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 mà 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 trong việc đảm bảo cập nhật thông tin các bài báo một cách đẩy đủ và cập nhật nhất.

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. Hệ thống sử dụng Web crawler để tìm kiếm và rút thông trích tin chỉ mục các bài báo khoa học được công bố trên các thư viện số (ACM, IEEE, CiteSeer) . Từ những thông tin chỉ mục này, hệ thống sẽ kết hợp với dữ liệu chỉ mục có sẵn trong DBLP[[1]](#footnote-1) để xây dựng lên dữ liệu chỉ mục các bài báo khoa học. Như vậy hệ thống đảm bảo được tính cập nhật với các bài báo mới được công bố trên các thư viện số cùng với đó hệ thống sử dụng được dữ liệu thông tin chỉ mục rất lớn có sẵn từ DBLP. Việc xây dựng dữ liệu chỉ mục các bài báo khoa học là rất cần thiết, thông qua dữ liệu xây dựng được, ta có thể phát triển các công cụ tìm kiếm bài báo khoa học cung cấp cho người dùng thông tin những bài báo mới đầy đủ và cập nhật nhất.

1. Mục tiêu và phạm vi đề tài.

2.1 Mục tiêu đề tài.

Mục tiêu của đề tài 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 và cập nhật của dữ liệu.

2.2 Phạm vi đề tài.

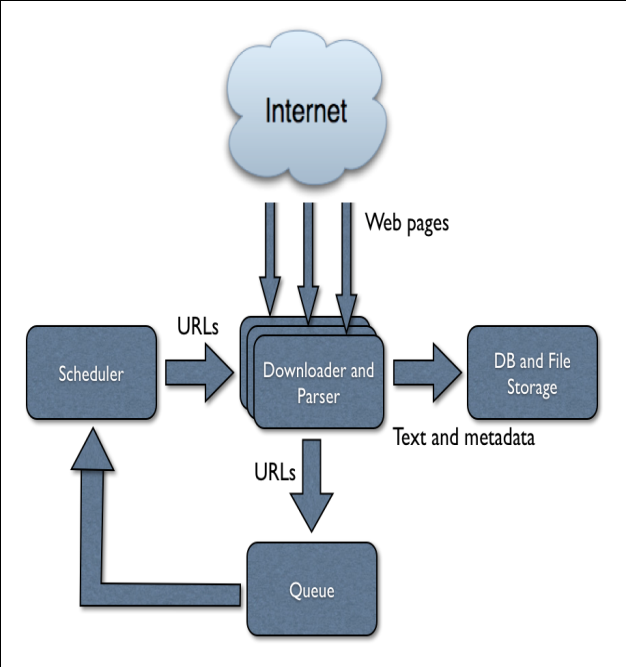
Hệ thống sử dụng Web Crawler để thu thập thông tin chỉ mục các bài báo khoa học thuộc lĩnh vực khoa học máy tính trên 3 thư viện số ACM, Citeseer, IEEEXplore. Cùng với đó hệ thống sử dụng dữ liệu chỉ mục có sẵn của DBLP, giúp dữ liệu thu thập được đảm bảo tính đầy đủ và cập nhật.

3.Các khái niệm và kiến thức cơ bản.

3.1 Web crawler[[2]](#footnote-2).

Web crawler, 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 Web Crawler Architecture

Web crawler bắt đầu từ danh sách các địa chỉ URL được gọi là seeds (hạt giống). Nó sẽ vào các địa chỉ này lọc thông tin rồi tìm ra các địa chỉ URL khác thêm chúng vào danh sách các địa chỉ đã duyệt qua gọi là crawl frontier. Sau đó nó lại lặp lại quá trình trước đó duyệt qua những URL mới. Cứ thế nó lần qua rất nhiều địa chỉ website và thu thập rất nhiều nội dung khác nhau.

3.2 Metadata[[3]](#footnote-3).

Siêu dữ liệu (metadata) 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. Vì vậy metadata là dữ liệu về dữ liệu.

3.3 Bibtex.

4. Khảo sát các đề tài liên quan.

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.

Từ đầu vào là file TOCs được nhập vào từ 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.



Hình : Hệ thống xây dựng cơ sở dữ liệu DBLP

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.

4.1.1 CompleteSearch DBLP[[4]](#footnote-4): cho phép người dùng tìm kiếm thông tin bài báo dựa trên các trường sau:

+Tìm kiếm theo từ.

+Tìm kiếm theo tên tác giả.

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

+Tìm kiếm theo năm xuất bản của bài báo.



Hình 14 : Hệ thống Complete Search

* + 1. [Faceted search](http://dblp.l3s.de)[[5]](#footnote-5): 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 15: Hệ thống FacetedDBLP

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



Hình 16: Hệ thống duyệt bài báo trong FacetedDBLP

* + 1. DBL – Browser: Là chương trình sử dụng để tìm kiếm trên file dữ liệu DBLP không cần kết nối internet. Chương trình cho phép nhìn một cách trực quan về thông tin của bài báo.



Hình 17: Chương trình DBL Brower

* 1. Lightweight Federated Digital Library (LFDL)

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 [].Hệ thống sử dung 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 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 . Sau khi hệ thống nhận diện bài báo là một bài báo về 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 , luật kết hợi 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.

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 [[6]](#footnote-6) cho phép người dùng tìm kiếm các bài báo được công bố trên các bài báo và hội nghị được tổ chức bởi ACM cũng như một số tổ chức khác.

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.

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…

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 các bài báo khoa học sử dụng web Crawler.

Phương pháp.Crawler trên thư viện số.

Rút trích thông tin bài báo.

Chương 3: Thực hiện.

1. Kiến trúc hệ thống.



Trong hình 1, 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.

1. Thiết kế cơ sở dữ liệu.

2.1 Mô tả cấu trúc dữ liệu của DBLP

Tính đến tháng 1/2011 DBLP chứa thông tin của 1,3 triệu bài báo, trong lĩnh vực công nghệ thông tin. DBLP không dùng một hệ quản trị cơ sở dữ liệu nào để lưu dữ liệu mà dữ liệu được ghi trong 125000 files. Dữ liệu của DBLP được xuất ra các dạng CDF, XML và SQL, người phát triển có thể download các file dữ liệu này từ trên web của chương trình.

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 từ file XML.

Dữ liệu được lưu trữ trong 3 bảng:

* dblp\_pub\_new: lưu thông tin bài báo.

Thông tin của bài báo 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 bao gồm tên tác giả, tác giả có phải là một người biên tập (editor) hay không. Trong bảng dữ liệu 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.



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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Trường (Field)** | **Kiểu**  **(Type)** | **Chú thích**  **(Comment)** | | Khóa chính  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 thông tin cấu trúc bảng** **dblp\_pub\_new** | | | |  |  |  |
| |  |  |  |  | | --- | --- | --- | --- | |  | **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 thông tin cấu trúc bảng dblp\_author\_ref\_new**   |  |  |  |  | | --- | --- | --- | --- | |  | **Trường**  **(Field)** | **Kiểu**  **(Type)** | **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 thông tin cấu trúc bảng dblp\_ ref\_new**  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) cũng như chủ đề của bài báo đề cập đến.  Từ cấu trúc này nhóm bổ xung thêm vào cấu trúc những bảng sau để bảng có thể lưu được các thông tin về chủ đề cũng như tóm tắt của bài báo. |  |  |
|  |  |  |
|  |  |  |

* dbsa\_sbj**:** lưu thông tinvề 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 xung 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, … ).

**Bảng dbsa\_sbj**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **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** **dbsa\_pub\_in\_dblp**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Trường (Field)** | **Kiểu**  **(Type)** | **Chú thích**  **(Comment)** |
| PRIMARY | Id | int(8) | Khóa tương ứng với id trong bảng dblp\_pub\_new |
|  | sbj\_id | int(8) | Khóa ngoại của dbsa\_sbj |
|  | Links | longtext | Những đường dẫn mà người dùng có thể tìm được bài báo. |

Bảng dbsa\_pub

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Trường (Field)** | **Kiểu**  **(Type)** | **Chú thích**  **(Comment)** |
| PRIMARY | Id | int(8) | Khóa chính của bảng |
|  | 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 “ , ”. |

1. Thu thập và rút trích thông tin từ các thư viện số
   1. Cách thức thu thập các bài báo từ thư viện số ACM

Phần này trình bày cách thức mà hệ thống DBSA thu thập các bài báo khoa học từ thư viện số ACM. Đầu vào của quá trình thu thập là từ hoặc cụm từ khóa do người dùng nhập.

Sau khi người dùng nhập xong từ khóa và yêu cầu tìm kiếm thì hệ thống bắt đầu tạo câu truy vấn dựa vào từ khóa do người dùng cung cấp và các thẻ (pattern) đã được định nghĩa. Một số thẻ (pattern) đã được sử dụng như bảng 1 bên dưới.

Kết quả của quá trình này là một danh sách các bài báo và các thông tin của mỗi bài báo đó được tìm thấy từ thư viện số ACM theo từ khóa trên.

|  |  |  |
| --- | --- | --- |
| 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>= | * Kết hợp 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 1: Các pattern sử dụng để thu thập các bài báo khoa học từ thư viện số ACM.

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

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

* Tạo một hàm có chức năng tạo ra một url để đưa lên trình duyệt web. Hàm này sẽ dựa vào từ khóa do người dùng cung cấp sau đó xử lý, gắn kết với các pattern đã được định nghĩa.
* 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/*](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 các bài báo từ thư viện số ACM.*

**Bước 2** :

* Gửi URL vừa tạo được lên Internet brower.
* Lấy nội dung trang web mà Internet brower trả về sau khi truy cập vào địa chỉ URL trên.
* Dựa vào các pattern 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, danh sách dường dẫn đến chi tiết mỗi bài báo cụ thể trong trang này.

🡪 *Mục tiêu : Danh sách các địa chỉ URL của các bài báo khoa học thu thập được từ địa chỉ URL ở bước 1.*

**Bước 3 :** Truy cập vào các bài báo

* Truy cập vào địa chỉ URL của một bài báo trong danh sách các URL bài báo đã thu được từ quá trình thu thập và phân tích ở bước 2. Sau đó chuyển sang bước 4 để lấy chi tiết các thông tin cần thiết.

🡪 *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 2.

* Sử dụng pattern có số thứ tự [5][6] trong bảng 1 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 nhưng 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 1 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 và mã số của bài báo (DOI). Lưu các thông tin đã tìm được và quay lại bước 3 để tiếp tục thu thập.*

3.2 Cách thức thu thập các bài báo từ thư viện số IEEExplore.

Tương tự như cách thức thu thập các bài báo khoa học từ thư viện số ACM. Sau khi nhận được từ khóa do người dùng cung cấp và yêu cầu tìm kiếm, hệ thống sẽ tạo câu truy vấn (hay địa chỉ URL) dựa vào từ khóa đó và các pattern đã được định nghĩa sẵn ở bảng 2 bên dưới.

Đầu ra của quá trình này cũng là danh sách các bài báo khoa học và các thông tin của mỗi bài báo đó.

|  |  |  |
| --- | --- | --- |
| 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 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 cách thức thu thập bài báo từ thư viện 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*](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 1.



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

**Bước 2:** Sử dụng câu truy vấn lấy về thông tin kết quả tìm kiếm

* Tương tự như cách thức thu thập bài báo từ thư viện số ACM, sau khi gửi câu truy vấn vừa tạo ở bước 1 lên Internet brower. Hệ thống sẽ lấy về trang kết quả tương thích với câu truy vấn, các bài báo liên quan tới từ khóa, số trang muốn thu thập và số kết quả tối đa trên trang đó theo yêu cầu.
* Sử dụng các pattern đã được định nghĩa để phân tích trang web kết quả vừa lấy về. Thu thâp và tạo ra danh sách các địa chỉ URL của các bài báo trong trang này.
* *Mục tiêu : gửi câu truy vấn lên Internet brower, thu thập thông tin và tạo* *danh sách các đường dẫn của các bài báo khoa học cần tìm.*

**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 các bài báo dựa vào trang web vừa lấy được ở bước 3.

* Sử dụng các pattern từ số [6] đến [17] trong bảng 2, và bộ phân tích cú pháp để tìm thông tin của một bài báo. Những thông tin bao gồm : Số của bài báo, tiêu đề, các đồng tác giả, tóm tắt, năm công bố, tổ chức công bố, DOI của một bài báo.
* 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. Cách thức thu thập các bài báo khoa học từ thư viện số citeseerx.

Đối với thư viện số citeseerx, khi người dùng nhập từ khoá tìm kiếm vào hệ thống sẽ tạo câu query theo cấu trúc :

[http://citeseerx.ist.psu.edu/search?q=**KEYWORD**&feed=atom&sort=rel](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 đường dẫn trên, hệ thống sẽ thực hiện kết nối với thư viện số citeseerx 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. Kết quả này trả về một link liên kết chứa các kết quả của dưới dạng XML được gọi là ATOM Link [[7]](#footnote-7) 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)"/>

. . .

Từ danh sách kết quả tìm kiếm được bởi citeseerx hệ thống sẽ dùng SAXPaser để lấy thông tin của từng bài báo .

SAXPaser [[8]](#footnote-8) – 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.

1. Hệ thống xây dựng dữ liệu chỉ mục (DBSA).

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 3: Màn hình cho phép xem kết quả tìm kiếm được.*

Chương 4: Kết luận - Đánh giá và hướng phát triển.

1. Thực nghiệm và đánh giá.

Để đánh giá công cụ DBSA, 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:

Từ khóa tìm kiếm 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 2: Kết quả của tìm kiếm đánh dấu chỉ mục của DBSA 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 3: Kết quả của tìm kiếm đánh dấu chỉ mục của DBSA với từ khóa là Database.

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 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 4: Kết quả bổ xung dữ liệu mới của hệ thống.

1. Kết luận 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 [1], [7], [8] 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. 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 ứng dụng 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] [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

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

[4] 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.

**[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] 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.

[7]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).

**[8] Rong Shi: “Automatic metadata discovery from noncooperative digital libraries”. in Proc. of IADIS international Conf. on e-Society 2003**

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

**[10]**[**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**

**[11] 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**

**[12] 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**

1. <http://www.informatik.uni-trier.de/~ley/db/> [↑](#footnote-ref-1)
2. <http://en.wikipedia.org/wiki/Webcrawler> [↑](#footnote-ref-2)
3. <http://en.wikipedia.org/wiki/Metadata> [↑](#footnote-ref-3)
4. <http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php> [↑](#footnote-ref-4)
5. <http://dblp.l3s.de/?q=&newQuery=yes&resTableName=query_resultmQ9GIx> [↑](#footnote-ref-5)
6. http://portal.acm.org/ [↑](#footnote-ref-6)
7. <http://www.xml.com/pub/a/2004/06/16/dive.html> [↑](#footnote-ref-7)
8. <http://en.wikipedia.org/wiki/Simple_API_for_XML> [↑](#footnote-ref-8)