Mục Lục

[I. Đề cương đề tài luận văn tốt nghiệp 1](#_Toc268022045)

[1) Tên đề tài . 1](#_Toc268022046)

[2) Nhóm thực hiện. 1](#_Toc268022047)

[3) Tính cấp thiết của đề tài/ Đặt vấn đề . 1](#_Toc268022048)

[4) Mục tiêu & phạm vi đề tài. 3](#_Toc268022049)

[5) Nội dung và kế hoạch triển khai. 3](#_Toc268022050)

[6) Kết quả dự kiến. 3](#_Toc268022051)

[7) Tài liệu tham khảo dự kiến 3](#_Toc268022052)

[II. Nghiên cứu liên quan 3](#_Toc268022053)

[1) Thu thập , tìm kiếm tài liệu. 4](#_Toc268022054)

[2) Phân loại tài liệu. 4](#_Toc268022055)

# ĐỀ CƯƠNG ĐỀ TÀI TỐT NGHIỆP

## Tên đề tài .

Xây dựng hệ thống tìm kiếm bài báo khoa học .

## Nhóm thực hiện.

Đỗ văn Tiến

Nguyễn Phước Cường

## Tính cấp thiết của đề tài/ Đặt vấn đề .

* ***Đặt Vấn đề :***

Trong quá trình học tập cũng như nghiên cứu khoa học, việc tham khảo các hướng đi cũng như những bài viết liên quan đến vấn đề mình nghiên cứu từ các bài báo khoa học là rất cần thiết. Hiện nay với sự phát triển của Internet cũng như những tiện ích chia sẻ trên mạng, các bài bào khoa học được nhiều người dùng upload lên trên web . Các bài báo này phần lớn chưa được phân loại (chưa chia theo chủ đề lĩnh vực mà bài báo đề cập đến ) mà chỉ được upload lên để chia sẻ cho người dùng , vì vậy khi một người dùng cần một bài báo khoa học liên quan đến lĩnh vực mình nghiên cứu thì sẽ rất khó khăn cho việc tìm kiếm . Thực tế này đặt ra cần xây dựng một hệ thống tìm kiếm các bài báo khoa học giúp người dùng có thể tìm kiếm các bài báo theo chủ đề hay lĩnh vực mà mình quan tâm, đồng thời với đó yêu cầu hệ thống thu thập các tài liệu từ Internet và phân loại những tài liệu theo những chủ đề để người dùng khi dễ dàng tìm kiếm từ tập dữ liệu này.

* ***Khái quát và kiến trúc của hệ thống :***

Hệ thống tìm kiếm các bài báo khoa học bao gồm chức năng sau :

***+ Thu thập và phân loại tài liệu từ một địa chỉ (link )*** mà người dùng cung cấp . Link này là địa chỉ của một web chia sẻ tài liệu mà tại đó người dùng có thể lấy được những bài báo khoa học . Các bài báo sau khi được tải xuống

***+ Thu thập các tài liệu theo một từ khóa của người dùng*** ,từ khóa này bao gồm

* Tên tác giả : Tên tác giả bài báo mà người dùng quan tâm.
* Tên hội nghị (Conference) : Tên của hội nghị mà người dùng muốn thu thập những bài báo.
* Tên lĩnh vực : Tên của lĩnh vực mà bài báo đề cập đến (Công nghệ thông tin, Y khoa, …)
* Tên chủ đề : tên chủ đề mà bài báo đề cập đến (Java, Data mining….) .
* Tên bài báo

Các bài báo khoa học tìm được có các định dạng theo file mà chương trình tìm kiếm được trên web (PDF,TXT,Doc ….)

* ***Kiến trúc hệ thống :***



Trong đó :

1 . Nếu người dùng nhập một đường link trực tiếp để hệ thống thu thập các bài báo tại link đó.

2. Nếu người dùng nhập vào hệ thống từ khóa .

*- Module SearchEngine :* Dùng các search engine có sẵn như Google, Yahoo! , Google Sholar API Hệ thống sẽ submit trực tiếp từ khóa của người dùng lên SearchEngine sau đó thu thập các link từ kết quả trả về để đưa vào hệ thống Crawl download tài liệu từ link đó .

*- Module Crawl :* sử dụng Crawler có sẵn giúp download các tài liệu có trong link được trả về .

*- Module Classification* : sử dụng để phân loại tài liệu khi hệ thống Crawl trả về . Trong module này có dùng các thuật toán như SVN để phân loại tài liệu.

## Mục tiêu & phạm vi đề tài.

***Mục Tiêu*** : Tìm và tải được tài liệu mà người dùng yêu cầu . Các bài báo thu thập được từ một link đòi hỏi phân chia được chuyên đề và lĩnh vực của bài báo.

***Phạm vi*** : Phạm vi tìm kiếm trong lĩnh vực công nghệ thông tin mà cụ thể là phần mềm , các chủ đề tìm kiếm về các môn học cũng như các lĩnh vực nghiên cứu .

## Nội dung và kế hoạch triển khai.

* Thời gian bắt đầu thực hiện đề tài 1/8/2010
* Nhóm sẽ cùng nhau nghiên cứu và tìm các giải pháp về vấn đề : Crawl, Search, Classification.
* Tham khảo các đề tài về Classification đã có trong dữ liệu của nhóm (bao gồm các bài luận văn Thạc Sĩ về vấn đề phân loại tài liệu ).
* Kế hoạch nghiên cứu và nội dung tìm hiểu , cài đặt được lên kế hoạch theo từng tuần . Vào cuối tuần là thời gian để các thành viên trong nhóm báo cáo và tổng kết lại các việc đã làm.

## Kết quả dự kiến.

* Xây dựng thành công hệ thống tìm kiếm các bài báo khoa học.
* Bước đầu phân loại được loại tài liệu khi download về .

## Tài liệu tham khảo dự kiến

* Sử dụng các API search Engine đã tìm hiểu (Google,Yahoo!) .
* Sử dụng một API của 1crawl .
* Thuật toán phân loại : SVN kết hợp KNN.
* Luận văn “Hệ thống nhận dạng và phân loại văn bản ” .

# NGHIÊN CỨU LIÊN QUAN

## Thu thập , tìm kiếm tài liệu.

Hiện nay trên thế giới có một số công cụ cho người dùng có thể tìm kiếm các bài báo trong đó có Google Scholar ( [http://scholar.google.com/](http://scholar.google.com/" \o "http://scholar.google.com/))

-    Google Scholar là Công cụ tìm kiếm dành cho học giả.  Google Scholar là công cụ tìm kiếm chuyên sâu giúp người dùng tìm kiếm các tài liệu học thuật (luận văn, luận án, sách, các bài báo khoa học,… ) về nhiều  lĩnh vực. Google Scholar có chức năng tìm kiếm đơn giản và chức năng tìm kiếm nâng cao để tăng tính chính xác và hiệu quả cho việc tìm thông tin. Ngoài ra người dùng còn có thể sử dụng chức năng ”Scholar preferences” để lựa chọn cách hiện thị kết quả tìm kiếm hay  chức năng “Library Links- online and offline” để tìm hiểu xem một thư viện đại học bạn đang quan tâm có tài liệu đó hay không.

- Về Google Sholar API hiện nay google không còn cung cấp trên Internet nữa nhưng chúng ta vẫn có thể sử dụng API của nó để thực hiện tìm kiếm một bài báo khoa học .

🡪 Đối với trong đề tài ta sử dụng được Google Sholar API ta sẽ được kết tốt hơn rất nhiều so với sử dụng Google Search Engine bình thường . Ví dụ khi ta tìm kiếm với từ khóa “Hoang Kiem” trên google Search Engine sẽ trả về các thông tin về Hoang Kiem nhưng không tập trung vấn đề ta cần tìm những bài báo của Hoang Kiem , ngược lại khi search trên Google Sholar kết quả trả về bao gồm những bài báo có tác giả Hoang Kiem . Như vậy trong việc giải quyết vấn đề tìm kiếm bài báo khoa học theo từ khóa,việc sử dụng Google Sholar API sẽ giúp kết quả tìm kiếm chính xác hơn .

## Phân loại tài liệu.

Trong phần này đã thực hiện việc khảo sát và nộp báo cáo cho Thầy ngày (17/4 )