MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

CHƯƠNG 1: TỔNG QUAN

1.1 Đặt vấn đề.

Cùng với sự phát triển của Internet, số lượng các bài báo khoa học được công bố trên các Web ngày càng tăng, điều này gây ra một số khó khăn khi người dùng muốn tìm kiếm các bài báo về vấn đề mà mình nghiên cứu, cũng như gây ra một thách thức lớn đối với các hệ thống đánh dấu, lưu trữ dữ liệu chỉ mục hỗ trợ tìm kiếm trong việc đảm bảo thông tin các bài báo được cập nhật đầy đủ, nhanh chóng và chính xác.

Hiện nay khi người nghiên cứu cần tìm kiếm một bài báo khoa học, thì họ có thể tìm kiếm trên các Search Engine như Google Scholar[1], và một số thư viện số phổ biến như: ACM[2] (thư viện số của tổ chức "Association for Computing Machinery"), IEEEXplore[3] (thư viện số của tổ chức "Institute of Electrical and Electronics Engineers"), thư viện mở CiteSeer[4] … hoặc từ cơ sở dữ liệu chỉ mục có sẵn như DBLP[5]. Vấn đề đặt ra ở đây là: đối với mỗi thư viện số thì việc cập nhật bài báo mới được thực hiện ngay khi có các cuộc hội thảo hay tạp chí mà tổ chức xuất bản, nhưng thư viện số không cập nhật ngay được những bài báo mới từ tổ chức khác - hay việc trao đổi dữ liệu giữa các thư viện số của các tổ chức khác nhau hiện nay còn rất hạn chế. Bên cạnh đó, những hệ thống đi đánh dấu, lưu trữ dữ liệu chỉ mục hiện nay như DBLP, hay hệ thống đi thu thập dữ liệu chỉ mục như ACI [3] của thư viện số CiteSeer chưa đảm bảo được tính cập nhật các bài báo mới, vì các nguồn lấy dữ liệu của các hệ thống phụ thuộc vào các thư viện số. Nhưng hiện nay, việc download tài liệu từ thư viện số bị giới hạn, cũng như các thuật toán sử dụng để rút

[1] http://scholar.google.com.vn/
[2] http://portal.acm.org
[3] http://ieeexplore.ieee.org
[4] http://citeseerx.ist.psu.edu/
[5] http://dblp.uni-trier.de/

trích thông tin chỉ mục từ các tài liệu download được chưa đạt được độ chính xác cao.

Xuất phát từ vấn đề trên cùng với sự định hướng của giáo viên hướng dẫn, chúng tôi phát triển một hệ thống dùng để xây dựng tích hợp làm giàu dữ liệu chỉ mục các bài báo khoa học, bằng cách rút trích thông tin bài báo trực tiếp từ các thư viện số, kết hợp với việc sử dụng dữ liệu chỉ mục có sẵn, để xây dựng lên dữ liệu chỉ mục các bài báo khoa học đảm bảo tính chính xác đầy đủ và cập nhật.

Hệ thống sử dụng Web Crawler để tìm kiếm và thu thập các bài báo khoa học được công bố trên các thư viện số (ACM, IEEEXplore, CiteSeer) sau đó sử dụng các luật cũng như các trình phân tích để rút trích thông tin chỉ mục - điều này đảm bảo dữ liệu thu thập có tính chính xác và cập nhật. Từ những thông tin chỉ mục thu thập được, hệ thống sẽ kết hợp với dữ liệu chỉ mục có sẵn trong DBLP để xây dựng lên một cơ sở dữ liệu chỉ mục các bài báo khoa học đảm bảo tính đầy đủ, chính xác và cập nhật.

Việc xây dựng dữ liệu chỉ mục các bài báo khoa học là rất cần thiết, thông qua dữ liệu chỉ mục xây dựng được, ta có thể phát triển các công cụ tìm kiếm bài báo khoa học đảm bảo nhu cầu tìm kiếm của người dùng.

1.2 Mục tiêu và phạm vi khóa luận.

1.2.1 Mục tiêu khóa luận.

- Mục tiêu của khóa luận là hướng tới xây dựng một hệ thống thu thập dữ liệu chỉ mục các bài báo khoa học đảm bảo được tính chất đầy đủ, chính xác và cập nhật của dữ liệu.

- Xây dựng một hệ thống có khả năng tự động cập nhật thông tin những bài báo mới nhất từ các thư viện số.

- Thông qua việc xây dựng hệ thống, các thành viên trong nhóm sẽ vận dụng những kiến thức của mình đã được học, cùng với đó trau dồi thêm các kỹ năng như: kỹ năng lập trình, kỹ năng làm việc nhóm …

1.2.2 Phạm vi khóa luận.

- Hệ thống sử dụng Web Crawler để thu thập thông tin chỉ mục các bài báo khoa học trên ba thư viện số ACM, CiteSeer, IEEEXplore.

- Hệ thống kết hợp dữ liệu thu thập được với dữ liệu có sẵn của DBLP, giúp thông tin thu thập được đảm bảo tính đầy đủ và cập nhật.

1.3 Kết quả dự kiến.

Có được cái nhìn tổng quan về các phương pháp xây dựng dữ liệu chỉ mục các bài báo khoa học hiện nay và kiến thức cụ thể về một số ứng dụng đã được xây dựng, để hỗ trợ cho việc xây dựng hệ thống cho riêng mình.

Xây dựng thành công hệ thống lưu trữ dữ liệu chỉ mục các bài báo khoa học bằng cách sử dụng Web Crawler trên các thư viện số, đồng thời kết hợp với việc sử dụng cơ sở dữ liệu chỉ mục có sẵn, để dữ liệu chỉ mục xây dựng được đảm bảo tính đầy đủ, chính xác và cập nhật.

1.4 Cấu trúc khóa luận

Chương 1 trình bày khái quát động cơ, mục tiêu và phạm vi của đề tài.

Chương 2 trình bày những nghiên cứu và hệ thống liên quan đến việc xây dựng dữ liệu chỉ mục các bài báo khoa học. Mục 2.2 trình bày sơ lược về các khái niệm liên quan, Mục 2.3 trình bày các nghiên cứu và các ứng dụng liên quan cùng với phần khảo sát các thư viện số mà hệ thống xây dựng trong khóa luận có sử dụng.

Chương 3 trình bày cách tiếp cận vấn đề xây dựng và làm giàu dữ liệu chỉ mục các bài báo khoa học sử dụng Web Crawler. Mục 3.2 trình bày phương pháp thu thập thông tin trên các thư viện số, Mục 3.3 trình bày cách thức phân tích dữ liệu để lấy thông tin bài báo khoa học. Cách kiểm tra trùng lặp dữ liệu được trình bày tại Mục 3.4. Trong mục 3.5 sẽ giới thiệu các luồng xử lý chính của hệ thống

Chương 4 Trình bày việc hiện thực hệ thống. Mục 4.2 trình bày kiến trúc hệ thống, Mục 4.3 trình bày thiết kế database, Mục 4.4 trình bày sơ đồ lớp của chương trình. Trong Mục 4.5 giới thiệu hệ thống mà khóa luận xây dựng được.

Chương 5 trình bày các thử nghiệm và đánh giá khi chạy hệ thống.

Chương 6 đưa ra kết luận và hướng phát triển hệ thống trong tương lai.

Phần phụ lục giới thiệu cách cài đặt hệ thống và hướng dẫn sử dụng chương trình và các chủ đề trong lĩnh vực khoa học máy tính được tham khảo từ Wikipedia.

CHƯƠNG 2: CÁC NGHIÊN CỨU VÀ HỆ THỐNG LIÊN QUAN

2.1 Mở đầu.

Trong chương 2, chúng tôi sẽ trình bày một số nghiên cứu và ứng dụng liên quan đến vấn đề thu thập, rút trích và xây dựng dữ liệu chỉ mục các bài báo khoa học. Phần đầu chúng tôi sẽ giới thiệu tổng quát về một số khái niệm trong vấn đề thu thập, rút trích dữ liệu, phần sau chúng tôi sẽ giới thiệu chi tiết về một số nghiên cứu, ứng dụng liên quan và những thư viện số có sử dụng trong hệ thống.

2.2 Một số khái niệm cơ bản.

2.2.1 Trích xuất thông tin (IE) và truy vấn thông tin (IR)

- **Trích xuất thông tin (Information Extraction[6])**

Theo tài liệu [19], trích xuất thông tin có nhiều định nghĩa được dùng phổ biến trên Internet:

- Theo (Jim Cowie and Yorick Wilks) [11]: IE là tên được đặt cho quá trình cấu trúc và kết hợp một cách có chọn lọc dữ liệu được tìm thấy, được phát biểu rõ ràng trong một hay nhiều tài liệu văn bản.
- Theo Line Eikvil [13]: IE là lĩnh vực nghiên cứu hẹp của xử lý ngôn ngữ tự nhiên và xuất phát từ việc xác định những thông tin cụ thể từ một tài liệu ngôn ngữ tự nhiên. Mục đích của trích xuất thông tin là chuyển văn bản về dạng có cấu trúc. Thông tin được trích xuất từ những nguồn tài liệu khác nhau và được biểu diễn dưới một hình thức thống nhất. Những hệ thống trích xuất thông tin văn bản không nhằm mục tiêu hiểu văn bản đưa vào, mà nhiệm vụ chính của nó là tìm kiếm các thông tin cần thiết liên quan, mà chúng ta mong muốn được tìm thấy.
- Cũng theo Line Eikvil [13], thành phần cốt lõi của các hệ thống trích xuất thông tin là một tập hợp các luật và mẫu dùng để xác định những thông tin liên quan cần trích xuất.

[6] http://en.wikipedia.org/wiki/Information_extraction

- Theo Tiến sĩ Alexander Yates ở trường đại học Washington [1] thì trích xuất thông tin là quá trình truy vấn những thông tin cấu trúc từ những văn bản không cấu trúc.
- Theo những chuyên gia về trích xuất thông tin của GATE[7] thì những hệ thống trích xuất thông tin sẽ tiến hành phân tích văn bản nhằm trích ra những thông tin cần thiết theo các dạng được định nghĩa trước, chẳng hạn như những sự kiện, các thực thể và các mối quan hệ.

Tóm lại, chúng ta có thể hiểu trích xuất thông tin (Information Extraction) là một kỹ thuật, lĩnh vực nghiên cứu có liên quan đến truy vấn thông tin (Information Retrieval), khai thác dữ liệu (Data mining), cũng như xử lý ngôn ngữ tự nhiên (Natural Language Processing). Mục tiêu chính của trích xuất thông tin là tìm ra những thông tin cấu trúc từ văn bản không cấu trúc hoặc bán cấu trúc. Trích xuất thông tin sẽ tìm cách chuyển thông tin trong văn bản không hay bán cấu trúc về dạng có cấu trúc và có thể biểu diễn hay thể hiện chúng một cách hình thức dưới dạng một tập tin cấu trúc XML hay một bảng cấu trúc (như bảng trong cơ sở dữ liệu chẳng hạn).

Một khi dữ liệu, thông tin từ các nguồn khác nhau, từ Internet có thể biểu diễn một cách hình thức, có cấu trúc. Từ đó chúng ta có thể sử dụng các kỹ thuật phân tích, khai thác dữ liệu (data mining) để khám phá ra các mẫu thông tin hữu ích. Chẳng hạn, việc cấu trúc lại các mẫu tin quảng cáo, mẫu tin bán hàng trên internet có thể giúp hỗ trợ tư vấn, định hướng người dùng khi mua sắm. Việc trích xuất và cấu trúc lại các mẫu tin tìm người, tìm việc sẽ giúp cho quá trình phân tích thông tin nghề nghiệp, xu hướng công việc, … hỗ trợ cho các người tìm việc, cũng như nhà tuyển dụng.

Rút trích thông tin không đòi hỏi hệ thống phải đọc hiểu nội dung của tài liệu văn bản, nhưng hệ thống phải có khả năng phân tích tài liệu và tìm kiếm các thông tin liên quan mà hệ thống mong muốn được tìm thấy. Các kỹ thuật rút trích thông

[7] http://gate.ac.uk/ie/

tin có thể áp dụng cho bất kỳ tập tài liệu nào mà chúng ta cần rút ra những thông tin chính yếu, cần thiết cũng như các sự kiện liên quan. Các kho dữ liệu văn bản về một lĩnh vực trên Internet là ví dụ điển hình, thông tin trên đó có thể tồn tại ở nhiều nơi khác nhau, dưới nhiều định dạng khác nhau. Sẽ rất hữu ích cho các khảo sát, ứng dụng liên quan đến một lĩnh vực nếu như những thông tin lĩnh vực liên quan được rút trích và tích hợp lại thành một hình thức thống nhất và biểu diễn một cách có cấu trúc. Khi đó thông tin trên Internet sẽ được chuyển vào một cơ sở dữ liệu có cấu trúc phục vụ cho các ứng phân tích và khai thác khác nhau.

- **Truy vấn thông tin (Information Retrieval[8])**

Theo [19], trích xuất thông tin là tìm ra các thông tin cấu trúc, thông tin cần thiết từ một tài liệu, trong khi truy vấn thông tin là tìm ra các tài liệu liên quan, hoặc một phần tài liệu liên quan từ kho dữ liệu cục bộ như thư viện số hoặc từ Internet để phản hồi cho người dùng tùy vào một truy vấn cụ thể.

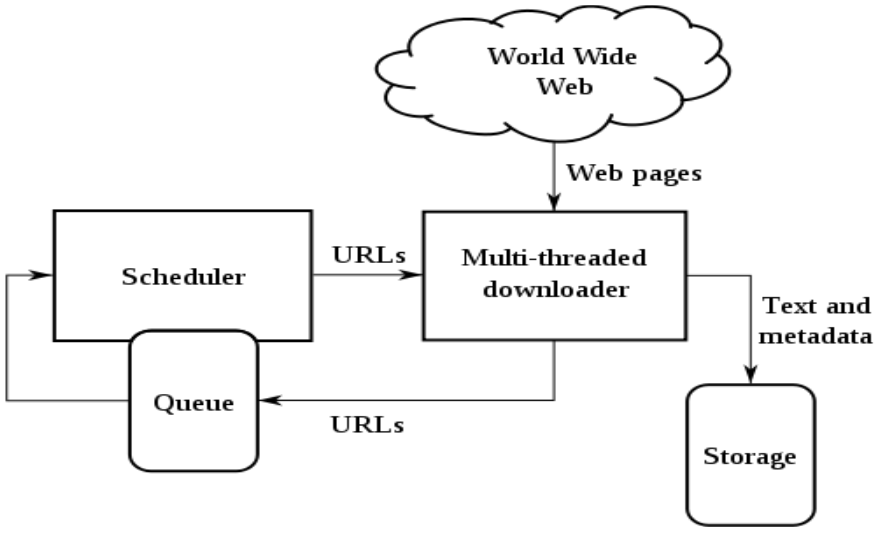
Truy vấn văn bản thông minh hướng tới tối ưu hay tìm kiếm các phương pháp nhằm cho kết quả phản hồi tốt hơn, gần đúng hoặc đúng với nhu cầu người dùng. Chẳng hạn tùy vào một truy vấn của người dùng, hệ thống có thể tìm ra những thành phần nào đó trong tài liệu phù hợp với câu truy vấn (chẳng hạn một đoạn, một câu trong tài liệu), thông minh hơn hệ thống có thể trả lời chính xác thông tin từ câu truy vấn hay câu hỏi của người dùng.

2.2.2 Web Crawler.

Theo định nghĩa trên Wikipedia [9], thì Web Crawler - Web Spider hay Web robot là một chương trình hoặc các đoạn mã có khả năng tự động duyệt các trang Web khác theo một phương thức tự động. Web Crawler thường được sử dụng để thu thập tài nguyên (như tin tức, hình ảnh, video …) trên Internet.

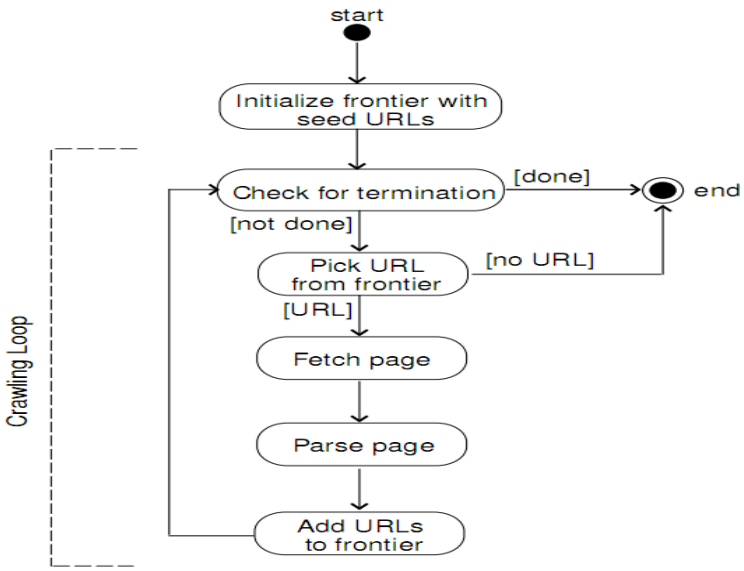
Quá trình thực hiện của Web Crawler là Web Crawling hay Web Spidering. Hầu hết các công cụ tìm kiếm online hiện nay đều sử dụng quá trình này để thu thập và cập nhập kho dữ liệu phục vụ nhu cầu tìm kiếm của người dùng.

[8] http://en.wikipedia.org/wiki/Information_retrieval
[9] http://en.wikipedia.org/wiki/Webcrawler



Hình 2.1- Kiến trúc Web Crawler (Wikipedia)

Web Crawler bắt đầu từ danh sách các địa chỉ URL được gọi là hạt giống (seeds), seeds được người dùng nhập vào - đây là những địa chỉ Web mà người dùng muốn thu thập thông tin. Hệ thống sẽ vào địa chỉ này, lọc thông tin rồi tìm ra các địa chỉ URL khác (dựa vào những liên kết có bên trong các seeds). Sau đó thêm chúng vào danh sách các địa chỉ đã được duyệt qua gọi là Crawl frontier. Hệ thống sẽ lặp lại quá trình trước đó để duyệt qua những URL mới. Quá trình Crawling sẽ qua rất nhiều địa chỉ Website và thu thập rất nhiều nội dung khác nhau từ địa chỉ thu thập đươc.



Hình 2.2 – Luồng xử lý quá trình Crawling (trích tài liệu [6])

Trong hệ thống của chúng tôi, Web Crawler được sử dụng để thu thập các đường dẫn chứa các bài báo phù hợp với nội dung đang thu thập từ các thư viện số. Từ những địa chỉ thu thập được hệ thống sẽ rút trích thông tin chỉ mục của bài báo bằng cách sử dụng các trình phân tích kết hợp với luật đã được định nghĩa trước.

2.2.3 Metadata.

- **Khái niệm Metadata**

Theo [19], Metadata (siêu dữ liệu) dùng để mô tả tài nguyên thông tin. Thuật ngữ "meta" xuất xứ là một từ Hy Lạp đùng để chỉ một cái gì đó có bản chất cơ bản hơn hoặc cao hơn. Một định nghĩa chung nhất và được dùng phổ biến trong cộng đồng những người làm công nghệ thông tin: "Metadata là dữ liệu về dữ liệu khác" (Metadata is data about other data) hay có thể nói ngắn gọn là dữ liệu về dữ liệu.

Trong các phạm vi cụ thể, những chuyên gia đưa ra các quan điểm khác nhau về Metadata:

- Theo Chris.Taylor giám đốc dịch vụ truy cập thông tin thư viện thuộc trường đại học Queensland[10] thì Metadata là dữ liệu có cấu trúc được dùng để mô tả những đặc điểm của tài nguyên. Một mẫu tin Metadata bao gồm một số lượng những phần tử được định nghĩa trước gọi là elements dùng mô tả đặc tính, thông tin tài nguyên. Mỗi elements có thể có 1 hay nhiều giá trị.
- Theo tiến sĩ Warwick Cathro thuộc thư viện quốc gia Australia[11] thì một phần tử Metadata hay còn gọi là Metadata elements mô tả tài nguyên thông tin, hay hỗ trợ truy cập đến một tài nguyên thông tin.

Tóm lại, ta có thể hiểu Metadata là thông tin dùng để mô tả tài nguyên thông tin.

- **Chuẩn Dublin Core Metadata**

Dublin Core Metadata[12] là một chuẩn Metadata được nhiều người biết đến và được dùng rộng rãi trong cộng đồng các nhà nghiên cứu, chuyên gia về thư viện số.

[10] http://www.library.uq.edu.au/iad/ctmeta4.html
[11] http://www.nla.gov.au/nla/staffpaper/cathro3.html
[12] http://dublincore.org/

Dublin Core Metadata lần đầu tiên được đề xuất năm 1995 bởi Dublin Core Metadata Element Initiative. Dublin là tên một địa danh Dublin, Ohio ở Mỹ nơi đã tổ chức hội thảo OCLC/NCSA Metadata Workshop năm 1995. Core có nghĩa là một danh sách các thành phần cốt lõi dùng mô tả tài nguyên (Element metadata), những thành phần này có thể mở rộng thêm.

Theo [20], tháng 9/2001 bộ yếu tố siêu dữ liệu Dublin Core Metadata được ban hành thành tiêu chuẩn Mỹ, gọi là tiêu chuẩn "The Dublin Core Metadata Element Set" ANSI/NISO Z39.85-2001.

Dublin Core Metadata bao gồm 15 yếu tố cơ bản (theo tài liệu [20]), được mô tả chi tiết trong bảng 2.1.

STT	Yếu tố	Mô tả
1	Title	Nhan đề hay tiêu đề của tài liệu
2	Creator	Tác giả của tài liệu, bao gồm cả tác giả cá nhân và tác giả tập thể
3	Subject	Chủ đề tài liệu đề cập dùng để phân loại tài liệu. Có thể thể hiện bằng từ, cụm từ/(Khung chủ đề), hoặc chỉ số phân loại/ (Khung phân loại).
4	Description	Tóm tắt, mô tả nội dung tài liệu. Có thể bao gồm tóm tắt, chú thích, mục lục, đoạn văn bản để làm rõ nội dung
5	Publisher	Nhà xuất bản, nơi ban hành tài liệu có thể là tên cá nhân, tên cơ quan, tổ chức, dịch vụ...
6	Contributor	Tên những người cùng tham gia cộng tác đóng góp vào nội dung tài liệu, có thể là cá nhân, tổ chức..
7	Date	Ngày, tháng ban hành tài liệu.
8	Type	Mô tả bản chất của tài liệu. Dùng các thuật ngữ mô tả phạm trù kiểu: trang chủ, bài báo, báo cáo, từ điển...
9	Format	Mô tả sự trình bày vật lý của tài liệu, có thể bao gồm; vật mang tin, kích cỡ độ dài, kiểu dữ liệu (.doc, .html, .jpg, xls, phần mềm....)

10	Identifier	Các thông tin về định danh tài liệu, các nguồn tham chiếu đến, hoặc chuỗi ký tự để định vị tài nguyên: URL (Uniform Resource Locators) (bắt đầu bằng http://), URN (Uniform Resource Name), ISBN (International Standard Book Number), ISSN (International Standard Serial Number), SICI (Serial Item & Contribution Identifier), ...
11	Source	Các thông tin về xuất xứ của tài liệu, tham chiếu đến nguồn mà tài liệu hiện mô tả được trích ra/tạo ra, nguồn cũng có thể là: đường dẫn (URL), URN, ISBN, ISSN...
12	Language	Các thông tin về ngôn ngữ, mô tả ngôn ngữ chính của tài liệu
13	Relation	Mô tả các thông tin liên quan đến tài liệu khác. Có thể dùng đường dẫn (URL), URN, ISBN, ISSN...
14	Coverage	Các thông tin liên quan đến phạm vi, quy mô hoặc mức độ bao quát của tài liệu. Phạm vi đó có thể là địa điểm, không gian hoặc thời gian, tọa độ...
15	Rights	Các thông tin liên quan đến bản quyền của tài liệu

Bảng 2.1 - Các yếu tố cơ bản của chuẩn Dublin Core Metadata

Trong hệ thống của chúng tôi, những thông tin Metadata sau được rút ra từ tài liệu (hay được gọi là những thông tin chỉ mục của bài báo):

- Creator (Author): thông tin tên của các tác giả tài liệu.
- Title: tựa đề tài liệu.
- Description (Abstract): tóm tắt nội dung của tài liệu.
- Publisher: nơi công bố, xuất bản tài liệu.
- Source (DOI): nơi download tài liệu hoặc địa chỉ chứa thông tin bài báo.
- Date (Year): năm công bố, xuất bản tài liệu.

2.2.4 Bibtex.

BibTeX[13] là một định dạng văn bản thô (text) cho các danh sách tài liệu tham khảo là sách, bài tạp chí khoa học, luận án, … do Oren Patashnik và Leslie Lamport

[13] http://en.wikipedia.org/wiki/BibTeX

đề xuất ra năm 1985. BibTeX cho phép tổ chức các thông tin về nguồn tài liệu (biểu ghi tài liệu) tham khảo một cách đồng bộ và ổn định (trích tài liệu [21]).

```
@INPROCEEDINGS {author:06,
 title    = {Some publication title},
 author   = {First Author and Second Author},
 crossref = {conference:06},
 pages    = {330—331},
}

@PROCEEDINGS {conference:06,
 editor    = {First Editor and Second Editor},
 title     = {Proceedings of the Xth Conference
on XYZ},
 booktitle = {Proceedings of the Xth Conference
on XYZ},
 year      = {2006},
 moth      =oct,
}
```

Hình 2.3- Ví dụ cấu trúc của file BibteX (nguồn Wikipedia)

Các tập tin BibTeX thường có đuôi .bib, cấu trúc của một file bibtex như sau:

- Từ khóa xác định loại tài liệu bao gồm: @article, @book, @thesis, …
- Nội dung của một trường trong file Bibtex được ghi trong hai dấu {…}.
- Các nội dung mô tả biểu ghi là những cặp [từ khóa mô tả = "nội dung mô tả"], được tách nhau bởi dấu ",".

Vì file Bibtex chứa thông tin của tài liệu (như bài báo, luận văn, …) do đó đối với mỗi tài liệu thì BibTex có kiểu lưu cấu trúc khác nhau nhận biết file BibTex này đang chứa nội dung của tài liệu nào.

Sau đây là các dạng file Bibtex của các loại tài liệu khác nhau (bảng 2.2), trong đó bao gồm các trường thông tin (field) yêu cầu mà file Bibtex đó bắt buộc phải lưu trữ, ngoài ra có thể có thêm những trường bổ sung:

Kiểu tài liệu (Entry Types)	Giải thích	Các trường yêu cầu có (Required fields)	Các trường có thể thêm (Optional fields)
article	Một bài báo từ một tạp chí.	author, title, journal, year	volume, number, pages, month, note, key
book	Cuốn sách từ một nhà xuất bản.	author/editor, title, publisher, year	volume, series, address, edition, month, note, key

booklet	Một ấn phẩm đã được in ấn nhưng không có nhà xuất bản hay cơ quan tài trợ.	title	author, owpublished, address, month, year, note, key
inbook	Một phần của cuốn sách nhưng không có tựa đề, có thể là một chương.	author/editor, title, chapter/pages, publisher, year	volume, series, address, edition, month, note, key
incollection	Một phần của cuốn sách có tiêu đề riêng của mình.	author, title, booktitle, year	editor, pages, organization, publisher, address, month, note, key
inproceedings	Bài báo trong kỷ yếu của hội nghị.	author, title, booktitle, year	editor, series, pages, organization, publisher, address, month, note, key
conference	Giống như inproceedings, bao gồm thông tin Scribe[14]	author, title, booktitle, year	editor, pages, organization, publisher, address, month, note, key
manual	Tài liệu kỹ thuật.	title	author, organization, address, edition, month, year, note, key
mastersthesis	Luận văn thạc sĩ	author, title, school, year	address, month, note, key
misc	Sử dụng khi tài liệu không xác định được loại.	none	author, title, howpublished, month, year, note, key
phdthesis	Luận văn tiến sĩ	author, title, school, year	address, month, note, key
proceedings	Kỷ yếu của hội nghị	title, year	editor, ublisher, organization, address, month, note, key

[14] http://en.wikipedia.org/wiki/Scribe

techreport	Một báo cáo được xuất bản bởi một trường học, hay cơ quan khác, thông thường được xuất bản theo số.	author, title, institution, year	type, number, ddress, month, note, key
unpublished	Một tài liệu chứa tựa đề và tên tác giả, nhưng chưa xuất bản.	author, title, note	month, year, key

Bảng 2.2 - Những kiểu file Bibtex (được tham khảo từ Wikipedia)

Trên các thư viện số ACM và IEEEXplore và CiteSeer, thông tin bài báo khoa học được xuất ra các file Bibtex, hệ thống sẽ phân tích nội dung trong đường dẫn trả về sau khi Crawl trên thư viện số để lấy file Bibtex, sau đó dùng trình phân tích file Bibtex để rút trích thông tin Metadata của bài báo. Trong phần 3.3 chương 3, chúng tôi sẽ trình bày chi tiết về cách thức sử dụng trình phân tích file Bibtex để lấy thông tin chỉ mục các bài báo.

2.3 Các nghiên cứu và ứng dụng liên quan.

2.3.1 Các nghiên cứu liên quan.

Xây dựng dữ liệu chỉ mục các bài báo khoa học hay việc rút trích thông tin Metadata của bài báo khoa học là một phần nghiên cứu trong lĩnh vực trích xuất thông tin (Information Extraction). Theo khảo sát được giới thiệu trong các bài báo [4][10] cũng như tìm hiểu của nhóm, hiện nay trong lĩnh vực trích xuất thông tin từ bài báo khoa học để xây dựng dữ liệu chỉ mục thì có một số nguồn dữ liệu thu thập và phương pháp tiếp cận mà từ đó có thể xây dựng dữ liệu như sau:

- **Nguồn dữ liệu thu thập.**

- Xây dựng dữ liệu chỉ mục các bài báo từ các file đề mục (tables of contents – TOCs) của các kỷ yếu hội thảo, tạp chí như hệ thống DBLP đã làm [14]. File TOCs chứa danh sách các bài báo được trình bày trong các hội nghị, cũng như danh

sách các bài viết được đăng trong các lần xuất bản của các tạp chí. Các hệ thống sử dụng các trình phân tích để thu thập thông tin chỉ mục các bài báo có trong file TOCs từ đó xây dựng lên cơ sở dữ liệu chỉ mục.

→ Như vậy: đối với các cơ sở chỉ mục có nguồn dữ liệu thu thập từ các file TOCs thì chúng ta thấy: nguồn dữ liệu này phụ thuộc vào khả năng thu thập những file TOCs từ các hội nghị, tạp chí. Hiện nay, với số lượng các cuộc hội nghị cũng như các tạp chí về khoa học máy tính ngày càng tăng, cùng với đó là vấn đề về bản quyền thì việc thu thập đầy đủ các file TOCs của tất cả các hội nghị, tạp chí là rất khó khăn. Từ đó dữ liệu thu thập được cũng khó đảm bảo được tính đầy đủ.

- Rút trích từ thông tin bài báo từ tài liệu dưới dạng file điện tử (sử dụng các file postscript hoặc file PDF), như các hệ thống được giới thiệu trong các bài báo [3][15]. Bằng việc phân tích nội dung các bài báo dưới dạng file điện tử thông qua việc sử dụng các luật, các thuật toán, kết hợp sử dụng máy học, các hệ thống sẽ thu được các thông tin chỉ mục từ nội dung của các bài báo.

→ Như vậy: với nguồn dữ liệu từ các bài báo dưới dạng file điện tử thì các hệ thống này đã tận dụng được nguồn dữ liệu có sẵn trong nội dung các bài báo. Nhưng việc sử dụng các luật, các thuật toán cũng như máy học trong việc trích xuất thông tin chỉ mục chưa đạt được độ chính xác cao và vẫn là một lĩnh vực đang nghiên cứu trong data mining, cùng với đó là những khó khăn trong việc thu thập tài liệu điện tử dưới dạng file điện tử hiện nay bị giới hạn trong việc download, do đó tính đúng đắn, đầy đủ của dữ liệu thu thập chưa được đảm bảo.

- Xây dựng dữ liệu chỉ mục bằng cách rút trích thông tin bài báo khoa học được công bố trên Internet. Những thông tin chỉ mục của bài báo có thể tồn tại trên các trang Website chia sẻ tài liệu, trên trang Website cá nhân của tác giả, hay thông tin chỉ mục có sẵn trên các thư viện số. Các hệ thống sử dụng các Search Engine hoặc Web Crawler tìm kiếm các bài báo trên Website sau đó sử dụng các luật, các thuật toán để rút ra thông tin bài báo như các hệ thống được giới thiệu trong các bài báo [5][17][20].

→ Với nguồn dữ liệu từ các bài báo được công bố trên Internet, thì các hệ thống đã tận dụng được nguồn dữ liệu khổng lồ. Nhưng các ứng dụng đã được xây dựng chưa tận dụng được những dữ liệu chỉ mục có sẵn.

- **Phương pháp tiếp cận rút trích thông tin chỉ mục.**

Theo [19], thì rút trích thông tin chỉ mục bài báo (hay rút trích thông tin Metadata) là lĩnh vực nghiên cứu thu hẹp thuộc lĩnh vực rút trích thông tin. Hầu hết các phương pháp rút trích Metadata hiện nay có thể chia làm 2 cách tiếp cận chính đó là: các phương pháp dựa trên học máy và phương pháp dựa trên luật kết hợp với sử dụng các từ điển, Ontologies.

Phương pháp rút trích thông tin dựa trên học máy (Machine Learning).

Bằng cách học từ tập huấn luyện (quan sát các đặc trưng của tập dữ liệu đã được xác định bởi chuyên gia), hệ thống sẽ phân tích nội dung dữ liệu mà người dùng đưa vào (thường là dạng text), để rút ra thông tin Metadata của tài liệu.

Theo [8], những phương pháp học máy để rút trích Metadata điển hình có thể kể đến như: lập trình logic, mô hình Markov ẩn (Hidden Markov Models), Support Vector Machince, và các phương pháp học thống kê khác. Trong [8], nhóm tác giả đã dùng SVM để rút trích metadata từ các bài báo khoa học. Quá trình rút trích của họ gồm hai bước: bước thứ nhất họ dùng SVM để phân lớp các dòng (lines) thuộc phần heading của các tài liệu (từ phần giới thiệu trở lên); bước thứ hai họ rút trích Metadata từ các dòng đã phân lớp trong bước thứ nhất dùng các luật dấu câu, ký tự viết hoa kết hợp với các từ điển.

Phương pháp rút trích thông tin dựa vào luật.

Các luật được các chuyên gia có kinh nghiệm đặt ra trước (ví dụ dựa vào từ khóa, font chữ để xác định vùng đặc biệt chứa dữ liệu). Dựa vào các luật, hệ thống sẽ rút ra thông tin Metadata ở vùng tương ứng.

Trong tài liệu [12], nhóm tác giả đã đề xuất một phương pháp rút trích cấu trúc logic (tiêu đề, các tác giả, các đề mục, các định nghĩa, định lý, …) từ các bài báo trong lĩnh vực toán học. Từ đó họ xây dựng đã xây dựng một trình duyệt giúp người dùng có thể dễ dàng đọc các bài báo toán học. Thuật toán học đề xuất gồm 2

bước: thứ nhất xác định những vùng đặc biệt trong tài liệu (số trang, đề mục, phần footnote cuối trang, tiêu đề của các bảng biểu và hình ảnh) dùng các từ khóa, kiểu dáng font chữ, khoảng cách không gian trình bày trong tài liệu; sau đó thông tin chi tiết sẽ được xác định từ các vùng này dựa vào kiểu dáng, vị trí và trình bày của từng vùng.

→ *Như vậy:* Mỗi cách tiếp cận đều có những ưu, nhược điểm riêng. Đối với các phương pháp máy học thì chúng ta cần phải tốn nhiều thời gian cho việc chọn mẫu, gán nhãn và để có kết quả tốt cần rất nhiều dữ liệu học. Bên cạnh đó các phương pháp dựa trên luật hay mẫu thì đơn giản và dễ dàng thực hiện hơn, nhưng để có kết quả tốt cũng tốn rất nhiều công sức cho việc khảo sát, định nghĩa luật của chuyên gia. Các luật cũng cần phải thay đổi khi xuất hiện các loại dữ liệu mới mà những luật hiện có không thể giải quyết được. Thông thường đối với từng bài toán cụ thể người ta sẽ đưa ra một cách tiếp cận và phương pháp giải quyết vấn đề tương ứng phù hợp với bài toán đặt ra.

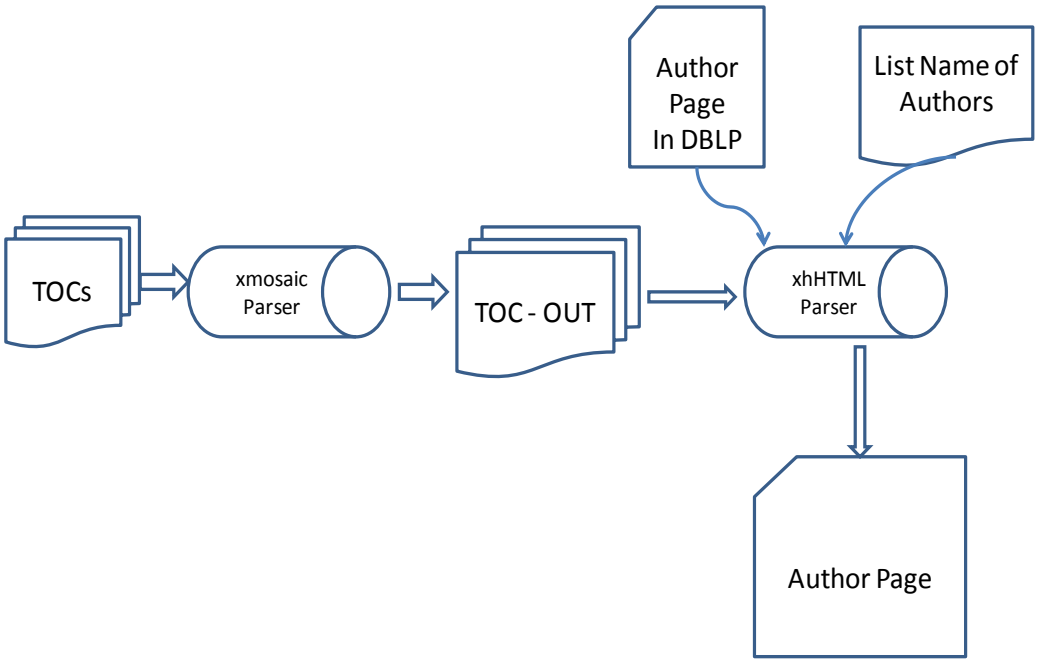
2.3.2 Các ứng dụng liên quan

Trong phần này, chúng tôi sẽ trình bày một số hệ thống dữ liệu chỉ mục đã tồn tại, các hệ thống sử dụng một trong các nguồn dữ liệu và phương pháp rút trích được giới thiệu ở phần 2.3.1. Cùng với đó chúng tôi sẽ trình bày khảo sát về các thư viện số mà hệ thống xây dựng trong khóa luận sẽ dựa trên đó để lấy thông tin chỉ mục các bài báo.

2.3.2.1 Digital Bibliography & Library Project (DBLP).

DBLP là một cơ sở dữ liệu 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 của Đức. Theo công bố trên trang Website chính của DBLP, thì tính đến tháng 1/2011 DBLP chứa thông tin chỉ mục của 1,5 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. Dữ liệu của DBLP được xuất ra các dạng CDF, XML và SQL, người phát triển có thể download các file này từ trên Website của hệ thống.

DBLP được xây dựng lên dựa vào việc phân tích và rút trích thông tin từ các file đề mục – mục lục (TOCs). Các file TOCs được tác giả sưu tầm từ các hội nghị, tạp chí.



Hình 2.4- Hệ thống xây dựng cơ sở dữ liệu DBLP

Các file TOCs được nhập vào bởi tác giả, hệ thống sẽ sử dụng các đoạn script và các parser để phân tích và rút trích thông tin các bài báo. Đồng thời với việc sử dụng các dữ liệu có sẵn của hệ thống như danh sách tên của tác giả, thông tin các bài báo đã có, hệ thống sẽ xây dựng lên trang thông tin của tác giả (Author Page). Author Page chứa thông tin về tác giả cũng như thông tin về các bài báo mà tác giả viết hoặc đồng tác giả, hình 2.4 là kiến trúc hệ thống của DBLP.

Hiện nay, có một số ứng dụng được xây dựng trên nguồn dữ liệu của DBLP, các ứng dụng này cung cấp chức năng cho phép người dùng tìm kiếm bài báo, như các hệ thống: Complete Search DBLP, Faceted search và DBL – Browser.

- **CompleteSearch DBLP**[15].

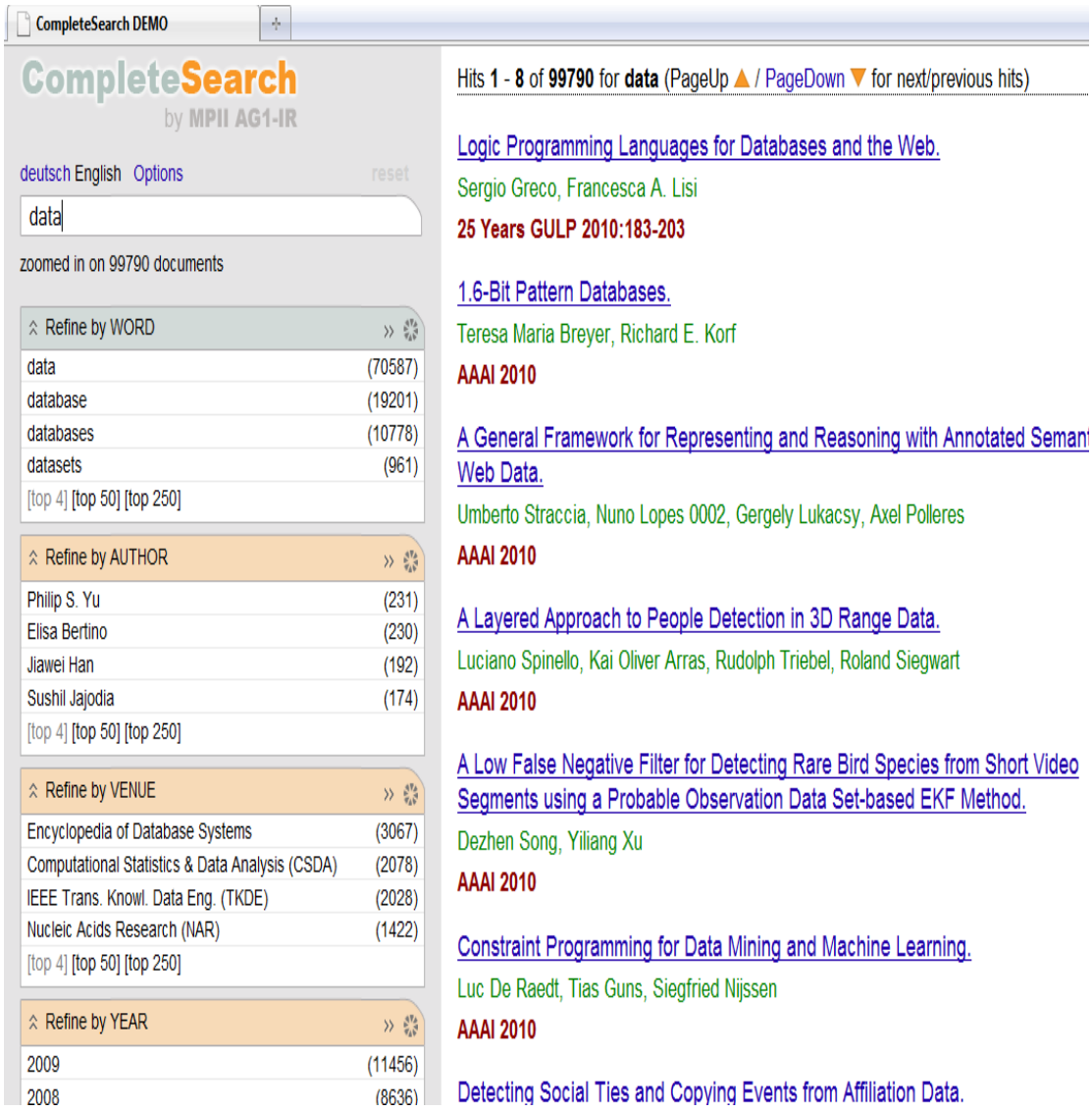
Đây là hệ thống cho phép người dùng tìm kiếm thông tin bài báo trên dữ liệu của DBLP, cách thực thi hệ thống được giới thiệu trong [7]. Bài báo trong hệ thống có thể được tìm kiếm theo các trường thông tin sau:

+Tìm kiếm theo từ khóa xuất hiện trong bài báo.

+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 2.5 - Hệ thống Complete Search

[15] http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php

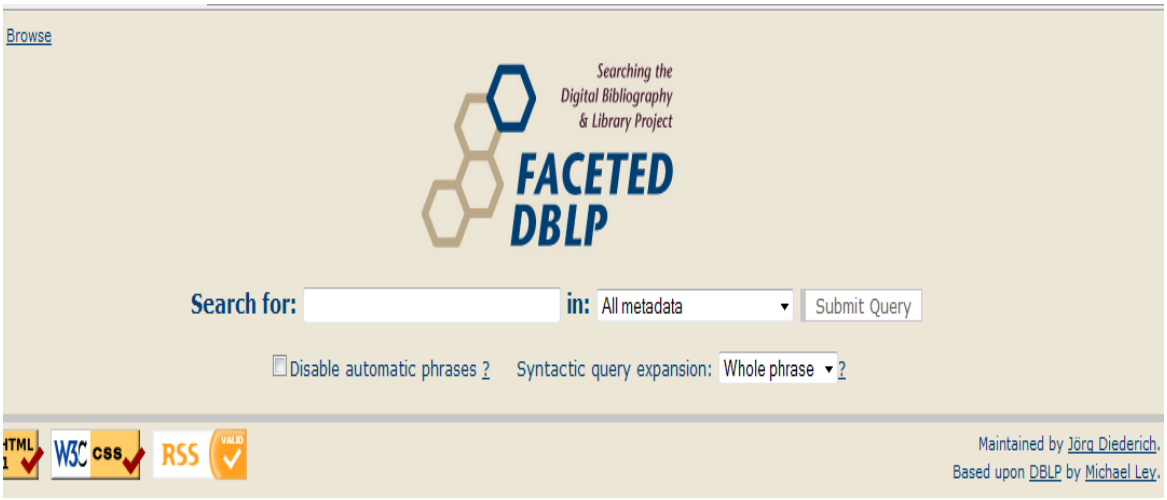
- **Faceted Search**[16].

Đây là hệ thống tìm kiếm bài báo trên dữ liệu của DBLP được giới thiệu trong [9], hệ thống cho phép người dùng tìm kiếm thông tin bài báo dựa trên các trường 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 2.6 - Hệ thống FacetedDBLP

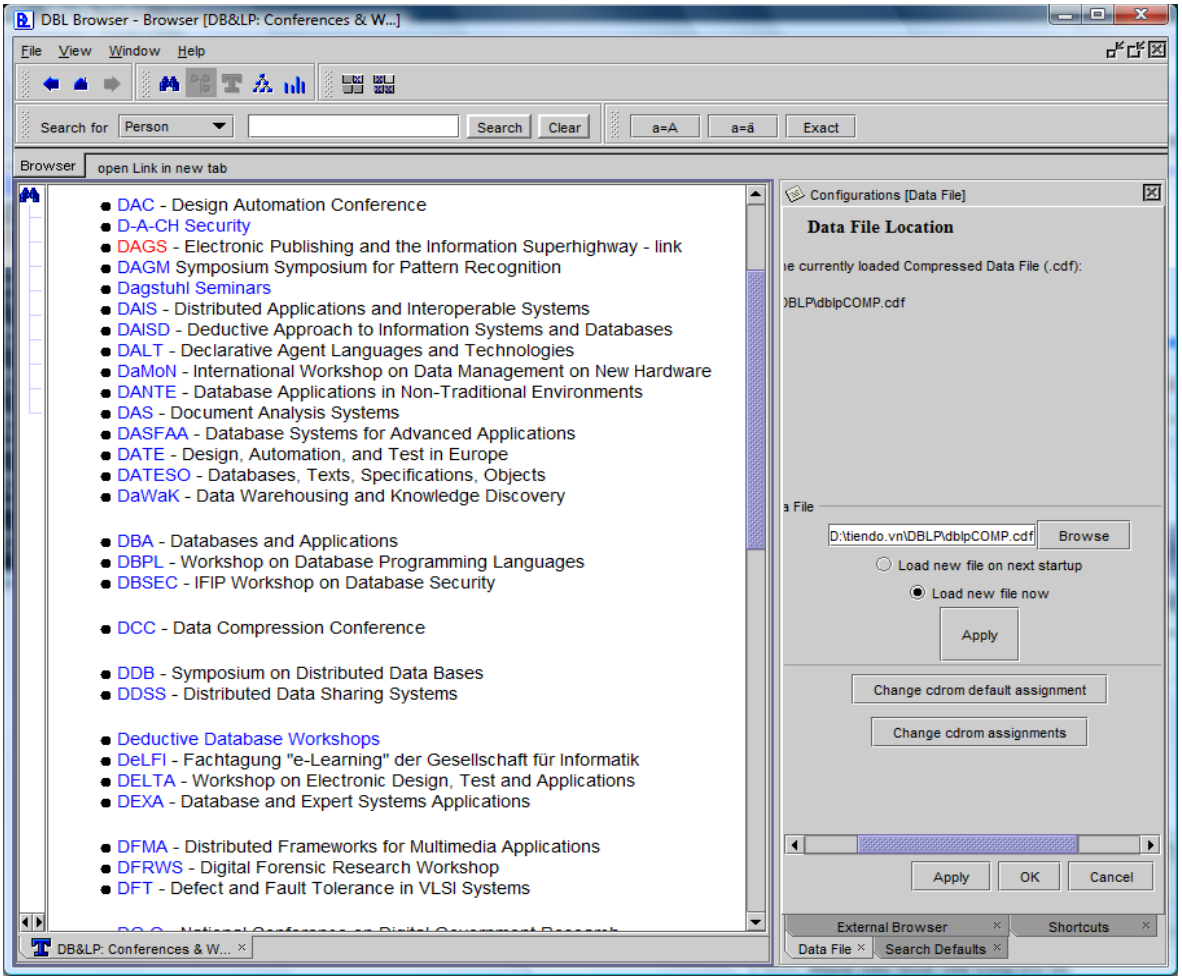
Ngoài ra hệ thống FaceTedDBLP còn cho phép người dùng duyệt tài liệu, bài báo trong DBLP theo danh sách dựa trên tên tác giả, tên hội nghị, tên tạp chí hay từ khóa mà người dùng tìm kiếm nhiều nhất trong hệ thống.

Hình 2.7 - Duyệt bài báo trong FacetedDBLP

[16] http://dblp.l3s.de/?q=&newQuery=yes&resTableName=query_resultmQ9GIx

➢ DBL – Browser

DBL – Browser, là chương trình sử dụng để tìm kiếm bài báo trên dữ liệu DBLP mà không cần kết nối Internet (dữ liệu của DBLP được tải về máy cục bộ). Chương trình cho phép hiển thị thông tin của bài báo một cách trực quan.



Hình 2.8 - Chương trình DBL Brower

➔ Như vậy: dữ liệu chỉ mục DBLP được thu thập bằng cách rút trích thông tin chỉ mục từ những file TOCs của kỷ yếu hội nghị, tạp chí được các tổ chức gửi về hoặc tác giả DBLP sưu tầm được. Vấn đề đặt ra ở đây là việc lấy được các file TOCs từ các hội nghị sẽ khó đảm bảo thông tin thu thập được sẽ đầy đủ và cập nhật nhất đối với các bài báo, để chứng minh cho điều này chúng tôi tiến hành khảo sát bằng cách tìm kiếm các bài báo trên các thư viện số với từ khóa là chủ đề trong lĩnh vực khoa học máy tính, sau đó kiểm tra tính tồn tại của thông tin bài báo trong DBLP.

Trong bảng 2.3 là kết quả được tính trung bình của 100 bài báo đầu tiên trên ba thư viện số ACM, Citeseer, IEEEXplore sau khi tìm kiếm với 2 từ khóa "Database" và "Data mining".

Từ khóa tìm kiếm	Phần trăm dữ liệu không tồn tại trong DBLP (%)	Phần trăm Dữ liệu trước năm 2010 không tồn tại trong DBLP (%)	Phần trăm dữ liệu trong năm 2010 không tồn tại trong DBLP (%)
Database	28,33	86,26	10,71
Data mining	43,67	77,45	14,51

Bảng 2.3 - Khảo sát tính cập nhật dữ liệu của DBLP

Trong đó kết quả được tính theo công thức sau:

+ Phần trăm dữ liệu không tồn tại trong DBLP: được tính bằng số các bài báo trong 100 bài báo tồn tại trên thư viện số (ACM, IEEEXplore, CiteSeer) nhưng không có trong DBLP.

+ Phần trăm dữ liệu trước năm 2010 không tồn tại trong DBLP: được tính bằng số bài báo trong 100 bài báo thu thập trên thư viện số có năm xuất bản trước năm 2010 tồn tại trong thư viện số (ACM, IEEEXplore, CiteSeer) nhưng không có trong DBLP.

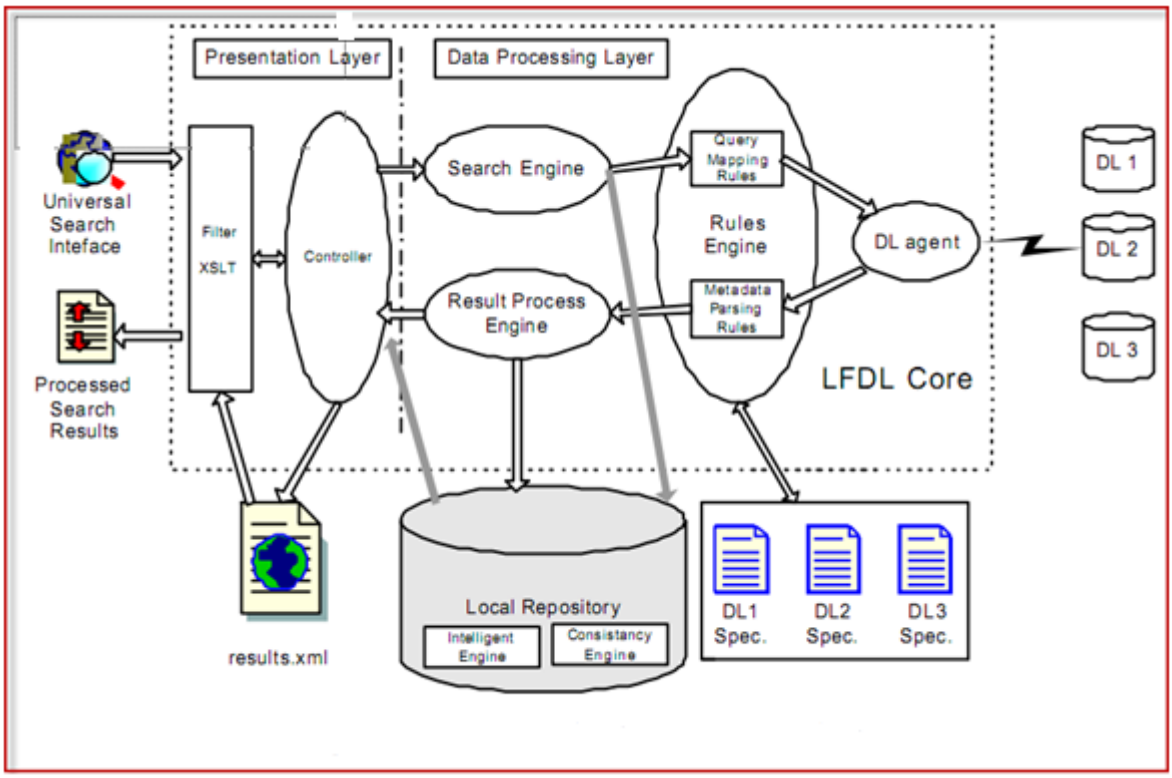
+ Phần trăm dữ liệu trong năm 2010 không tồn tại trong DBLP: được tính bằng số bài báo trong 100 bài báo thu thập trên thư viện số có năm xuất bản trong năm 2010 không có trong DBLP.

Dựa vào bảng 2.3 ta thấy dữ liệu của DBLP chưa đảm bảo được tính đầy đủ và cập nhật dữ liệu mới một cách nhanh chóng.

Mặt khác như khảo sát ở trên, phần lớn các hệ thống phát triển trên nguồn dữ liệu của DBLP là sử dụng dữ liệu chỉ mục của DBLP chứ chưa có hệ thống nào đưa ra phương pháp bổ sung dữ liệu còn thiếu cho DBLP ngoài cách cập nhật dữ liệu của tác giả DBLP.

2.3.2.2 Lightweight Federated Digital Library (LFDL)

Đây là một hệ thống tìm kiếm và thu thập dữ liệu chỉ mục các bài báo khoa học từ thư viện số. Hệ thống được giới thiệu trong bài báo [16], tác giả bài báo chỉ ra rằng hiện nay thông tin các bài báo trong các thư viện số đều không tuân theo một chuẩn lưu trữ thông tin Metadata nhất định, mà mỗi thư viện số có cách lưu khác nhau, từ đó tác giả đề xuất cách xây dựng một hệ thống có khả năng tổng hợp thông tin từ các thư viện số thành một khối dữ liệu có cấu trúc đồng nhất. Trong hình 2.9 giới thiệu kiến trúc của hệ thống LFDL (Lightweight Federated Digital Library), ứng với mỗi thư viện số hệ thống sẽ có các đặc tả và đưa ra các luật để rút thông tin Metadata khác nhau (các luật này có cấu trúc XML), những thông tin rút được từ các thư viện số sẽ được lưu xuống Database.



Hình 2.9 – Kiến trúc LFDL

Các luồng dữ liệu và tương tác giữa các thành phần của hệ thống LFDL như sau:

- Đầu tiên khi khởi tạo, hệ thống sẽ đọc tất cả các đặc tả (như ở hình 2.10, 2.11) của thư viện số (DL 1, DL 2, DL 3) bao gồm: luật liên kết truy vấn và

luật rút trích Metadata (phần rules Engine ở trên hình 2.9). Các đặc tả ở đây được hệ thống rút ra trên thư viện số thông qua quá trình phân tích cấu trúc hiển thị thông tin bài báo cho người dùng (phân tích nội dung HTML hiển thị kết quả tìm kiếm cho người dùng).

- Khi người dùng tương tác với hệ thống và yêu cầu tìm kiếm thông tin bài báo (yêu cầu được gửi thông qua Seach Interface), hệ thống sử dụng bộ xử lý trung tâm để tối ưu câu tìm kiếm sau đó chuyển sang công cụ tìm kiếm.
- Công cụ tìm kiếm sẽ dựa vào các đặc tả đầu vào (phần phân tích ở trên) đồng thời sử dụng các luật liên kết tạo câu truy vấn. Sau đó gửi câu truy vấn lên thư viện số để lấy kết quả phù hợp về.
- Sau khi nhận kết quả trả về từ thư viện số hệ thống sẽ xử lý để rút ra thông tin Metadata của bài báo dựa vào các luật trong phần đặc tả tương ứng mỗi thư viện số (phần data Processing Layer trong kiến trúc hình 2.9). Thông tin Metadata đã rút ra sẽ được lưu vào cơ sở dữ liệu (Local Repositiry) ở máy cục bộ.
- Để hiển thị thông tin bài báo thu thập được cho người dùng, hệ thống còn lưu thông tin bài báo dưới dạng XML (file results.xml). Khi cần hiển thị, hệ thống sẽ sử dụng bộ xử lý XSLT[17] để chuyển nội dung XML sang HTML hoặc XHTML.

➔ Đây là một hệ thống thu thập dữ liệu chỉ mục các bài báo từ các thư viện số bằng cách phân tích nội dung trong thư viện số kết hợp với việc sử dụng các luật để rút trích thông tin. Theo thông tin của tác giả được công bố trong [16], hiện hệ thống mới thu thập được tựa đề (title) và đường dẫn (hyperlink) của bài báo trong các thư viện số ACM, NEEDS, NACA, COGPRINTS, CSTC, LTRS, và WCR. Hệ thống LFDL mới chỉ được giới thiệu trong nội dung bài báo, chưa có ứng dụng chạy trực tuyến (online) cùng với đó hệ thống chưa tận dụng được những cơ sở dữ liệu chỉ mục có sẵn, cũng như thông tin chỉ mục của các bài báo có sẵn trên thư viện số, để dữ liệu chỉ mục thu thập được đảm bảo tính đầy đủ và chính xác.

[17] http://en.wikipedia.org/wiki/XSLT

```
<RESULT-METADATA Title="Result page metadata parsing:" hasRecordLevel="false">
    <MATCH-START enforced="true" Title="the beginning of matching string of result metadata"><div class="authors"></MATCH-START>
    <MATCH-END enforced="true" Title="the end of matching string of result metadata"></div></MATCH-END>
    <EXCLUDE Title="the string should be excluded or removed when parsing"></EXCLUDE>
    <EXCLUDE Title="the string should be excluded or removed when parsing"></EXCLUDE>
    <EXCLUDE Title="the string should be excluded or removed when parsing"></EXCLUDE>
    <METADATA-FIELD order="1" multiple="true" delimeter="," Title="information about a particular metadata field">CREATOR</METADATA-FIELD>
</RESULT-METADATA>
```

Hình 2.10 – Đặc tả cho thư viện số ACM portal (Trích tài liệu [16])

```
<RESULT-METADATA Title="Result page metadata parsing:" hasRecordLevel="true">
    <MATCH-START Title="the beginning of matching string of result metadata">null</MATCH-START>
    <MATCH-END Title="the end of matching string of result metadata">null</MATCH-END>
</RESULT-METADATA>
<RECORD-METADATA Title="Record page metadata parsing:">
    <MATCH-START Title="the beginning of matching string of result metadata">name="DC.title"</MATCH-START>
    <MATCH-END isLastIndex="true" Title="the end of matching string of result metadata">" name="DC.creator"</MATCH-END>
    <EXCLUDE Title="the string should be excluded or removed when parsing">/><meta content="</EXCLUDE>
    <REPLACE Title="replace old string with new string">
        <OLD-STRING Title="the old string to be replaced">" name="DC.creator"</OLD-STRING>
        <NEW-STRING Title="replace with the new string">;</NEW-STRING>
    </REPLACE>
    <METADATA-FIELD order="1" multiple="true" delimeter=";" Title="information about a particular metadata field">CREATOR</METADATA-FIELD>
</RECORD-METADATA>
```

Hình 2.11 – Đặc tả cho thư viện số Cogprints (Trích tài liệu [16])

2.3.2.3 Autonomous Citation Indexing (ACI).

ACI là hệ thống thu thập và đánh dấu chỉ mục các bài báo khoa học được sử dụng trong thư viện số Citeseer và được giới thiệu trong bài báo của Giles [3].Hệ thống sử dụng các Web Search Engines (như Alta vista, Hotbot, Excite) đồng thời kết hợp với sử dụng các thuật toán Heuristic để tìm kiếm những bài báo bằng những từ khóa như "publications", "paper", "postscript" … những bài báo tìm kiếm được có định dạng file PDF hoặc PostScript đươc download về. Sau đó các bài báo được chuyển sang file text. Hệ thống sử dụng chương trình (PreScript[18]) để xác định xem nội dung bài báo có phải là một tài liệu nghiên cứu không.

Khi bài báo download được là một tài liệu nghiên cứu, hệ thống sẽ thực hiện việc phân tích để nhận diện, rút trích các thành phần chính của bài báo bao gồm:

- URL: Rút trích từ đường dẫn download tài liệu.
- Header: phần tựa đề (title), thông tin tác giả (author) của bài báo.
- Phần tóm tắt của bài báo (abstract).
- Phần giới thiệu (introduction).
- Phần tham khảo (citations): danh sách các tài liệu mà bài báo tham khảo.
- Nội dung bài báo.

Phần thông tin của bài báo mà hệ thống ACI chú ý tới là phần Citations (hay references) của bài báo. Sau khi nhận được thông tin các trích dẫn thì việc tiếp theo mà hệ thống sẽ thực hiện là phân tích các trích dẫn để lấy thông tin chỉ mục các bài báo. Một vấn đề đặt ra là cùng một tài liệu nhưng khi được trích dẫn ở những bài báo khác nhau thì có thể có những định dạng khác nhau như ví dụ hình 2.12 là thông tin của cùng 1 bài báo được trích dẫn trong nhiều tài liệu khác nhau:

[18] http://www.nzdl.org/technology/prescript.html

```
[7] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J.
        Stone. Classification and Regression Trees.
        Wadsworth, Pacific Grove, California, 1984.
6. L. Breiman, J. Friedman, R. Olshen and C. Stone,
        Classification and Regression Trees, Wadsworth
        and Brooks, 1984.
[1] L. Breiman et al. Classification and Regression
        Trees. Wadsworth, 1984.
```

Hình 2.12 – Ví dụ nội dung Citations của cùng 1 tài liệu

Để giải quyết vấn đề này, hệ thống ACI đã đưa ra các bước sau để nhận diện các chỉ mục của cùng một tài liệu:

+ Chuyển định dạng chữ của Citations về dạng thường.

+ Loại bỏ các dấu nối trong phần Citations.

+ Loại bỏ các dấu thứ tự.

+ Chuyển những từ viết tắt có trong Citations thành từ đầy đủ nội dung ví dụ như "pro." thành "proceedings" …

+ Loại bỏ một số những ký tự như vol., volume, no. …

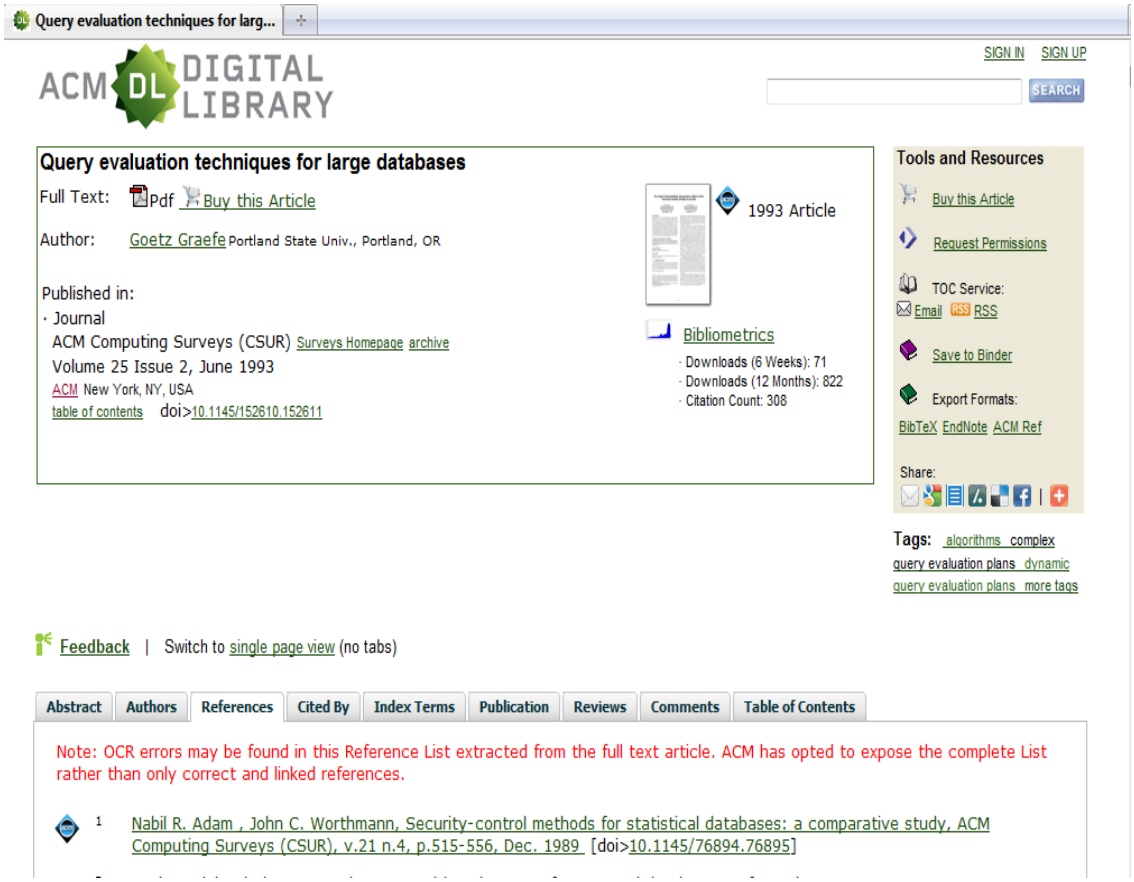
Sau khi qua các bước tiền xử lý bên trên, hệ thống sẽ sử dụng thuật toán LikeIT để so sánh thông tin trích dẫn đang xét với những dữ liệu chỉ mục có sẵn trong hệ thống để xác định xem thông tin tài liệu này đã có trong hệ thống chưa. Nếu thông tin chưa tồn tại thì dữ liệu sẽ được thêm vào database của hệ thống. Với dữ liệu thu thập được hiện nay hệ thống đã xây dựng lên thư viện số CiteSeer cho phép người dùng tìm kiếm thông tin bài báo.

➔ Như vậy với việc sử dụng các Search Engine để đi tìm kiếm và download các bài báo trên Internet thì hệ thống đã tận dụng được nguồn dữ liệu khổng lồ được chia sẻ trên mạng, nhưng hiện nay đối với các thư viện số thì việc download các bài báo bị giới hạn. Mặt khác, độ chính xác của việc rút trích thông tin chỉ mục bài báo từ các file điện tử hiện nay vẫn chưa cao và vẫn là một vấn đề lớn đang được nghiên cứu trong lĩnh vực rút trích thông tin. ACI cũng chưa tận dụng được nguồn dữ liệu chỉ mục đã được đánh dấu sẵn trên các thư viện số cũng như các cơ sở dữ liệu chỉ mục có sẵn.

2.3.2.4 Thư viện số ACM, CiteSeer, IEEEXplore.

- **Thư viện số ACM.**

ACM (Association for Computing Machinery) là một tổ chức hoạt động trong lĩnh vực đào tạo và nghiên cứu khoa học liên quan đến máy tính, ACM cung cấp một thư viện số ACM Portal cho phép người dùng tìm kiếm các bài báo được công bố trong các hội nghị, tạp chí được tổ chức và xuất bản bởi ACM cũng như một số tổ chức khác có phối hợp, liên kết với ACM.



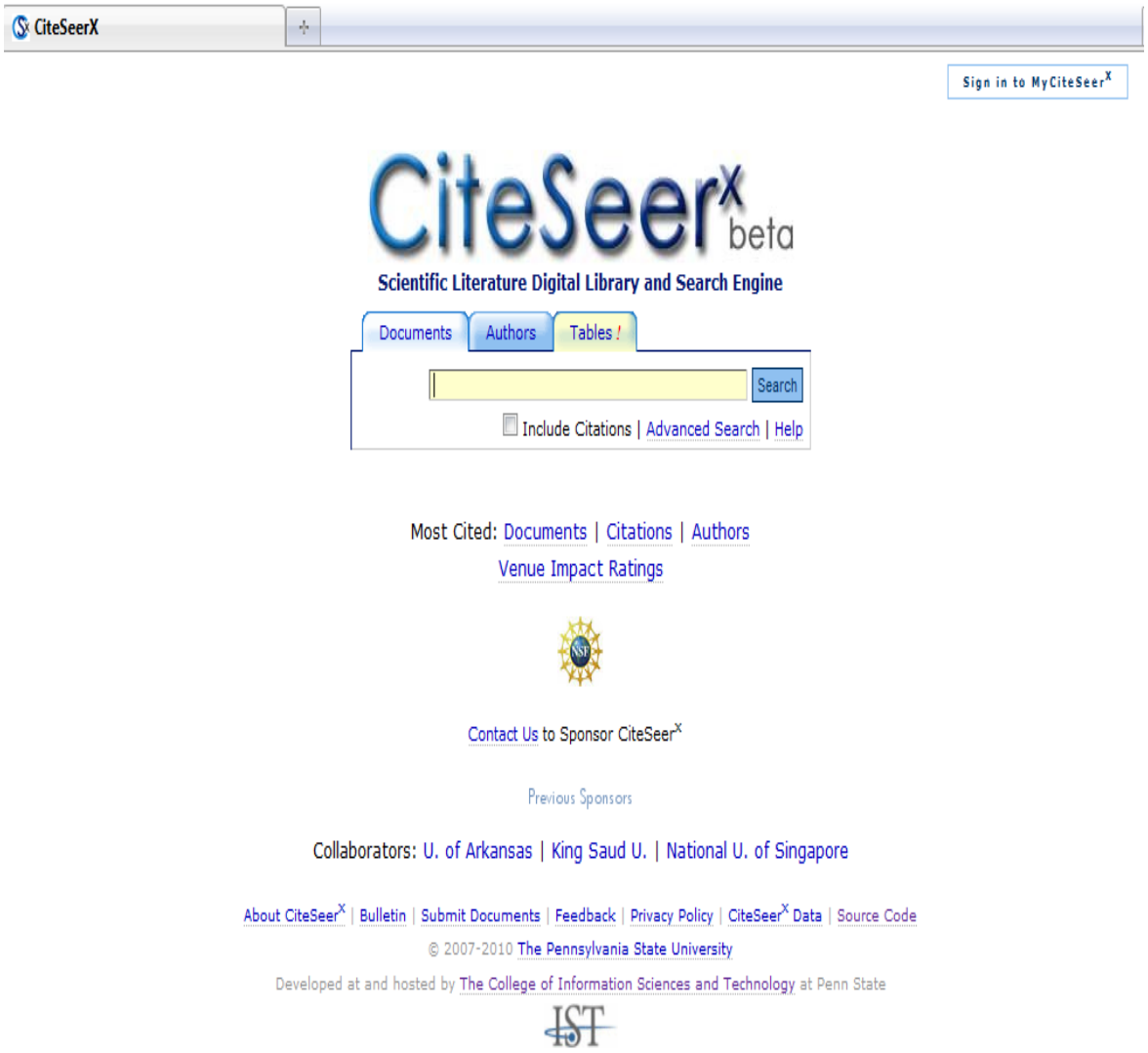
Hình 2.13 – Thư viện số ACM

Theo thông tin được công bố trên trang chủ của ACM[19], tính đến tháng 1 năm 2011 dữ liệu của ACM chứa thông tin khoảng 1,6 triệu bài báo trong nhiều lĩnh vực khác nhau của công nghệ thông tin. Hệ thống không cho phép người dùng không có tài khoản download các bài báo từ thư viện số về, thư viện chỉ cho phép người dùng thông thường tra cứu và xem thông tin chỉ mục của bài báo.

[19] http://portal.acm.org/

- **Thư viện số mở CiteSeer**.

CiteSeer là một thư viện số cho phép người dùng tìm kiếm thông tin của các bài báo thuộc lĩnh vực khoa học máy tính của nhiều tổ chức khoa học khác nhau. Hệ thống sử dụng ACI để đi đánh dấu và lưu trữ chỉ mục các bài báo trên Internet.



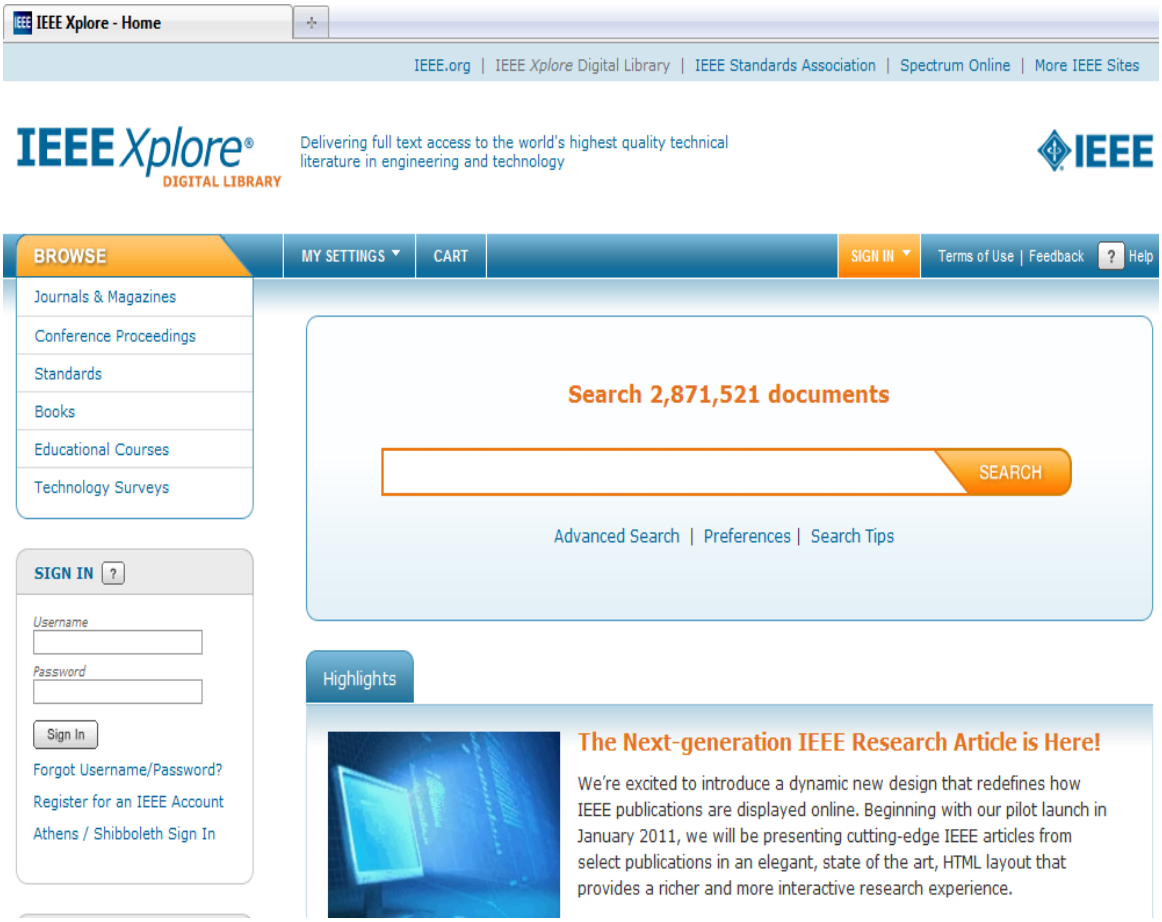
Hình 2.14 – Thư viện số CiteSeer

Theo thông tin từ trang chủ của thư viện số CiteSeer[20] thì tính đến tháng 1 năm 2011 dữ liệu của CiteSeer chứa thông tin của khoảng 1,6 triệu bài báo khoa học trong các hội nghị, tạp chí của nhiều tổ chức khác nhau được thu thập bằng hệ thống ACI. Hệ thống cho phép người dùng download bài báo về máy cá nhân.

[20] http://citeseerx.ist.psu.edu/

➢ **Thư viện số IEEEXplore.**

Thư viện số IEEEXplore 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…



Hình 2.15 – Thư viện số IEEEXplore

Theo thông tin từ trang chủ của hệ thống, tính đến tháng 1 năm 2011 dữ liệu của thư viện số này chứa thông tin của khoảng 2,8 triệu bài báo từ các hội nghị cũng như các tạp chí. Hệ thống không cho phép người dùng thông thường download và xem dữ liệu chỉ mục của bài báo có trong thư viện.

CHƯƠNG 3: XÂY DỰNG VÀ LÀM GIÀU DỮ LIỆU CHỈ MỤC VỚI WEB CRAWLER.

3.1 Mở đầu

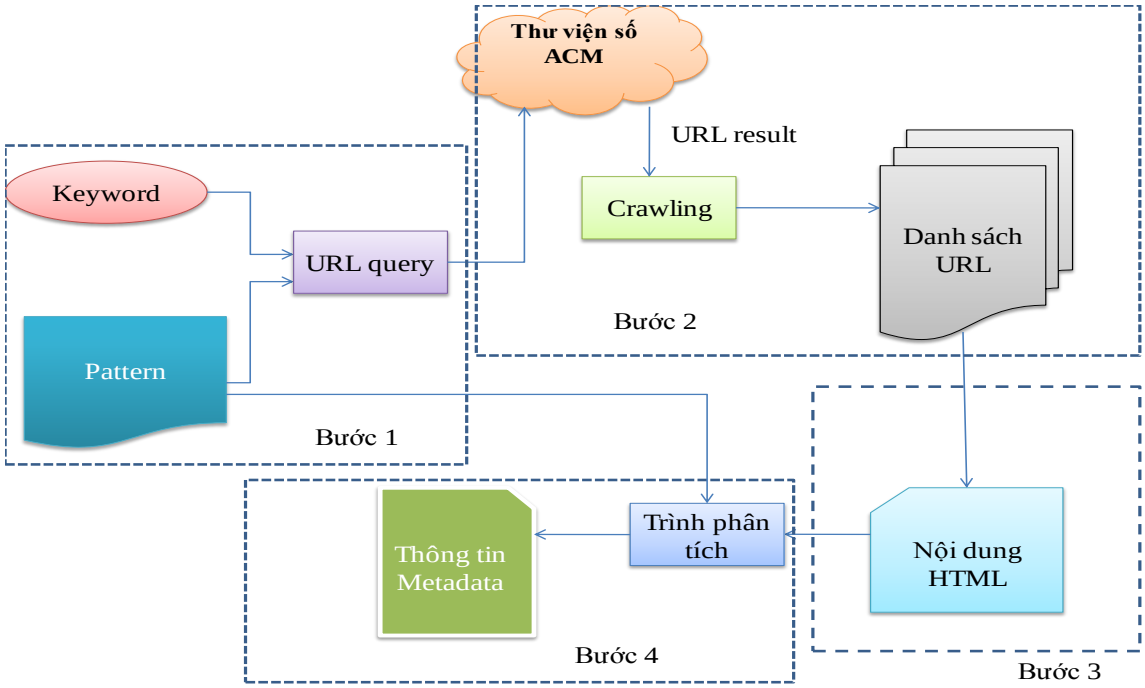
Hệ thống xây dựng và làm giàu dữ liệu chỉ mục sẽ thu thập thông tin các bài báo khoa học từ thư viện số, sau đó kết hợp những dữ liệu chỉ mục các bài báo thu thập được với thông tin các bài báo cáo trong cơ sở dữ liệu chỉ mục có sẵn trong DBLP, từ đó xây dựng lên một dữ liệu chỉ mục đầy đủ, cập nhật và chính xác.

Trong chương 3, chúng tôi sẽ trình bày cách thức mà hệ thống đi thu thập các bài báo dữ liệu từ thư viện số, cách rút trích thông tin chỉ mục của bài báo, cùng với các luồng xử lý chính của hệ thống.

3.2 Phương pháp thu thập trên thư viện số.

Đầu vào của quá trình thu thập là từ khóa được người dùng nhập vào hoặc chọn từ danh sách tên tác giả được lấy từ DBLP hay tên chủ đề được hệ thống lấy từ Wikipedia (phụ lục C).

3.2.1 Cách thức thu thập các bài báo từ thư viện số ACM



Hình 3.1 – Các bước thu thập trên ACM

Trong phần này, chúng tôi sẽ trình bày cách thức mà hệ thống thu thập bài báo khoa học từ thư viện số ACM. Bằng việc sử dụng các thẻ (pattern) đã định nghĩa sẵn kết hợp với từ khóa, hệ thống sẽ tiến hành tạo câu truy vấn (URL Query) trên thư viện số ACM và tiến hành Crawl để lấy các đường dẫn tới các bài báo được trả về từ thư viện số (các đường dẫn chứa trong nội dung của URL query).

Sau quá trình Crawl, kết quả trả về là danh sách các đường dẫn đến các bài báo. Hệ thống sẽ sử dụng các luật, các pattern để rút thông tin Metadata của bài báo.

STT	Tên thẻ (pattern)	Giải thích
1	http://portal.acm.org/	Các pattern dùng để tạo câu truy vấn gửi lên thư viện số ACM để tìm kiếm. Ví dụ: câu truy vấn được tạo với từ khóa là computer như sau: http://portal.acm.org/results.cfm?query=computer&dl=ACM&coll=Portal&short=0
2	results.cfm?query=	
3	&dl=	
4	&coll=Portal&short=0	
5	http://portal.acm.org/exportformats.cfm?id=	Dựa vào URL result kết hợp Hai pattern này để tìm ra ID của mỗi bài báo, từ đó lấy file bibtex của bài báo này thông qua tìm kiếm trên nội dung HTML pattern 6 .
6	&expformat=bibtex	
7	http://portal.acm.org/tab_abstract.cfm?id=	Pattern lấy phần tóm tắt của bài báo dựa vào ID của bài báo đã được tìm thấy ở trên.
8	ACMEndGetAbstract &usebody=tabbody	
9	.*Found(\d+,*\d*) of.*	Regular expression để tìm kiếm thông tin tổng số lượng bài báo tìm được từ thư viện số trên URL result.
10	.*Results \d+ - \d+ of (\d+,*\d*).*	Pattern Tìm Số kết quả trả về từ nội dung URL query trong một trang

11	(exportformats.cfm[.]+bibtex)	
12	<A HREF="(citation.cfm.*)" class.*	
13	(exportformats[.]cfm.+bibtex)	
14	\d+&	

Bảng 3.1 - Các pattern sử dụng để thu thập các bài báo khoa học từ thư viện số ACM.

Quá trình Crawl, thu thập bài báo từ thư viện số ACM có thể chia thành 4 bước nhỏ như sau:

Bước 1: Tạo câu truy vấn dựa vào từ khóa và các pattern

Như đã trình bày ở trên, để bắt đầu quá trình thu thập bài báo khoa học từ thư viện số ACM portal thì đầu tiên chúng ta phải tạo câu truy vấn (URL query) để gửi lên thư viện số. Câu truy vấn sẽ được tạo ra dựa vào từ khóa kết hợp một số thẻ đã được định nghĩa trong bảng 3.1.

Nếu từ khóa là một cụm từ thì các khoảng trắng giữa các từ phải chuyển thành "20%" trước khi kết hợp với các pattern để tạo URL query. Tại vì trên thư viện số ACM các khoảng trắng giữa các từ trên câu truy vấn được thay thế bằng "20%".

Ví dụ: Tạo URL khi người dùng nhập từ khóa là "computer vision"

- Chuyển "Computer vision" → "computer20%vision"
- Kết hợp các pattern lại và thêm từ khóa vào: http://portal.acm.org/ + results.cfm?query= + từ khóa + &dl= + ACM + &coll=Portal&short=0
- Với từ khóa là "computer vision" ta sẽ nhận được URL là:

http://portal.acm.org/results.cfm?query=computer20%vision&dl=ACM&coll=Portal&short=0

→ *Mục tiêu*: Tạo ra URL query để thu thập bài báo từ thư viện số ACM.

Bước 2: Gửi URL vừa tạo được lên trình duyệt Web. Sau đó lấy nội dung trang Web mà trình duyệt trả về khi truy cập vào địa chỉ URL trên (lấy nội dung HTML của trang).

Hệ thống dựa vào các pattern để phân tích và rút ra một số thông tin cần thiết từ trang Web này như: tổng số kết quả được tìm thấy, số kết quả trong một trang, Sau đó hệ thống bắt đầu Crawl để thu thập danh sách đường dẫn đến thông tin chi tiết mỗi bài trong trang có đường dẫn URL query.

→ *Mục tiêu*: Thu thập danh sách các địa chỉ URL của các bài báo khoa học trong địa chỉ URL query ở bước 1.

Bước 3: Truy cập vào các bài báo

Từ danh sách các địa chỉ URL của mỗi bài báo đã thu thập được ở bước trên, hệ thống sẽ truy cập vào đường dẫn của mỗi bài báo để lấy toàn bộ nội dung trang HTML.

→ *Mục tiêu*: Lấy về nội dung chi tiết của một trang Web từ một địa chỉ URL chứa nội dung bài báo.

Bước 4: Lấy các thông tin của một bài báo từ nội dung đã thu được từ bước 3.

Sử dụng pattern có số thứ tự 5, 6 trong bảng 3.1 để tìm ID của mỗi bài báo sau đó lấy về file Bibtex của bài báo đó theo ID của cừa rút được.

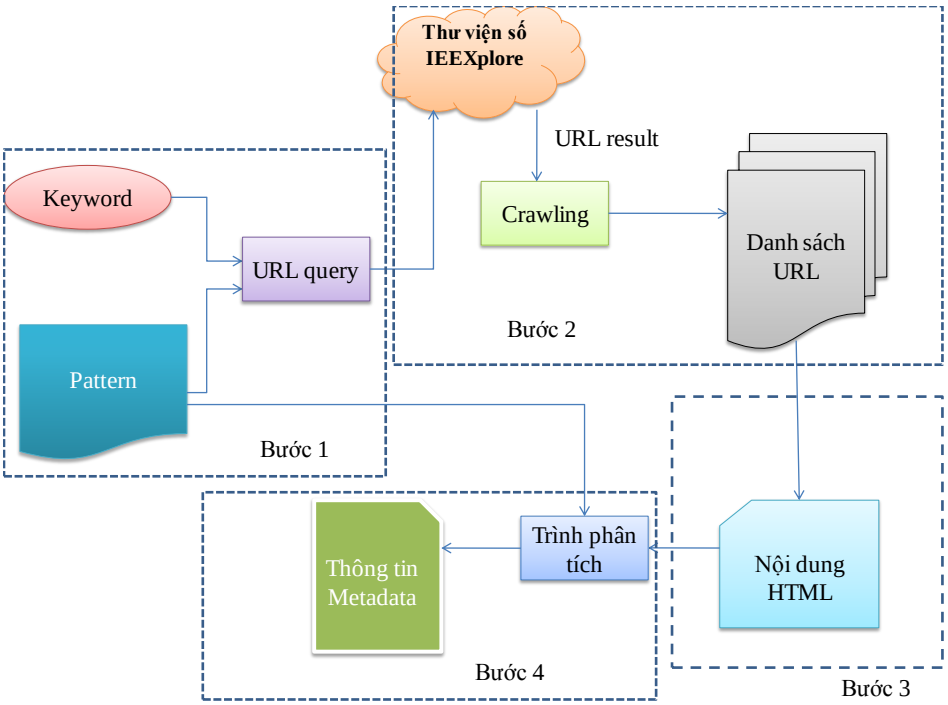
Sau khi thu thập được file Bibtex của một bài báo, hệ thống sử dụng công cụ phân tích cú pháp file Bibtex để lấy ra các thông tin cần thiết của bài báo chứa trong file Bibtex này.

Riêng phần tóm tắt của bài báo do không được lưu trong file Bibtex như những thông tin khác. Vì vậy, để lấy được phần tóm tắt của một bài báo cần phải sử dụng thêm hai pattern 7, 8 trong bảng 3.1 để tìm kiếm phần tóm tắt (abstract) trong nội dung trang Web thu được ở bước 3.

→*Mục tiêu*: Thu thập các thông tin theo yêu cầu đặt ra là: Tên bài báo, các đồng tác giả, đường dẫn URL, năm công bố, tóm tắt, tổ chức công bố. Sau đó hệ thống lưu các thông tin bài báo đã tìm được và quay lại bước 3 để tiếp tục thu thập.

➔ Như vậy với việc sử dụng kết hợp giữa các pattern, các trình phân tích và phương pháp Crawl hệ thống sẽ thu thập được thông tin Metadata của các bài báo. Một vấn đề đặt ra ở đây là khi cấu trúc trang thay đổi chương trình phải thay đổi các pattern tương ứng. Để khắc phục tình huống này, hệ thống có thêm chức năng cho phép người dùng đổi các pattern tương ứng với những thay đổi của cấu trúc trang.

3.2.2 Cách thức thu thập các bài báo từ thư viện số IEEEXplore.



Hình 3.2 – Các bước thu thập trên IEEEXplore

Tương tự như chức năng thu thập các bài báo khoa học từ thư viện số ACM như đã trình bày ở phần trên, hệ thống sẽ tạo câu truy vấn (URL query) dựa vào từ khóa và các pattern đã được định nghĩa sẵn ở bảng 3.2 bên dưới. Đầu ra của quá trình này cũng là danh sách các bài báo khoa học và các thông tin Metadata của mỗi bài báo.

STT	Tên thẻ (Pattern)	Giải thích
1	http://ieeexplore.ieee.org/search/freesearchresult.jsp?queryText=	Chuỗi này kết hợp với từ khóa để tạo câu truy vấn vào thư viện số IEEEXplore để thu thập các bài báo.
2	&rowsPerPage= &pageNumber=	Các điều kiện được thêm vào sau câu truy vấn để điều chỉnh số trang được lấy về và số kết quả trong một trang.
3	([0-9,]+) results	Tổng số kết quả tìm được

4	<span class=\"type\">\s*(.+)	Kiểu của tài liệu
5	.*(.+) "+ "\s+(.+)	Pattern tìm bài báo trong trang Web chứa danh sách các link.
6	(.*), \d*\.*\s?(.*	Pattern tìm tổ chức công bố
7	(.*?)\.?\s?Proceedings\\s?(.*)	Pattern xác định tài liệu được công bố trên các kỷ yếu hội thảo (Proceeding).
8	<a href=".*arnumber=(\d+).*">	Mã số của bài báo
9	<a\\s*href=[^<]+>\\s*(.+)\\s*	Pattern lấy tiêu đề của bài báo
10	<p>\\s+(.+)	Pattern lấy các đồng tác giả
11	Abs tract	Các pattern dùng để lấy phần tóm tắt của bài báo từ nội dung trả về từ đường dẫn tới bài báo.
12	.*[^,] '?\d+\)?	
13	<p>\\s*(.+)	
14	<div class=\"abstract RevealContent	
15	Publication Year:\\s*(\\d{4})	Pattern lấy năm công bố của bài báo
16	Page\\(s\\):\\s*(\\d+)\\s*- \\s*(\\d*)	Số trang của bài báo
17	Digital Object Identifier:\\s*(.+)	Pattern lấy số DOI của một bài báo

Bảng 3.2 - Các pattern sử dụng để thu thập các bài báo khoa học từ thư viện số IEEEXplore.

Quá trình thu thập bài báo từ thư viện số IEEEXplore có thể chia thành 4 bước nhỏ sau:

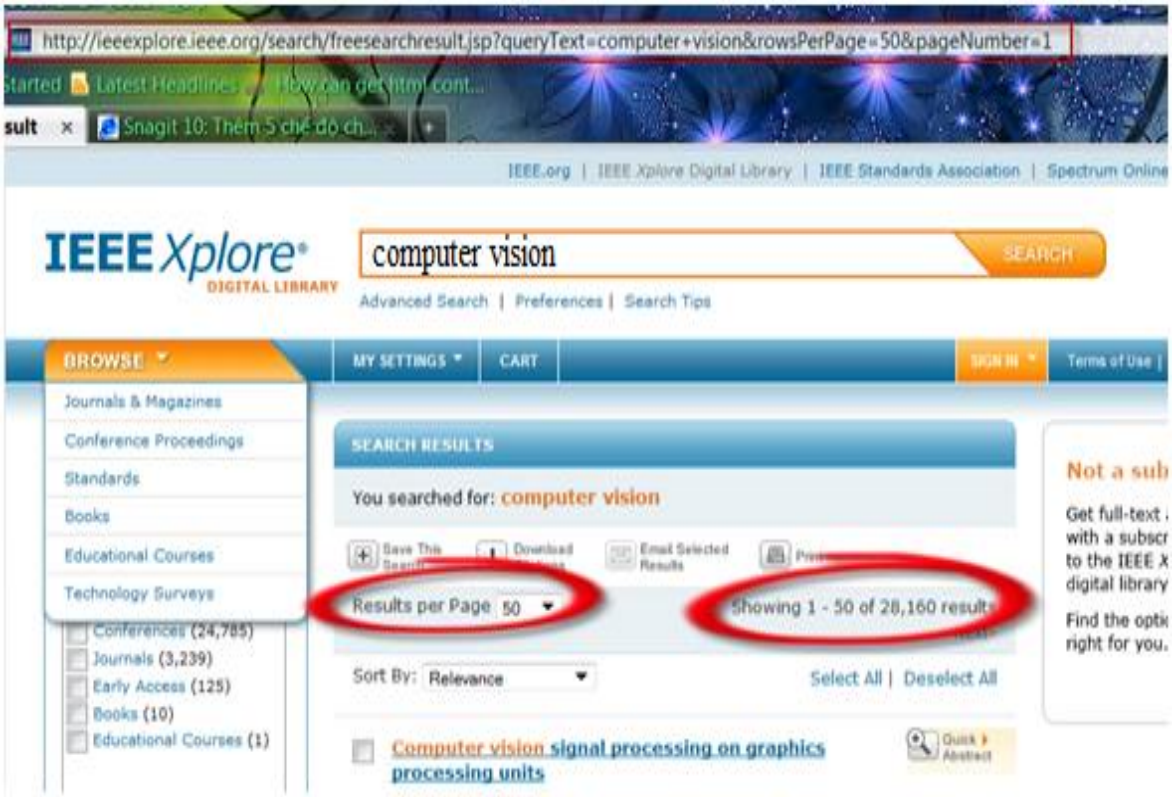
Bước 1: Tạo câu truy vấn dựa vào từ khóa và các pattern

Tương tự như bước 1 ở chức năng thu thập bài báo từ thư viện số ACM, nhưng trên thư viện số IEEEXplore thì sẽ thay thế các thẻ (pattern) khác và khi từ khóa là cụm từ thì khoảng trắng giữa hai từ sẽ được thay bằng dấu cộng "+".

Ví dụ: câu truy vấn sinh ra khi người dùng nhập từ khóa là "computer vision" thì câu truy vấn được tạo ra là:
http://ieeexplore.ieee.org/search/freesearchresult.jsp?queryText=computer+vision&rowsPerPage=50&pageNumber=1

Trong đó "&rowsPerpage= 50" là số kết quả tối đa trong một trang Web và "&pageNumber=1" là kết quả tìm kiếm và hiển thị là ở trang 1, như ở hình 3.3.

→*Mục tiêu:* tạo ra câu truy vấn (URL query) để thu thập các bài báo từ thư viện số IEEEXplore.



Hình 3.3 – Kết quả tìm kiếm từ thư viện số IEEEXplore

Bước 2: sử dụng câu truy vấn để lấy về kết quả thông tin cần tìm kiếm

Sau khi nhận được yêu cầu tìm kiếm, hệ thống sẽ gửi câu truy vấn vừa được tạo ở bước 1 lên trình duyệt Web. Dựa vào thông tin của câu truy vấn, trình duyệt sẽ tìm kiếm trong thư viện số IEEEXplore và trả về kết quả phù hợp với từ khóa yêu

cầu. Hệ thống sử dụng các pattern đã được định nghĩa ở bảng 3.2 để Crawl thu thập danh sách các địa chỉ URL tới các bài báo từ nội dung trả về của URL query trên bước 1.

→*Mục tiêu:* gửi câu truy vấn lên thư viện số, Crawl trên URL query để lấy về danh sách các đường dẫn đến nội dung các bài báo liên quan.

Bước 3: truy cập để lấy nội dung của các bài báo.

Tương tự bước 3 của cách thu thập từ thư viện số ACM. Hệ thống sẽ lấy về nội dung trong đường dẫn chứa bài báo trả về từ thư viện số.

→*Mục tiêu*: truy cập vào địa chỉ URL của một bài báo cụ thể lấy về nội dung trang Web chứa thông tin chi tiết của bài báo đó.

Bước 4: lấy thông tin chi tiết của mỗi bài báo dựa vào URL thông tin chi tiết bài báo vừa lấy được ở bước 3.

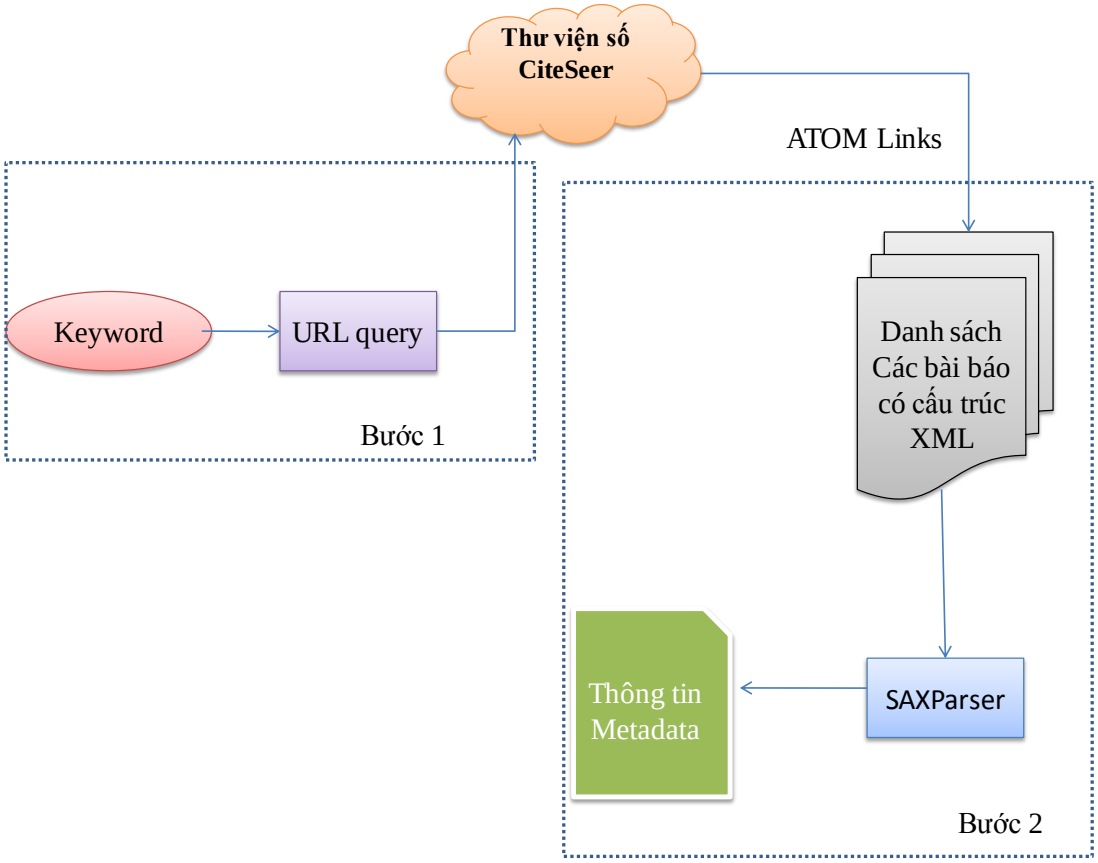
Hệ thống sử dụng các pattern từ số 6 đến 17 trong bảng 3.2, và bộ phân tích Bibtext để phân tích file Bibtext lấy được từ nội dung bên trong trang Web của bài báo để tìm thông tin chỉ mục của bài báo. Những thông tin thu thập bao gồm: số của bài báo trả về từ thư viện số, tiêu đề, các đồng tác giả, tóm tắt, năm công bố, tổ chức công bố của một bài báo.

Sau đó hệ thống lưu thông tin của bài báo vừa thu thập và quay lại bước 3 truy cập tới bài báo tiếp theo.

→*Mục tiêu*: thu thập thông tin chi tiết của một bài báo và lưu lại thành danh sách kết quả cần tìm theo yêu cầu của người dùng.

➔ Với cách kết hợp các pattern và trình phân tích Bibtex hệ thống có thể rút ra thông tin chỉ mục các bài báo trên thư viện số IEEEXplore. Trong trường hợp cấu trúc trang của IEEEXplore thay đổi thì chương trình có chức năng cho phép người dùng thay đổi các pattern tương ứng đảm bảo việc rút trích thông tin không bị ảnh hưởng.

3.2.3 Cách thức thu thập các bài báo từ thư viện số CiteSeer.



Hình 3.4 – Các bước thu thập trên thư viện số CiteSeer

Quá trình thu thập thông tin bài báo đối với thư viện số CiteSeer có thể chia làm 3 bước như sau:

Bước 1: tạo URL query

Khi người dùng nhập từ khoá tìm kiếm vào, hệ thống sẽ tạo URL query theo cấu trúc:

http://citeseerx.ist.psu.edu/search?q=***KEYWORD***&feed=atom&sort=rel

Trong đó keyword sẽ được thay thế bằng từ khoá tìm kiếm. Sau khi thực hiện câu query bằng URL query, hệ thống sẽ thực hiện kết nối với thư viện số CiteSeer để lấy kết quả trả về, ở đây kết quả trả về là một danh sách các bài báo mà hệ thống thư viện số CiteSeer tìm kiếm được.

Bước 2: Lấy thông tin bài báo từ kết quả trả về từ thư viện số.

Kết quả trả về từ thư viện số CiteSeer là một đường dẫn chứa thông tin bài báo dưới dạng XML được gọi là ATOM Link[21], mỗi bài báo được thể hiện trong kết quả trả về có dạng cấu trúc như sau:

```
<entry>
  <title>The Courtship of Atom</title>
  <summary>The Atom syndication specification may
   move to a new home at the W3C.</summary>
  <link rel="alternate"
</entry>
```

Hình 3.5 - Cấu trúc file XML của dữ liệu trả về từ CiteSeer

Từ danh sách kết quả tìm kiếm được bởi CiteSeer hệ thống sẽ dùng SAXPaser để lấy thông tin chỉ mục của từng bài báo. SAXParser[22] – Simple API For XML Paser là bộ phân tích nội dung XML theo cơ chế đọc từng ký tự một cách tuần tự (từ trên xuống dưới, từ trái qua phải). SAX cung cấp một số phương thức (callback), mà dựa vào các phương thức này giúp cho việc xác định các thông tin của bài báo từ nội dung XML.

Cách thức sử dụng trình phân tích SAX (parser) để lấy thông tin bài báo như sau:

+ Phương thức báo hiệu cho parser bắt đầu và kết thúc phân tích một tài liệu XML: ở đây tài liệu là một file XML chứa danh sách các bài báo. Mỗi bài báo được gọi là một Element (thể hiện) trong tài liệu XML .

startElement(): bắt đầu phân tích một thể hiện của tài liệu XML

endElement() : kết thúc phân tích một thể hiện của tài liệu XML

+ Phương thức báo hiệu bắt đầu và kết thúc phân tích một bài báo, trong phương thức này các thông tin của bào báo sẽ được lấy ra (được gọi là từng attributes). Thông tin của một bài báo bao gồm: tiêu đề (title), tác giả (authors),

[21] http://www.xml.com/pub/a/2004/06/16/dive.html

[22] http://en.wikipedia.org/wiki/Simple_API_for_XML

năm xuất bản (year), tóm tắt (abstract), đường dẫn tới bài báo (links). Những thông tin này sẽ được bộ phân tích lấy lần lượt từ trên xuống, giúp hệ thống thu thập được thông tin chỉ mục của bài báo.

startDocument() : bắt đầu phân tích một tài liệu XML

endDocument(): kết thúc phân tích một tài liệu XML

Như vậy sử dụng bộ phân tích SAX ta có thể lấy được thông tin của từng bài báo trong danh sách trả về từ thư viện số CiteSeer.

➔ Với việc sử dụng ATOM link chúng ta tận dụng được thông tin mà hệ thống cung cấp ngay trong kết quả trả về từ URL query. Khác với cách lấy thông tin của hai thư viện nêu ở trên, hệ thống không cần truy cập tới từng nội dung bài báo để lấy file Bibtex. Kết quả thực nghiệm cho thấy thời gian thu thập bài báo từ CiteSeer nhanh rất nhiều so với ACM và IEEEXplore.

3.3 Bộ phân tích Bibtex (Bibtex Parser).

Như đã trình bày ở phân thu thập thông tin từ thư viện số, hệ thống sử dụng bộ phân tích Bibtex parser để phân tích file .bib thu thập được từ thư viện số ACM và IEEEXplore để lấy thông tin chỉ mục của các bài báo. Bibtex là định dạng kiểu cấu trúc dùng để biểu diễn thông tin của tài liệu. Trong các thư viện số, các file Bibtex lưu thông tin Metadata của bài báo. Hệ thống sử dụng Bibtex parser trong chương trình Jabref[23] để thực hiện việc phân tích các file Bibtex thu được để lấy thông tin chỉ mục các bài báo.

Quá trình phân tích file Bibtex thực thi khi hệ thống rút được file .bib từ thư viện số. Dựa vào cấu trúc đã định nghĩa sẵn của file Bibtex dưới dạng XML.

Trình biên dịch cũng sẽ sử dụng SAX để phân tích nội dung file BibTex để xác định file Bibtex chứa nội dung của kiểu tài liệu nào (thông tin các loại tài liệu và các trường thông tin mà file Bibtex chứa có thể xem tại bảng 2.2). Từ việc xác định được loại tài liệu, thì hệ thống sẽ dựa vào các trường thông tin của tài liệu đó chứa, để lấy thông tin Metadata của tài liệu.

[23] http://jabref.sourceforge.net/index.php

Sau đây là ví dụ về cấu trúc và các hàm lấy thông tin của loại tài liệu là ARTICLE:


```
public static final BibtexEntryType ARTICLE =
    new BibtexEntryType()
    {
        public String getName()
        {
            return "Article";
        }

        public String[] getOptionalFields()
        {
            return new String[]
            {
                "number", "month", "part", "eid", "note"
            };
        }

        public String[] getRequiredFields()
        {
            return new String[]
            {
                "author", "title", "journal", "year", "volume", "pages"
            };
        }

        public String describeRequiredFields()
        {
            return "AUTHOR, TITLE, JOURNAL and YEAR";
        }

        public boolean hasAllRequiredFields(BibtexEntry entry, BibtexDatabase database)
        {
            return entry.allFieldsPresent(new String[]
                {
                    "author", "title", "journal", "year", "bibtexkey", "volume", "pages"
                }, database);
        }
    };
```

Hình 3.6 – Ví dụ cấu trúc của BibTex dạng Article

3.4 Kiểm tra dữ liệu trùng lặp.

Sau đây chúng tôi xin trình bày cách lưu thông tin của một bài báo được thu thập về từ các hệ thống thư viện và cách xử lý trùng lặp dữ liệu.

Để đảm bảo dữ liệu thu thập không bị trùng lặp với các dữ liệu đã có trong hệ thống cũng như trong cơ sở dữ liệu chỉ mục có sẵn trong DBLP thì hệ thống sẽ tiến hành kiểm tra tính tồn tại của bài báo thu thập được. Khi một bài báo được lấy về từ hệ thống thu thập (bài báo đã được rút các thông tin), hệ thống sẽ dựa vào các trường thông tin sau để kiểm tra sự trùng lặp dữ liệu đã có trong database:

+ Tựa đề bài báo (title).

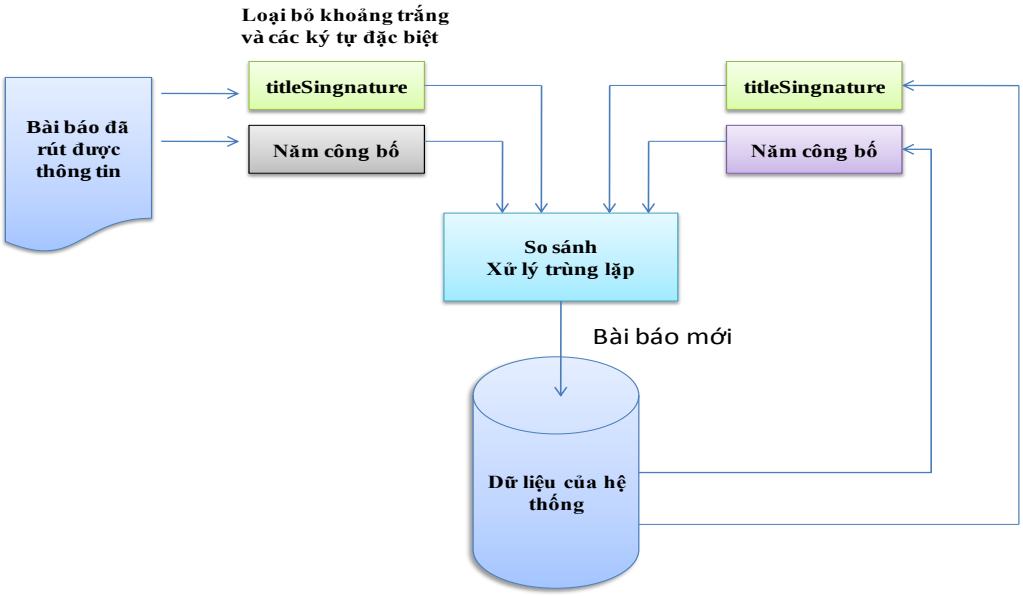
+ Năm công bố bài báo.

Để kiểm tra sự trùng lặp trên trường tựa đề bài báo, hệ thống sẽ so sánh tựa đề bài báo thu thập được trên trường tựa đề bài báo có trong cơ sở dữ liệu. Đầu tiên hệ thống sẽ tiền xử lý nội dung tựa đề của bài báo thu thập. Việc tiền xử lý tựa đề bài báo bao gồm loại bỏ khoảng trắng và một số ký tự đặc biệt như "!?,.". Sau đó chuỗi này sẽ được chuyển về cùng dạng chữ thường.

Đối với dữ liệu đã có trong database (trường title có trong database) hệ thống cũng xử lý tương tự như tựa đề của bài báo thu thập được. Như vậy, việc so sánh tựa đề bài báo là việc truy vấn tựa đề của bài báo vừa thu thập trên trường thông tin title của bài báo đã có trong database.

- Nếu tựa đề của bài báo thu thập tồn tại trong database thì hệ thống sẽ đi so sánh trường thứ 2 của bài báo đó là năm xuất bản. Nếu năm xuất bản của bài báo vừa thu thập và bài báo có title giống với bài báo thu thập có trong database khác nhau thì chứng tỏ bài báo thu thập được đã có trong dữ liệu của chương trình.

- Nếu bài báo thu thập được có title không giống title của bài báo nào trong database hoặc title giống trong database nhưng năm xuất bản khác nhau thì bai báo thu thập được là mới.



Hình 3.7 - Xử lý dữ liệu trùng lặp

3.5 Các luồng xử lý dữ liệu trong hệ thống

3.5.1 Luồng xử lý chung của hệ thống

Với chức năng cập nhật và xây dựng dữ liệu chỉ mục. Hệ thống cho phép người dùng tìm kiếm và cập nhật các bài báo mới trên thư viện số cũng như duyệt các bài báo đã có trong hệ thống vì vậy khi hệ thống được khởi động sẽ có hai luồng khác nhau.

- Thứ nhất, nếu người sử dụng muốn quản lý các bài báo có trong cơ sở dữ liệu thì chọn quá trình 3- Quản lý cơ sở dữ liệu (Quá trình này được mô tả rõ trong phần 4.4.4).

- Thứ hai, nếu muốn thực hiện chức năng thu thập các bài báo mới từ trên các thư viện số. Theo luồng xử lý này, việc thực hiện cập nhật có thể tiến hành theo quyết định của người dùng là: do người dùng thu thập, hay đặt chế độ cho hệ thống tự động thu thập theo thời gian định trước. Để thu thập, trước tiên hệ thống phải kiểm tra xem máy tính đã được kết nối Internet hay chưa.

+ Nếu hệ thống chưa kết nối với Internet, thì hệ thống sẽ gửi thông báo cho người dùng và kết thúc xử lý.

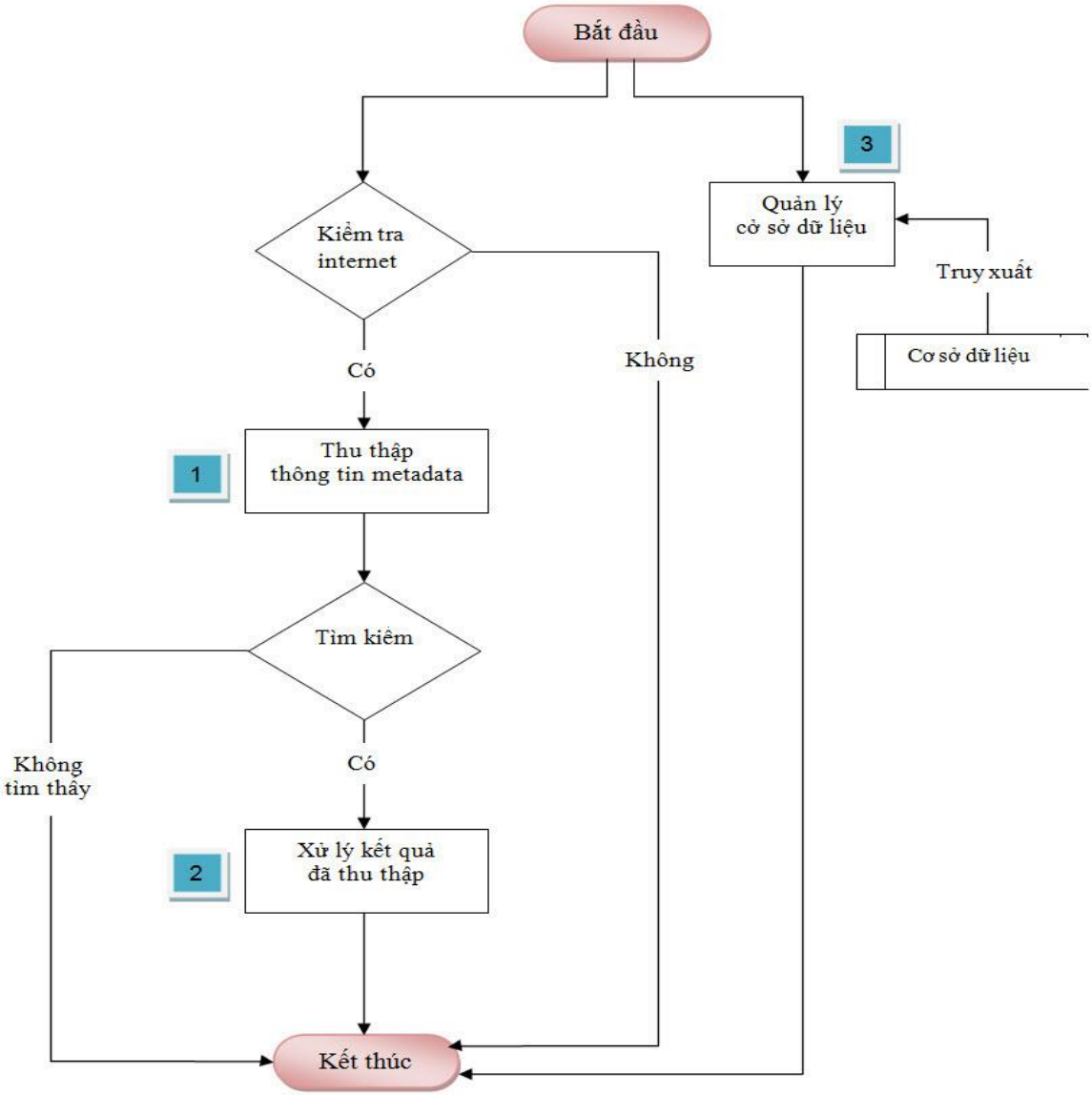
+ Nếu hệ thống có kết nối với Internet, thì hệ thống chuyển sang quá trình 2 thu thập Metadata (Quá trình này được mô tả rõ trong phần 4.4.2). Sau khi kết thúc quá trình thu thập hệ thống sẽ kiểm tra có kết quả trả về hay không.

- Nếu không có kết quả thì xuất thông báo và kết thúc hoạt động.
- Nếu có kết quả trả về thì chuyển sang quá trình 3 xử lý kết quả thu thập (Quá trình này được mô tả rõ trong phần 4.4.3).

Sau đó kết thúc luồng xử lý.

Hình 3.8 mô tả các luồng chính của hệ thống. Trong đó bao gồm các luồng phụ sau:

- Luồng 1: Quá trình thu thập Metadata từ các thư viện số.
- Luồng 2: Quá trình xử lý những kết quả đã được thu thập.
- Luồng 3: Quá trình quản lý các cơ sở dữ liệu của hệ thống.



Chú thích:
1. Quá trình thu thập Metadata từ các thư viện số
2. Quá trình xử lý những kết quả đã được thu thập
3. Quá trình quản lý cơ sở dữ liệu của hệ thống

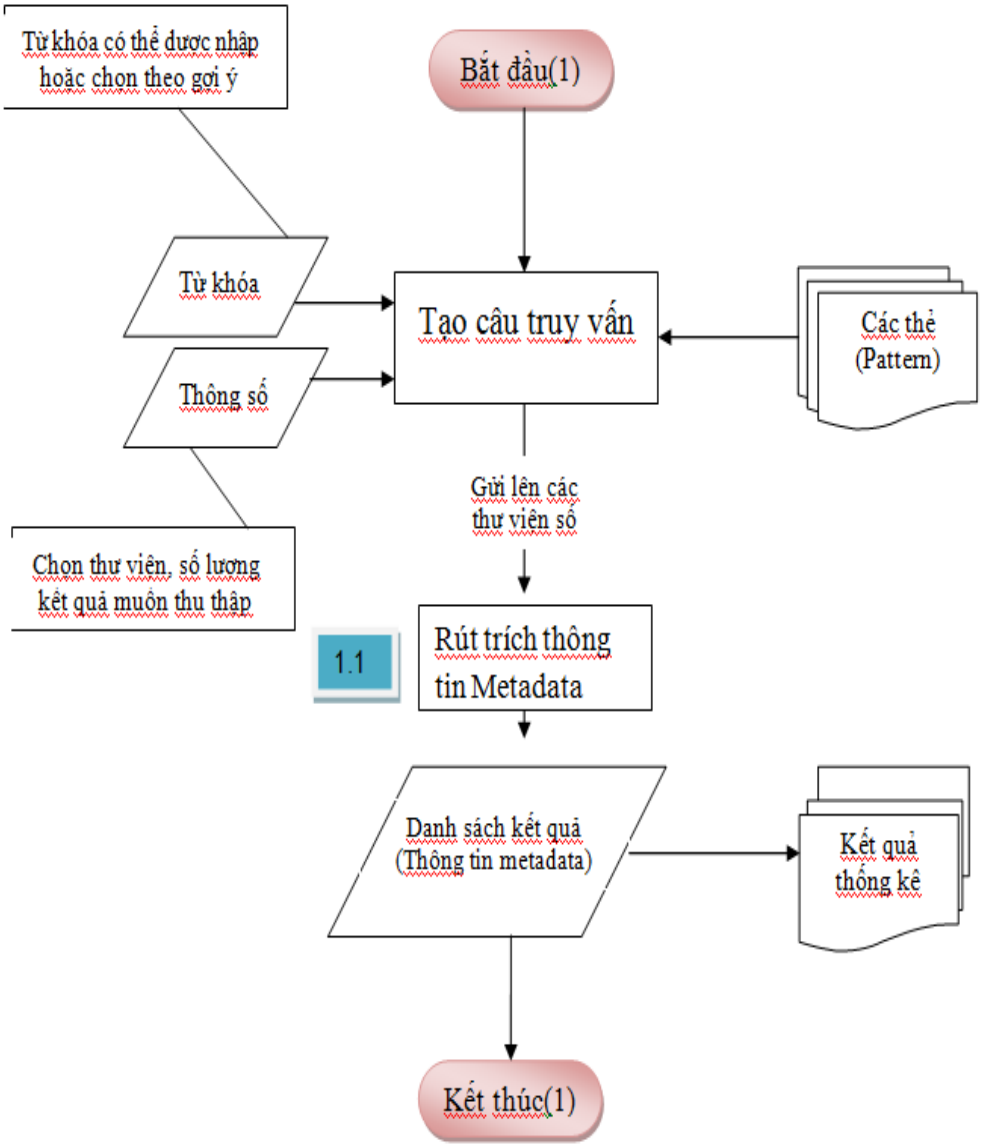
Hình 3.8- Các luồng xử lý chính của chương trình.

3.5.2 Quá trình thu thập thông tin Metadata từ thư viện số

Quá trình thu thập thông tin từ các thư viện số gồm các bước xử lý sau:

Người dùng tương tác với hệ thống yêu cầu tìm kiếm các bài báo theo các thông số đầu vào như: từ khóa, chọn thư viện số, số lượng kết quả muốn thu thập. Từ khóa do người dùng nhập hoặc chọn từ danh sách gợi ý của hệ thống. Dựa vào

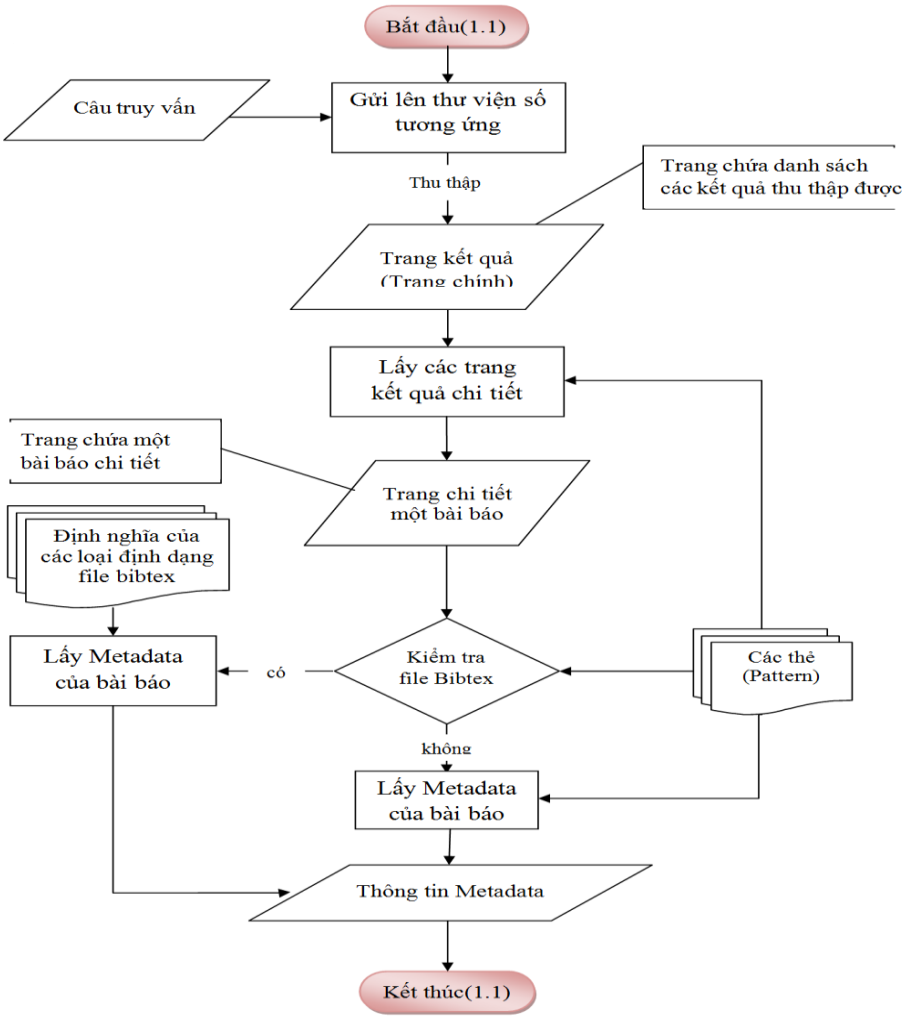
các thông số trên và các thẻ (pattern) đã được hệ thống định nghĩa tạo ra câu truy vấn. Sau đó câu truy vấn sẽ được gửi lên các thư viện số và lấy các kết quả trả về, đưa ra thống kê và kết thúc quá trình xử lý.



Hình 3.9 – Luồng xử lý thu thập thông tin Metadata.

Trong luồng xử lý này có một luồng xử lý phụ là 1.1 luồng này dùng để rút trích thông tin Metadata được trình bày chi tiết trong phần kế tiếp.

3.5.3 Rút trích thông tin Metadata



Hình 3.10- Luồng xử lý rút trích thông tin Metadata

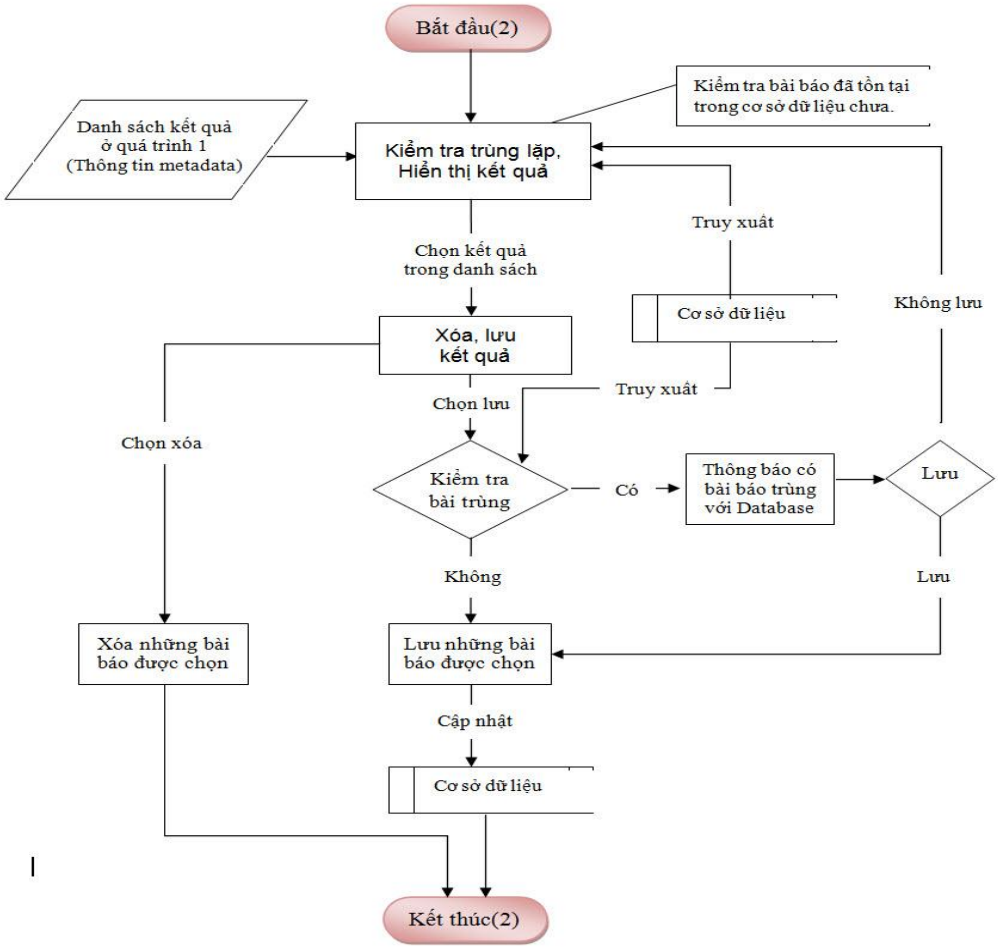
Quá trình lấy thông tin Metadata – thông tin chỉ mục bài báo từ các thư viện số gồm các bước sau:

- Lấy các câu truy vấn được tạo ở quá trình tạo câu truy vấn, gửi lên các thư viện số tương ứng. Từ câu truy vấn sẽ lấy về nội dung trang kết quả trả về từ thư viện số, trang này chứa danh sách các đường dẫn tới các bài báo phù hợp nội dung tìm kiếm trong dữ liệu của thư viện số.
- Dựa vào các thẻ (pattern) đã định nghĩa trong hệ thống và nội dung kết quả trả về vừa lấy được để tìm ra đường dẫn tới trang thông tin chi tiết tới mỗi bài báo trong kết quả trả về từ thư viện số.

- Truy cập vào từng trang chi tiết, kiểm tra xem trong trang đó có chứa file Bibtex hay không. Nếu có thì truy cập vào file Bibtex đó dựa vào các định nghĩa các loại định dạng file Bibtex của hệ thống (dùng bộ phân tích file Bibtex), từ đó rút ra thông tin Metadata của bài báo. Nếu không tồn tại file Bibtex thì truy cập vào mỗi đường dẫn chứa bài báo, sau đó dựa vào các thẻ (pattern) rút ra thông tin Metadata của bài báo trong mỗi trang kết quả đó.

3.5.4 Xử lý kết quả thu thập.

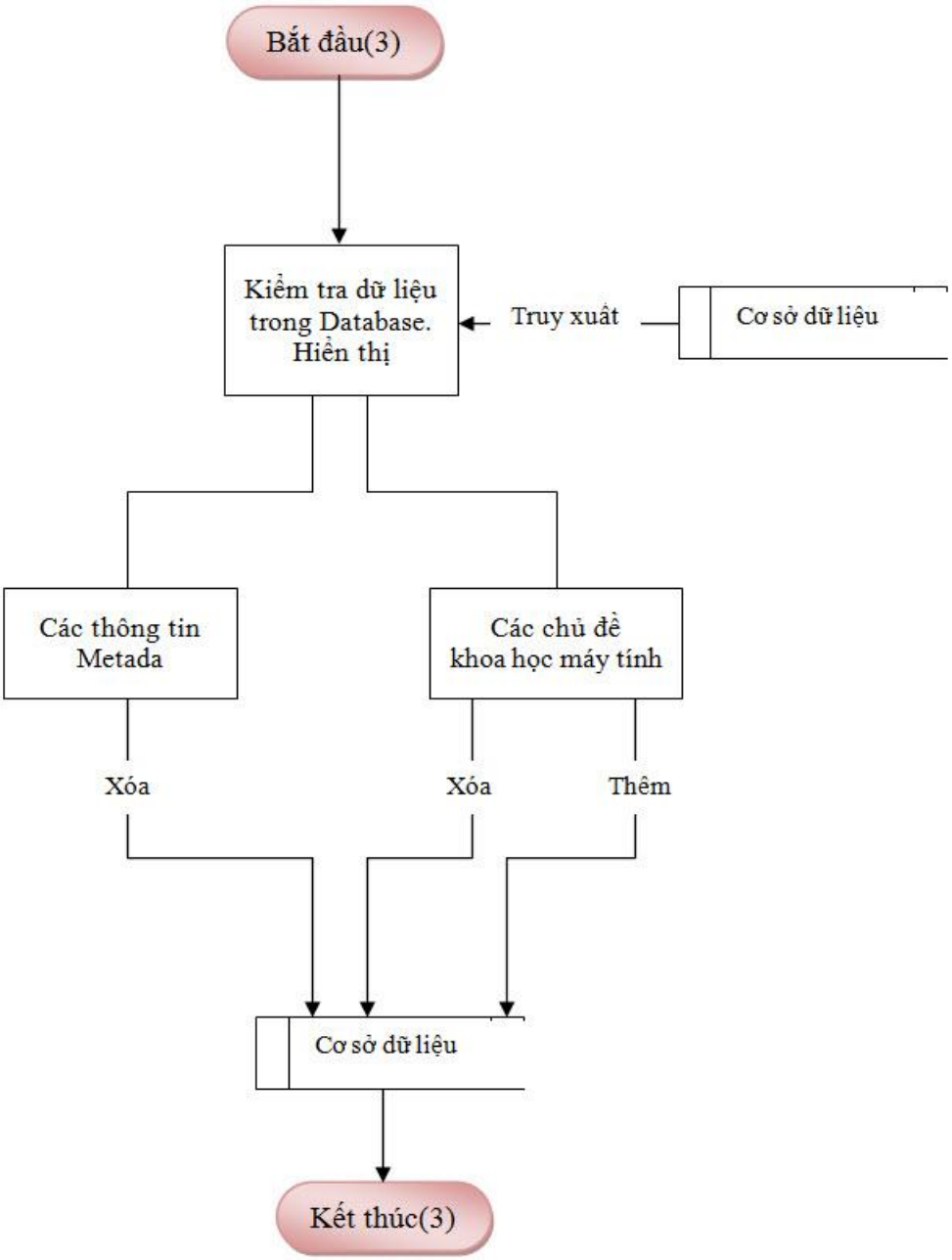
Từ danh sách kết quả ở quá trình trên, hệ thống sẽ kiểm tra xem trong số các bài báo mới thu thập có bài nào trùng với dữ liệu có trong database, sau đó hiển thị danh sách kết quả lên cho người dùng xem. Người dùng có thể chọn các bài báo trong danh sách để xóa khỏi danh sách hoặc lưu xuống cơ sở dữ liệu.



Hình 3.11 – Luồng xử lý kết quả thu thập được

3.5.5 Quản lý cơ sở dữ liệu

Ở quá trình này, hệ thống cho phép người sử dụng có thể quản lý cơ sở dữ liệu của mình. Có thể xem hoặc xóa các bài báo có trong cơ sở dữ liệu. Xem, xóa hoặc thêm chủ đề dùng làm từ khóa tìm kiếm cho hệ thống.



Hình 3.12 – Luồng xử lý quản lý cơ sở dữ liệu hệ thống

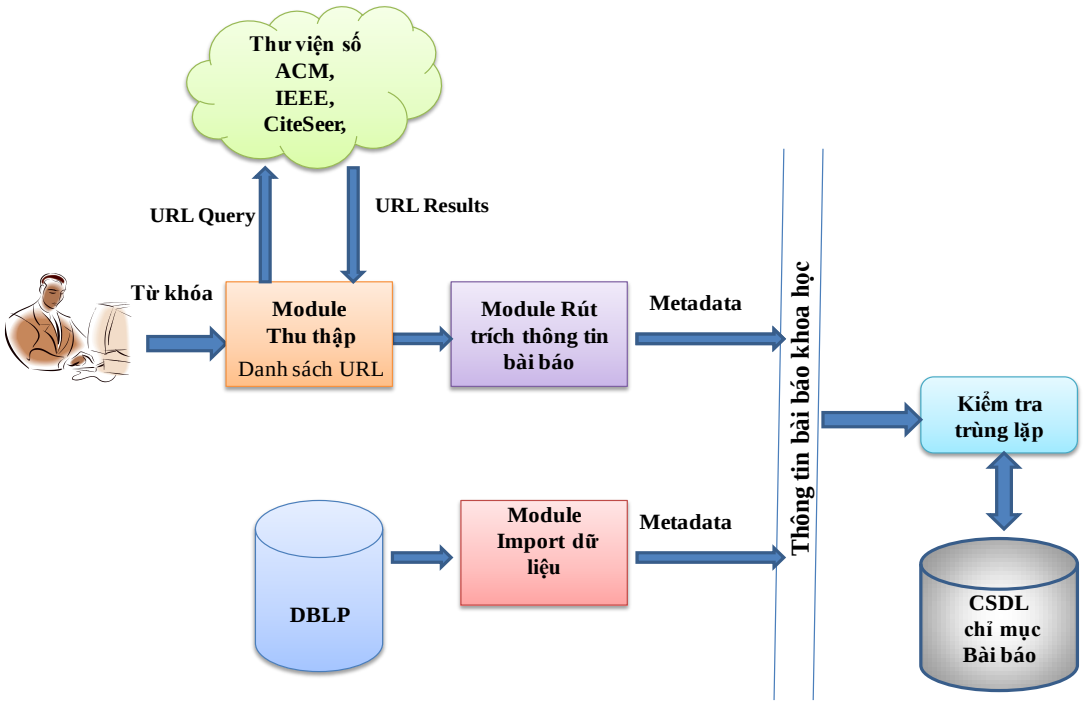
CHƯƠNG 4: HIỆN THỰC HỆ THỐNG.

4.1 Mở đầu

Trong chương 4, chúng tôi sẽ giới thiệu về hệ thống được xây dựng trong khóa luận. Sau đây là các thông số về công cụ phát triển ứng dụng:

+ Công cụ phát triển: Eclipse
+ Trình quản lý cơ sở dữ liệu: MySQL
+ Ngôn ngữ phát triển ứng dụng: Java
+ Môi trường ứng dụng: Desktop
+ Công nghệ: Hibernate, Web Crawler.

4.2 Kiến trúc hệ thống.



Hình 4.1 – Kiến trúc hệ thống

Trong hình 4.1, chúng tôi giới thiệu kiến trúc của hệ thống, dữ liệu đầu vào của hệ thống là từ khóa được nhập vào từ người dùng hoặc hệ thống tự động chọn theo cài đặt định sẵn từ danh sách các chủ đề trong lĩnh vực khoa học máy tính được lấy từ Wikipedia (phục lục C), hay danh sách tên các tác giả được lấy từ DBLP. Dựa vào từ khóa được nhập vào, hệ thống sẽ Crawl các bài báo trên các thư viện số, kết quả trả về từ các thư viện số là các đường dẫn tới các bài báo phù hợp với từ khóa tìm kiếm tương ứng, module rút trích thông tin bài báo sẽ sử dụng các trình phân tích kết hợp luật đã được định nghĩa trước, để nhận diện và rút ra các thông tin chỉ mục bài báo. Từ thông tin chỉ mục của bài báo, module kiểm tra trùng lặp dữ liệu sẽ kiểm tra tính tồn tại của bài báo trong DBLP sau đó lưu kết quả tìm kiếm xuống cơ sở dữ liệu. Module import dữ liệu DBLP có chức năng kết nối và cập nhật dữ liệu từ DBLP.

4.3 Thiết kế cơ sở dữ liệu.

4.3.1 Mô tả cấu trúc dữ liệu của DBLP

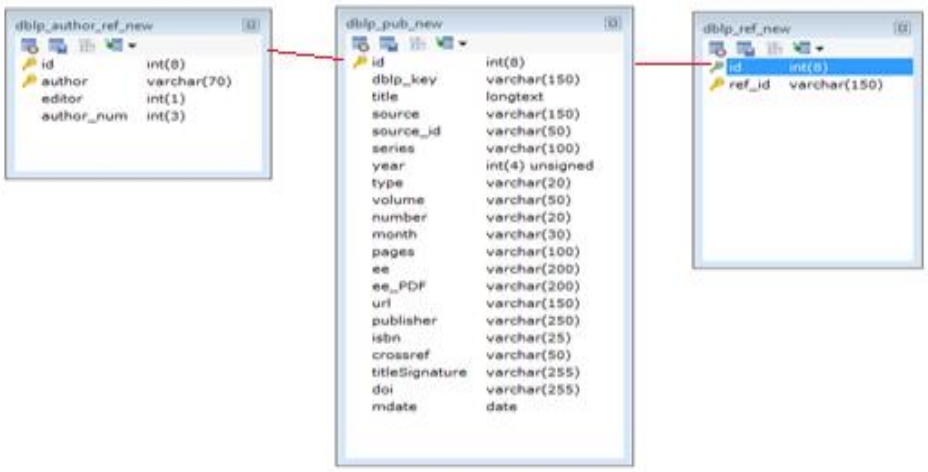
Sau đây là cấu trúc bảng SQL của DBLP được Tiến sĩ Jörg Diederich[24] xây dựng lên từ file XML của DBLP. Dữ liệu này được sử dụng trong hệ thống tìm kiếm Faceted DBLP và được cập nhật mỗi tuần một lần bằng cách sử dụng một đoạn script đọc dữ liệu trực tiếp từ file XML.

Dữ liệu được lưu trữ trong 3 bảng:

- **dblp_pub_new**: lưu thông tin bài báo.Thông tin trong bảng bao gồm: tựa đề bài báo, năm xuất bản, số trang, tên nhà xuất bản, và một số thông tin định danh bài báo trong file XML của DBLP được giải thích chi tiết trong phần mô tả phía dưới. Trong cơ sở dữ liệu này mỗi bài báo có một id riêng được sinh tự động và dùng chung cho các bảng có liên quan đến bài báo.
- **dblp_author_ref_new**: lưu thông tin về tác giả bài báo. Thông tin trong bảng bao gồm: tên tác giả, tác giả có phải là một người biên tập (editor) hay không. Trong bảng này, những tác giả viết cùng một bài báo thì có id giống nhau và giống id tương ứng của bài báo có trong bảng dblp_pub_new.

[24] http://www.l3s.de/~diederich/

- **dblp_ref_new**: lưu thông tin về các tham chiếu (reference) giữa các bài báo. Bài báo có id tương ứng trong bảng dblp_pub_new được tham chiếu bởi những bài báo nào được xác định bằng khóa dblp key.



Hình 4.2 – Mô hình dữ liệu của DBLP

Chi tiết về các trường trong các bảng được giải thích như sau:

Khóa chính	Trường (Field)	Kiểu (Type)	Chú thích (Comment)
PRIMARY	id	int(8)	Khóa chính (được sinh tự động, tăng dần).
	dblp_key	varchar(150)	Khóa trong file XML dữ liệu DBLP – Cấu trúc: tên loại tài liệu (book, conf)/ nhà xuất bản/tên tác giả đầu tiên trong tài liệu và năm công bố.
	title	longtext	Tựa đề của tài liệu.
	source	varchar(150)	Tên của hội nghị nơi bài viết được công bố: tên hội nghị, tạp chí, …
	source_id	varchar(50)	Tham chiếu đến các nguồn xuất bản

			(phần đầu tiên của dblp_key) - cấu trúc: tên tạp chí/tênviết tắt hội nghị + Năm diễn ra hội nghị.
	series	varchar(100)	Series của tài liệu (chỉ áp dụng với sách và proceeding).
	year	int(4) unsigned	Năm xuất bản của tài liệu.
	type	varchar(20)	Thể loại của tài liệu ví dụ article, proceedings, …
	volume	varchar(50)	Tập của nơi phát hành tài liệu. (tài liệu thuộc tập mấy trong cuốn phát hành)
	number	varchar(20)	Số tập của nơi phát hành tài liệu. (nơi phát hành có bao nhiêu tập)
	month	varchar(30)	Tháng tài liệu được xuất bản.
	pages	varchar(100)	Tài liệu thuộc trang bao nhiêu trong cuốn xuất bản.
	ee	varchar(200)	Địa chỉ URL tới bản điện tử của tài liệu.
	ee_PDF	varchar(200)	Địa chỉ URL tới bản PDF của tài liệu.
	url	varchar(150)	Địa chỉ của tài liệu trong dữ liệu của DBLP (bắt đầu bằng db/ ...).
	publisher	varchar(250)	Tên của nhà xuất bản; tên trường đối với tài liệu là luận văn; hoặc trang chủ nơi xuất bản.
	Isbn	varchar(25)	International Standard Book Number - mã số tiêu chuẩn quốc tế có tính chất thương mại duy nhất để xác định một

			quyển sách.
	crossref	varchar(50)	Tham chiếu chéo đến một tài liệu khác. Các tài liệu trong cùng một hội nghị, cùng một năm thì có crossref giống nhau.
	titleSignature	varchar(255)	Tựa đề của tài liệu không bao gồm khoảng trắng và các ký tự đặc biệt.
	doi	varchar(255)	digital object identifier – cung cấp thông tin giúp người dùng có thể tìm được tài liệu trên Internet.
	mdate	Date	Lần cuối cùng chỉnh sửa thông tin tài liệu.

Bảng 4.1 - Thông tin cấu trúc bảng dblp_pub_new

Khóa chính	Trường (Field)	Kiểu (Type)	Chú thích (Comment)
PRIMARY	id	int(8)	Khóa tương ứng với id trong bảng dblp_pub_new.
PRIMARY	author	varchar(70)	Tên của tác giả.
	editor	int(1)	Giá trị trả về giá trị là đúng khi tác giả cũng là một người biên tập (editor).
	author_num	int(3)	Số thứ tự của tác giả (tương ứng trong file gốc XML) . Một bài báo có 5 tác giả thì số tương ứng bắt đầu từ 0, tác giả có số tương ứng như thế nào thì có author_num tương tự vậy.

Bảng 4.2 - Thông tin cấu trúc bảng dblp_author_ref_new

Khóa Chính	Trường (Field)	Kiểu (Type)	Chú thích (Comment)
PRIMARY	id	int(8)	Khóa tương ứng với id trong bảng dblp_pub_new
	ref_id	varchar(150)	Khóa dblp_key của những bài báo được trích dẫn.

Bảng 4.3 - Thông tin cấu trúc bảng dblp_ref

Như vậy trong cấu trúc bảng của dblp được trình bày ở trên, hệ thống không chứa thông tin phần tóm tắt của bài báo (abstract).

4.3.2 Cơ sở dữ liệu hệ thống.

Từ cấu trúc các bảng của dblp ở trên, nhóm xây dựng thêm vào cấu trúc những bảng sau, để đảm bảo việc có thể cập nhật được dữ liệu mới của DBLP và lưu được các thông tin của các bài báo mà hệ thống thu thập được bao gồm phần tóm tắt của bài báo.

- **dbsa_sbj:** lưu thông tin về chủ đề của lĩnh vực khoa học máy tính.
- **dbsa_pub:** lưu thông tin bài báo được thu thập về từ các thư viện số.
- **dbsa_pub_in_dblp**: bảng lưu thông tin bổ sung của các bài báo trong dữ liệu DBLP bao gồm: chủ đề, những đường dẫn mở rộng (nơi mà bài báo có thể được tìm thấy – trang cá nhân của tác giả …).

Khóa Chính	Trường (Field)	Kiểu (Type)	Chú thích (Comment)
PRIMARY	id	int(8)	Khóa chính của chủ đề
	sbj_name	varchar(150)	Tên của chủ đề.

Bảng 4.4 – Thông tin cấu trúc bảng dbsa_sbj

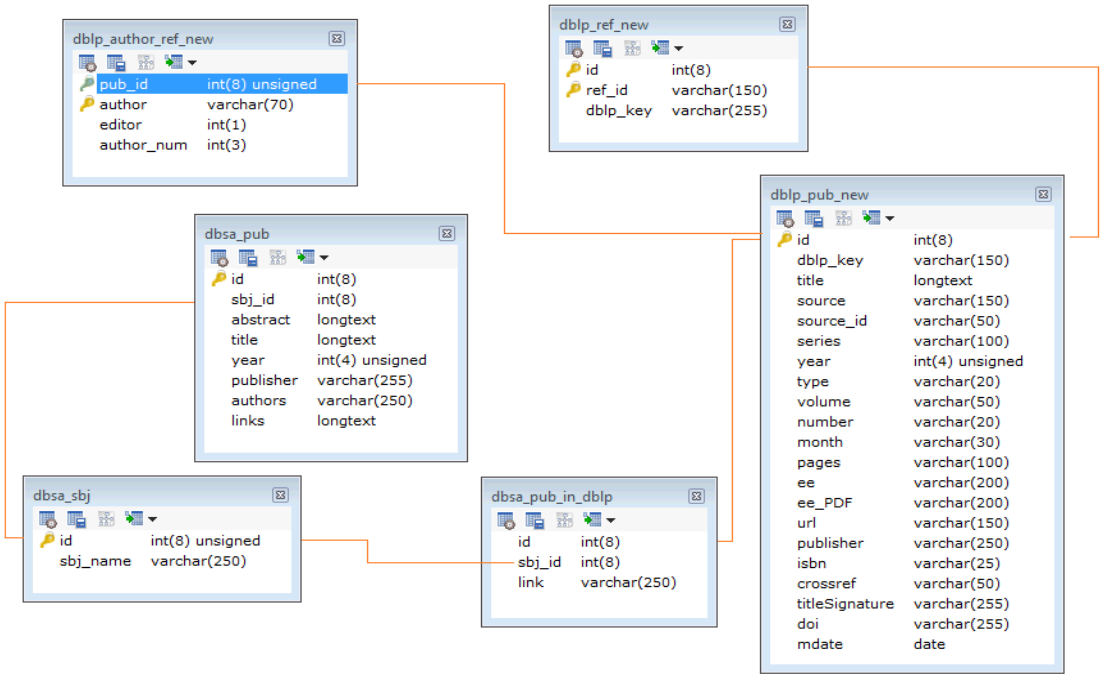
Khóa Chính	Trường (Field)	Kiểu (Type)	Chú thích (Comment)
PRIMARY	id	int(8)	Khóa tương ứng với id trong bảng dblp_pub_new
FK	sbj_id	int(8)	Khóa ngoại của dbsa_sbj
	links	longtext	Những đường dẫn mà người dùng có thể tìm được bài báo.

Bảng 4.5 – Thông tin cấu trúc bảng dbsa_pub_in_dblp

Khóa Chính	Trường (Field)	Kiểu (Type)	Chú thích (Comment)
PRIMARY	id	int(8)	Khóa chính của bảng
FK	sbj_id	int(8)	Khóa ngoại của dbsa_sbj
	abstract	longtext	Tóm tắt của bài báo
	title	longtext	Tựa đề bài báo
	year	int(4)	Năm xuất bản của tài liệu.
	publisher	varchar(250)	Tên nhà xuất bản
	authors	Varchar(250)	Tên các tác giả của bài báo, mỗi tên được cách nhau bằng dấu “ ,”
	links	longtext	Những đường dẫn mà người dùng có thể tìm được bài báo. Mỗi link khác nhau được cách nhau bằng dâu “,”.

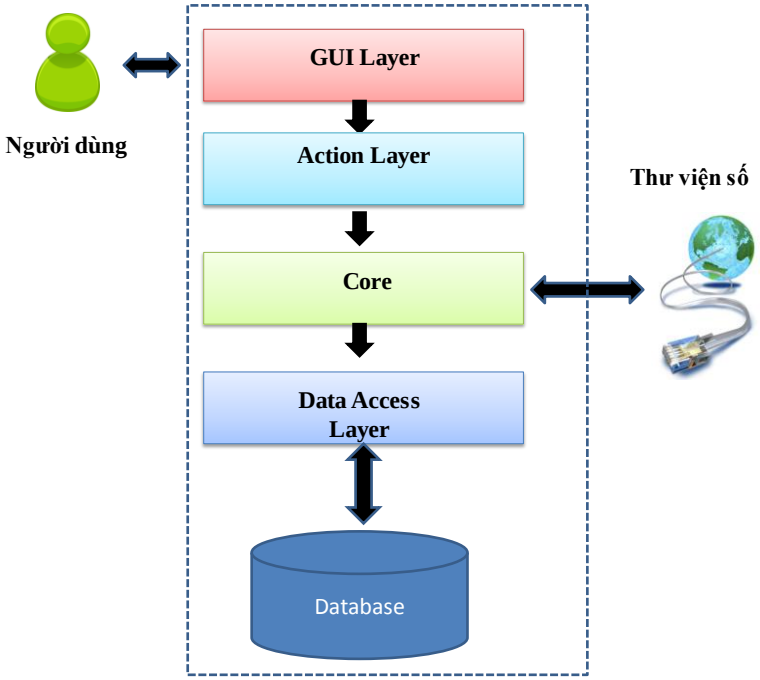
Bảng 4.6 - Thông tin cấu trúc bảng dbsa_pub

Dưới đây là mô hình các bảng có trong cơ sở dữ liệu của hệ thống.



Hình 4.3 – Mô hình dữ liệu hệ thống.

4.4 Kiến trúc phân lớp của hệ thống



Hình 4.4 - Kiến trúc phân tầng của hệ thống.

Kiến trúc của chương trình được chia làm 3 tầng trong đó:

+Tầng GUI: là tầng quản lý giao diện của chương trình.

+Tầng Action: chứa các lớp xử lý các sự kiện trong hệ thống.

+ Tầng Data Access: đảm nhận việc kết nối và xử lý dữ liệu.

+ Core: chứa các phương thức để kết nối với thư viện số để thu thập dữ liệu và phân tích các file Bibtex để lấy thông tin chỉ mục của bài báo.

Chương trình bao gồm 7 gói (package), mỗi gói có nhiều lớp xử lý các sự kiện trong chương trình. Sau đây là danh sách các package và một số lớp chính trong package.

**** GUI:*** Các lớp giao diện của chương trình.

*uit.tkorg.dbsa.gui.**

.main : giao diện chính của chương trình

.fetcher : chức năng thu thập và xử lý kết quả.

.databasemanagement : quản lý database của chương trình.

.statistic: thống kê kết quả thu thập.

**** Action:*** Các lớp xử lý sự kiện của người dùng.

*uit.tkorg.dbsa.actions.**

.fetchers : Xử lý các sự kiện thu thập thông tin Metadata.

.databasemanagement : Sự kiện quản lý cơ sở dữ liệu.

**** Core:***

*uit.tkorg.dbsa.core.**

.fetchers: xử lý trong chức năng thu thập.

.database: xử lý tương tác với cơ sở dữ liệu và quản lý dữ liệu.

.hibernate: tương tác với cơ sở dữ liệu sử dụng Hibernate.[25]

**** Mode:***

*uit.tkorg.dbsa.model.**

.Author đối tượng tác giả của bài báo khoa học.

[25] http://www.hibernate.org/

.*Publication* đối tượng bài báo khoa học của DBLP.

.*DBSAPublication:* đối tượng bài báo của chương trình.

.*subject* đối tượng chủ đề bài báo khoa học.

.*Author.hbm.xml* file maping với cơ sở dữ liệu bảng author.

.*Publication.hbm.xml* file maping với cơ sở dữ liệu bảng publicaction.

.*DBSAPublication.hbm.xm*

.*subject.hbm.xml*

**** Documentation:*** Tài liệu của chương trình.

*uit.tkorg.dbsa.documentation.**

.*doc*

.*references*

.*presentations*

...

**** Resources:*** Hình ảnh, biểu tượng đã được sử dụng trong chương trình.

*uit.tkorg.dbsa.resources.**

.*images*

.*icon*

...

**** Properties:***

*uit.tkorg.dbsa.properties.files.**

.*DBSA_Resources_EN.properties:* Tập tin ngôn ngữ tiếng Anh.

.*DBSA_Resources_VN.properties:* Tập tin ngôn ngữ tiếng Việt.

.*DBSAApplicationConst :* định nghĩa các biến hằng số trong chương trình.

.*DBSAModulesProperties :* Lớp định nghĩa các module của chương trình.

.*FileLocationProperties :* Lớp định nghĩa các đường dẫn được sử dụng.

.*GUIProperties:* Lớp định nghĩa những hình ảnh trong chương trình.

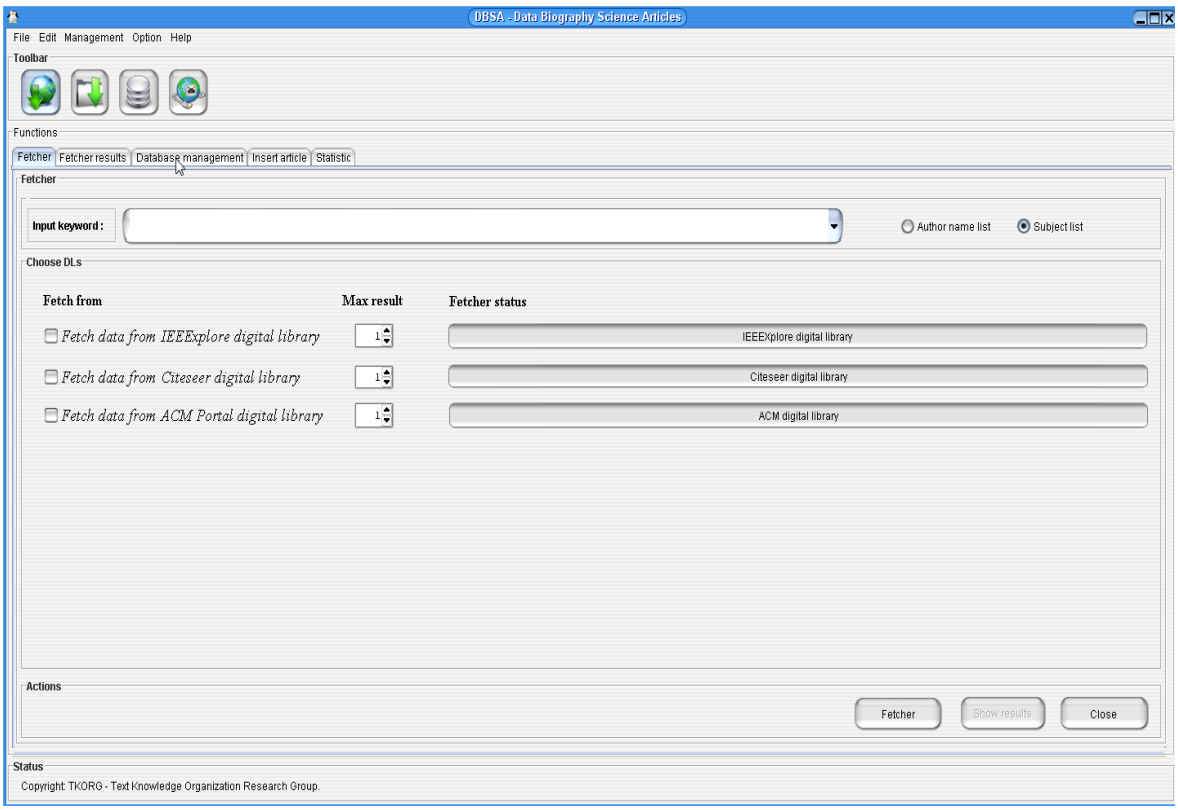
.*DBSA_IEEE_Default_Pattern:* Những thẻ (Pattern) mặc định để rút trích thông tin Metadata từ thư viện số IEEE.

.DBSA_ACM_Default_Pattern: Những thẻ (Pattern) mặc định để rút trích thông tin Metadata từ thư viện số ACM.

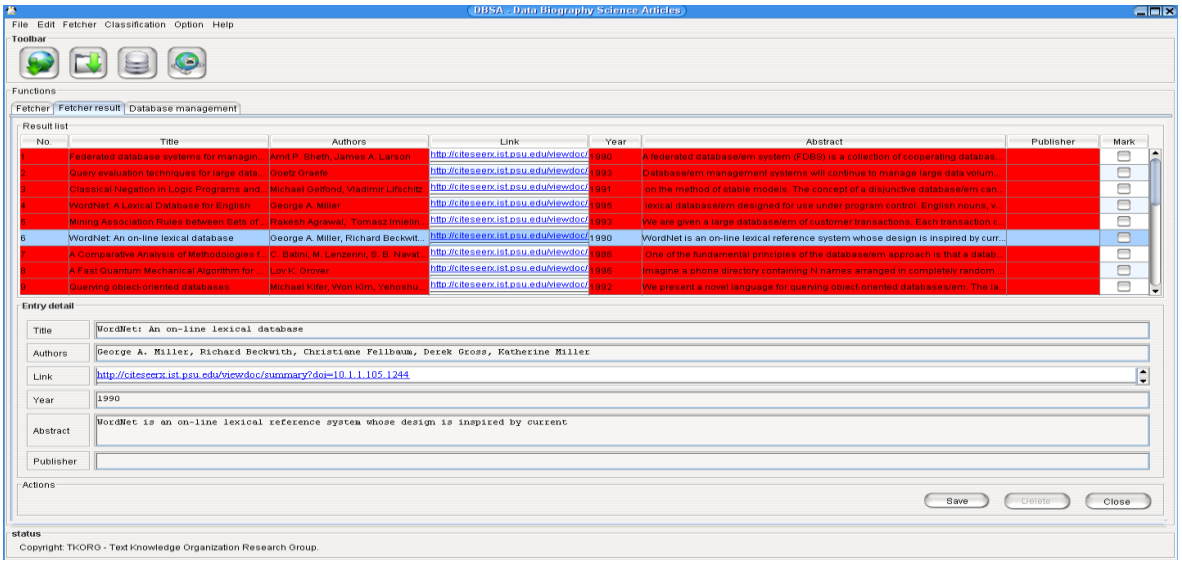
.DBSA_Define_Pattern: Những thẻ (Pattern) để rút trích thông tin Metadata từ hai thư viện số ACM và IEEE mà chương trình dang sử dụng.

4.5 Hệ thống xây dựng dữ liệu chỉ mục.

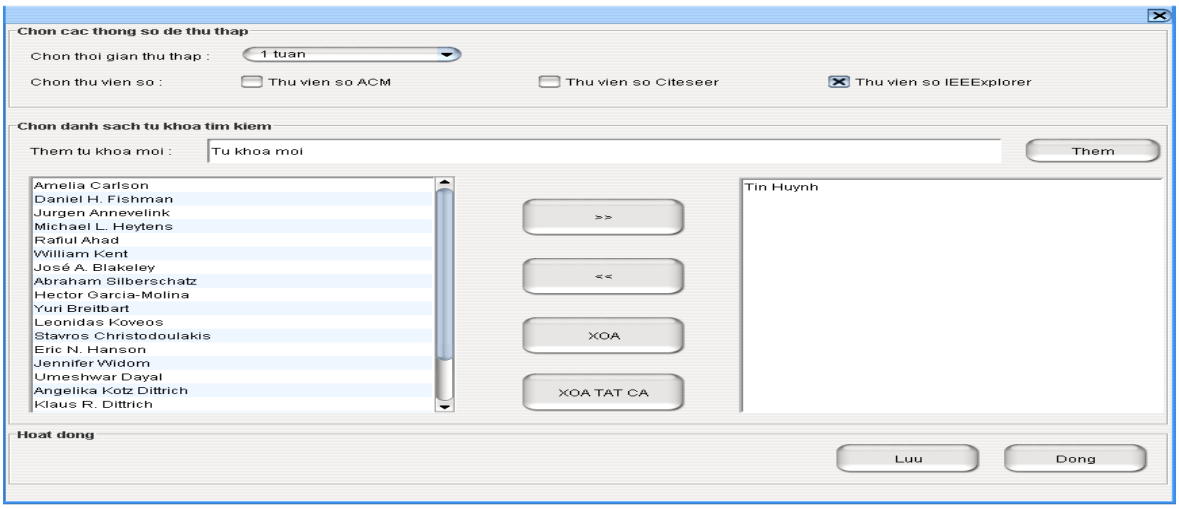
Hệ thống của chúng tôi quản lý cơ sở dữ liệu bằng MySQL và được xây dựng bằng Java do đó có thể chạy tốt trên bất cứ hệ điều hành nào như Windows, Linux. Hệ thống có giao diện bao gồm tiếng Anh và tiếng Việt, có cách hiển thị dữ liệu trực quan giúp người dùng chỉnh sửa các thông tin của bài báo hoặc thêm bớt dữ liệu trực tiếp và tương tác tốt với người dùng. Sau đây là một số hình ảnh của chương trình.



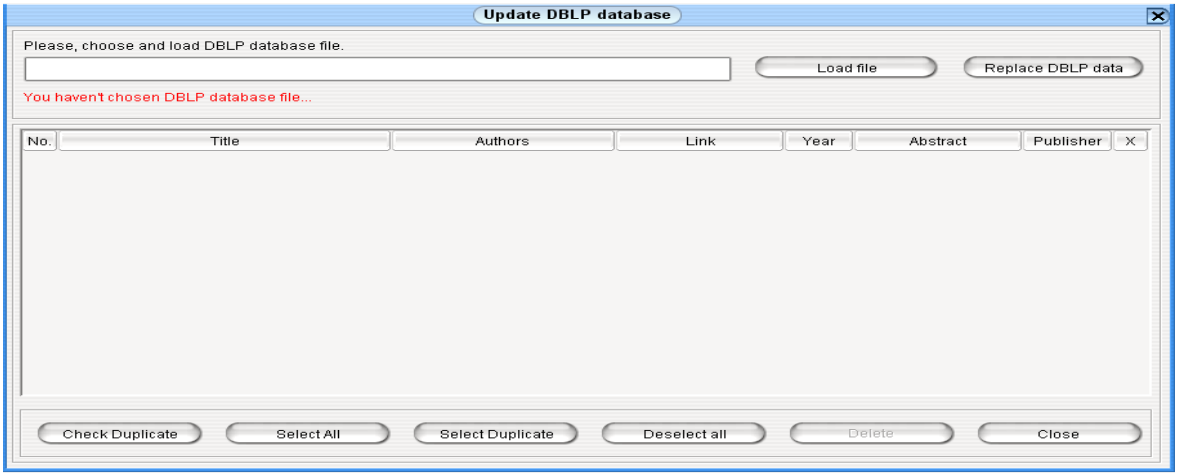
Hình 4.5 – Giao diện chính của hệ thống



Hình 4.6– Kết quả thu thập từ hệ thống



Hình 4.7 – Cài đặt tự động cập nhật bài báo mới



Hình 4.8 – Chức năng cập nhật dữ liệu DBLP

CHƯƠNG 5: THỰC NGHIỆM ĐÁNH GIÁ

5.1 Kết quả thực nghiệm.

Dữ liệu DBLP đưa vào hệ thống được tác giả công bố tháng 1 năm 2011 chứa khoảng 1,5 triệu bài báo (dblp-2011-1-26.sql.gz [26]).

Để đánh giá tính cập nhật, đầy đủ dữ liệu của hệ thống xây dựng được, chúng tôi tiến hành truy vấn trên hệ thống đã xây dựng với đầu vào là từ khóa chủ đề trong lĩnh vực khoa học máy tính (bảng dưới thống kê kết quả khi tìm kiếm 2 từ khóa database và data mining). Sau đó chúng tôi xét trên 100 kết quả trả về lần lượt từ 3 thư viện số ACM, IEEEXplore và CiteSeer thì kết quả thu được như sau:

Với từ khóa nhập vào là: database

Thư viện số	Thời gian thu thập (phút)	Số bài tồn tại trong DBLP (%)	Số bài trước năm 2005 mà trong DBLP không chứa (%)
ACM	33	93	85,71
CiteSeer	0.5	78	90,91
IEEEXplore	1.02	44	48,21

Bảng 5.1 - Kết quả của hệ thống với từ khóa là Database

Với từ khóa nhập là: data mining

Thư viện số	Thời gian thu thập (phút)	Số bài tồn tại trong DBLP (%)	Số bài trước năm 2005 mà trong DBLP không chứa (%)
ACM	32	52	66,67
CiteSeer	0.25	71	72,41
IEEEXplore	1	46	12,96

Bảng5.2 - Kết quả của hệ thống với từ khóa là Data mining.

[26] http://dblp.l3s.de/dblp++.php

+ Thời gian thu thập: tính khi hệ thống đã được nhập vào từ khóa và người dùng yêu cầu thu thập thông tin.

+ Số bài tồn tại trong DBLP: được tính bằng số bài trong 100 bài trả về từ thư viện số và tồn tại trong dữ liệu DBLP.

+ Số bài trước năm 2005 mà trong DBLP không chứa: được tính bằng số bài có năm xuất bản trước năm 2005 trong 100 bài báo trả về từ thư viện số mà không chứa trong dữ liệu DBLP.

Như vậy, theo các đánh giá trên trung bình hệ thống đã cập nhật được các bài báo khi được công bố trên thư viện số cũng như bổ sung những dữ liệu còn thiếu trong trong DBLP (kết quả thể hiện ở bảng 5.3).

Thư viện số	Dữ liệu của bài báo được bổ sung vào DBLP (%)
ACM	27,5
Citeseer	25,2
IEEEXplore	55

Bảng 5.3 - Kết quả bổ sung dữ liệu mới của hệ thống.

(Dữ liệu của bài báo được bổ sung vào DBLP được tính bằng trung bình số lượng bài báo được bổ sung trên các thư viện số với 2 từ khóa là database và data mining.)

Để đánh giá tính cập nhật dữ liệu của hệ thống xây dựng, chúng tôi tìm kiếm bài báo được xuất bản năm 2010 ví dụ như bài báo: "Gate framework based metadata extraction from scientific papers" của tác giả Tin Huynh, Kiem Hoang [18] được công bố tháng 12 năm 2010, chúng tôi thấy chỉ trên thư viện số của tổ chức công bố bài báo là IEEEXplore tồn tại thông tin bài báo này, còn trên các thư viện số khác hoặc trong dữ liệu chỉ mục DBLP chưa có thông tin chỉ mục của bài báo này. Như vậy, đối với hệ thống chúng tôi đã có thể cập nhật được thông tin bài báo mới được công bố trên thư viện số.

5.2 Đánh giá

Sau khi thực hiện khóa luận chúng tôi đã đạt được những kết quả sau:

Về mặt kiến thức:

- Chúng tôi đã có được những kiến thức về các hệ thống xây dựng đánh dấu dữ liệu chỉ mục hiện nay.
- Chúng tôi đã có được kiến thức chung về việc rút trích thông tin Metadata, Bibtex.
- Với việc xây dựng hệ thống chúng tôi đã có kiến thức trong việc sử dụng các công nghệ như Web Crawler, Hibernate, BibTex parser, …

Về mặt kinh nghiệm:

- Chúng tôi đã có được những kinh nghiệm về kỹ năng lập trình, làm việc nhóm. Những kinh nghiệm này sẽ giúp ích cho chúng tôi cho quá trình làm việc tại các công ty sau khi ra trường.
- Có được kinh nghiệm trong việc viết báo cáo, trình bày báo cáo và những kỹ năng mềm cần thiết cho một kỹ sư ngành công nghệ phần mềm.

Về chương trình xây dựng trong khóa luận:

- Dựa vào kết quả các thực nghiệm được trình bày bên trên, hệ thống đã đảm được những mục tiêu mà chúng tôi đã đưa ra là xây dựng thành công hệ thống thu thập thông tin sử dụng WebCrawler đồng thời kết hợp dữ liệu chỉ mục có sẵn từ DBLP. Dữ liệu thu thập từ hệ thống xây dựng đảm bảo được tính chính xác và cập nhật.
- Mặc dù vậy chương trình còn một số hạn chế như: cần bổ sung thêm nhiều thư viện để kết quả thu thập được là đầy đủ nhất.

CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.

6.1 Kết luận.

Trong khóa luận này, chúng tôi hiện thực một hệ thống dùng để xây dựng và cập nhập dữ liệu chỉ mục các bài báo khoa học sử dụng Web Crawler trên các thư viện số đồng thời kết hợp dữ liệu chỉ mục có sẵn. Như vậy, so với hệ thống DBLP hay các hệ thống được giới thiệu trong các bài báo [14][5][16] chúng tôi đã cải thiện được độ chính xác cũng như hiệu năng khi Crawl trên các thư viện số. Khác với giải pháp được giới thiệu trong [3] của hệ thống ACI, thì hệ thống chúng tôi đã tận dụng được dữ liệu có sẵn từ DBLP cũng như thu thập được dữ liệu chỉ mục có sẵn trực tiếp từ thư viện số mà không cần download tài liệu về. Khóa luận cũng như đề ra một hướng tiếp cận để bổ sung và cập nhật dữ liệu DBLP bằng cách lấy thông tin trực tiếp từ thư viện số. Ngoài ra với chức năng tự động cập nhật theo thời gian định sẵn hệ thống đảm bảo được tính cập nhật các bài báo mới được công bố trên thư viện số.

6.2 Hướng phát triển.

Bước tiếp theo trong tương lai chúng tôi sẽ hoàn thiện hệ thống với các chức năng:

- Nâng cao hiệu năng thu thập cũng như rút ngắn thời gian phân tích kết quả trên thư viện số.
- Thu thập các bài báo từ nhiều nguồn khác nhau. Bao gồm những thư viện số khác và từ các trang cá nhân của tác giả.
- Phân loại chủ đề cho các bài báo khoa học đã được thu thập dựa trên những thông tin về chỉ mục của bài báo.
- Xây dựng công cụ tìm kiếm các bài báo khoa học dựa trên dữ liệu mà thu thập được

TÀI LIỆU THAM KHẢO.

1. Tài liệu tiếng Anh

[1] Alexander Yates. *"Information Extraction from the Web: Techniques and Applications"*. Phd thesis, University of Washington, 2007.

[2] Badawia M. Albassuny. *"Automatic metadata generation applications: a survey study*". International Journal of Metadata, Semantics and Ontologies . Volume 3, Number 4 / 2008. pp 260 – 282.

[3] C.L. Giles, K. Bollacker, S. Lawrence,CiteSeer: *"An Automatic Citation Indexing System"*.Digital Libraries 98: Third ACM Conf. Digital Libraries, ACM Press,New York, 1998, pp. 89-98.

[4] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan, "*A Survey of Web Information Extraction Systems*" IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 10, pp. 1411-1428, Oct. 2006.

[5]G. Pant, K. Tsioutsiouliklis, J. Johnson, C.L. Giles: "*Panorama: Extending Digital Libraries with Topical Crawlers*". Proc. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004).

[6] Gautam Pant , Padmini Srinivasan, Filippo Menczer."*Crawling the Web*". 2004.

[7] Holger Bast, Ingmar Weber: *"The CompleteSearch Engine: Interactive, Efficient, and Towards IR&DB Integration"*, CIDR 2007: 3rd Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 2007, 88-95.

[8] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E.A. Fox, *"Automatic document metadata extraction using support vector machines"*. In: Proceedings of the 3rd ACM/IEEECS Joint Conference on Digital Libraries, International Conference on Digital Libraries, pages 37–48. IEEE Computer Society Press, Washington, DC, 2003.

[9] J. Diederich and W.-T. Balke: "*FacetedDBLP - Navigational Access for Digital Libraries"*. Bulletin of IEEE Technical Committee on Digital Libraries, Volume 4 Issue 1, Spring 2008, ISSN 1937-7266.

[10] Jane Greenberg, Kristina Spurgin , Abe Crystal. *"Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions"*. Int. J. Metadata, Semantics and Ontologies, Vol. 1, No. 1, 2006.

[11] Jim Cowie and Yorick Wilk. *"Information Extraction"*, 1996.

[12] K. Nakagawa, A. Nomura, and M. Suzuki, *"Extraction of Logical Structure from Articles in Mathematics"*, MKM, LNCS 3119, pages 276-289, Springer Berlin Heidelberg from Articles in Mathematics, 2004.

[13] Line Eikvil. "*Information Extraction from World Wide Web – A Survey"*. Norwegian Computing Center, PB, Citeseer. July 1999.

[14] Michael Ley, *"The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspective"*. Lecture Notes in Computer Science, 2002, Volume 2476/2002, 481-486.

[15]Nicola Zeni, Nadzeya Kiyavitskaya, Luisa Mich, John Mylopoulos, James Cordy. "*A Lightweight Approach to Semantic Annotation of Research Papers"*. Natural Language Processing and Information Systems (2007), pp. 61-72.

[16] Rong Shi, Kurt Maly, Mohammad Zubair: *"Automatic metadata discovery from noncooperative digital libraries"*. in Proc. of IADIS international Conf. on e-Society 2003

[17] Roth, D.L. *"The emergence of competitors to the Science Citation Index and the Web of Science",* Current Science, Vol. 89 (2005), 1531 – 1536.

[18] Tin Huynh, Kiem Hoang . *"GATE framework based metadata extraction from scientific papers"*. Dept. of Comput. Sci., Univ. of Inf. Technol., Ho Chi Minh City, Vietnam. 03 December 2010.

2. Tài liệu tiếng Việt

[19] Huỳnh Ngọc Tín, "*Báo cáo chuyên đề rút trích thông tin*", Đại Học Công Nghệ Thông Tin, Năm 2010.

3. Tài liệu Internet

[20] http://www.nlv.gov.vn/nlv/index.php/en/2008060697/DUBLIN-CORE/XML-Metadata-va-Dublin-Core-Metadata.html

[21] http://hanoi.centre-linux.org/article.php3?id_article=99

PHỤ LỤC A: HƯỚNG DẪN CÀI ĐẶT HỆ THỐNG.

1. Các bước tạo database cho chương trình:

Đối với hệ thống chạy lần đầu chưa có CSDL, việc cài đặt cơ sở dữ liệu bao gồm việc import dữ liệu DBLP bằng tay và thêm các bảng của hệ thống bằng script SQL đi theo của chương trình. Sau đây chúng tôi xin giới thiệu cách cài đặt cơ sở dữ liệu trong trường hợp này, đối với trường hợp đã có CSDL sẵn thì việc import CSDL đơn giản là việc restore CSDL vào database.

Thông số hệ thống:

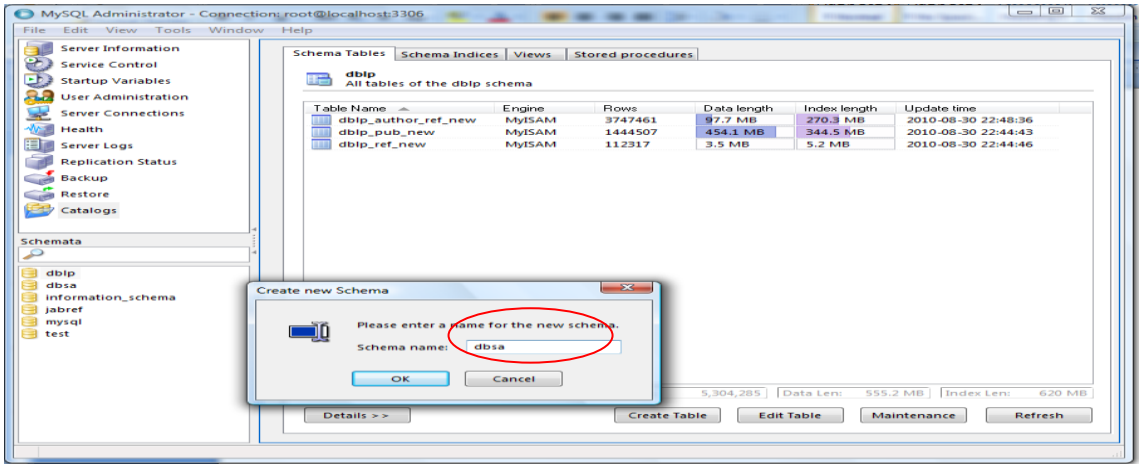
- Các phần mềm: Mysql Server
- Tên database : **dbsa**
- Hệ quản trị cơ sở dữ liệu: MySQL
- Tên truy cập : root
- Mật khẩu : root

Bước 1:

- Tải cơ sở dữ liệu mới của DBLP tại địa chỉ: http://dblp.l3s.de/dblp++.php
- Cài đặt đầy đủ các phần mềm môi trường và tương tác: bộ MySQL, Java…

Bước 2:

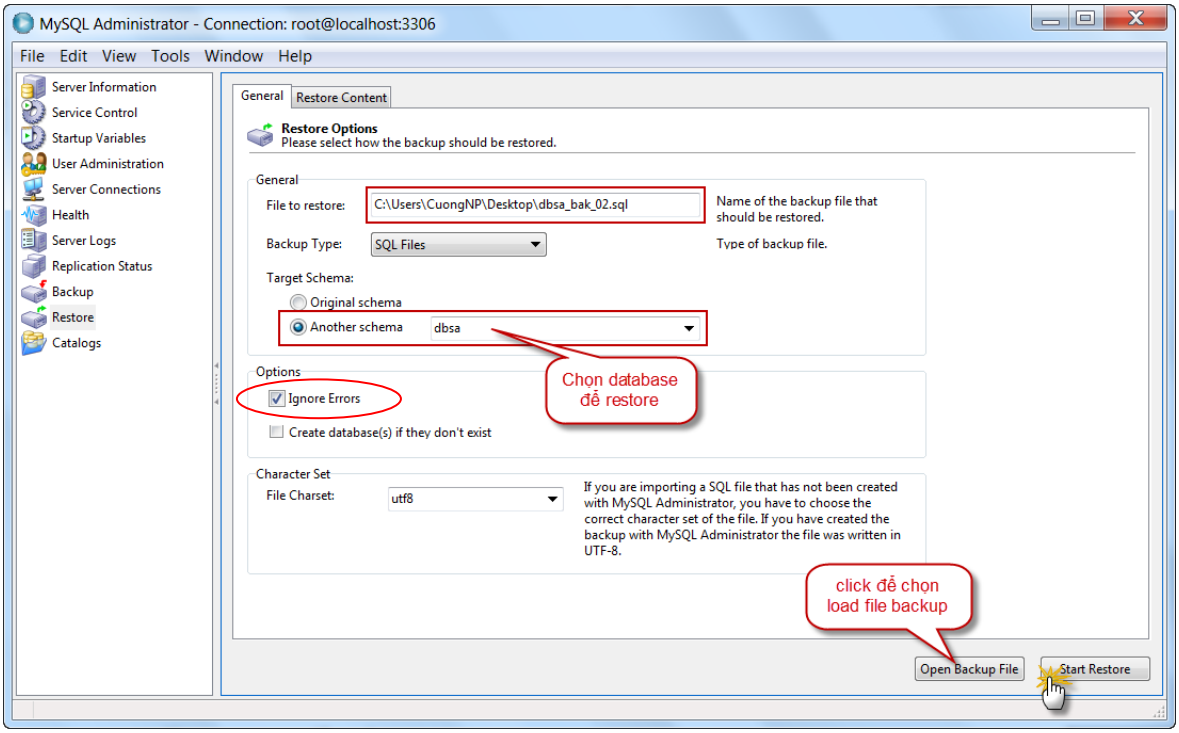
- Kiểm tra cơ sở dữ liệu dbsa đã tồn tại hay chưa.
  - Nếu đã tồn tại thì xóa bảng cũ sau đó tạo lại database mới.
  - Nếu chưa có thì tạo database mới có tên là ‘dbsa’.
- Tạo database mới: Mở MySQL Administrator → Catalogs→ Create New Schema (hoặc Ctrl+N) → Nhập tên database, như hình 1.



Hình 1- Tạo database “DBSA” trong MySQL

Bước 3: Restore lại database dblp từ file script vào database dbsa.

Mở MySQL Administrator → Open Backup file (Chọn file cơ sở dữ liệu DBLP mới vừa tải về). → Nhập và chọn các thông số như hình 2 → Start restore.



Hình 2 – Restore database 'dbsa' từ tập tin dblp_databse.sql

Bước 4: Thêm các bảng mới vào cơ sở dữ liệu "dbsa"

Sau khi đã tạo được database "dbsa" trong cơ sở dữ liệu MySQL, tiếp theo chúng ta tiến hành chỉnh sửa database cho phù hợp với hệ thống. Thêm 3 bảng mới vào cơ sở dữ liệu: **dbsa_pub, dbsa_pub_in_dblp**, **dbsa_sbj**. Sửa tên cột **id** trong bảng **dblp_author_ref_new** thành **pub_id.**

- Thêm bảng **dbsa_pub:**

```
CREATE TABLE `dbsa_pub` (
  `id` int(8) NOT NULL AUTO_INCREMENT COMMENT 'Id cua bai
bao duoc thu thap ve tu he thong DBSA',
  `sbj_id` int(8) DEFAULT NULL COMMENT 'Id cua tua de bai
bao sau khi phan loai',
  `astract` longtext COMMENT 'Tom tat cua bai bao',
  `title` longtext COMMENT 'Tua de cua bai bao',
```

```
  `year` int(4) unsigned DEFAULT NULL COMMENT 'Nam xuat ban
cua bai bao',
  `publisher` varchar(255) DEFAULT NULL COMMENT 'Nha xuat
ban tai lieu',
  `authors` varchar(250) DEFAULT NULL COMMENT 'Ten cac tac
gia cua bai bao',
  `links` longtext COMMENT 'Cac duong dan mo rong cua bai
bao',
  UNIQUE KEY `id` (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

- Thêm bảng **dbsa_pub_in_dblp:**

```
CREATE TABLE `dbsa_pub_in_dblp` (
  `id` int(8) DEFAULT NULL COMMENT 'id cua bai bao trong du l
ieu dblp',
  `sbj_id` int(8) DEFAULT NULL COMMENT 'id cua bang chu de ba
i bao',
  `link` varchar(250) DEFAULT NULL COMMENT 'Cac lien ket mo r
ong cua bai bao'
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

- Thêm bảng **dbsa_sbj:**

```
CREATE TABLE `dbsa_sbj` (
  `id` int(8) unsigned NOT NULL AUTO_INCREMENT COMMENT 'Id cu
a chu de bai bao',
  `subj_name` varchar(250) DEFAULT NULL COMMENT 'Ten cua chu
de bai bao',
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='Luu thong tin
 chu de bai bao'
```

Chú ý: Khi đã có dữ liệu thì file backup cũng sẽ có dữ liệu.

- Chỉnh sửa tên cột 'id' trong bảng **dblp_author_ref_new**:

```
alter table dblp_author_ref_new change id pub_id int(8)
unsigned;
```

Hoàn tất quá trình tạo database cho chương trình.

2. Các bước backup dữ liệu của chương trình:

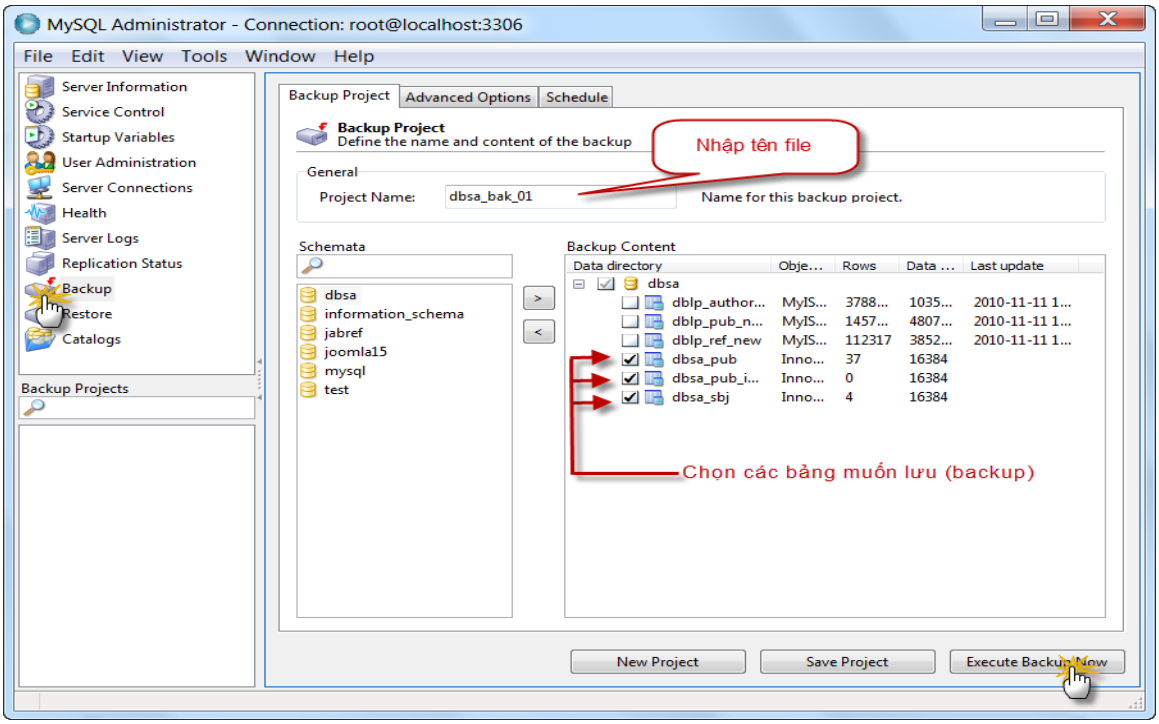
Khi muốn cập nhật database mới nhất của DBLP, việc dầu tiên là phải lưu các bảng hiện có của hệ thống đang có để tránh việc mất mát dữ liệu. Các bảng cần phải lưu (backup) là:

- **dbsa_pub**
- **dbsa_pub_in_dblp**
- **dbsa_sbj**

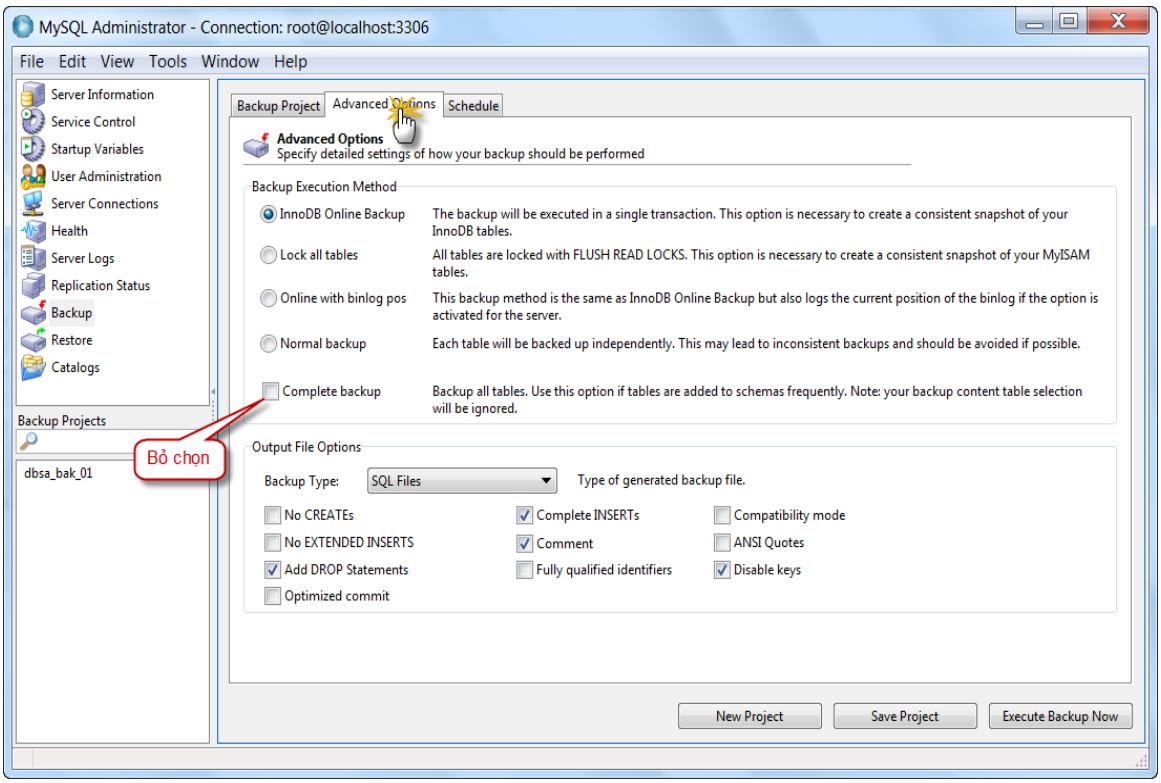
Sau đây là hướng dẫn các bước thực hiện lưu các bảng trên:

Bước 1:

Mở My Administrator → chọn Backup → chọn New Project → Nhập các thông số như hình 3, chọn cơ sở dữ liệu là dbsa, tên tập tin muốn lưu, và nhấn chọn 3 bảng: **dbsa_pub, dbsa_pub_in_dblp**, **dbsa_sbj**. → Nhấn Execute backup now.



Hình 3 – Lựa chọn backup dữ liệu (1)



Hình 4 – Lựa chọn backup dữ liệu (2)

Lưu ý: Mặc định MySQL sẽ chọn lưu tất cả các bảng có trong database. Để lưu theo lựa chọn ở trên phải hủy chọn "complete backup" như hình 4.

Bước 3: phục hồi dữ liệu đã backup.

Đầu tiên chọn file backup từ máy đã được backup như ở bước trước, chọn đúng các thông số như ở hình 2. Sau đó nhấn "Start restore" để bắt đầu quá trình phục hồi dữ liệu.

Như vậy chúng ta đã hoàn thành quá trình tạo cơ sở dữ liệu cho hệ thống xây dựng dữ liệu chỉ mục sử dụng Webcrawler.

PHỤ LỤC B: HƯỚNG DẪN SỬ DỤNG CHƯƠNG TRÌNH.

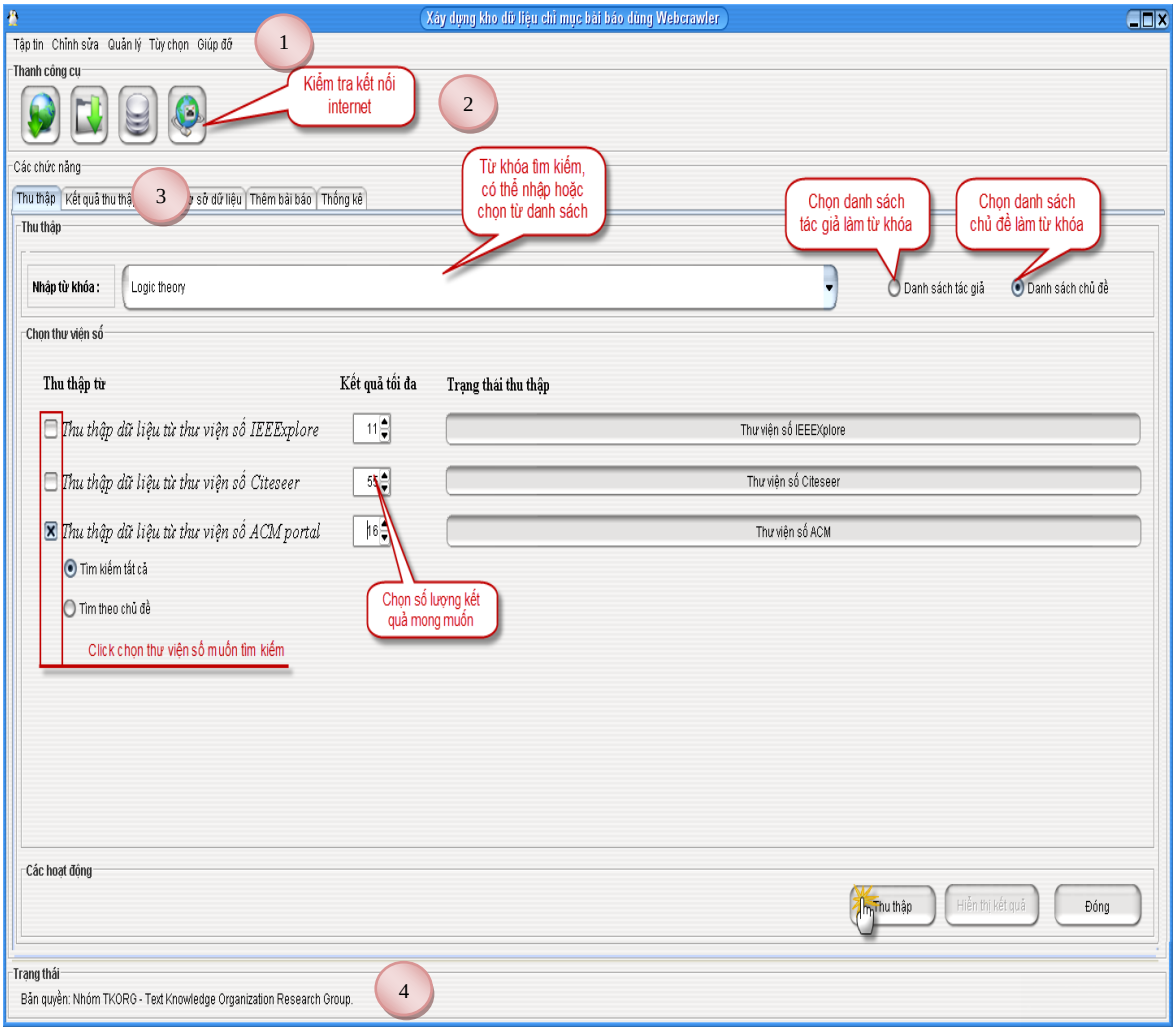
1. Giao diện chức năng thu thập thông tin Metadata từ các thư viện số.

Đây là giao diện đầu tiên khi người sử dụng khởi động chương trình.

Các thành phần trong giao diện như hình 1 gồm:

- Thanh trình đơn (1). (Menu bar)
- Thanh công cụ (2) (Tool bar)
- Các tab chức năng của chương trình (3)
- Thanh trang thái, hướng dẫn (4) (Status bar)

Các chức năng chính của hệ thống nằm trong phần các tab chức năng.

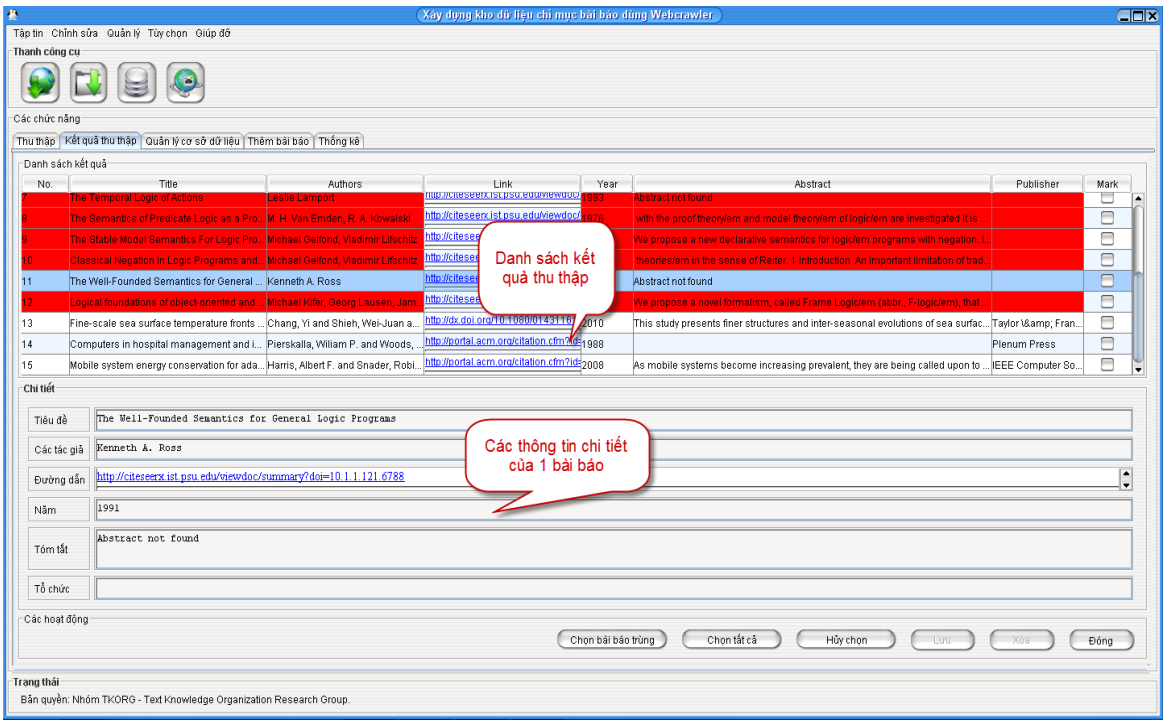


Hình 1 – Giao diện chính của hệ thống

2. Mô tả chức năng thu thập

Như hình 1, các thành phần và xử lý chính của chức năng này như sau:

- Phần từ khóa: Có thể nhập một từ khóa mới hoặc chọn từ khóa theo gợi ý của hệ thống. Nhấn vào nút danh sách tác giả hoặc danh sách chủ đề để thay đổi danh sách từ khóa gợi ý.
- Phần lựa chọn các thông số đầu vào cho quá trình: chọn thư viện số muốn thu thập, số lượng kết quả trả về ứng với mỗi thư viện số đó.
- Phần thứ ba chứa các sự kiện nhấn nút "Thu thập" để bắt đầu quá trình thu thập, sau khi thu thập xong thì có thể nhấn nút "Hiển thị kết quả" để chuyển sang tab kết quả.

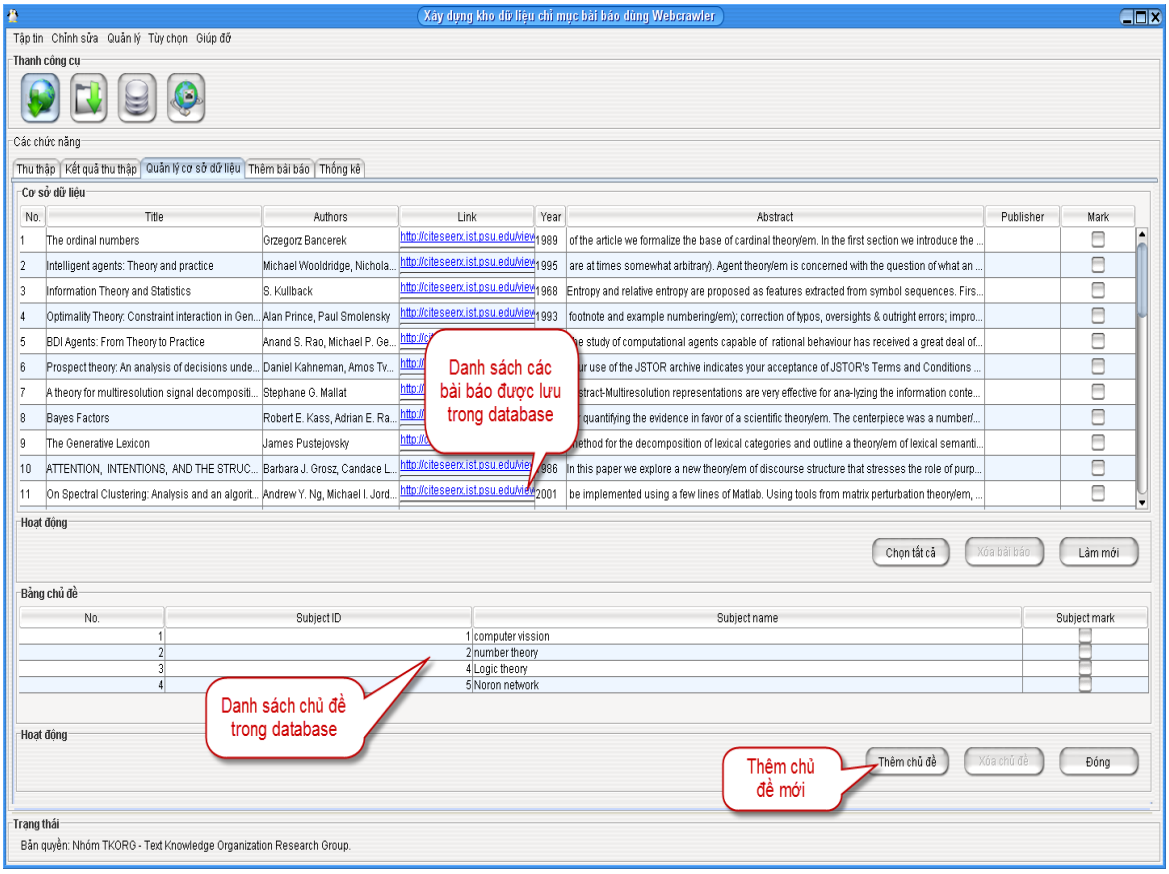


Hình 2 – Tab kết quả thu thập

- Sau khi kết thúc quá trình thu thập tab kết quả thu thập sẽ hiển thị danh sách kết quả như hình 2. Các bài báo bị tô màu là bài báo bị trùng với bài báo trong cơ sở dữ liệu. Khi chọn xem một bài báo, thông tin bài báo đó sẽ hiển thị ở phần thông tin chi tiết. Người dùng có thể nhấn vào đường dẫn liên kết để truy cập trực tiếp bài báo đó trên thư viện số.

- Ở phần này người dùng có thể chọn các bài báo bằng cách nhấn vào các nút ở cột bên phải bảng danh sách kết quả. Hoặc nhấn nút "chọn tất cả", "chọn tất cả các bài trùng" để chọn nhanh. Sau đó người dùng có thể lưu hoặc xóa các bài đã chọn.

3. Chức năng quản lý cơ sở dữ liệu



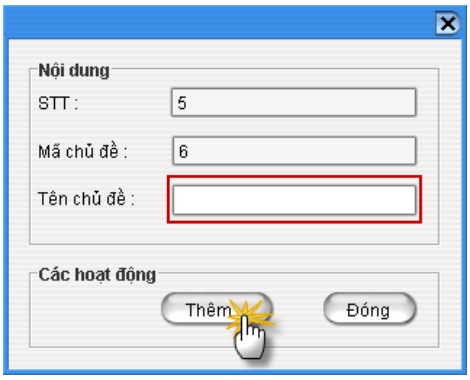
Hình 5 – Quản lý cơ sở dữ liệu

Đây là chức năng cho phép người sử dụng quản lý cơ sở dữ liệu của mình.

Chức năng này gồm hai thành phần chính như sau:

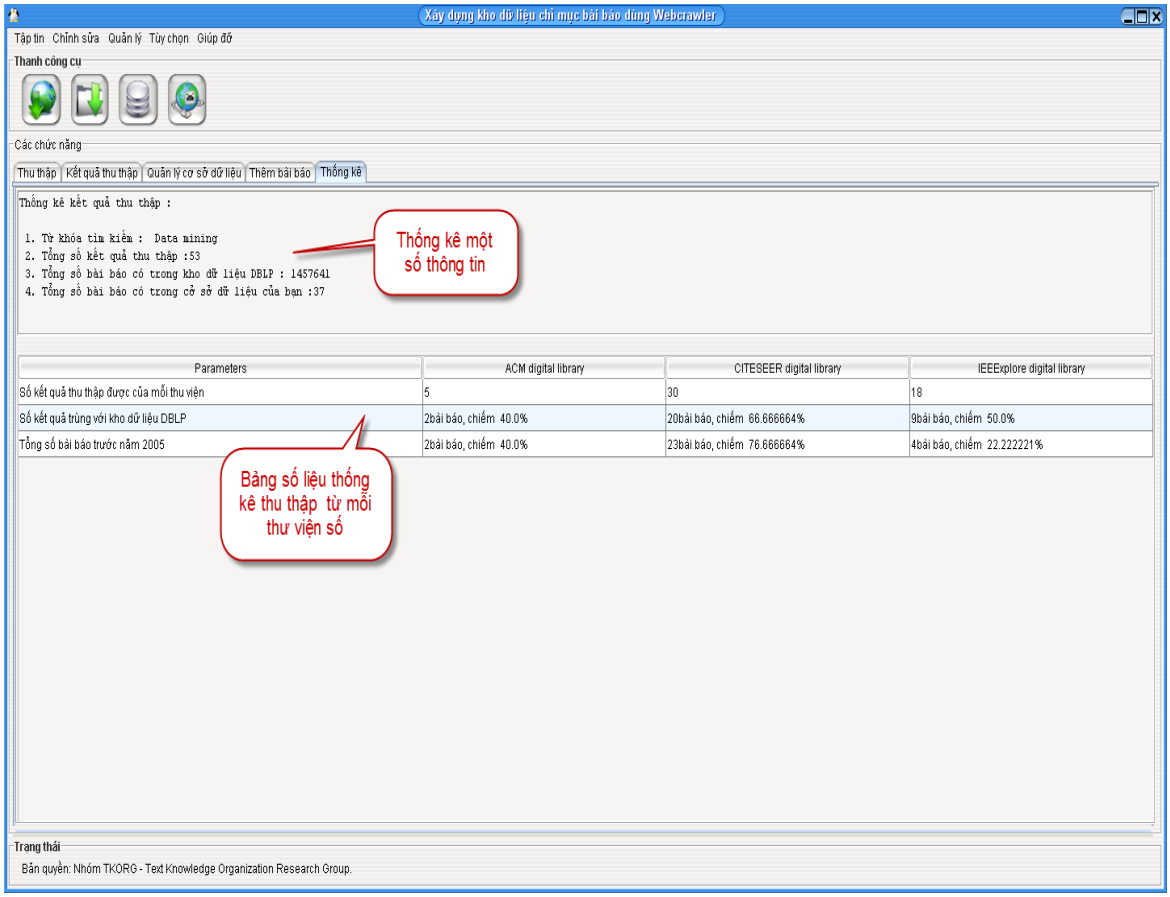
- Quản lý các bài báo khoa học đã được người dùng lưu trong cơ sở dữ liệu. Người dùng có thể xem danh sách bài báo và chọn xóa các bài báo theo ý muốn.
- Quản lý danh sách các chủ đề, tương tự như chức năng trên người dùng có thể xem và xóa các chủ đề theo ý muốn. Ngoài ra người dùng có thể thêm một chủ đề mới bằng cách nhấn vào nút "Thêm chủ đề". Khi nhấn vào nút

này hệ thống sẽ hiển thị lên chức năng thêm chủ đề như hình 6, sau đó người dùng nhập tên chủ đề và nhấn "Thêm mới".



Hình 6 – Thêm chủ đề mới

4. Thông kê kết quả thu thập



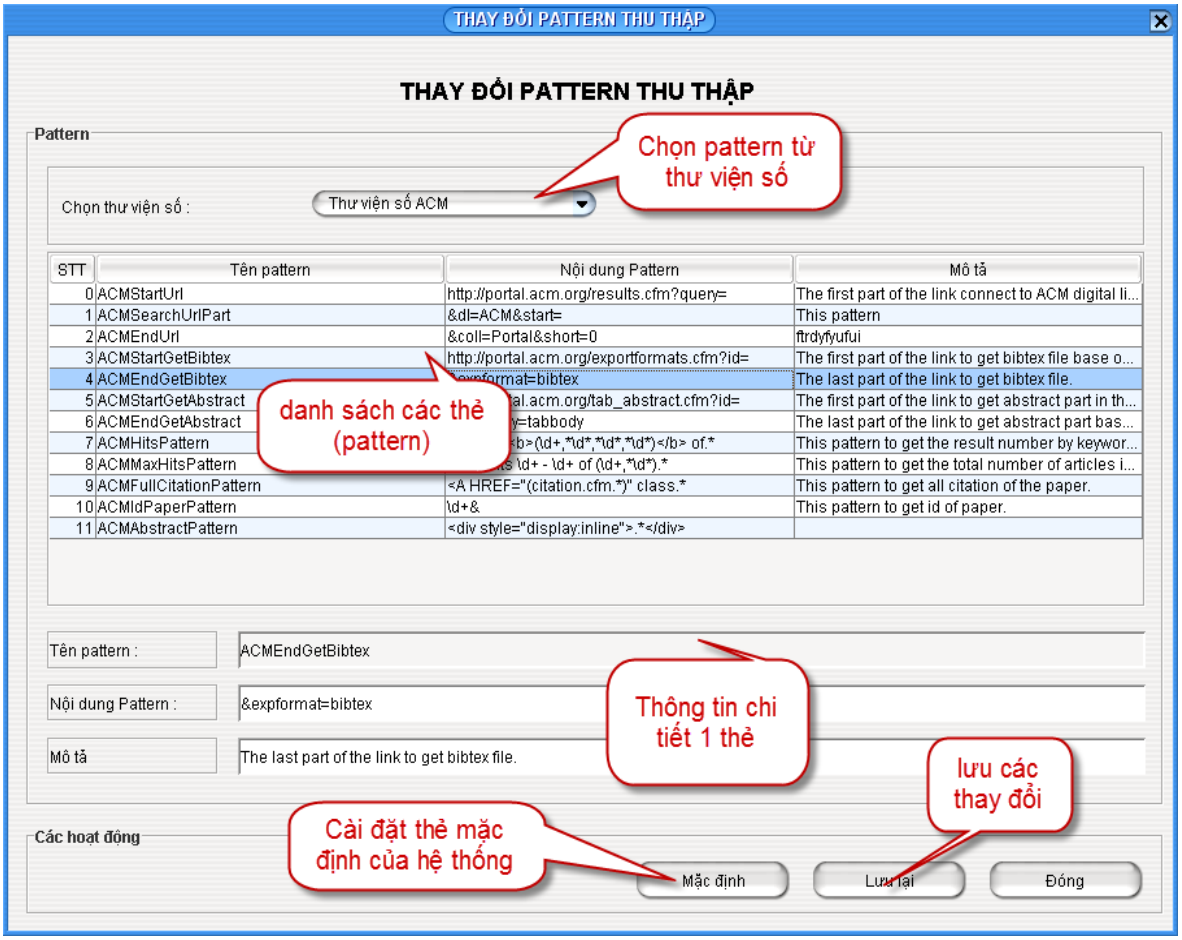
Hình 7- Bảng thống kê kết quả sau khi thu thập hoàn thành

Sau khi chức năng thu thập hoàn tất hệ thống sẽ tự động thống kê các kết quả từ mỗi thư viện số, kiểm tra, so sánh… và đưa ra kết quả như hình 7

5. Chức năng thay đổi các thẻ rút trích (Pattern)

- Chức năng đổi các thẻ (pattern). Đây là chức năng định nghĩa các thẻ để truy xuất các thư viện số và thu thập các thông tin từ các thư viện số. Người dùng có thể chọn một thẻ và thay đổi nội dung sau đó lưu lại, hoặc cài đặt các thẻ theo mặc định của hệ thống như hình 8 đã mô tả rõ.

Lưu ý: *khi thay đổi nội dung các thẻ sẽ dễ dẫn tới hệ thống không hoặc động tốt và làm việc không chính xác.*

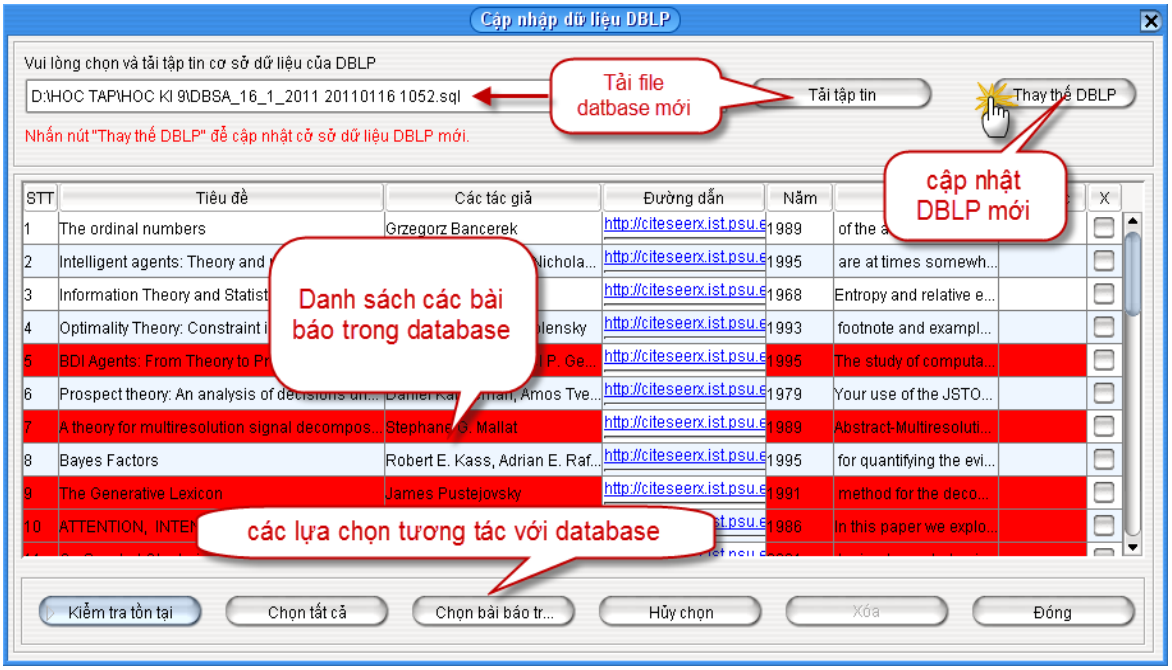


Hình 8 – Chỉnh sửa thông tin các thẻ của các thư viện số(pattern)

6. Chức năng cập nhật cơ sở dữ liệu mới của DBLP

- Để thực hiện chức năng này, đầu tiên bạn phải tải về bộ cơ sở dữ liệu mới nhất của DBLP từ địa chỉ http://dblp.l3s.de/dblp++.php. Sau đó bạn chọn chức năng cập nhật cơ sở dữ liệu DBLP từ thanh trình đơn. Từ giao diện của hệ thống chọn load tập tin

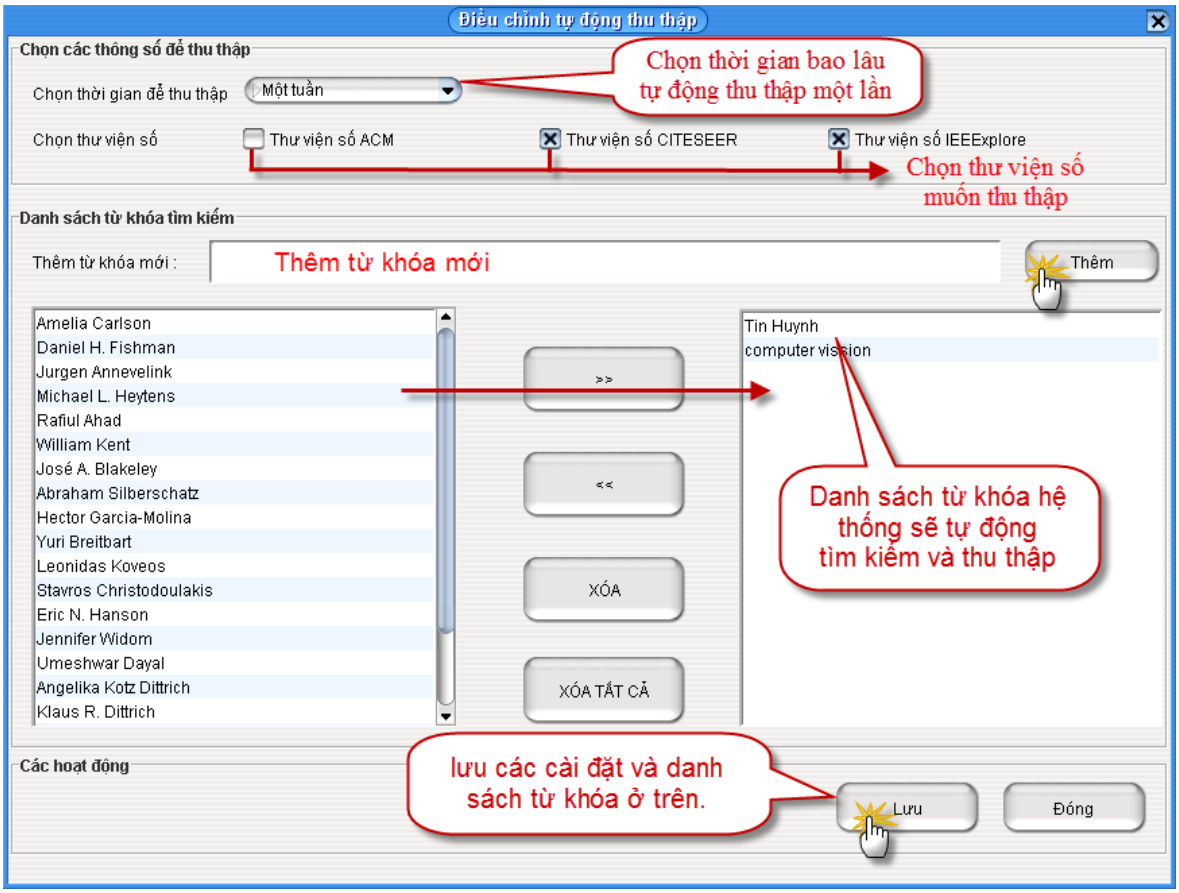
vừa tải về tiếp theo nhấn nút "Thay thế DBLP". Hệ thống sẽ nhận sự kiện và thay thế cơ sở dữ liệu mới thay cho cơ sở dữ liệu cũ. Sau khi cập nhật xong hệ thống sẽ kiểm tra so sánh dữ liệu của DBLP mới với dữ liệu các bài báo khoa học đã được bạn lưu từ trước. Hệ thống sẽ hiển thị ra danh sách các bài báo như hình 9, những bài báo bị trùng sẽ được tô màu. Tại đây cho phép người dùng có thể chọn các bài báo và xóa theo ý muốn.



Hình 9 – Chức năng cập nhật cơ sở dữ liệu DBLP

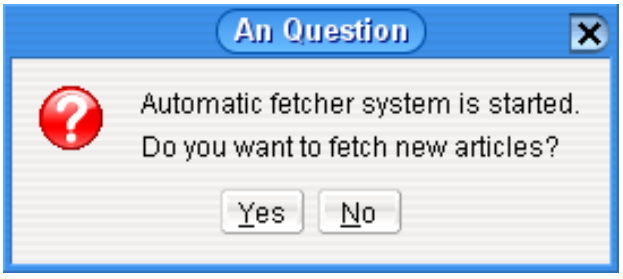
7. Chức năng tự động thu thập dữ liệu:

- Chức năng này cho phép người dùng có thể tự động thu thập dữ liệu theo định kì trên các thư viện và danh sách từ khóa do người dùng chọn như hình 10. Sau khi người dùng nhấn nút lưu thì hệ thống sẽ lưu lại tất cả các thông tin mà người dùng đã chọn. Danh sách từ khóa người dùng có thể thêm mới hoặc chọn các gợi ý của hệ thống ở bên cột trái.



Hình 10 – Cài đặt thông tin cho chức năng tự động thu thập dữ liệu.

- Sau khi cài đặt xong như ở trên, mỗi khi hệ thống được khởi động lên thì sẽ kiểm tra xem đã đến thời gian tự động thu thập hay chưa. Nếu đã đến thời gian tự động theo như cài đặt trước, hệ thống sẽ bật chức năng lên như hình-11. Xác nhận xem người dùng có muốn thu thập không, nếu có thì hệ thống sẽ tự động thu thập tất cả các bài mới, loại bỏ các bài trùng với cơ sở dữ liệu đã có và lưu xuống cơ sở dữ liệu.



Hình 11 – Xác nhận người dùng có muốn tự động thu thập hay không

PHỤC LỤC C: CÁC CHỦ ĐỀ TRONG KHOA HỌC MÁY TÍNH

Các chủ đề trong lĩnh vực khoa học máy tính được tham khảo từ Wikipedia

1. **Theoretical computer science**
    - 1.1 *Mathematical logic*
    - 1.2 *Automata theory*
    - 1.3 *Number theory*
    - 1.4 *Graph theory*
    - 1.5 *Type theory*
    - 1.6 *Category theory*
    - 1.7 *Computational geometry*
    - 1.8 *Quantum computing theory*
2. **Algorithms and data structures**
    - 2.1 Analysis of algorithms
    - 2.2 Algorithms
    - 2.3 Data structures
3. **Computer elements and architecture**
    - 3.1 *Digital logic*
    - 3.2 *Microarchitecture*
    - *3.3 Multiprocessing*

4. Computational science

4.1 Numerical analysis

4.2 Computational physics

4.3 Computational chemistry

4.4 Bioinformatics

5. Artificial Intelligence

5.1 Machine learning

5.2 Computer vision

5.3 Natural language processing/Computational linguistics

5.4 Robotics

5.5 Image Processing

5.6 Pattern Recognition

5.7 Congnitive science

5.8 Evolutionary computation

5.9 Information retrieval

5.10 Knowledge Representation

6. Software Engineering

6.1 Operating systems

6.2 Computer networks

6.3 Databases

6.4 Computer security

6.5 Ubiquitous computing

6.6 Systems architecture

6.7 Compiler design

6.8 Programming languages