếm được đưa vào dưới dạng các từ khóa, các từ khóa này được dịch ra thành các từ khóa trong ngôn ngữ đích (muốn tìm) nhờ vào tự điển.