CHƯƠNG 1: TỔNG QUAN

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 (Metadata[[1]](#footnote-1)) trong việc đảm bảo thông tin các bài báo đẩy đủ và cập nhật nhất.

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[[2]](#footnote-2), các thư viện số như ACM[[3]](#footnote-3), IEEEXplore[[4]](#footnote-4), CiteSeer[[5]](#footnote-5) … hoặc dữ liệu chỉ mục có sẵn như DBLP[[6]](#footnote-6). Vấn đề đặt ra ở đây là: đối với mỗi thư viện số của tổ chức tương ứng sẽ cập nhật những bài báo mới khi có các hội thảo, hay khi các tạp chí do tổ chức đó xuất bản (ví dụ: thư viện số ACM với các cuộc hội thảo, tạp chí từ tổ chức “Association for Computing Machinery”; thư viện số IEEEXplore tương ứng với các bài báo trong các hội thảo, tạp chí từ tổ chức “Institute of Electrical and Electronics Engineers”) việc cập nhật những bài báo mới từ các tổ chức khác hay việc trao đổi dữ liệu giữa các thư viện số hiện nay là rất hạn chế. Trong khi đó các dữ liệu chỉ mục hiện tại như DBLP, hay hệ thống đi thu thập dữ liệu chỉ mục như ACI [2] 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 trích thông tin chỉ mục từ các tài liệu download được chưa đạt được độ chính xác cao.

Xuất phát từ vấn đề trên, chúng tôi phát triển một hệ thống dùng để xây dựng dữ liệu chỉ mục các bài báo khoa học bằng việc kết hợp giữa rút trích thông tin bài báo trực tiếp từ các thư viện số, kết hợp với sử dụng dữ liệu chỉ mục có sẵn để dữ liệu xây dựng được 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 thông trích 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 từ đó 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.

2. Mục tiêu và phạm vi khóa luận.

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.

- Hệ thống có khả năng thể tự động cập nhật những bài báo mới nhất từ các thư viện số.

2.2 Phạm vi khóa luận.

- Hệ thống sử dụng Web Crawler để thu thập thông tin chỉ mục các bài báo khoa học 3 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.

3. Cấu trúc khóa luận

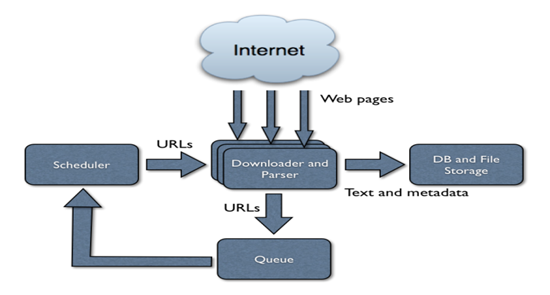
CHƯƠNG 2 CÁC NGHIÊN CỨU VÀ HỆ THỐNG LIÊN QUAN

CƠ SỞ LÝ THUYẾT.

1.1 Web crawler.

Web crawler[[7]](#footnote-7), web spider hay web robot là 1 chương trình hoặc các đoạn mã có khả năng tự động duyệt các trang web khác theo 1 phương thức một cách thức tự động, web crawler thường được sử dụng để thu thập tài nguyên (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 cập nhập kho dữ liệu phục vụ tìm kiếm, giúp người dùng có thể tìm kiếm thông tin nhanh hơn.



Hình 1- Kiến trúc Web Crawler (Wikiperdia)

Web crawler bắt đầu từ danh sách các địa chỉ URL được gọi là seeds (hạt giống), hệ thống sẽ vào địa chỉ này lọc thông tin rồi tìm ra các địa chỉ URL khác (các URL có trong nội dung của trang seeds) sau đó thêm chúng vào danh sách các địa chỉ đã duyệt qua gọi là crawl frontier. Sau đó hệ thống lặp lại quá trình trước đó để duyệt qua những URL mới. Cứ thế Crawling qua rất nhiều địa chỉ Website và thu thập rất nhiều nội dung khác nhau từ địa chỉ thu thập đươc.

Trong hệ thống của chúng tôi, Web crawler được sử dụng để thu thập các đường dẫn chứa nội dung 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 metadata của bài báo bằng cách sử dụng các trình phân tích kết hợp với luật định nghĩa trước.

3.2 Metadata.

3.2.1 Khái niệm Metadata

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[[8]](#footnote-8) 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[[9]](#footnote-9) 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.

3.2.2 Chuẩn Dublin Core Metadata

Dublin Core Metadata[[10]](#footnote-10) 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. Dublin Core Metadata lần đầu tiên được 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 [12], 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 [12] được mô tả chi tiết trong bảng bên dưới

|  |  |  |
| --- | --- | --- |
| 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 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:

- 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.

3.3 Bibtex.

BibTeX[[11]](#footnote-11) do Oren Patashnik và Leslie Lamport tạo ra năm 1985. BibTeX 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, … 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 từ bài giới thiệu tại Document Freedom Day).

@Book{abramowitz+stegun,

author = "Milton {Abramowitz} and Irene A. {Stegun}",

title = "Handbook of Mathematical Functions with

Formulas, Graphs, and Mathematical Tables",

publisher = "Dover",

year = 1964,

address = "New York",

edition = "ninth Dover printing, tenth GPO printing"

}

Hình 2- Ví dụ cấu trúc của file BibteX (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 “,”.

Trên các thư viện số ACM và IEEEXplore, 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.

Các hệ thống liên quan.

Theo khảo sát được giới thiệu trong các bài báo [3,7] cũng như tìm hiểu của nhóm, hiện nay có một số nguồn mà từ đó có thể xây dựng dữ liệu chỉ mục bài báo bao gồm: xây dựng dữ liệu chỉ mục từ các file đề mục (tables of contents– TOCs của các cuộc hội thảo) như hệ thống DBLP [8]; rút trích từ nội dung bài báo (sử dụng các file postscript hoặc file PDF của các bài báo được công bố trên mạng) như hệ thống được giới thiệu trong các bài báo [2,9] hay các hệ thống sử dụng crawler để thu thập thông tin các bài báo trên web [4,11,12].

Trong phần 4 này, nhóm trình bày một số hệ thống dữ liệu chỉ mục đã tồn tại đồng thời khảo sát các thư viện số mà hệ thống sẽ dựa trên đó để lấy thông tin các bài báo.

4.1 Digital Bibliography & Library Project (DBLP).

DBLP cung cấp thông tin về chỉ mục các bài báo trong lĩnh vực khoa học máy tính, hệ thống được phát triển bởi trường đại học [Universität Trier](http://en.wikipedia.org/wiki/University_of_Trier) của Đức. 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 web 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 (tables of contents– TOCs). Các file TOCs được tác giả sưu tầm từ các hội nghị, tạp chí.



Hình 3- Hệ thống xây dựng cơ sở dữ liệu DBLP

Từ đầu vào là 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 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.

Hiện nay có một số ứng dụng được xây dựng trên nguồn dữ liệu của DBLP cung cấp chức năng cho phép người dùng tìm kiếm bài báo, trong đó có các hệ thống: Complete Search DBLP, [Faceted search](http://dblp.l3s.de) và DBL – Browser.

CompleteSearch DBLP[[12]](#footnote-12).

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 [5], bài báo có thể được tìm kiếm theo các trường 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 4- Hệ thống Complete Search

[Faceted search](http://dblp.l3s.de)[[13]](#footnote-13):

Hệ thống tìm kiếm bài báo trên dữ liệu của DBLP được giới thiệu trong [6], 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 5- Hệ thống FacetedDBLP

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



Hình 6 - Duyệt bài báo trong FacetedDBLP

DBL – Browser:

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



Hình 7 - Chương trình DBL Brower

🡺 Như vậy dữ liệu chỉ mục DBLP thu thập bằng cách rút trích thông tin chỉ mục từ những file TOCs của các hội nghị, tạp chí được các tổ chức gửi về hoặc tác giả DBLP sưu tầm được. Với phương pháp này dữ liệu không đảm bảo được tính cập nhật và đầy đủ của dữ liệu để minh chứng cho điều này nhóm 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 1 là kết quả được tính trung bình của 100 bài báo đầu tiên trên 3 thư viện số ACM, Citeseer, IEEEXplore sau khi tìm kiếm với 2 từ khóa “Database” và “Data mining”. Dựa vào bảng 2 ta thấy dữ liệu của DBLP chưa đảm bảo được tính đầy đủ và cập nhật dữ liệu.

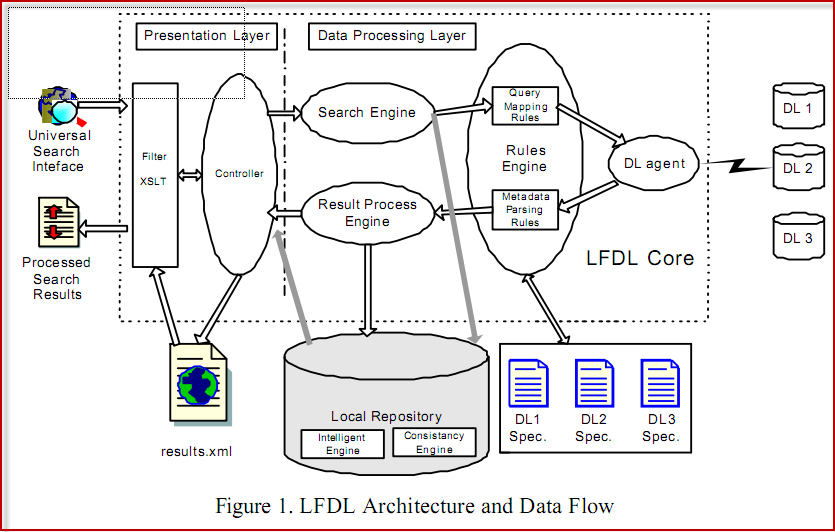
|  |  |  |  |
| --- | --- | --- | --- |
| Từ khóa tìm kiếm | Phần trăm dữ liệu không tồn tại trong DBLP  (%) | Phần trăm Dữ liệu trước năm 2010 không tồn tại trong DBLP  (%) | Phần trăm dữ liệu trong năm 2010 không tồn tại trong DBLP  (%) |
| Database | 28,33 | 86,26 | 10,71 |
| Data mining | 43,67 |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

Bảng 2: Khảo sát tính cập nhật dữ liệu của DBLP

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 cho DBLP.

4.2 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 [10], trong 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 khà năng tổng hợp thông tin từ các thư viện số thành một khối dữ liệu có cấu trúc đồng nhất. Trong hình 8 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.



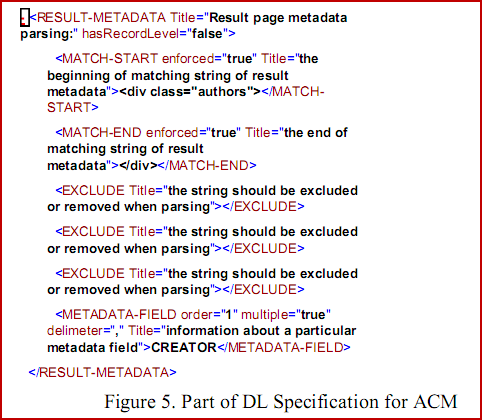
Hình 8 – Kiến trúc LFDL

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

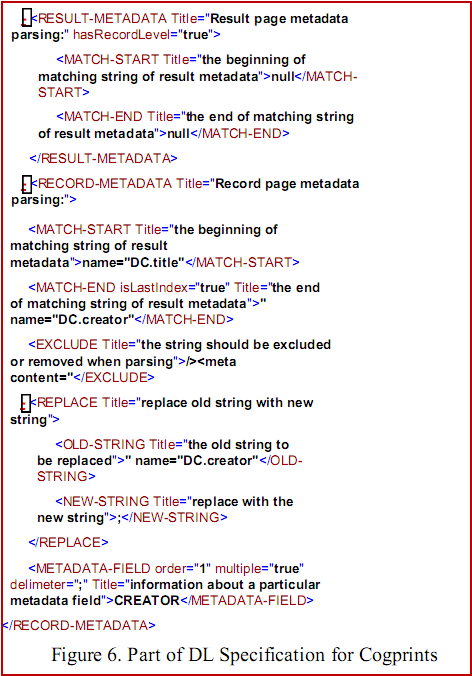
Đầu tiên khi khởi tạo hệ thống sẽ đọc tất cả các đặc tả (như ở hình 9, 10) của thư viện số bao gồm: luật liên kết truy vấn và luật rút trích metadata. Các đặc tả ở đây được hệ thống rút ra trên thư viện số thông qua quá trình phân tích cấu trúc hiển thị thông tin bài báo cho người dùng (phân tích nội 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, 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 đặt 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.



Hình 9 – Phần đặt tả cho thư viện số ACM portal



Hình 10 – Phần đặt tả cho thư viện số Cogprints

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. Thông tin metadata đã rút ra sẽ được lưu vào cơ sở dữ liệu ở máy local.

Để hiển thị thông tin bài báo thu thập được cho người dùng, hệ thống còn lưu thông tin bài báo dưới dạng XML. Khi cần hiển thị, hệ thống sẽ sử dụng bộ xử lý XSLT[[14]](#footnote-14) 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 đánh giá trong bài báo [10], hiện hệ thống mới thu thập được tựa đề (title) và đường dẫn (hyperlink) của bài báo, cùng với đó hiệu năng của hệ thống chưa được cao. Dữ liệu của hệ thống chưa tận dụng được các dữ liệu chỉ mục có sẵn để đảm bảo tính đầy đủ cho thông tin các bài báo.

4.3 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 [2].Hệ thống sử dụng các web search engines (như Alta vista, Hotbot, Excite) kết hợp với các sử dụng các thuật toán heuristic để tìm kiếm và download các bài báo dưới định dạng file PDF hoặc PostScript từ trên Web về. Hệ thống sẽ chuyển các bài báo download được sang file text sau đó sử dụng một chương trình (PreScript) 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 nhận diện, rút trích các thành phần chính của bài báo bằng cách sử dụng các thuật toán heuristic, các luật, kết hợp với việc sử dụng máy học để rút các trường thông tin của bài báo.Các trường thông tin 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.

Từ nội dung rút trích được hệ thống sẽ tổ chức thành cơ sở dữ liệu chỉ mục và hiện nay dữ liệu này dùng để tra cứu trên hệ thống thư viện số CiteSeer.

🡺 Như vậy với việc sử dụng các search engine để đi tìm kiếm và download các bài báo trên Internet thì hệ thống đã sử dụng được nguồn dữ liệu khồng 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 việc rút trích thông tin chỉ mục bài báo từ các các định dạng khác nhau độ chính xác vẫn chưa cao và vẫn là một vấn đề lớn 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.

4.4 Các thư viện số ACM, CiteSeer, IEEEXplore.

4.4.1 Thư viện số ACM.

ACM (Association for Computing Machinery) là một tổ chức họat động trong lĩnh đào tạo và nghiên cứu khoa học trong lĩnh vực máy tính, ACM cung cấp một thư viện số ACM Portal [[15]](#footnote-15) cho phép người dùng tìm kiếm các bài báo được công bố trong hội nghị, tạp chí được tổ chức và xuất bản bởi ACM cũng như một số tổ chức khác.



Hình 11 – Thư viện số ACm

Theo thông tin được công bố trên trang trang chủ của ACM, tính đến tháng 1 năm 2011 dữ liệu của ACM chứa thông tin khoảng 1,6 triệu bài báo trong nhiều lĩnh vực của công nghệ thông tin. Hệ thống có giới hạn trong việc cho phép người dùng download tài liệu.

4.4.2 Thư viện số 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.

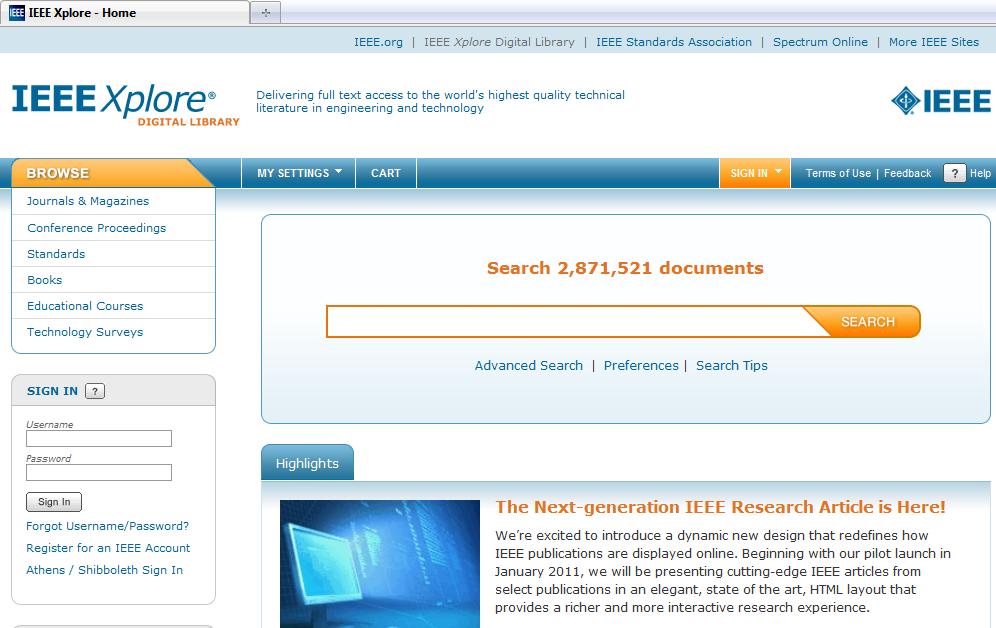


Hình 12 – Thư viện số CiteSeer

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.

4.4.3 Thư viện số IEEEXplore.

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



Hình 13 – Thư viện số IEEEXplore

Theo thông tin từ trang chủ của hệ thống, tính đến tháng 1 năm 2011 dữ liệu của hệ thống chứa thông tin của khoảng 2,8 triệu bài báo từ khác hội nghị cũng như các tạp chí. Hệ thống giới hạn việc download và xem dữ liệu chỉ mục đối với người dùng chưa có tài khoản từ hệ thống.

CHƯƠNG 2: XÂY DỰNG DỮ LIỆU CHỈ MỤC SỬ DỤNG WEB CRAWLER.

Phương pháp Crawler trên thư viện số.

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

Phần này sẽ trình bày cách thức mà hệ thống thu thập bài báo khoa học từ thư viện số ACM. Đầu vào của quá trình thu thập là từ khóa được người dùng nhập vào hoặc chọn từ danh sách tên tác giả hay chủ đề có trong hệ thống. Bằng việc sử dụng các thẻ (pattern) đã định nghĩa sẵn kết hợp với từ khóa vừa được nhập vào hệ thống sẽ tiến hành tạo câu truy vấn trên thư viện số ACM và tiến hành Crawl lấy các đường dẫn tới các bài báo được trả về từ thư viện số.

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>= | Hai pattern này để tìm ra ID của mỗi bài báo, từ đó lấy file bibtex của bài báo này. |
| 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<b>(\d+,\*\d\*)</b> of.\* | Tổng số lượng bài báo tìm được từ thư viện số. |
| 10 | .\*Results \d+ - \d+ of (\d+,\*\d\*).\* | Số kết quả trong một trang |
| 11 | (exportformats.cfm[.]+bibtex) |  |
| 12 | <A HREF="(citation.cfm.\*)" class.\* |  |
| 13 | (exportformats[.]cfm.+bibtex) |  |
| 14 | \d+& |  |

Bảng 3 - 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 nói 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 dựa vào từ khóa kết hợp một số thẻ đã được định nghĩa trong bảng 3.

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.

Ví dụ 1: 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](http://portal.acm.org/results.cfm?query=computer20%25vision&dl=ACM&coll=Portal&short=0)

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

Bước 2:

Gửi URL vừa tạo được lên trình duyệt Web.

Lấy nội dung trang Web mà trình duyệt trả về sau khi truy cập vào địa chỉ URL trên.

Dựa vào các pattern để phân tích và rút ra một số thông tin cần thiết từ trang web này. Một số thông tin như tổng số kết quả được tìm thấy, số kết quả trong 1 trang, Crawl danh sách dường dẫn đến thông tin chi tiết mỗi bài báo cụ thể trong trang này.

🡪 Mục tiêu: Thu thập danh sách các địa chỉ URL của các bài báo khoa học trong địa chỉ URL ở bước 1được trả về từ thư viện số.

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 về thông tin metadata của mỗi bài báo.

🡪 Mục tiêu: Lấy về nội dung chi tiết của một trang web từ một địa chỉ URL.

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 để tìm ID của mỗi bài báo sau đó lấy về file bibtex của bài báo đó theo ID của bài báo.

Sau khi thu thập được file Bibtex của một bài báo, 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 thì 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 để tìm kiếm trong nội dung trang web thu được ở bước 2.

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.

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

Tương tự như chức năng thu thập các bài báo khoa học từ thư viện số ACM như trình bày ở phần trên, hệ thống sẽ tạo câu truy vấn (tạo URL query) dựa vào từ khóa và các pattern đã được định nghĩa sẵn ở bảng 4 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 IEEE để thu thập các bài báo. |
| 2 | &rowsPerPage= | 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. |
| &pageNumber= |
| 3 | ([0-9,]+) results | Tổng số kết quả tìm được |
| 4 | <span class=\"type\">\s\*(.+) | Kiểu của tài liệu |
| 5 | .\*<strong>(.+)</strong><br>"+ "\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?(.\*) | Proceedings pattern |
| 8 | <a href=".\*arnumber=(\d+).\*"> | Mã số của bài báo |
| 9 | <a\\s\*href=[^<]+>\\s\*(.+)\\s\*</a> | Pattern lấy tiêu đề của bài báo |
| 10 | <p>[\\s+(.+)](file:///\\s+(.+)) | Pattern lấy các đồng tác giả |
| 11 | <a href=\"(.+)\" class=\"bodyCopySpaced\">Abstract</a> | Các pattern dùng để lấy phần tóm tắt của 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\*)](file:///\\d+)\\s*-\\s*(\\d*)) | Số trang của bài báo |
| 17 | Digital Object Identifier:\\s\*<a href=.\*>(.+)</a> | Pattern lấy số DOI của một bài báo |

Bảng 4 - 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 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”. [http://ieeexplore.ieee.org/search/freesearchresult.jsp?queryText= + computer + vision](http://ieeexplore.ieee.org/search/freesearchresult.jsp?queryText=%20+%20computer + vision) + &rowsPerPage= + 50 + &pageNumber= + Câu truy vấn được tạo ra với từ khóa “computer vision” 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 có thể hiển thị là 50. Và “&pageNumber=1” là kết quả tìm kiếm và hiển thị là ở trang 1, như ở hình 14.

Mục tiêu: Tạo ra câu truy vấn để thu thập các bài báo từ thư viện số IEEE.



Hình 14 – Kết quả tìm kiếm từ thư viện số IEEEXplore

Bước 2: Sử dụng câu truy vấn 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 ở bước 1. Hệ thống sử dụng các pattern đã được định nghĩa ở bảng 2 để để Crawl thu thâp 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 bài báo.

Bước 3 : Tương tự bước 3 của cách thu thập từ thư viện số ACM.

Mục tiêu: truy cập vào địa chỉ URL của một bài báo cụ thể lấy về 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.

Sử dụng các pattern từ số 6 đến 17 trong bảng 4, và bộ phân tích cú pháp để tìm thông tin của một 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.

1.3 Cách thức thu thập các bài báo khoa học từ thư viện số CiteSeer.

Đối với thư viện số CiteSeer, 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 tìm kiếm được.

Trong thư viện số CiteSeer kết quả này trả về là một đường dẫb liên kết chứa thông tin bài báo dưới dạng XML được gọi là ATOM Link [[16]](#footnote-16) 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"

type="text/html"

href="[**http://www.xml.com/pub/a/2004/05/19/deviant.html**](http://www.xml.com/pub/a/2004/05/19/deviant.html)"/>

. . .

Hình 15 - Cấu trúc file HTML 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 của từng bài báo .

SAXPaser [[17]](#footnote-17) – 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.

Phương thức báo hiệu cho paser 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 .

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

endDocument(): kết thúc phân tích một 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), năm xuất bản (year), tóm tắt (abstract), đường dẫn tới bài báo (links). Những thông tin này sẽ được bộ phân tích lấy lần từ trên xuống giúp hệ thống lấy được thông tin bài báo.  
 startElement(): bắt đầu phân tích một thể hiện của tài liệu XML

endElement() :kết thúc phân tích một thể hiện của 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 Citeseerx.

Bibtex Paser.

Như đã trình bày ở phần 3.3 trong chương tổng quan, Bibtex là định dạng kiểu cấu trúc dùng để biểu diễn thông tin của tài liệu. Trong các thư viện số, các file Bibex lưu thông tin metadata của bài báo. Hệ thống sử dụng Bibtex parser trong chương trình Jabref[[18]](#footnote-18) để thực hiện việc phân tích các file Bibtex thu được để lấy thông tin.

Sau đây là các dạng file Bibtex:

|  |  |  |  |
| --- | --- | --- | --- |
| 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](http://en.wikipedia.org/wiki/Scribe_%28markup_language%29)[[19]](#footnote-19) | 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 | Kỹ yếu của hội nghị | title, year | editor, ublisher, organization, address, month, note, key |
| 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 tựa đề và tên tác giả nhưng chưa xuất bản. | author, title, note | month, year, key |

Bảng 5 - Những kiểu file Bibtex được tham khảo từ Wikipedia

Quá trình phân tích file Bibtex thực thi khi hệ thống rút được file từ thư viện số. Sau khi bằng trình phân tích xác định được file bibtex chứa thông tin của kiểu tài liệu nào, trình phân tích sẽ sử dụng luật kết hợp với dữ liệu các trường chứa thông tin của file Bibtex để lấy các dữ liệu chứa trong file.

Kiểm tra dữ liệu trùng lặp.

Sau đây nhóm 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 việc dữ liệu thu thập được không bị trùng lặp với các dữ liệu đã có trong hệ thống cũng như trong dữ liệu chỉ mục có sẵn trong DBLP thì hệ thống sẽ tiến hành kiểm tra tính tồn tại của bài báo. 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 thông tin sau để kiểm tra sự trùng lặp dữ liệu có đã có trong database:

+ Tựa đề bài báo (title). Trong dữ liệu có sẵn có trường titleSingnatere, đây là tựa đề của bài báo sau khi bỏ khoảng trắng và một số ký tự đặc biệt như “!?,.”. Bài báo sau khi rút được tựa đề hệ thống sẽ bỏ khoảng trắng và các ký tự đặc biệt và so sánh với trường titleSingnatere có trong hệ thống để kiểm tra xem bài báo có trong hệ thống chưa.

+ Năm công bố bài báo.

+ Tên tác giả bài báo.

Dữ liệu của DBLP được đưa vào hệ thống nhờ module import, từ dữ liệu đưa vào kết hợp với các thông tin mà bài báo vừa thu thập được hệ thống sẽ tiến hành kiểm tra lần lượt các trường được chọn phía trên để kiểm tra tính tổn tại của bài báo.

Nếu dữ liệu được đã tồn tại - ba trường trên của bài báo được thu thập giống với dữ liệu trong hệ thống thì chương trình sẽ đưa ra thông báo trên giao diện cho người dùng biết.



Hình 15 - Luồng xử lý dữ liệu trùng lặ

CHƯƠNG 3: HIỆN THỰC HỆ THỐNG.

Kiến trúc hệ thống.



Hình 16 – Kiến trúc hệ thống

Trong hình 16, chúng tôi giới thiệu kiến trúc của ứng dụng, dữ liệu đầu vào của hệ thống là từ khóa được nhập vào từ người dùng hoặc chọn từ danh sách các chủ đề trong lĩnh vực khoa học máy tính được lấy từ Wikipedia hoặc 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 link tới các bài báo phù hợp với từ khóa tìm kiếm, module rút trích thông tin bài báo sẽ sử dụng các trình phân tích kết hợp luật 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 lắp 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.

Thiết kế cơ sở dữ liệu.

2.1 Mô tả cấu trúc dữ liệu của DBLP

Sau đây là cấu trúc bảng SQL của DBLP được Tiến sĩ [Jörg Diederich](http://www.l3s.de/%7Ediederich) 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 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 cótrong bảng dblp\_pub\_new.

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 17 – Các bảng trongcơ sở dữ liệu 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. | |  | 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 xuât 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 6 - 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 7 - 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 8 – Thông tin cấu trúc bảng dblp\_ref

2.2 Cơ sở dữ liệu hệ thống thống.

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

Từ cấu trúc này nhóm bổ sung thêm vào cấu trúc những bảng sau để đảm bảo việc có thể cập nhật được dữ liệu mới của DBLP và có thể lưu được các thông tin về chủ đề, cũng như 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. |

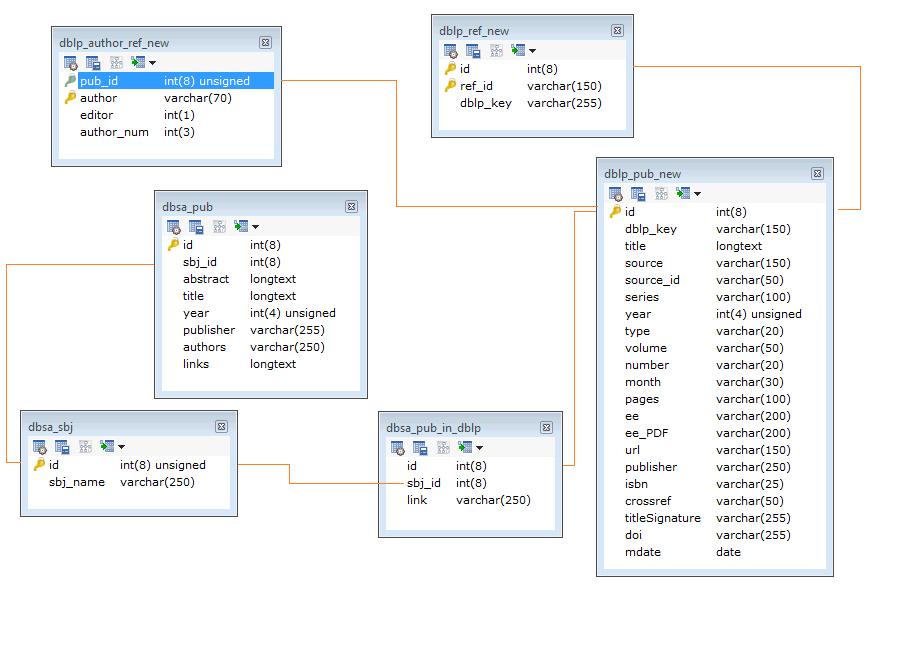
Bảng 9 – Thông tin cấu trúc bảng dbsa\_sbj

|  |  |  |  |
| --- | --- | --- | --- |
| 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 10 – 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 “,”. |

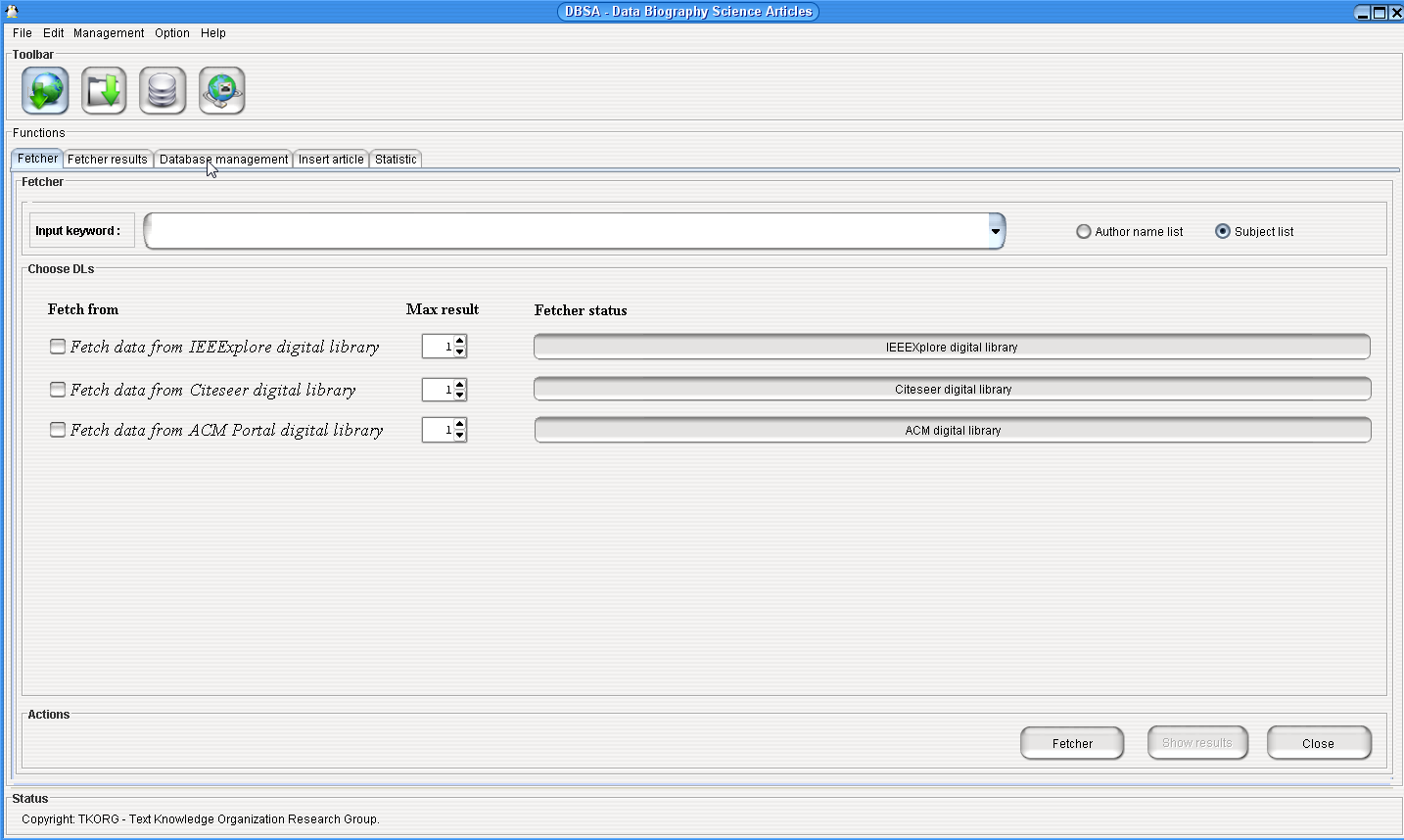
Bảng 11 - Thông tin cấu trúc bảng dbsa\_pub



Hình 18: Các bảng trong cơ sở dữ liệu hệ thống

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 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.



Hình 19 – Giao diện của hệ thống



Hình 10 – Kết quả thu thập từ hệ thống

CHƯƠNG 4: KẾT LUẬN – ĐÁNH GIÁ VÀ HƯỚNG PHÁT TRIỂN

Kết quả thực nghiệm.

Để đánh kết quả của hệ thống , chúng tôi thực hiện việc truy vấn, tìm kiếm và thu thập 100 bài báo với các chủ đề tương ứng trong các thư viện số thì kết quả thu được như sau:

Với từ khóa nhập vào là: database

|  |  |  |  |
| --- | --- | --- | --- |
| Thư viện số | Thời gian tìm kiếm  (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 12 - Kết quả của hệ thống với từ khóa là Database

Từ khóa tìm kiếm là: data mining

|  |  |  |  |
| --- | --- | --- | --- |
| Thư viện số | Thời gian tìm kiếm  (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 |

Bảng 13 - Kết quả của hệ thống với từ khóa là Data mining.

Như vậy theo các đánh giá trên trung bình hệ thống đã cập nhật được các bài mới ngay khi được công bố trên thư viện số cũng như bổ sung những dữ liệu còn thiếu trong trong DBLP.

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

Bảng 14: Kết quả bổ sung dữ liệu mới của hệ thống.

Kết luận, đánh giá và hướng phát triển.

Trong bài báo này, chúng tôi đã giới thiệu một hệ thống xây dựng và cập nhập dữ liệu chỉ mục các bài báo khoa học sử dụng Web Crawler trên các thư viện số. Như vậy với so với các nghiên cứu [8, 4, 10] 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 [2], 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. Cũng như đề ra một hướng tiếp cận để bổ sung và cập nhật dữ liệu DBLP. 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.

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

PHỤ LỤC 1 HƯỚNG DẪN CÀI ĐẶT HỆ THỐNG.

PHỤ LỤC 2 HƯỚNG DẪN SỬ DỤNG CHƯƠNG TRÌNH.

Tài liệu tham khảo.

[1] 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.

[2] 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.

[3] 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.

[4]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).

[5] Holger Bast, Ingmar Weber: “The CompleteSearch Engine: Interactive, Efficient, and Towards IR&DB Integration”, [CIDR 2007: 3rd Biennial Conference on Innovative Data Systems Research](http://www-db.cs.wisc.edu/cidr/cidr2007/index.html), Asilomar, CA, USA, 2007, 88-95.

[6] 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.

[7] 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

[8] [Michael Ley](http://www.springerlink.com/content/?Author=Michael+Ley), “The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspective”. [Lecture Notes in Computer Science](http://www.springerlink.com/content/0302-9743/), 2002, Volume 2476/2002, 481-486.

[9][Nicola Zeni](http://www.citeulike.org/user/wnpx/author/Zeni:N), [Nadzeya Kiyavitskaya](http://www.citeulike.org/user/wnpx/author/Kiyavitskaya:N), [Luisa Mich](http://www.citeulike.org/user/wnpx/author/Mich:L), [John Mylopoulos](http://www.citeulike.org/user/wnpx/author/Mylopoulos:J), [James Cordy](http://www.citeulike.org/user/wnpx/author/Cordy:J). “A Lightweight Approach to Semantic Annotation of Research Papers”. Natural Language Processing and Information Systems (2007), pp. 61-72.

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

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

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

1. <http://en.wikipedia.org/wiki/Metadata> [↑](#footnote-ref-1)
2. <http://scholar.google.com> [↑](#footnote-ref-2)
3. <http://portal.acm.org/> [↑](#footnote-ref-3)
4. <http://ieeexplore.ieee.org/> [↑](#footnote-ref-4)
5. <http://citeseerx.ist.psu.edu/> [↑](#footnote-ref-5)
6. <http://dblp.uni-trier.de/> [↑](#footnote-ref-6)
7. <http://en.wikipedia.org/wiki/Webcrawler> [↑](#footnote-ref-7)
8. http://www.library.uq.edu.au/iad/ctmeta4.html [↑](#footnote-ref-8)
9. http://www.nla.gov.au/nla/staffpaper/cathro3.html [↑](#footnote-ref-9)
10. http://dublincore.org/ [↑](#footnote-ref-10)
11. <http://en.wikipedia.org/wiki/BibTeX> [↑](#footnote-ref-11)
12. <http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php> [↑](#footnote-ref-12)
13. <http://dblp.l3s.de/?q=&newQuery=yes&resTableName=query_resultmQ9GIx> [↑](#footnote-ref-13)
14. http://en.wikipedia.org/wiki/XSLT [↑](#footnote-ref-14)
15. http://portal.acm.org/ [↑](#footnote-ref-15)
16. <http://www.xml.com/pub/a/2004/06/16/dive.html> [↑](#footnote-ref-16)
17. <http://en.wikipedia.org/wiki/Simple_API_for_XML> [↑](#footnote-ref-17)
18. <http://jabref.sourceforge.net/index.php> [↑](#footnote-ref-18)
19. <http://en.wikipedia.org/wiki/Scribe> [↑](#footnote-ref-19)