ĐỀ CƯƠNG ĐỀ TÀI TỐT NGHIỆP

Mục lục

[1. Tên đề tài 2](#_Toc274316451)

[2. Đặt vấn đề: 2](#_Toc274316452)

[3. Mục tiêu và phạm vi của đề tài: 4](#_Toc274316453)

[3.1. Mục tiêu của đề tài: 4](#_Toc274316454)

[3.2. Phạm vị của đề tài: 4](#_Toc274316455)

[4. Các ứng dụng và nghiên cứu liên quan: 5](#_Toc274316456)

[4.1 Thư viện số ACM 5](#_Toc274316457)

[4.2 Thư viện số IEEExplore 7](#_Toc274316458)

[4.3 Thư viện số Citeseer 11](#_Toc274316459)

[4.4 Thư viện số ScienceDriect 12](#_Toc274316460)

[4.5 Công cụ Jabref 14](#_Toc274316461)

[4.6 Digital Bibliography & Library Project (DBLP). 16](#_Toc274316462)

[5. Khung phân loại tại liệu theo lĩnh vực khoa học máy tính: 21](#_Toc274316463)

[6. Nội dung kế hoạch triển khai. 24](#_Toc274316464)

[6.1. Nội dung 24](#_Toc274316465)

[6.2. Kế hoạch triển khai: 29](#_Toc274316466)

[7. Kết quả dự kiến: 31](#_Toc274316467)

[8. Các tài liệu tham khảo chính dự kiến: 32](#_Toc274316468)

# Tên đề tài

Xây dựng dữ liệu chỉ mục các bài báo khoa học (Science Article Bibliography Database - DBSA).

# 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 và được nhiều người dùng chia sẻ trên mạng internet làm cho khối lượng tài liệu ngày càng phong phú và đa dạng. Song song với chức năng cung cấp chia sẻ tài liệu, các thư viện số cũng như các dữ liệu chỉ mục còn cho phép người dùng khả năng tìm kiếm tài liệu theo từ khóa, mặc dù vậy khi người dùng tìm kiếm bài báo theo những chủ đề mình đang quan tâm thì các hệ thống đáp ứng được một phần hoặc không cho phép tìm kiếm theo chủ đề vì những bài báo trong hệ thống không được phân loại theo chủ đề hoặc nếu được phân loại thì các chủ đề này chỉ khái quát không nói lên được trọng tâm của bài báo đề cập đến, khiến người dùng gặp nhiều khó khăn trong quá trình tìm kiếm và sử dụng nguồn tài liệu này.

Nhóm xin giới thiệu một số thư viện (ACM[[1]](#footnote-1), Citeseer[[2]](#footnote-2), IEEE Xplore[[3]](#footnote-3) , ScienceDriect[[4]](#footnote-4)) và một cơ sở dữ liệu chỉ mục (DBLP[[5]](#footnote-5)) mà người dùng có thể tìm kiếm hầu hết các bài báo khoa học liên quan đến lĩnh vực công nghệ thông tin

Thư viện số ACM sử dụng một khung phân loại chủ đề các bài báo đó là ACM Computing Classification[[6]](#footnote-6), khung phân loại này được áp dụng cho các bài báo của tổ chức ACM (**Association for Computing Machinery**), cũng như bài báo của tổ chức khác được thư viện ACM cung cấp. Khi mỗi bài báo được thêm vào thư viện thì việc xác định chủ đề của bài báo được làm thủ công dựa trên hướng dẫn và khung phân loại của tổ chức. Kết quả tìm kiếm của người dùng từ thư viện số không bao gồm chủ đề của bài báo.

Thư viện số Citeseer chứa một số lượng lớn các bài báo khoa học của rất nhiều các tổ chức, người dùng có thể tìm kiếm các bài báo theo các thông tin về tác giả và tựa đề bài báo, hệ thống không thể tìm kiếm bài báo theo chủ đề cũng như không có một khung phân loại chủ đề riêng.

Thư viện số IEEE Xplore của tổ chức “Institute of Electrical and Electronics Engineers” cung cấp các bài báo khoa học liên quan đến lĩnh vực máy tính, tại thư viện số này người dùng có thể duyệt theo khung phân loại của thư viện số bao gồm 16 chủ đề, những chủ đề này khái quát các lĩnh vực chứ không chi tiết khiến cho người dùng gặp khó khăn trong việc tìm kiếm tài liệu theo chủ đề.



*Hình 1 : Khung phân loại của thư viện số IEEXplore*

Thư viện số ScienceDriect – một thư viện số cung cấp các bài báo, tạp chí cũng như sách thuộc rất nhiều lĩnh vực, tại đây người dùng có thể duyệt tài liệu theo chủ đề của bài báo, nhưng cũng như thư viện số IEEEXplore khung phân loại chủ đề của ScienceDriect cũng rất khái quát khiến cho việc tìm kiếm của người dùng gặp khó khăn.

Cơ sở dữ liệu chỉ mục DBLP chứa rất nhiều thông tin chỉ mục của các bài báo của rất nhiều hội nghị. Từ cơ sở dữ liệu này người dùng có thể tìm kiếm bài báo theo tác giả, theo thông tin metadata cũng như tựa đề bài báo. Nhưng thông tin về chủ đề của bài báo thì trong cơ sở dữ liệu chỉ mục không có.

Như vậy vấn đề đặt ra ở đây là cần xây dựng một cơ sở dữ liệu lưu chỉ mục các bài báo, thông tin lưu trữ bao gồm cả chủ đề của bài báo đề câph đến, 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 thông tin bài báo khoa học từ các cơ sở chỉ mục có sẵn và các thư viện số, 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 cho phép người dùng tìm kiếm 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: từ các cơ sở dữ liệu chỉ mục có sẵn như DBLP và các thư viện số ACM, Citeseer, IEEExplore , ScienceDriect.
* 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-7). Khung phân loại này được sử dụng vì đây là nguồn tham khảo đáng tin cậy được cộng đồng công nhận và sử dụng. Đối với lĩnh vực khoa học máy tính khung phân loại này chi tiết hơn chứ không khái quát như những khung phân loại của thư viện số cũng như ACM Computing Classification.

# Các ứng dụng và nghiên cứu liên quan:

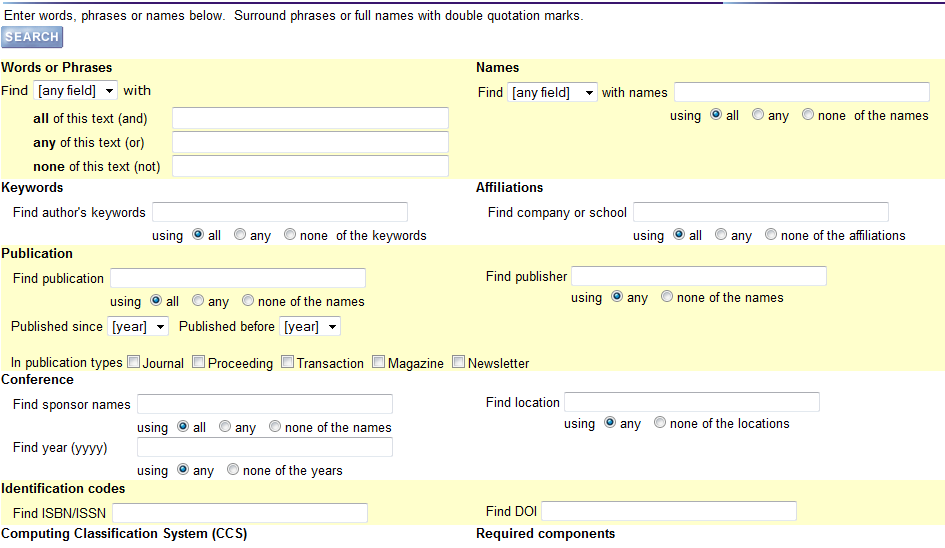
Trong phần này nhóm xin trình bày chi tiết các khảo sát về dữ liệu chỉ mục DBLP, các thư viện số:ACM DL(**A**ssociation for **C**omputing **M**achinery **D**igital **L**ibrary), **IEEE**xplore, **C**iteseer, Scient Driect và một công cụ cho phép tạo cơ sở chỉ mục và quản lý thông tin chỉ mục các bài báo Jabref [[8]](#footnote-8).

## 4.1 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-9) 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 mà trong tựa đề (title), tóm tắt bài báo (abstract) hoặc từ khóa bài báo (keyword) chứa những với từ hoặc cụm từ người dùng nhập vào ô 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.



*Hình 2: Một số tùy chọn trong tìm kiếm nâng caocủa ACM*

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-10) - 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 xung: năm 1964, 1991 và 1998). Khi một bài báo đưa lên thư viện số, tác giả bài báo sẽ dựa vào hướng dẫn của ACM và CCS để chọn đúng chủ đề cho bài được đăng lên.Sau đó ACM sẽ có một đội ngũ các chuyên gia (editor) sẽ kiểm lại tính chính xác giữa nội dung bài báo và chủ đề được tác giả xác định.

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 (theo phiên bản năm 1998). 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 |  |

*Bảng 1: Các chủ đề trong CCS.*

## Thư viện số IEEExplore

Thư viện số IEEE Xplore của tổ chức “Institute of Electrical and Electronics Engineers” cung cấp các bài báo khoa học liên quan đến lĩnh vực máy tính. Tại thư viện số này người dùng có thể tìm được các bài báo công bố bởi tổ chức IEEE và các tổ chức khác như AIP, IET, IBM, AVS…

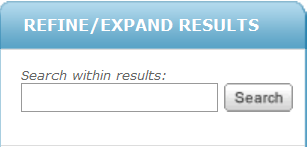
Từ khóa tìm kiếm được nhập từ người dùng là một từ hoặc cụm từ, hệ thống sẽ dựa vào những nội dung nhập vào và xem xét chúng có tồn tại trong các phần sau: phần tên bài báo, tên tác giả, tên hội thảo, hoặc trong phần tóm tắt (abstract) của bài báo. Nếu bài báo nào có chứa từ hoặc cụm từ này ở một trong các phần đã nêu ở trên thì sẽ được hiển thị ra ở trang kết quả tìm kiếm.

Thông tin lấy được trong bài báo từ kết quả trả về từ hệ thống bao gồm:

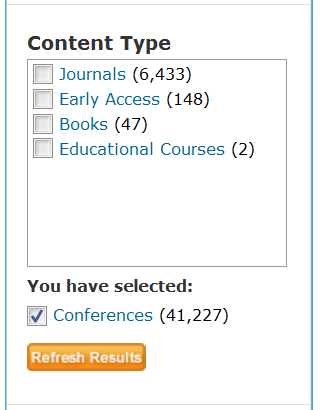
* Tên bài báo và đường dẫn của bài báo đó.
* Các đồng tác giả.
* Hội thảo của bài báo đó kèm theo đường dẫn của hội thảo đó.
* Mã số của bài báo.
* Thời gian công bố.
* Tóm tắt sơ lược của bài báo.

Sau khi hiển thị các kết quả tìm kiếm, hệ thống sẽ hỗ trợ người lọc các bài báo dựa vào các thông tin sau:

* + Tìm kiếm theo một từ khóa mới từ danh sách kết quả vừa tìm được.

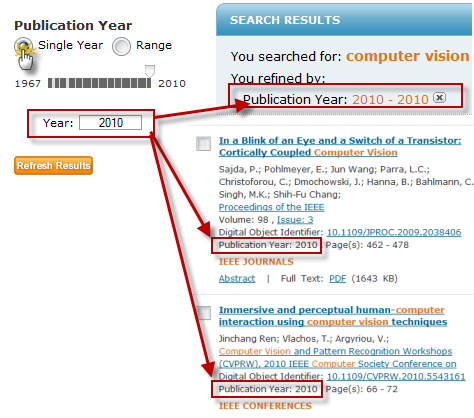


*Hình 3.Tìm kiếm theo từ khóa mới trong IEEE*

* Kiểu chứa (Content type): bài báo nằm trong: hội nghị (Conferences), tạp chí (journals), sách (books), khoa học giáo dục (Educational courses)… 

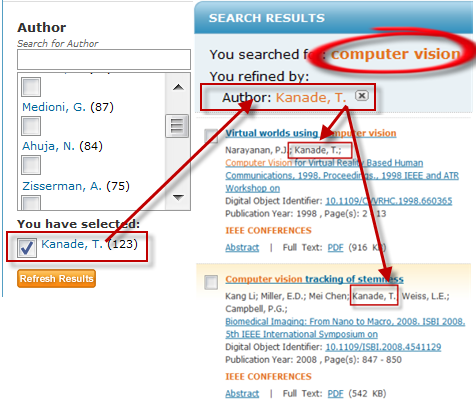
*Hình 4. Lọc lại kết quả tìm kiếm theo kiểu văn bản trong IEEEXplore.*

* Năm công bố của bài báo (Publication year): năm xác định, hoặc trong một giai đoạn từ năm nào đến năm nào. Ví dụ như ở hình bên dưới chọn tất cả các bài báo theo từ khóa “Computer vision” được công bố năm 2010.



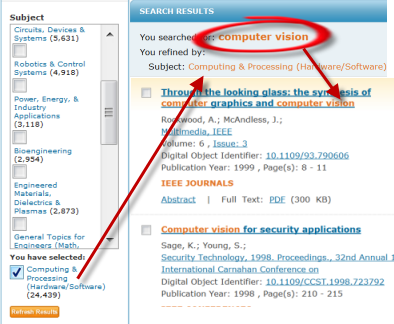
*Hình 5. Lọc lại kết quả tìm kiếm theo năm xuất bản trong IEEEXplore.*

* Tên tác giả (Author):



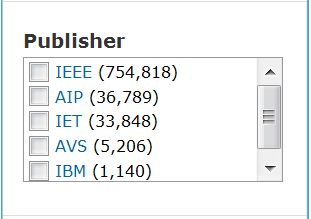
*Hình 6 Llọc lại kết quả tìm kiếm theo tác giả trong IEEEXplore.*

* Chủ đề (Subject).



*Hình 7. Lọc lại kết quả tìm kiếm theo chủ đề trong hệ thống IEEEXplore.*

* Tổ chức công bố. (Publisher)



*Hình 8. Lọc lại kết quả tìm kiếm theo tổ chức công bốtrong hệ thống IEEEXplore*

* Ngoài ra còn cho lọc lại theo thông tin về tên hội thảo, tên quốc gia, tên thành phố….

Chỉ có thành viên của trang web mới được phép xem toàn bộ bài báo hoặc tải bài báo về máy.

Các bài báo trong hệ thống được phân loại theo 16 chủ đề, những chủ đề này khái quát các lĩnh vực chứ không chi tiết khiến cho người dùng gặp khó khăn trong việc tìm kiếm tài liệu theo chủ đề.

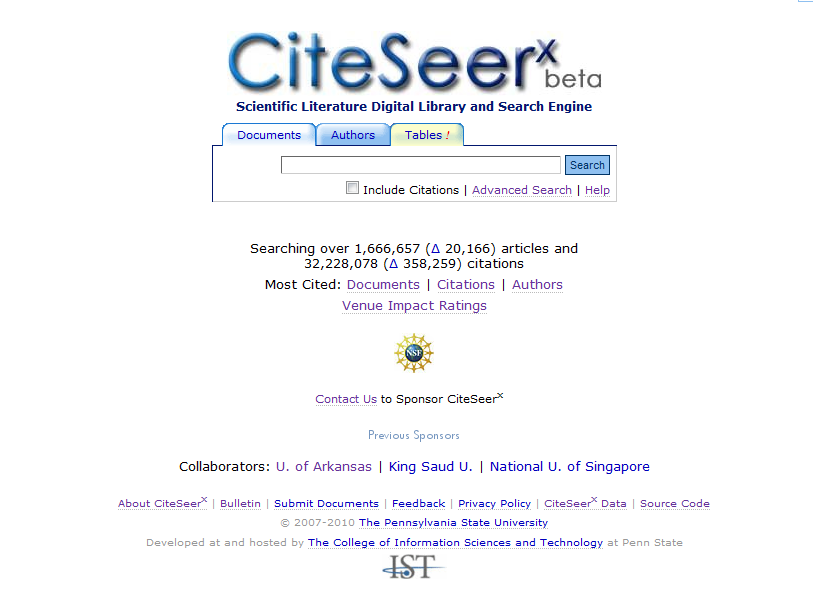


*Hình 9: Khung phân loại của thư viện số IEEXplore*

## 4.3 Thư viện số Citeseer

Citeseer là một thư viện số cung cấp các tài liệu khoa học, các bài báo khoa học được công bố bởi nhiều tổ chức khoa học. Các tài liệu này chủ yếu thuộc về lĩnh vực máy tính.

Hệ thống cho phép người dùng tìm kiếm theo từ khóa và tên tác giả bài báo. Hệ thông dựa vào sự xuất hiện của các từ khóa trên tiêu đề cũng như tom tắt (astract) của bài báo và tên tác giả để trả về kết quả tìm kiếm cho người dùng.



*Hình 10 Hệ thống tìm kiếm CiteSeer*

Mỗi bài báo trong danh sách trả về chứa các thông tin sau:

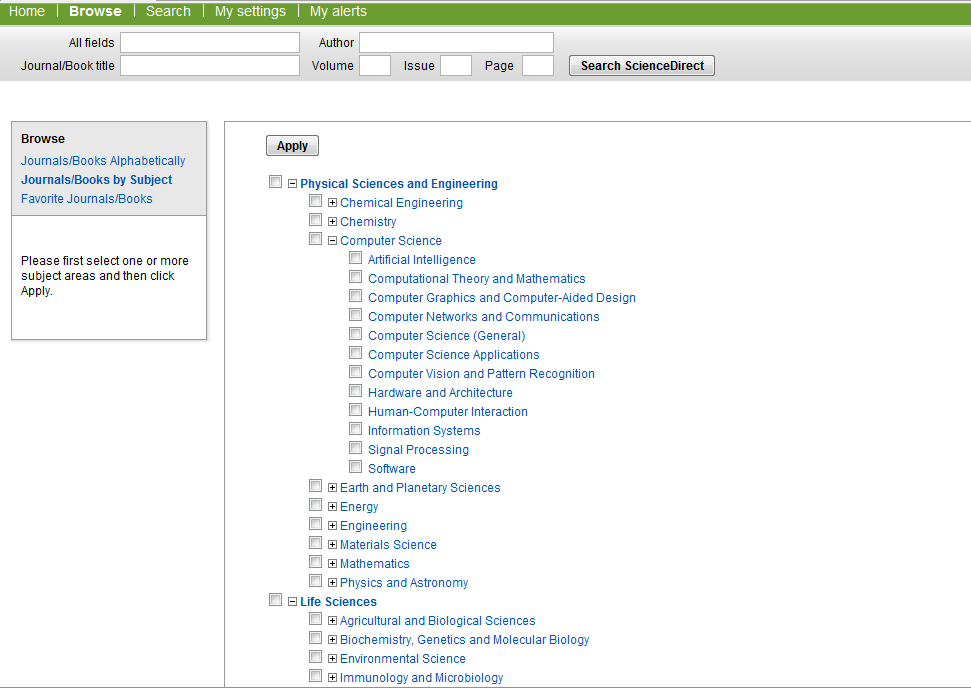
* Tên bài báo.
* Các đồng tác giả của bài báo đó
* Năm công bố.
* Tóm tắt của bài báo
* Các trích dẫn.
* Đường dẫn cho phép tải tài liệu về máy.

## Thư viện số ScienceDriect [[11]](#footnote-11)

Sciencedriect là một thư viện sốcó thu phí**,** cung cấp cho người dùng sách báo và tài liệu thuộc rất nhiều lĩnh vực trong đó có lĩnh vực công nghệ thông tin. Hiện nay trong cơ sở dữ liệu của nó chứa hơn 10 triệu bài báo, tạp chí và các chương sách. Trong đó có khoảng 2.500 tạp chí và 10 nghìn quyển sách, trung bình một năm có nửa triệu thông tin bài báo,tạp chí,sách được đưa thêm vào cơ sở dữ liệu.

Khi người dùng tìm kiếm một bài báo trên thư viện số này, hệ thống cho phép người dùng tìm kiếm theo các trường sau:

* Tên tác giả (Author).
* Tên bài báo, tựa đề sách báo, tạp chí (Journal/Book title).
* Thông tin về ấn bản bài báo (volume, page).

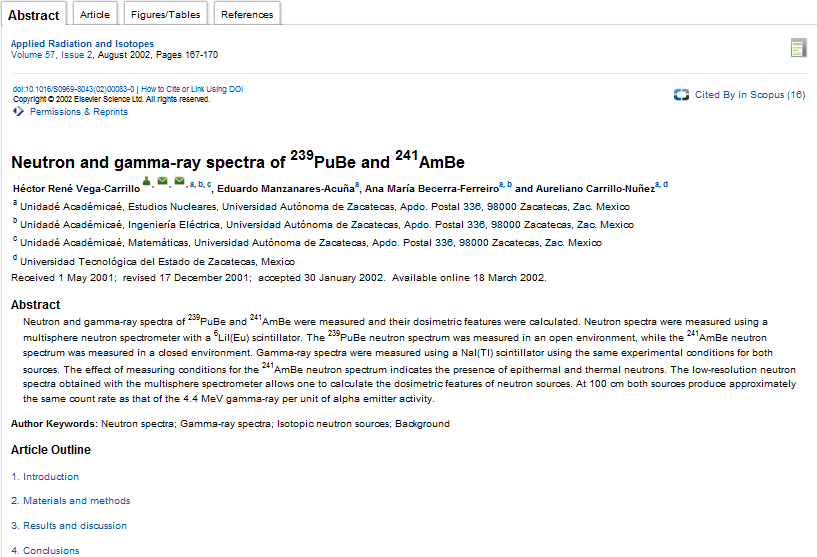


*Hình 11: Các trường tìm kiếm từ hệ thống ScienceDriect*

Kết quả tìm kiếm trả về từ thư viện số dựa vào từ tìm kiếm từ người dùng và sự xuất hiện từ tìm kiếm trong trường tương ứng của bài báo.

Thông tin bài báo hệ thống tìm kiếm trả về báo bao gồm:

* Tên tác giả bài báo.
* Tên tựa đề bài báo.
* Phần tóm tắt của bài báo (Abstract).
* Keywords của bài báo.
* Phác thảo về cấu trúc của bài báo.

****

*Hình 12 : Nội dung kết quả trả về của một bài báo trong hệ thống ScienceDriect*

## Công cụ Jabref

* + 1. Jabref là gì?

Jabref [[12]](#footnote-12) 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 lưu trữ ở đây bao gồm các phần tham chiếu (reference) của một bài báo khoa học cùng với: tên bài báo, tên tác giả, tên hội nghị, 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 và Nizar Batada, Jabref là viết tắt của “**J**ava, **A**lver, **B**atada, **Ref**erence”.

* + 1. Chức năng của chương trình Jabref.
* Jabref được sử dụng để quản lý tài liệu bao gồm: sách, báo, tạp chí… ở trên máy tính cá nhân cũng như thu thập từ hệ thống tìm kiếm trên các thư viện số của chương trình.
* Jabref cho phép tìm kiếm các tài liệu và gom nhóm tài liệu nhằm dễ quản lý hơn.
* Ngoài ra Jabref còn cho phép người dùng tạo file tài liệu tham khảo Bibtex để sử dụng trong quá trình viết báo cáo.
  + 1. Một số đặc điểm của Jabref [[13]](#footnote-13)
* Sau khi tìm kiếm trên mạng chương trình sẽ hiện thị các kết quả tìm được như sau:



*Hình 13: Kết quả tìm kiếm từ trên internet của Jabref*

* + Các tài liệu được đánh dấu chữ “D” 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.
* Đây là trình soạn thảo nâng cao định dạng file bibtex: cho phép chỉnh sửa các chi tiết của file Bibtex như tên tác giả, tiêu đề, năm xuất bản, tổ chức…
* Tìm kiếm thông tin bài báo từ các thư viện số trên internet như: Meline, ACM, Citeseer, IEEExplore và arXIV…
* Phân loại các thực thể: cho phép người dùng tìm kiếm 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. Nhận xét về Jabref với một số chương trình khác có cùng chức năng[[14]](#footnote-14)
* 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.

## 4.6 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 và 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[[15]](#footnote-15): cho phép tìm kiếm các thông tin từ người dùng đưa vào theo các loại sau:

+Tìm kiếm theo từ.

+Tìm kiếm theo tên tác giả.

+Tìm kiếm theo tên tổ chức công bố bài báo.

+Tìm kiếm theo năm xuất bản của bài báo.



*Hình 14 : Hệ thống Complete Search*

* [Faceted search](http://dblp.l3s.de) [[16]](#footnote-16) : là hệ thống tìmn kiếm trên dữ liệu của DBLP cho phép người dùng tìm kiếm thông tin theo những cách sau:

+Tìm kiếm dựa vào thông tin metadata bài báo.

+Tìm kiếm theo tên tác giả.

+Tìm kiếm theo nơi công bố bài báo.



*Hình 15: Hệ thống FacetedDBLP*

Ngoài ra chương trình còn cho phép người dùng duyệt tài liệu theo danh sách dựa trên tên tác giả, tên hội nghị, tên tạp chí hay từ khóa mà người dùng tìm kiếm nhiều nhất trong hệ thống.



*Hình 16: Hệ thống duyệt bài báo trong FacetedDBLP*

* DBL – Browser: Là chương trình sử dụng để tìm kiếm trên file dữ liệu DBLP không cần kết nối internet. Chương trình cho phép nhìn một cách trực quan về thông tin của bài báo.



*Hình 17: Chương trình DBL Brower*

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 bài báo hay tạp chí mà ta nhập vào. Chương trình sử dụng kết hợp một bộ phân tích cú pháp (parser) tên Mkhtml và file chứa tất cả tên của các tác giả tiến hành tạo thành file của chương trình.



*Hình 18: Hệ thống xây dựng cơ sở dữ liệu DBLP*

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

Khung được tham khảo tại Wiki, khung phân loại này được sử dụng vì đây là nguồn tham khảo đáng tin cậy được cộng đồng công nhận và sử dụng. Đối với lĩnh vực khoa học máy tính khung phân loại này chi tiết hơn chứ không khái quát như những khung phân loại của thư viện số cũng như ACM Computing Classification.

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



*Hình 19: 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.



*Hình 20: Kiến trúc module thu thập.*

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.



*Hình 21: Kiến trúc module phân lớp.*

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 [[17]](#footnote-17) 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.



*Hình 22: Kiến trúc module import dữ liệu.*

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.



*Hình 23: Kiến trúc module rút trích thông tin bài báo.*

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 giao diện 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 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 . 2010

[2] Aixin Sun and Ee-Peng Lim. *Hierarchical Text Classification and Evaluation*. Center for Advanced Information SystemsNanyang Technological University, IEEE International Conference on Data Mining (ICDM 2001), Califonia USA.

[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.

[7] Gui-Rong Xue. Dikan Xing. Qiang Yang.Yong Yu, *Deep* *Classification in Large-scale Text Hierarchies,* SIGIR'08, July 20-24, 2008, Singapore.

[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/>

[9*]* *A probabilistic Model of Redundancy in Information Extraction* – Doung Downey Etzini, and Stephen Soderland - 2005.

[10] Steve Lawrence, C. Lee Giles, Kurt Bollacker,*Digital Libraries and Autonomous Citation Indexing*, Computer, vol. 32, no. 6, pp. 67-71, June 1999

1. <http://portal.acm.org/portal.cfm> [↑](#footnote-ref-1)
2. <http://citeseerx.ist.psu.edu/> [↑](#footnote-ref-2)
3. <http://ieeexplore.ieee.org/Xplore/guesthome.jsp> [↑](#footnote-ref-3)
4. <http://www.sciencedirect.com> [↑](#footnote-ref-4)
5. <http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php> [↑](#footnote-ref-5)
6. <http://www.acm.org/about/class/ccs98-html> [↑](#footnote-ref-6)
7. http://en.wikipedia.org/wiki/Computer\_science [↑](#footnote-ref-7)
8. http://jabref.sourceforge.net/ [↑](#footnote-ref-8)
9. http://portal.acm.org/ [↑](#footnote-ref-9)
10. http://www.acm.org/about/class/1998 [↑](#footnote-ref-10)
11. <http://www.sciencedirect.com> [↑](#footnote-ref-11)
12. <http://jabref.sourceforge.net/> [↑](#footnote-ref-12)
13. <http://jabref.sourceforge.net/> [↑](#footnote-ref-13)
14. <http://en.wikipedia.org/wiki/Comparison_of_reference_management_software> [↑](#footnote-ref-14)
15. <http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php> [↑](#footnote-ref-15)
16. <http://dblp.l3s.de/?q=&newQuery=yes&resTableName=query_resultaE1pW9> [↑](#footnote-ref-16)
17. [http://en.wikipedia.org/wiki/ Support\_vector\_machine](http://en.wikipedia.org/wiki/%20Support_vector_machine) [↑](#footnote-ref-17)