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 5](#_Toc270602852)

[4.3. Thư viện số Citeseer 7](#_Toc270602853)

[4.4. Công cụ Jabref 7](#_Toc270602854)

[4.5 Digital Bibliography & Library Project (DBLP). 11](#_Toc270602855)

[4.6. Khung phân loại tại liệu theo lĩnh vực khoa học máy tính: 13](#_Toc270602856)

[5. Nội dung kế hoạch triển khai. 16](#_Toc270602857)

[5.1. Nội dung 16](#_Toc270602858)

[5.2. Kế hoạch triển khai: 20](#_Toc270602859)

[6. Kết quả dự kiến: 23](#_Toc270602860)

[7. Các tài liệu tham khảo chính dự kiến: 23](#_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 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 nạp lên trên mạng internet làm cho khối lượng tài liệu ngày càng phong phú và đa dạng.

Khi cần tìm kiếm một bài báo khoa học người dùng có thể vào các thư viện số như ACM [[1]](#footnote-2), Citeseer [[2]](#footnote-3), IEEE Xplore[[3]](#footnote-4) hoặc các dữ liệu chỉ mục như DBLP[[4]](#footnote-5) để tìm kiếm các bài báo. Nhưng hạn chế của các thư viện số cũng như dữ liệu chỉ mục là những thông tin của bài báo không bao gồm loại chủ đề mà bài báo đề cập đến hoặc nếu có thì chủ đề này được xác định một cách thủ công khi đưa lên thư viện số theo một khung phân loại có sẵn ACM Computing Classification[[5]](#footnote-6). Vì vậy cần xây dựng một cơ sở dữ liệu lưu các cơ sở chỉ mục các bài báo bao gồm cả thông tin về chủ đề của bài báo từ đó có thể xây dựng các ứng dụng cho phép người dùng tìm kiếm bài báo theo chủ đề.

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 và các thư viện chỉ mục có sẵn, sau đó dựa vào thông tin phần tóm tắt (abstract), tiêu đề (title) của mỗi bài báo để phân loại bài báo theo chủ đề. Các thông tin thu được lưu xuống 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 cơ sở dữ liệu chỉ mục có sẵn như DBLP [[6]](#footnote-7) và các thư viện số.
* 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 [[7]](#footnote-8).

# Các ứng dụng và nghiên cứu liên quan:

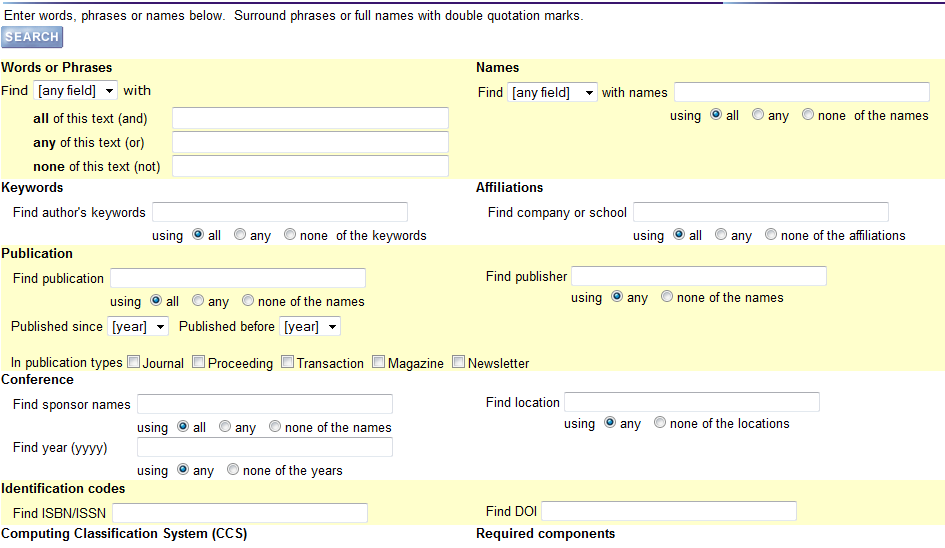
Hiện nay có một số ứng dụng và nghiên cứu liên quan về vấn đề tạo cở sở dữ liệu thông tin chỉ mục các bài báo bao gồm các thư viện số và chương trình quản lý thư mục tham khảo. Sau khi tham khảo và tìm hiểu nhóm quyết định chọn khảo sát ba thư viện số sau: ACM DL(**A**ssociation for **C**omputing **M**achinery **D**igital **L**ibrary), **IEEE**xplore, **C**iteseer và một công cụ là Jabref [[8]](#footnote-9).

## Thư viện số ACM

ACM (**Association for Computing Machinery**) là một tổ chức họat động trong lĩnh vực máy tính, tổ chức này cung cấp một thư viện số ACM Portal [[9]](#footnote-10) mà tại đây người dùng có thể tìm kiếm các bài báo được công bố bởi ACM và các tổ chức khác.

Người dùng có thể nhập vào ô tìm kiếm từ hoặc cụm từ, hệ thống sẽ tự động tìm kiếm và trả về kết quả tương ứng với từ hoặc cụm từ tìm kiếm.

Hệ thống thư viện cũng cung cấp chức năng tìm kiếm nâng cao, với chức năng này người dùng có thể hiệu chỉnh tìm kiếm của mình khi thêm một số thông tin như năm xuất bản, từ khóa trong bài báo, nhà xuất bản, tên hội nghị.



*Một số tùy chọn trong tìm kiếm nâng cao*

Thông tin lấy được từ kết quả trả về của một bài báo bao gồm:

* Tóm tắt của bài báo (abstract).
* Những bài báo tham chiếu đến bài báo này (Citings).
* Những bài báo được bài báo này tham chiếu tới (References).
* Người dùng không được tải tài liệu về nếu không có tài khoản.
* Trong hệ thống các bài báo được phân chia chuyên đề theo khung phân loại được định nghĩa bởi ACM (ACM Computing Classification System [[10]](#footnote-11) - CCS), khung phân loại có cấu trúc dạng cây và được định nghĩa, bổ xung khi một phiên bản mới ra (từ khi được đưa ra cho tới năm 2010 đã có 3 lần bổ sung năm 1964, 1991 và 1998). 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 chuyên gia sẽ kiểm lại bài của người đăng.

Cấu trúc của CCS gồm 1473 node được chia trên cây phân lớp thành 3 cấp, cấp đầu tiên bao gồm 11 node, cấp sau đó gồm 81 node và còn lại thuộc cấp số 3 (1998 version). Bảng dưới đây thể hiện các node trong cấp đầu tiên của khung lớp CCS.

|  |  |  |
| --- | --- | --- |
| **ID Node đầu tiên** | **Tên Node** | **Chú thích** |
| A. | General Literature | Tài liệu tổng quát |
| B. | Hardware | Phần cứng |
| C. | Computer systems organization | Các tổ chức máy tính |
| D. | Sofware | Phần mềm |
| E. | Data | Dữ liệu |
| F. | Theory of computation | Nguyên lý tính toán. |
| G. | Mathematics of computing | Toán trong máy tính. |
| H. | Information Systems | Hệ thống thông tin. |
| I. | Computing Methodologies |  |
| J. | Computer Applications | Cácứng dụng máy tính. |
| K. | Computing Milieux |  |

## 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 chứa 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), tên tổ chức công bố(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: một 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 người dùng không phải là thành viên của website thì người dùng chỉ được phép tìm kiếm và xem một phần về các bài báo như đã nêu ở trên và không được xem toàn bộ bài báo cũng như tải bài báo đó.

Dữ liệu của thư viện số được các tác giả của các bài báo nhập vào, hệ thống sẽ đưa ra một mẫu và yêu cầu tác giả hoặc tổ chức điền những thông tin cần thiết.

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 [[11]](#footnote-12) :**

* Thư viện số này chia ra nhiều loại chủ đề khác nhau như: communocation, networking & broadcasting, computing & processing, engineering profession…

## 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. ACI đánh dấu các trích dẫn của bài báo dựa vào một số đặc trưng và một vài thuật toán.

Ví dụ như: Trong một bài báo thì thông tin tác giả hầu như là tiếp theo sau phần tiêu đề của bài báo. Citeseer còn sử dụng database về tên tác giả, tên tổ chức… Để so khớp và rút thông tin.

Ngoài ra bài báo Digital libraries and autonomous citation Indexing, tác giả đã nói bốn phương pháp nhận dạng nhóm các trích dẫn của bài báo.

* String distance or edit distance (khoảng cách chuỗi)
* Word frequency and word occurrence (phương pháp dựa vào tuần suất xuất hiện của từ trong văn bản).
* Knowledge about subfields or the structure (tri thức về phần con hoặc cấu trúc). Trong trường hợp là các trích dẫn, subfields như tên tác giả, tiêu đề, năm công bố, …
* Probabilistic model (mô hình xác xuất) mô hình rút thông tin trong bài báo. (A Probabilistic Model of Redundancy in Information Extraction.)

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”.

Lưu ý khi tài liệu đã được thêm vào cơ sở dữ liệu của Jabref thì không được thay đổi vị trí cũng như tên file. Nếu thay đổi thì Jabref sẽ không tìm thấy và không quản lý được tài liệu đó.

* + 1. Chức năng của chương trình Jabref.

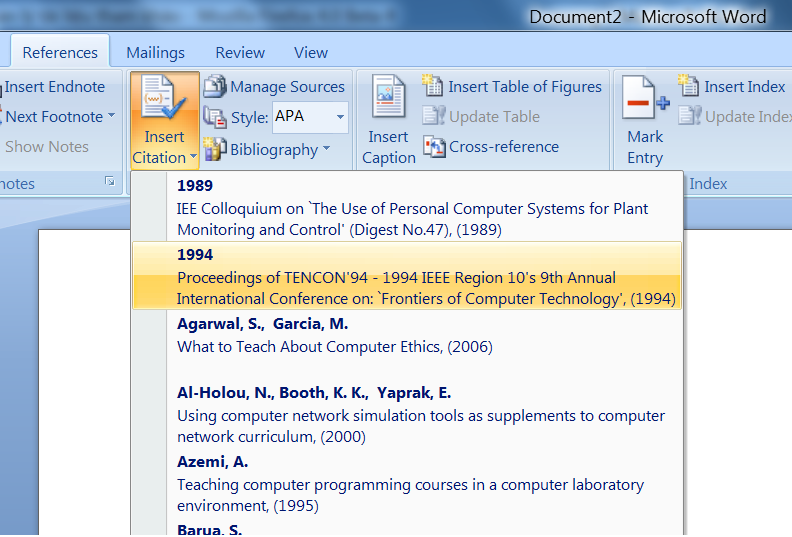
Chức năng của Jabref là tạo cơ sở dữ liệu thông tin các bài báo khoa học được lưu trong máy hoặc các thông tin bài báo được tìm và lấy về từ các thư viện số. Cơ sở dữ liệu này có thể dùng để tra cứu và điền thông tin trích dẫn trong quá trình viết báo cáo.

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



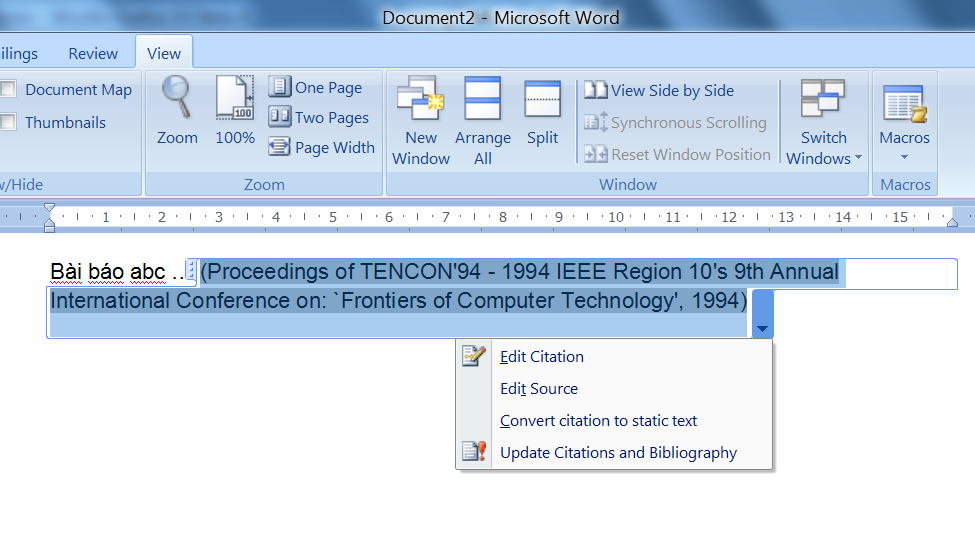
Hình 4.1: Kết quả tìm kiếm từ trên internet của Jabref

* + 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. Jabref tạo file Bibtex cho tham khảo:
  + Có thể kết hợp Jabref để tạo tài liệu tham khảo rất tốt. Trong các bài báo, sách, báo cáo... khi khối lượng tài liệu tham khảo lớn và không muốn phải thao tác bằng tay.
  + Jabref còn hỗ trợ chèn tài liệu tham khảo vào trong văn bản word.



Hình 4.2: Chèn tài liệu tham khảo vào word.

* + Chọn bài báo cần trích dẫn.



Hình 4.3: Trích dẫn của bài báo sau khi được thêm vào.

* + 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 Đức. Ban đầu trang web chỉ tập trung vào cung cấp các thông tin về lĩnh vực cơ sở dữ liệu (DataBase systems) và lập trình logic (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 dạng CDF, XML và SQL, 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 Complete Search DBLP , [Faceted search](http://dblp.l3s.de) và DBL – Browser.

* CompleteSearch DBLP: cho phép tìm kiếm các thông tin từ người dùng đưa vào theo các loại sau:

+ Tìm kiếm theo từ.

+ Tìm theo tên tác giả.

+ Tìm theo tên tổ chức công bố bài báo.

+ Theo năm xuất bản của bài báo.



* [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 – Browser: Là chương trình sử dụng để tìm kiếm trên file dữ liệu DBLP offline Chương trình cho phép nhìn một cách trực quan về thông tin của 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

DBLP được xây dựng lên từ các file sách các đề mục – mục lục (tables of contents– TOCs) của các hội nghị và tạp chí. 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

1. Cơ sở toán học **(Theoretical computer science)**.
   1. Logic toán (Mathematical logic).
   2. Lý thuyết số (Number theory).
   3. Lý thuyết đồ thị (Graph theory).
   4. Lý thuyết kiểu (Type theory).
   5. Lý thuyết phân loại (Catepory theory).
   6. Hình học tính toán (Computational geometry).
   7. Giải tích số (Numerical analysis).
   8. Lý thuyết Ôtômat (Automata theory)
2. **Cấu trúc dữ liệu và giải thuật (Algorithms and data structures).**
   1. Phân tích thuật toán (Analysis of algorithms).
   2. Thuật toán (Algorithms).
   3. Cấu trúc dữ liệu (Data structures).
3. **Kiến trúc máy tính (Computer** **architecture).**
   1. Logic số (Digital logic).
   2. Kiến trúc vi mô (Micro architecture).
   3. Đa xử lý (Multiprocessing).
4. **Khoa học tính toán (Computational science).**
   1. Phân tích số học (Numerical analysis).
   2. Vật lý tính toán (Computational physics).
   3. Hóa học tính toán (Computational chemistry).
   4. Tin sinh học (Bioinformatics).
5. **Trí tuệ nhân tạo (Artificial Intelligence).**
   1. Máy học (Machine learning).
   2. Thị giác máy (Computer vision).
   3. Xử lý ngôn ngữ tự nhiên (Natural language processing / Computational linguistics ).
   4. Rôbô học (Robotics).
   5. Xử lý ảnh (Image processing).
   6. Nhận dạng mẫu (Pattern recognition).
   7. Khoa học nhận thức (Congnitive science).
   8. Sự sống nhân tạo (Evolutionary computation).
   9. Rút trích thông tin (Information retrieval).
   10. Biểu diễn tri thức (Knowledge representation).
   11. Giao tiếp người máy (Human–computer interaction)
   12. Khai phá dữ liệu (Data Mining)
6. **Kỹ thuật phần mềm (Software engineering).**
   1. Hệ điều hành (Operating systems).
   2. Mạng máy tính (Computer networks).
   3. Cơ sở dữ liệu (Database).
   4. Bảo mật máy tính (Computer security).
   5. Ubiquitous computing.
   6. Kiến trúc hệ thống (Systems architecture).
   7. Thiết kế trình biên dịch (Compiler design).
   8. Ngôn ngữ lập trình (Programming languages).

# Nội dung kế hoạch triển khai.

## Nội dung

* + 1. Kiến trúc chương trình



Mô tả kiến trúc hệ thống:

* Người dùng tương tác với module thu thập tài liệu của hệ thống bằng cách nhập từ khóa vào sau đó yêu cầu hệ thống tìm kiếm. Hệ thống sẽ tìm các bài báo khoa học trên các thư viện số sau đó kết quả trả về được xử lý trong module lấy thông tin.
* Module lấy thông tin sẽ phân tích kết quả trả về để lấy những thông tin về metadata của bài báo từ đó người dùng có thể lưu lại kết quả hoặc chọn chức năng phân lớp chủ đề của bài báo. Hệ thống cho phép chỉnh sửa,bổ xung thông tin cần thiết về dữ liệu vừa rút ra được
* Module phân lớp theo chủ đề dựa vào tiêu đề hoặc tóm tắt của các bài báo vừa được thu thập được xác định chủ đề của bài báo theo khung phân loại của chương trình.
* Khi lưu dữ liệu hệ thống sẽ kiểm tra các bài báo vừa tìm kiếm được xem có trùng lặp với các dữ liệu đã có trong cơ sở dữ liệu không. Nếu trùng thì loại bỏ tài liệu đó ra.
* Cuối cùng lưu các kết quả về thông tin bài báo tìm kiếm được xuống cơ sở dữ liệu sau khi đã kiểm tra tính trùng lặp của dữ liệu.



Mô tả module thu thập:

* Khi người dùng đưa ra yêu cầu tìm kiếm các bài báo khoa học theo tên tác giả, tiêu đề hoặc hội nghị (các thông tin cho phép người dùng tìm kiếm được quy định bởi các thư viện số). Hệ thống sẽ tìm kiếm thông tin các bài báo liên quan từ ba thư viện số là ACM, Citeseer, IEEExplore.
* Kết quả trả về từ hệ thống thư viện số là đường dẫn tới các bài báo mà hệ thống tìm được. Những đường link này là đầu vào cho module lấy thông tin các bài báo khoa học.



Mô tả module phân lớp:

* Đầu tiên cho hệ thống học từ tập dữ liệu học gồm có các bài báo khoa học và khung phân loại chủ đề thuộc lĩnh vực khoa học máy tính đã nói ở trên.
* Thông tin của bài báo vừa được thu thập được đưa vào module phân loại. Từ tập dữ liệu học và thông tin đưa vào hệ thống sẽ sử dụng thuật toán SVM [[12]](#footnote-13) để phân loại các bài báo ra theo từng chủ đề.
* Thông tin về chủ đề của bài báo cùng với những thông tin khác được ghi xuống cơ sở dữ liệu sau khi kiểm tra tính trùng lắp.



Module import dữ liệu :

* Từ cơ sở dữ liệu các bài báo khoa học có sẵn hệ thống này sẽ lấy ra các thông tin metadata như tiêu đề, tác giả, hội nghị, năm công bố, tham khảo.
* Các bài báo từ dữ liệu cơ sở có sẵn sẽ được xác định chủ đề sau đó thêm vào cơ sở dữ liệu của chương trình.



Mô tả module lấy thông tin:

Mỗi bài báo tìm được sẽ được rút các thông tin metadata bằng các parser và hiển thị lên cho người dùng. Thông tin metadata gồm tiêu đề, tác giả, hội nghị, năm công bố, tóm tắt, phần tham khảo.

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 tương ứng hỗ trợ như khảo sát ở trên.
* Cập nhật dữ liệu từ các cơ sở chỉ mục có sẵn trong đề tài tà DBLP.
* Phân loại bài báo khoa học theo khung phân loại dựa trên abstract, title được rút ra từ thông tin của bài báo.
* 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 3/1/ 2010 trong 20 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 nhóm khác. |
| 6/9 – 20/9 | 2 | * Tìm kiếm và khảo sát các chương trình, đề tài tương tự (Expert Finder System for DBLP). * Liên hệ với các tác giả có liên quan để xin tài liệu, source code. * Đọc các bài báo tham khảo về phần xây dựng dữ liệu về thông tin chỉ mục các bài báo. * Phân tích thiết kế cơ sở dữ liệu (với nhóm có liên quan). * Thiết kế màn hình chương trình. | * Viết báo cáo trình bày về các hệ thống tương tự. * Có bảng thiết kế cơ sở dữ liệu. * Có thiết kế về màn hình chương trình các yêu cầu input, output của mỗi module. * Viết được phần các nghiên cứu liên quan về những hệ thống tương tự trong báo cáo chính |
| 20/9 – 4/10 | 2 | * 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. * Thiết kế module thu thập dữ liệu | * 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. |
| 4/10- 18/10 | 2 | * 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 bài báo từ thư viện số. |
| 18/10 – 1/11 | 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. |
| 1/11 – 29/11 | 4 | * Cài đặt và kiểm thử môdun phân loại tài liệu | * Cài đặt xong các thuật toán phân loại tài liệu dựa trên abstract và title của bài báo.   🡪 Viết vào báo cáo phần phân loại tài liệu của hệ thống. |
| 29/11 – 13/12 | 2 | * Import dữ liệu từ DBLP vào hệ thống, cho hệ thống chạy kết hợp các module | * Import được dữ liệu từ DBLP vào hệ thống (dữ liệu của 1 số tác giả). |
| 13/12 – 20/ 12 | 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. |
| 20/12 - 27/12 | 1 | * Hoàn thiện báo cáo | Gửi Thầy báo cáo góp ý cho nhóm |
| 27/12 – 3/1 | 1 | * Hoàn thiện báo cáo và chương trình |  |
| Tổng | 20 | Tổng kết và đánh giá. |  |

## Kết quả dự kiến:

* Thu thập được cơ sở dữ liệu từ các thư viện điện tử.
* Cập nhật dữ liệu được từ cơ sở dữ liệu chỉ mục DBLP.
* Phân loại được tài liệu thuộc lĩnh vực khoa học máy tính dựa trên abtract hoặc title 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*

[8] PhilipFisher-Ogden Chương trình Expert Finder System for [DBLP](http://www.informatik.uni-trier.de/%7Eley/db/) <http://code.google.com/p/dblp-expert-finder/>

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://dblp.mpi-inf.mpg.de/dblp-mirror/index.php> [↑](#footnote-ref-5)
5. <http://www.acm.org/about/class/ccs98-html> [↑](#footnote-ref-6)
6. http://www.informatik.uni-trier.de/~ley/db/index.html [↑](#footnote-ref-7)
7. http://en.wikipedia.org/wiki/Computer\_science [↑](#footnote-ref-8)
8. http://jabref.sourceforge.net/ [↑](#footnote-ref-9)
9. http://portal.acm.org/ [↑](#footnote-ref-10)
10. http://www.acm.org/about/class/1998 [↑](#footnote-ref-11)
11. http://ieeexplore.ieee.org/xpl/booksBySubjectCategory.jsp [↑](#footnote-ref-12)
12. [http://en.wikipedia.org/wiki/ Support\_vector\_machine](http://en.wikipedia.org/wiki/%20Support_vector_machine) [↑](#footnote-ref-13)