**HỆ THỐNG XÂY DỰNG VÀ CẬP NHẬT DỮ LIỆU CHỈ MỤC CÁC BÀI BÁO KHOA HỌC**

*Huỳnh Ngọc Tín* [*tinhn@uit.edu.vn*](mailto:tinhn@uit.edu.vn) *, Đỗ Văn Tiến* [*tiendo.vn@gmail.com*](mailto:tiendo.vn@gmail.com) *, Nguyễn Phước Cường* [*cuongnp.uit.se@gmail.com*](mailto:cuongnp.uit.se@gmail.com)

***Khoa công nghệ phần mềm – Trường Đại Học Công Nghệ Thông Tin – Đại học Quốc Gia Thành Phố Hồ Chí Minh***

***Khu Phố 6 – Linh Trung – Thủ Đức Thành Phố Hồ Chí Minh***

**Tóm tắt:**

Cùng với sự phát triển của internet, số lượng các bài báo khoa học, số tạp chí điện tử, cũng như số các tác giả viết báo khoa học ngày càng tăng, điều này gây ra một số khó khăn khi mà người người dùng muốn tìm kiếm các bài báo về một vấn đề mà mình nghiên cứu - cũng như gây ra một thách thức lớn với các hệ thống đánh dấu,lưu trữ chỉ mục trong việc cập nhật thông tin các bài báo mới một cách đẩy đủ và nhanh nhất. Trong bài báo này chúng tôi đề cập đến phát triển một hệ thống dùng để xây dựng và cập nhật thông tin dữ liệu chỉ mục các bài báo khoa học từ các thư viện số và dữ liệu chỉ mục có sẵn. Hệ thống có chức năng tìm kiếm các bài báo khoa học được công bố trên các thư viện số (ACM[[1]](#footnote-1), IEEE[[2]](#footnote-2), CiteSeer[[3]](#footnote-3)) sau đó rút các thông tin về chỉ mục của bài báo, bao gồm các thông tin tên tác giả, tựa đề, tóm tắt, năm công bố, nơi công bố và địa chỉ đường dẫn của bài báo. Từ những thông tin chỉ mục này, hệ thống sẽ kết hợp với dữ liệu chỉ mục có sẵn trong DBLP[[4]](#footnote-4) 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 và 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.

Từ khóa: Rút trích thông tin, Thư viện số, DBLP, Webmining

1. **Giới thiệu**

Trong quá trình học tập cũng như nghiên cứu khoa học, việc tham khảo các hướng đi cũng như những bài viết liên quan đến vấn đề mình nghiên cứu từ các bài báo khoa học đã được công bố là rất cần thiết. Trong lĩnh vực khoa học máy tính thì số lượng các tác giả, hội thảo, tạp chí ngày càng nhiều từ đó số lượng các bài báo được công bố ngày càng tăng khiến người nghiên cứu gặp khó khăn trong việc tìm kiếm cũng như xác định được các bài báo có liên quan đến vấn đề của mình tiếp cận. Hiện nay khi người nghiên cứu cần tìm kiếm một bài báo khoa học, thì họ có tìm kiếm trên các thư viện số như ACM , IEEE, CiteSeer … hoặc dữ liệu chỉ mục có sẵn như DBLP. Vấn đề đặt ra ở đây là:đối với các thư viện số thì các bài báo được cập nhật một cách thường xuyên khi có các hội thảo, nhưng các thư viện số này thường tương ứng với một tổ chức (nơi tổ chức các cuộc hội thảo), như vậy việc cập nhật các bài báo của tổ chức khác của thư viện số gặp hạn chế. Trong khi đó các dữ liệu chỉ mục hiện tại chưa đảm bảo được tính cập nhật các bài báo mới vì nó phụ thuộc vào nguồn từ các thư viện số cũng như các cuộc hội thảo này. Như vậy ta có thể kết hợp việc thu thập, đánh chỉ mục các bài báo từ nhiều thư viện số khác nhau và kết hợp một dữ liệu chỉ mục có sẵn điều này đảm bảo dữ liệu chỉ mục các bài báo được cập nhật một cách nhanh nhất và đầy đủ nhất.

Chức năng chính của hệ thống chúng tối giới thiệu là tìm kiếm và đánh dấu chỉ mục các bài báo từ các thư viện số sau đó kết hợp với dữ liệu chỉ mục có sẵn để tạo lên cơ sở dữ liệu chỉ mục các bài báo khoa học. Hệ thống sẽ tìm kiếm những bài báo theo từ khóa được nhập vào từ người dùng (từ khóa bao gồm cả tên tác giả, tên hội nghị) hoặc người dùng có thể chọn những chủ đề mà hệ thống gợi ý. Những chủ đề của hệ thống gợi ý cho người dùng được lấy từ tên các chủ đê thuộc lĩnh vực khoa học máy tính trên Wikipedia[[5]](#footnote-5), điều này đảm bảo tính đúng đắn của dữ liệu tìm kiếm.

Thư viện số là nơi công bố những bài báo của hội nghị, tạp chí vì thế kết quả trả về từ các thư viện số sau khi tìm kiếm bao gồm cả những bài báo mới nhất. Ứng dụng sử dụng dữ liệu của DBLP, một thư viện dữ liệu chỉ mục dạng mở, dữ liệu chỉ mục này cung cấp thông tin khoảng 1,5 triệu bài báo trong lĩnh vực khoa học máy tính nhưng dữ liệu này chưa đầy đủ và việc cập nhật dữ liệu là rất hạn chế (phần 2 sẽ trình bày rõ vấn đề này). Hệ thống cũng sẽ kiểm tra tính tồn tại của bài báo trả về thư viện số trong dữ liệu DBLP, sau đó hệ thống sẽ lưu các bài báo không có trong DBLP xuống cơ sở dữ liệu điều này đảm bảo dữ liệu của hệ thống là đầy đủ và chính xác.

Trong phần 2 chúng tôi sẽ giới thiệu một số nghiên cứu về cách xây dựng dữ liệu chỉ mục các bài báo khoa học cũng như giới thiệu về dữ liệu chỉ mục DBLP. Trong phần 3 sẽ giới thiệu cấu trúc và các module trong hệ thống của chúng tôi. Phần 4 sẽ đưa ra một số đánh giá khi chạy ứng dụng. Phần cuối chúng tôi sẽ đưa ra kết luận và hướng phát triển của hệ thống.

1. **Những nghiên cứu- ứng dụng liên quan.**

Xây dựng dữ liệu chỉ mục các bài báo khoa học là một ứng dụng trong vấn đề rút trích thông tin. Hệ thống sẽ rút trích các thông tin chỉ mục các bài báo từ thư viện số.Từ dữ liệu chỉ mục các bài báo ta có thể xây dựng lên các thư viện số cũng như các công cụ tìm kiếm giúp cho người dùng rút ngắn thời gian trong quá trình tìm kiếm các bài báo khoa học. Hiện nay theo khảo sát của chúng tôi, có một số phương pháp và hướng tiếp cận đến các nguồn mà từ đó có thể xây dựng dữ liệu chỉ mục bao gồm: xây dựng dữ liệu chỉ mục từ các file đề mục (tables of contents– TOCs của các cuộc hội thảo) [1] và rút trích từ nội dung bài báo được công bố trên mạng (sử dụng các file postscript hoặc file PDF của các bài báo được công bố trên mạng) [5],[6].

Theo [1] -Digital Bibliography & Library Project (DBLP), DBLP cung cấp thông tin chỉ mục các bài báo trong lĩnh vực khoa học máy tính, hệ thống này 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 tạp chí và hội nghị. DBLP được xây dựng lên từ các file danh sách các đề mục – mục lục (tables of contents– TOCs), các file TOCs được tác giả nhận từ các tạp trí, hội nghị sau đó qua quá trình xử lý bằng các bộ phân tích hệ thống sẽ đánh chỉ mục thông tin các bài báo, từ đó lưu xuống và hình thành lên cơ sở dữ liệu chỉ mục DBLP. Dữ liệu của DBLP là dữ liệu mở, dữ liệu được xuất ra các định dạng CDF, XML và MySQL, người phát triển có thể tải các file dữ liệu từ trên trang chủ của DBLP. Dựa trên cơ sở này đã có nhiều công trình phát triển các công cụ tìm kiếm như giới thiệu trong các bài báo[7],[8].

Để đánh giá tính cập nhật dữ liệu của DBLP, chúng tôi đã sử dụng phiên bản mới nhất của DBLP (Tháng 12 năm 2010). Để kiểm chứng,chúng tôi đã thực hiện bằng cách tìm kiếm các bài báo trong thư viện số với từ khóa là chủ đề trong lĩnh vực khoa học máy tính, sau đó kiểm tra 100 bài báo đầu tiên được trả về từ các thư viện số bao gồm cả các bài trước năm 2010 và sau năm 2010, sau đó tìm kiếm các bài báo này trong dữ liệu của DBLP thì được kết quả bảng 1 (tỉ lấy trung bình khảo sát trên 3 thư viện số ACM, Citeseer, IEEEXplore). Dựa vào bảng 1 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  (%) | Dữ liệu không xác định được năm không tồn tại trong DBLP  (%) |
| Database | 28,33% | 86,26% | 10,71% | 3,03% |
| Data mining | 43,67% | 77,45% | 14,51% | 8,04% |

*Bảng 1: Khảo sát tính cập nhật dữ liệu của DBLP*

Trong các bài báo [5], [6], tác giả giới thiệu về Autonomous Citation Indexing – một hệ thống đánh dấu chỉ mục các bài báo sử dụng trong thư viện số CiteSeer. ACI sẽ sử dụng một search engine để đi thu thập, tải các bài báo trên Internet về với các định dạng file PostScript hoặc PDF; sau đó ACI sẽ sử dụng các thuật toán và các bộ phân tích để rút trích các thông tin chỉ mục bài báo trên các file này và tổ chức thành dữ liệu chỉ mục các bài báo từ đó xây dựng lên thư viện số Citeseer, hệ thống cho phép người dùng tìm kiếm thông tin các bài báo trên dữ liệu chỉ mục này. Như vậy, đối với các bài báo mà đòi hỏi người dùng phải trả phí khi tải về như các bài báo trên thư viện số ACM, IEEEXplore, Spinger … thì hệ thống ACI chưa thể thu thập được dữ liệu chỉ mục, trong khi đó các thư viện số của các tổ chức đã đánh dấu dữ liệu chỉ mục của bài báo mới này.

Jabref [[6]](#footnote-6) là một công cụ quản lý dữ liệu chỉ mục, một ứng dụng xây dựng bằng ngôn ngữ Java và có mã nguồn mở. Công cụ này có chức năng tổ chức dữ liệu chỉ mục các bài báo khoa học của người dùng thành cơ sở dữ liệu dưới dạng các file BibTeX[[7]](#footnote-7), từ đó dễ dàng cho phép người dùng tìm kiếm, tra cứu các bài báo đã tham khảo. Ngoài ra chương trình này cũng có chắc năng thu thập các bài báo từ thư viện số như ACM, JSTOR, … nhưng tại thời điểm chúng tôi thực hiện, chức năng này không thể tìm kiếm dữ liệu khi người dùng yêu cầu, mặt khác các kết quả tìm kiếm về bao gồm tất cả kết quả trả về từ thư viện số không tận dụng được những dữ liệu chỉ mục có sẵn như DBLP.

1. **Phương pháp thực hiện.**

Khi người dùng nhập từ khóa tìm kiếm, hệ thống sẽ gửi từ khóa tìm kiếm lên các thư viện số. Kết quả trả về là đường dẫn tới các bài báo tồn tại trong thư viện số. Chúng tôi sẽ sử dụng các bộ phân tích và các luật để rút trích các thông tin chỉ mục của bài báo. Sau khi rút trích được thông tin bài báo sẽ kiểm tra tính tồn tại của bài báo trong dữ liệu có sẵn của và dữ liệu chỉ mục DBLP và cuối cùng sẽ lưu thông tin bài báo vào cơ sở dữ liệu của hệ thống.

* 1. Kiến trúc hệ thống.



*Hình 1: Kiến trúc DBSA*

Trong hình 1, chúng tôi giới thiệu kiến trúc của ứng dụng, từ khóa tìm được nhập vào từ người dùng hoặc chọn từ danh sách các chủ đề trong lĩnh vực máy tính được lấy từ Wikipedia [10], điều này đảm bảo được sự chính xác cho xác định các bài báo thuộc lĩnh vực khoa học máy tính. Sau khi nhận được kết quả trả về từ các thư viện số - là các đường link tới các bài báo phù hợp với từ khóa tìm kiếm trong thư viện số, module rút trích thông tin bài báo sẽ sử dụng các trình phân tích và áp dụng các luật để nhận điện 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 sau quá trình kiểm tra thì tùy vào quyết định của người dùng xem có lưu thông tin bài trùng lặp không mà sẽ lưu kết quả vào dữ liệu chỉ mục. Module import dữ liệu DBLP có chức năng kết nối với dữ liệu của DBLP và cập nhật dữ liệu của khi DBLP xuất một bản mới.

* 1. Rút trích thông tin chỉ mục.

Kết quả trả về từ thư viện số là danh sách các đường dẫn chứa thông tin các bài báo trong dữ liệu của thư viện số. Hệ thống sử dụng một số luật và trình phân tích để lấy nội dung bài báo từ các đường dẫn này.

Hệ thống sẽ sử dụng bộ phân tích HTMLParser và các luật (Regular Expression), để xác định số các kết quả trả về từ thư viện số.Các thư viện số ACM và IEEEXplore, Citeseer thì tương ứng với mỗi bài báo thì thư viện số sẽ cung cấp file Bibtex, việc dùng các luật để rút ra các file Bibtex này giúp hệ thống rút trích được thông tin chỉ mục các bài báo.

Bibtex là một định dạng dùng để biểu diễn thông tin chỉ mục các bài báo, sách. Khi các thông tin bài báo được lưu dưới định dạng này ứng dụng sẽ sử dụng bộ phân tích Bibtex Parser để rút ra các thông tin như tên bài báo, tác giả, tóm tắt, năm xuất bản, nhà xuất bản, đường dẫn dùng để tải tài liệu.

@article{Gettys90,

author = {Jim Gettys and Phil Karlton and Scott McGregor},

title = {The {X} Window System, Version 11},

journal = {Software Practice and Experience},

volume = {20},

number = {S2},

year = {1990},

abstract = {A technical overview of the X11 functionality. This is an update

of the X10 TOG paper by Scheifler \& Gettys.}

}

*Hình 2: Ví dụ về nội dung một file Bibtex*

* 1. Kiểm tra tính trùng lặp của bài báo trong dữ liệu DBLP

Để các bài báo mà hệ thống thu thập về không bị trùng lặp với những bài báo đã có trong cơ sở dữ liệu cũng như trong DBLP thì hệ thống sẽ kiểm tra tính tồn tại của bài báo báo dựa vào tiêu đề cũng như năm xuất bản của bài báo.

Đầu tiên, hệ thống sử dụng tiêu đề của bài báo để so sách với tiêu đề của các bài báo đã có trong cơ sở dữ liệu. Tiêu đề của bài báo được loại bỏ khoảng trắng và các đấu,ký tự không có nghĩa như “!”, “?” … và được chuyển về chữ thường. Sau đó hệ thống sẽ thực hiện việc kiểm tra bằng cách tìm trong dữ liệu của DBLP và cơ sở dữ liệu DBSA xem có bài báo nào có tựa đề như kết quả vừa tìm kiếm được không (tất cả các tiêu đề của các bài báo trong DBLP cũng đã được xử lý theo bước trên và lưu vào một bảng trong cơ sở dữ liệu). Nếu như dữ liệu về tựa đề bài báo tồn tại trong database, hệ thống sẽ kiểm tra năm xuất bản của bài báo với năm xuất bản bài báo trùng trong dữ liệu điều này nhằm đảm bảo tính chính xác của kết quả kiểm tra. Kết quả so sánh sẽ được hiển thị chon người dùng lựa chọn phương án lưu hay không lưu kết quả tìm kiếm được.

* 1. DBSA – Ứng dụng xây dựng dữ liệu chỉ mục các bài báo khoa học.

Chúng tôi xây dựng một ứng dụng đánh dấu chỉ mục các bài báo từ các thư viện số, kết hợp sử dụng thư viện chỉ mục có sẵn để xây dựng lên một dữ liệu chỉ mục đảm bảo tính đầy đủ và cập nhật các bài báo.Hệ thống 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 với người dùng một cách trực quan.



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

1. **Thực Nghiệm và Đánh giá.**

Để đánh giá công cụ DBSA, chúng tôi thực hiện việc truy vấn và tìm kiếm 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 | Số bài tồn tại trong DBLP  (%) | Số bài trước năm 2005 mà trong DBLP không chứa  (%) |
| ACM | 33 phút | 93% | 85,71% |
| CiteSeer | 33 giây | 78% | 90,91% |
| IEEEXplore | 62 giay | 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 | Số bài tồn tại trong DBLP  (%) | Số bài trước năm 2005 mà trong DBLP không chứa  (%) |
| ACM | 32 phút | 52% | 66,67% |
| CiteSeer | 15 giây | 71% | 72,41% |
| IEEEXplore | 58 giây | 46% | 12,96% |

*Bảng 3: Kết quả của tìm kiếm đánh dấu chỉ mục của DBSA với từ khóa là Database*

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.Hệ thống sử dụng các trình phân tích và các luật để rút trích thông tin chỉ mục các bài báo được thu thập về từ thư viện số. Với việc sử dụng cơ sở dữ liệu có sẵn DBLP đảm bảo được tính đầy đủ các bài báo khoa học cũng như cập nhật một cách nhanh nhất những bài báo mới được công bố từ các thư viện số.

Bước tiếp theo trong tương lai chúng tôi sẽ hoàn thiện ứng dụng với các chức năng

1. 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.
2. 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.
3. Xây dựng tìm kiếm các bài báo khoa học dựa trên dữ liệu mà thu thập được.
4. **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).

1. <http://www.acm.org/> [↑](#footnote-ref-1)
2. <http://ieeexplore.ieee.org/search/> [↑](#footnote-ref-2)
3. <http://citeseerx.ist.psu.edu/> [↑](#footnote-ref-3)
4. <http://www.informatik.uni-trier.de/~ley/db/> [↑](#footnote-ref-4)
5. <http://en.wikipedia.org/wiki/Computer_science> [↑](#footnote-ref-5)
6. <http://jabref.sourceforge.net/> [↑](#footnote-ref-6)
7. <http://www.bibtex.org/> [↑](#footnote-ref-7)