ĐỀ CƯƠNG ĐỀ TÀI TỐT NGHIỆP

Mục lục

[1. Tên đề tài 2](#_Toc270602845)

[2. Đặt vấn đề: 2](#_Toc270602846)

[3. Mục tiêu và phạm vi của đề tài: 3](#_Toc270602847)

[3.1. Mục tiêu của đề tài: 3](#_Toc270602848)

[3.2. Phạm vị của đề tài: 3](#_Toc270602849)

[4. Các ứng dụng và nghiên cứu liên quan: 3](#_Toc270602850)

[4.1. Thư viện số ACM 3](#_Toc270602851)

[4.2. Thư viện số IEEExplore 4](#_Toc270602852)

[4.3. Thư viện số Citeseer 5](#_Toc270602853)

[4.4. Công cụ Jabref 6](#_Toc270602854)

[4.5 Digital Bibliography & Library Project (DBLP). 8](#_Toc270602855)

[4.6. Khung phân loại tại liệu theo lĩnh vực khoa học máy tính: 9](#_Toc270602856)

[5. Nội dung kế hoạch triển khai. 12](#_Toc270602857)

[5.1. Nội dung 12](#_Toc270602858)

[5.2. Kế hoạch triển khai: 14](#_Toc270602859)

[6. Kết quả dự kiến: 16](#_Toc270602860)

[7. Các tài liệu tham khảo chính dự kiến: 17](#_Toc270602861)

# Tên đề tài

Xây dựng dữ liệu chỉ mục các bài báo khoa học được thu thập và phân loại từ các thư viện số.

# Đặt vấn đề:

Trong quá trình học tập cũng như nghiên cứu khoa học việc tìm kiếm và tham khảo tài liệu khoa học là điều rất cần thiết và quan trọng. Nhưng hiện nay với tốc độ phát triển mạnh mẽ của công nghệ thông tin và truyền thông, cùng với những tiện ích chia sẽ tài liệu trên mạng Internet. Dẫn tới khối lượng tài liệu khá phong phú và đa dạng nhưng hầu hết chưa được phân loại rõ ràng và đôi lúc người khó khăn trong việc tìm kiếm. Vấn đề đặt ra cần có một công cụ tiện ích có khả năng tìm kiếm, thu thập thông tin chỉ mục của những bài báo khoa học từ trên mạng. Sau đó dựa vào nội dung của mỗi bài báo để phân loại các bài báo khoa học theo các chủ đề khác nhau. Giúp người dùng dễ dàng tìm kiếm và sử dụng hơn.

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 các thư viện điện tử (thư viện số) cung cấp hoặc được nhiều người dùng upload lên trên mạng internet làm cho khối lượng tài liệu ngày càng phong phú và đa dạng. Cùng với đó hiện nay có Vì vậy việc tìm kiếm tài liệu đặt ra yêu cầu: cần có một cơ sở dữ liệu về các bài báo khoa học được phân loại theo chủ đề mà bài báo đề cập để người dùng có thể dễ dàng tra cứu và tìm kiếm thông tin.

Các thư viện viện số phổ biến hiện nay như ACM [[1]](#footnote-2), Citeseer [[2]](#footnote-3), IEEE Xplore[[3]](#footnote-4) khi người dùng tra cứu một bài báo khoa học thì kết quả trả về từ hệ thống bao gồm thông tin về tác giả, tiêu đề, nơi công bố, nhà xuất bản, tóm tắt của bài báo, phần trích dẫn và đường dẫn để truy cập hoặc tải tài liệu đó về (đối với hai thư viện số ACM và IEEEXplore đòi hỏi người dùng phải đăng ký làm thành viên và phải trả phí trước khi tải về). Nhưng những thông tin mà thư viện số cung cấp không bao gồm loại chuyên đề của bài báo, như vậy những thông tin thu thập từ các thư viện số về cần thiết phải phân loại để xác định chuyên đề của bài báo.

Hệ thống xây dựng cơ sở chỉ mục các bài báo khoa học có chức năng thu thập các thông tin bài báo khoa học từ các thư viện số, sau đó dựa vào thông tin phần tóm tắt của mỗi bài báo để phân loại bài báo theo chủ đề. Các thông tin thu được lưu thành dữ liệu chỉ mục tạo cơ sở để xây dựng các hệ thống tìm kiếm thông tin bài báo.

# 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 thư viện số.
* Phân loại các bài báo sau khi hệ thống thu thập về theo từng chuyên mục, chuyên đề đã được định nghĩa trước.

## Phạm vị của đề tài:

* Thông tin về các bài báo được thu thập từ các thư viện số ACM, IEEExplore, Citeseer.
* Khung phân loại của bài báo khoa học được sử dụng trong đề tài thuộc lĩnh vực khoa học máy tính.

# Các ứng dụng và nghiên cứu liên quan:

Hiện nay có một số ứng dụng và nghiên cứu liên quan về vấn đề tạo cở sở dữ liệu thông tin chỉ mục các bài báo bao gồm các thư viện số và chương trình quản lý thư mục tham khảo. Sau khi tham khảo và tìm hiểu nhóm quyết định chọn khảo sát ba thư viện số sau: ACM DL(**A**ssociation for **C**omputing **M**achinery **D**igital **L**ibrary), **IEEE**xplore, **C**iteseer và một công cụ là Jabref.

## Thư viện số ACM

Thư viện số **ACM DL** là một phần của **ACM portal**, hệ thống lưu trữ các thông tin ngắn gọn dễ hiểu của các bài báo khoa học, các tạp chí và báo cáo hội nghị. Giúp cho người dùng dễ dàng tìm kiếm các tài liệu, bài báo theo yêu cầu. Ngoài ra, người dùng còn có thể tìm các bài báo, tạp chí và bài báo ở các hội nghị khác thông qua The Guide.

Đối với The ACM Digital Library người dùng có thể tìm kiếm bằng một từ khóa bao gồm:

* Thông tin về tác giả bài báo (Names, Institutions, Authors, Editors, Reviewers).
* Thông tin về nơi công bố (Publication Year, Publication Names, ACM Publications, All publications, Content Formats, Publishers).
* Thông tin về hội nghị (Sponsors, Events, Proceeding Series).

Thông tin lấy được từ bài báo bao gồm:

* Tóm tắt của bài báo (abstract).
* Citings (những bài báo tham chiếu đến bài báo này).
* References (những bài báo tham chiếu tới).
* Người dùng không được tải tài liệu về nếu không có tài khoản.

Các bài báo được phân chia chuyên đề theo khung phân loại được định nghĩa bởi ACM, khung phân loại dạng cây và được định nghĩa và bổ xung bởi người dùng. Khi một bài báo đưa lên thư viện số người dùng sẽ dựa vào hướng dẫn của ACM để chọn đúng chủ đề cho bài được đăng lên. Khi bài báo được đưa lên trên trang ACM sẽ có một đội ngũ các editor kiểm lại bài của người đăng.

## Thư viện số IEEExplore

* <http://ieeexplore.ieee.org> đây là trang web hỗ trợ tìm kiếm các bài báo khoa học. Hệ thống sẽ tìm kiếm các bài báo trong thư viện số IEEExplore dựa vào các khóa do người dùng nhập.
* Từ khóa là một từ hoặc cụm từ có trong một phần hoặc nhiều phần sau của một bài báo: phần tên bài báo, tên tác giả, tên hội thảo, hoặc một cụm từ trong phần tóm tắt (abstract) của bài báo. Nếu bài báo nào có chứa từ hoặc cụm từ này ở một trong các phần đã nêu ở trên thì sẽ được hiển thị ra ở trang kết quả tìm kiếm.
* Kết quả mà trang web này đưa ra sau khi tìm kiếm là một danh sách các bài báo, mỗi bài báo chưa các thông tin sau:
* Tên bài báo và đường dẫn của bài báo đó.
* Các đồng tác giả.
* Hội thảo của bài báo đó kèm theo đường dẫn của hội thảo đó.
* Mã số của bài báo.
* Thời gian công bố.
* Số trang.
* Tóm tắt sơ lược của bài báo.
* Định dạng của các bài báo hầu hết là các file PDF.
* Sau khi hiển thị các kết quả tìm kiếm ra, hệ thống này sẽ hỗ trợ người lọc các bài báo dựa vào các thông tin sau:
* Nơi lưu trữ từ khóa (Content type): Hội nghị, buổi hội thảo (conference), Journals.
* Năm công bố của bài báo: một năm xác định, hoặc trong một giai đoạn từ năm nào đến năm nào.
* Tên tác giả: 1 tên hoặc nhiều tên.
* Tên hội thảo.
* Tên nhà xuất bản.
* Môn học hay chuyên đề.
* Tên quốc gia, tên thành phố.

Chỉ có thành viên của trang web mới được phép xem toàn bộ bài báo hoặc tải các bài báo về máy. Nếu bạn không phải là thành viên thì bạn chỉ được tìm kiếm và xem một phần về các bài báo như đã nêu ở trên và không được xem toàn bộ bài báo cũng tải bài báo đó.

Thư viện số thêm dữ liệu do các tác giả của các bài báo đó nhập vào, hệ thống sẽ đưa ra một mẫu và yêu cầu tác giả hoặc tổ chức điền thông tin vào.

Thư viện số IEEExplore phân loại các bài báo khoa học theo tên bài báo và chủ đề của bài báo.

**Phân loại theo tiêu đề:**

* Các bài báo được sắp xếp tên bài báo dựa vào từ đầu tiên theo thứ tự abc trong bảng chữ cái. Giúp người dùng có thể tìm kiếm nhanh bài báo theo từ đầu tiên đó.
* Người dùng còn có thể tìm nhanh bằng cách nhập từ khóa hoặc một đoạn nhỏ trong tiêu đề.

**Phân loại theo chủ đề:**

* Thư viện số này chia ra nhiều loại chủ đề khác nhau như: communocation, networking & broadcasting, computing & processing, engineering profession… (http://ieeexplore.ieee.org/xpl/booksBySubjectCategory.jsp)

## Thư viện số Citeseer

Là một thư viện số mà tài liệu được thư viện cung cấp chủ yếu là về lĩnh vực máy tính. Thư viện số này dùng hệ thống Autonomous Citation Indexing (ACI) để đánh chỉ mục và tìm kiếm tài liệu từ đó tạo cơ sở để người dùng có thể tìm kiếm được các bài báo. Khi người dùng tra cứu thông tin về một bài báo thì hệ thống sẽ trả về các thông tin sau: Link download, các thông tin metadata (abtract, title, year).

## Công cụ Jabref

* + 1. Jabref là gì?

Jabref là một phần mềm quản lý tài liệu tham khảo của các bài báo khoa học, bằng cách sử dụng định dạng file Bibtex để lưu trữ thông tin. Các thông tin ở đây bao gồm các phần reference của một bài báo khoa học như: tên bài báo, tên tác giả, tên hội thảo của bài báo, năm công bố, tóm tắt của bài báo …

Chương trình được viết trên nền java có thể chạy được tốt trên hầu hết các hệ điều hành khác nhau.

Phiên bản đầu tiên của Jabref được công bố vào năm 2003 bởi Morten O. Alver and Nizar Batada và Jabref là viết tắt của “**J**ava, **A**lver, **B**atada, **Ref**erence”.

* + 1. Chức năng của chương trình Jabref.

Chức năng chính của Jabref là tìm kiếm và tạo cơ sở dữ liệu về từ các tài liệu tham khảo của các bài báo trên mạng.

* Sau khi tìm kiếm trên mạng chương trình sẽ hiện thị các kết quả tìm được như sau:



* + Các tài liệu được đánh dấu chữ “D” màu da cam là các tài liệu đã tồn tại trong cở sở dữ liệu.
  + Mặc định chương trình sẽ đánh dấu các tài liệu tìm được, nhưng chúng ta có thể bỏ chọn tất cả bằng “Deselect all”, hoặc chọn “Deselect all duplicates” để bỏ tất cả các tài liệu đã trùng trong cơ sở dữ liệu trước khi thêm chúng vào dữ liệu của bạn.
    1. Một số đặc điểm của Jabref
* Trình soạn thảo nâng cao định dạng file bibtex: cho phép chỉnh sửa các chi tiết của file Bibtex.
* Chức năng tìm kiếm: dùng để tìm kiếm các thẻ mục trong các bài báo khoa học.
* Tìm kiếm với Meline, ACM, Citeseer, IEEExplore và arXIV
* Phân loại các thực thể: cho phép bạn gom nhóm các đối tượng một cách rõ ràng, dựa vào các từ khóa.
* Có thể import database từ bên ngoài vào với nhiều loại định dạng file khác nhau như: Bibtex, CSA, endnote, refer, ISI, Medline, ovid, pubMed, RIS… không thể nhập dữ liệu từ file SQL.
* Có thể xuất ra nhiều loại định dạng file khác nhau như: bibtex, rdf, xml, txt, ods, ris, csv…
  + 1. So sánh Jabref với một số chương trình có cùng chức năng:

Xem thêm tại:

<http://en.wikipedia.org/wiki/Comparison_of_reference_management_software>

* Đây là chương trình miễn phí, mã nguồn mở.
* Chạy trên nền java và chạy tốt trên hầu hết các hệ điều hành.
* Các định dạng file có thể xuất ra như: Bibtex, Endnode, refer, Bibix, Medline, RIS, SQL…
* Các định dạng file có thể nhập vào: Bibtex, CSA, endnote, refer, ISI, Medline, ovid, pubMed, RIS… không thể nhập dữ liệu từ file SQL.
* Có thể tương tác với các chức năng xử lý từ như: microsoft word office, open office, Kile/lyX [10]...
* Có thể kết nối và lấy thông tin bài báo từ các thư viện số như: Meline, ACM, Citeseer, IEEExplore và arXIV.

## 4.5 Digital Bibliography & Library Project (DBLP).

DBLP cung cấp thông tin về chỉ mục các bài báo trong lĩnh vực khoa học máy tính, hệ thống được phát triển bởi trường đại học [Universität Trier](http://en.wikipedia.org/wiki/University_of_Trier) của [Germany](http://en.wikipedia.org/wiki/Germany). Ban đầu trang web chỉ tập trung vào cung cấp các thông tin về lĩnh vực DataBase systems và Logic Programming, sau đó hệ thống được mở rộng sang các mục khác trong khoa học máy tính.

Tính đến tháng 1/2010 DBLP chứa thông tin của 1,3 triệu bài báo trong lĩnh vực khoa học máy tính được thu thập từ các thư viện số, các hội nghị à các tạp chí.

DBLP không dùng một hệ quản trị cơ sở dữ liệu nào để lưu dữ liệu mà dữ liệu được ghi trong 125000 files.

Dữ liệu của DBLP được xuất ra các dang CDF, XML, SQL và người phát triển có thể download các file dữ liệu này từ trên web của chương trình.

Hiện nay có một số ứng dụng được xây dựng cho phép tìm kiếm các thông tin về các bài báo được rút ra từ dữ liệu của DBLP trong đó có CompleteSearch DBLP , [Faceted search](http://dblp.l3s.de) và DBL – Brown phổ biến và nhiều người dùng.

* CompleteSearch DBLP : cho phép tìm kiếm các thông tin từ người dùng đưa vào.
* [Faceted search](http://dblp.l3s.de) cho phép người dùng tìm kiếm thông tin theo những cách sau: theo metadata, theo tác giả và nơi công bố.
* DBL – Brown : Là chương trình sử dụng để tìm kiếm trên file dữ liệu DBLP offline Chương trình cho phép nhìn một cách trực quan về dữ liệu của một bài báo.

Ban dầu DBLP là một tập các danh sách các đề mục – mục lục (tables of contents– TOCs ) của các hội nghị và tạp chí trong lĩnh vực dữ liệu và lập trình logic . Các file TOCs được nhập bằng tay và theo định dạng HTML được link tới trang giới thiệu về dữ liệu đang thêm vào bằng thủ công .

Bước tiếp theo hệ thống sẽ phân tích file TOCs để xây dựng một "author pages". Một Author Page của một tác giả chứa danh sách các bài viết của tác giả và các bài viết có tác giả tham gia (đồng tác giả). Kết quả thu được từ sau bước này là file TOC\_OUT chứa thông tin chỉ mục của page hay tạp chí mà ta nhập vào. Chương trình sử dụng kết hợp một bộ parser tên mkhtml và file chứa tất cả tên của các tác giả tiến hành tạo thành file của chương trình.



*Hệ thống DBLP*

## Khung phân loại tại liệu theo lĩnh vực khoa học máy tính:

Khung được tham khảo tại Wiki[[4]](#footnote-5)

1. [**Theoretical computer science**](http://en.wikipedia.org/wiki/Computer_science#Theoretical_computer_science).
   1. Mathematical logic.
   2. Automatical theory.
   3. Number theory.
   4. Graph theory.
   5. Type theory.
   6. Catepory theory.
   7. Computational geometry.
   8. Quantum computing theory.
2. **Algorithms and data structures.**
   1. Analysis of algorithms.
   2. Algorithms.
   3. Data structures.
3. **Computer elements and architecture.**
   1. Digital logic.
   2. Microarchitecture.
   3. Multiprocessing.
4. **Computational science.**
   1. Numerical analysis.
   2. Computational physis.
   3. Computational chemistry.
   4. Bioinformatics.
5. **Artificial Intelligence.**
   1. Machine learning.
   2. Computer vision.
   3. Natural language processing/ Computational linguistics.
   4. Robotics.
   5. Image processing.
   6. Pattern recognition.
   7. Congnitive science.
   8. Evolutionary computation.
   9. Information retrieval.
   10. Knowledge representation.
6. **Software engineering.**
   1. Operating systems.
   2. Computer networks.
   3. Database.
   4. Computer security.
   5. Ubiquitous computing.
   6. Systems architecture.
   7. Compiler design.
   8. Programming languages.

# Nội dung kế hoạch triển khai.

## Nội dung

* + 1. Kiến trúc chương trình







5.1.2 Chức năng của chương trình:

- Thu thập thông tin các bài báo khoa học trên các thư viện số theo từ khóa của người dùng nhập vào. Từ khóa ở đây bao gồm các thông tin được các thư viện số hỗ trợ như khảo sát ở trên.

- Phân loại bài báo khoa học theo khung phân loại dựa trên abtract nà module thu thập về.

- Kiểm tra và lưu thông tin của các bài báo xuống database của chương trình

## Kế hoạch triển khai:

Thời gian dự kiến thực hiện trong 2/8/2010 đến 27/12/2010 trong 18 tuần.

|  |  |  |  |
| --- | --- | --- | --- |
| **Thời gian** | **Tuần** | **Công việc cần thực hiện** | **Kết quả dự kiến** |
| 2/8 -> 6/9 | 3 | * Viết đề cương của đề tài. * Tìm các bài báo tham khảo chính của đề tài. * Khảo sát các thư viện số và các chương trình, đề tài có cùng chức năng. * Tìm các đề tài, nghiên cứu về phân lớp tài liệu. | * Thảo luận đề cương từ đó chỉnh sửa và định ra hướng đi của đề tài. * Có được bản khảo sát về các thư viện số, khung phân loại của các thư viện số. * Tìm được các bài báo key để thực hiện trong đề tài. * Đưa ra được khung phân loại trong lĩnh vực khoa học máy tính. * Viết phần dẫn nhập của báo cáo. Và phần khảo sát về các thư viện số. * Tìm được hướng đi của đề tài sau khi nghe đóng góp ý kiến của Thầy và các nhón khác. |
| 6/9 – 20/9 | 2 | * Đọc các bài báo tham khảo chính để đưa ra các việc cần thực hiện trong chương trình. * Tham khảo cách rút trích các thông tin từ thư viện số của chương trình Jabref để từ đó thiết kế module rút trích các thông tin từ các thư viện số này. * Phân tích thiết kế cơ sở dữ liệu với nhóm có liên quan. | * Phân tích thiết kế phần lấy thông tin từ thư viện số. * Xem phần giải pháp tránh trùng lắp dữ liệu trong cơ sở dữ liệu. * Thiết kế cơ sở dữ liệu cho chương trình. * Viết phần phân tích thiết kế cơ sở dữ liệu cho chương trình vào báo cáo. |
| 20/9 – 4/10 | 2 | * Thiết kế giao diện chương trình. * Cài đặt và kiểm thử môdun rút trích thông tin từ các thư viện số. * Đọc các bài báo và bước đầu cài đặt thử phần phân loại tài liệu, rút đặc trưng. | * Hoàn thiện môdun rút trích thông tin từ thư viện số và cập nhật vào cơ sở dữ liệu. * Viết phần báo cáo về môdun rút trích thông tin. |
| 4/10- 18/10 | 2 | * Đọc tài liệu và các hệ thống phân loại tài liệu đã có (phương pháp, souce code chạy thử) để đưa ra hướng đi của nhóm. * Thu thập tài liệu để làm tài liệu thử nghiệm cho phần phân loại. | * Phân tích thiết kế phần phân loại tài liệu. * Có dữ liệu cho phần kiểm thử phân loại tài liệu. * Viết phần nghiên cứu các đề tài phân lọai tài liệu, thuật toán và hướng đi của đề tài. |

|  |  |  |  |
| --- | --- | --- | --- |
| 18/10 – 2/11 | 4 | * Cài đặt và kiểm thử môdun phân loại tài liệu | * Cài đặt xong các thuật toán phân loại tài liệu dựa trên tài liệu tham khảo. |
| 22/11 – 9/11 | 1 | * Test và kiểm tra toàn bộ chương trình | Đưa ra thống kê được tỉ lệ dựa trên kết quả trả về của chương trình. |
| 29/11 - 13/12 | 2 | Hoàn thiện báo cáo | Gửi Thầy báo cáo góp ý cho nhóm |
| 13/12 – 27/12 | 2 | Hoàn thiện báo cáo và chương trình |  |
| Tổng | 18 |  |  |

## Kết quả dự kiến:

* Thu thập được cơ sở dữ liệu từ các thư viện điện tử.
* Phân loại được tài liệu thuộc lĩnh vực khoa học máy tính dựa trên abtract mà module thu thập về (đã được định nghĩa trước trong khung phân loại).

# Các tài liệu tham khảo chính dự kiến:

[1]Ashwin Pulijala. Susan Gauch. *Hierarchical Text Classification*. Department of Electrical Engineering and Computer ScienceUniversity of Kansas .

[2] Aixin Sun and Ee-Peng Lim. *Hierarchical Text Classification and Evaluation*. Center for Advanced Information SystemsNanyang Technological University

[3] Koller D. and Sahami M. (1997). *Hierarchically Classifying Documents using Very Few Words*. International Conference on Machine Learning, pp.170-178, Volume 14, Morgan-Kauffman.

[4] Thorsten Joachims. *Text categorization with SVM lear with many relevant*

[5] Mladenic D. Grobelnik M. (1998). Feature Selection for Classification Based on Text Hierarchy.Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD-98.

[6] Tao Wang. *Document Classification with ACM Subject Hierarchy.* Electrical and Computer Engineering, 2007. CCECE 2007. Canadian Conference on

[7] Gui-Rong Xue. Dikan Xing. Qiang Yang.Yong Yu Deep. *Classification in Large-scale Text Hierarchies*

1. http://portal.acm.org/portal.cfm [↑](#footnote-ref-2)
2. <http://citeseerx.ist.psu.edu/> [↑](#footnote-ref-3)
3. <http://ieeexplore.ieee.org/Xplore/guesthome.jsp> [↑](#footnote-ref-4)
4. http://en.wikipedia.org/wiki/Computer\_science [↑](#footnote-ref-5)