ĐỀ CƯƠNG ĐỀ TÀI TỐT NGHIỆP

Mục lục

[1. Tên đề tài 1](#_Toc270589827)

[2. Đặt vấn đề: 2](#_Toc270589828)

[3. Mục tiêu và phạm vi của đề tài: 3](#_Toc270589829)

[3.1. Mục tiêu của đề tài: 3](#_Toc270589830)

[3.2. Phạm vị của đề tài: 3](#_Toc270589831)

[4. Các ứng dụng và nghiên cứu liên quan: 3](#_Toc270589832)

[4.1. Thư viện số ACM 3](#_Toc270589833)

[4.2. Thư viện số IEEExplore 4](#_Toc270589834)

[4.3. Thư viện số Citeseer 5](#_Toc270589835)

[4.4. Công cụ Jabref 5](#_Toc270589836)

[4.5. Khung phân loại tại liệu theo lĩnh vực khoa học máy tính: 8](#_Toc270589837)

[5. Nội dung kế hoạch triển khai. 11](#_Toc270589838)

[5.1. Nội dung 11](#_Toc270589839)

[5.2. Kế hoạch triển khai: 11](#_Toc270589840)

[6. Kết quả dự kiến: 13](#_Toc270589841)

[7. Các tài liệu tham khảo chính dự kiến: 13](#_Toc270589842)

# Tên đề tài

Tìm kiếm và quản lý các bài báo khoa học.

# Đặt vấn đề:

Tìm kiếm và quản lý tài liệu là một vấn đề không nhỏ. Ngày nay, theo đà phát triển của công nghệ thông tin và Internet, khối lượng tài liệu được chia sẻ trên mạng ngày càng phong phú và đa dạng hơn. Vì vậy vấn đề tìm kiếm và quản lý tài liệu nói chung trong đó tìm kiếm và quản lý các bài báo khoa học nói riêng nhầm phục vụ quá trình học tập cũng như nghiên cứu khoa học càng trở nên cần thiết hơn. Vấn đề đặt ra làm sao sắp xếp và phân loại các tài liệu một cách chi tiết rõ ràng giúp cho người dùng có thể tìm kiếm được nhanh chóng và chính xác các tài liệu, các bài báo khoa học mà họ cần.

Những thư 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, nhan đề, nơi công bố, nhà xuất bản, tóm tắt của bài báo, phần tham khảo và đường dẫn để truy cập hoặc tải tài liệu đó về (ACM và IEEEXplore đòi hỏi người dùng phải đăng ký làm thành viên và phải có trả phí). Vấn đề đặt ra của đề tài là xây dựng cơ sở dữ liệu về thông tin của các bài báo từ các thư viện trực tuyến này qua đó làm cơ sở để xây dựng các hệ thống tìm kiếm tra cứu thông tin bài báo khoa học.

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 sẽ được phân loại bài báo theo chủ đề.

# 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.
* 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:

* Xây dựng thư viện chỉ mục của các bài báo khoa học trong lĩnh vực khoa học máy tính.

# Các ứng dụng và nghiên cứu liên quan:

Hiện nay có rất nhiều ứng dụng khác nhau và nhiều nghiên cứu liên quan về vấn đề này. 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**) [4], IEEExplore[5], Citeseer [6].** Và một công cụ là **Jabref.**

## Thư viện số ACM

Thư viện số **ACM DL** là một phần của **ACM portal** [7], nó 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  chúng ta có thể tìm kiếm bằng một từ khóa bao gồm :

* Thông tin về tác giả bài báo ([Names](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=4), [Institutions](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=3208), [Authors](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=2), [Editors](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=3), [Reviewers](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=3209))
* Thông tin về nơi công bố ([Publication Year](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3221), [Publication Names](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=1), [ACM Publications](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3224), [All publications](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3225), [Content Formats](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3210), [Publishers](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3213))
* Thông tin về hội nghị ([Sponsors](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=5&dim=3207), [Events](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=5&dim=3198), [Proceeding Series](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=5&dim=3206). )

Thông tin lấy được từ bài báo bao gồm :

* Tóm tắt của bài báo (abstract)
* citings (where the paper has been referenced by other papers)
* references (by the paper to other papers)
* index terms from ACM's [Computing Classification System (CCS)](http://www.acm.org/class/)
* reviews from ACM's Computing Reviews.

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 xem lại bài của người đăng.

## Thư viện số IEEExplore

* Đây là trang web hỗ trợ tìm kiếm các bài báo khoa học. Hệ thống sẽ tìm kiếm các bài báo dựa trên từ khóa do người dùng nhập vào.
* 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. Thông tin của mỗi bài báo bao gồm:
* 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 lượt 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 dùng bỏ bớt kết quả hiển thị của các bài báo. Bằng cách đưa ra thêm các điều kiện nhằm tìm chính xác hơn bài báo mà người dùng mong muốn, bỏ bớt các bài báo mà người dùng cần. Bằng cách cho người dùng chọn thêm một số điều kiện 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 tải các bài báo này 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, không được xem toàn bộ bài báo và tải bài báo đó.

## 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 ngườ 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 [8]là một phần mềm quản lý tài liệu tham khảo của các bài báo khoa học, bằng cách sử dụng định dạng file Bibtex [9] 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[11].

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

Khung dựa trên wiki[12]

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. Cơ sở dữ liệus
   4. Computer security
   5. Ubiquitous computing
   6. Systems architecture
   7. Compiler design
   8. Programming languages

# Nội dung kế hoạch triển khai.

## Nội dung

* + 1. Mô hình hệ thống:
    2. Kiến trúc chương trình

## Kế hoạch triển khai:

Thời gian thực hiện trong 2/8/2010 đến

|  |  |  |  |
| --- | --- | --- | --- |
| **Thời gian** | **Tuần** | **Công việc cần thực hiện** | **Kết quả dự kiến** |
| 2/8 -> 30/8 | 2 | * 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. |
| 30/8 – 13/9 | 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. * Viết phần phân tích thiết kế cơ sở dữ liệu cho chương trình vào báo cáo. |
| 13/9 – 27/9 | 2 | * Cài đặt và kiểm thử môdun rút trích thông tin từ các thư viện số. | * Hoàn thiện môdun rút trích thông tin từ thư viện số vào cơ sở dữ liệu. * Viết phần báo cáo về môdun rút trích thông tin. |
| 27/9– 1/10 | 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. |
| 11/10 – 8/11 | 4 | * Cài đặt và kiêm thử môdun phân loại tài liệu | * Cài đặt xong các thuật toán phân loại tài liệu dựa trên |
|  |  |  |  |

## Kết quả dự kiến:

# Các tài liệu tham khảo chính dự kiến:

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)