**Xây dựng và cập nhập 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 dùng muốn tìm kiếm bài báo về một vấn đề nào đó cũng như gây ra một thách thức lớn với các hệ thống đánh dấu, lưu trữ chỉ mục các bài báo trong việc cập nhật các bài báo một cách đẩy đủ và nhanh nhất. Trong bài báo này chúng tôi đề cập đến phát triển một ứng dụng 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. Ứng dụng có chức năng tìm kiếm các bài báo khoa học được công bố trên các thư viện số (ACM, IEEE, CiteSeer) 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[1] để xây dựng lên dữ liệu chỉ mục các bài báo khoa học. Như vậy ứng dụng đảm bảo được tính cập nhật với các bài báo mới được công bố trên các thư viện số và ứng dụng sử dụng được dữ liệu 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.

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 cuộc hội thảo, tạp trí ngày càng nhiều từ đó số lượng các bài báo công bố ngày càng tăng khiến người nghiên cứu khoa học khó khăn trong việc tìm kiếm cũng như xác định được các bài báo có liên quan đến vấn đề của mình nghiên cứu. Hiện nay khi cần tìm kiếm một bài báo khoa học, người nghiên cứu có thể tìm kiếm trên các search engine như googleschoolar, các thư viện số như ACM [2], IEEE [3] CiteSeer[4]… hoặc dữ liệu chỉ mục có sẵn như DBLP [1]. Vấn đề đặt ra ở đây là số lượng các hội thảo, tạp chí trong lĩnh vực khoa học máy tính ngày càng nhiều làm cho việc tìm kiếm bài báo của người dùng gặp khó khăn cùng với đó đặt ra một thách thức lớn với những hệ thống này trong việc thu thập và cập nhật một cách đầy đủ nhất các bài báo khoa học từ đó cung cấp cho người dùng thông tin chính xác về các bài báo theo sự tìm kiếm của người dùng.

Chức năng chính của ứng dụng là tím kiếm và đánh dấu chỉ mục các bài báo từ các thư viện số, hệ thống sẽ tìm kiếm những bài báo theo chủ đề được người dùng nhập vào. Chủ đề tìm kiếm được lấy từ tên các chủ đê thuộc lĩnh vực khoa học máy tính, điều này đảm bảo tính đúng đắn của dữ liệu tìm kiếm được. Thư viện số là nơi công bố những bài báo của hội nghị, tạp chí vì thế kết quả trả về 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). Ứng dụng sẽ kiểm tra tính tồn tại của bài báo trả về thư viện số trong dữ liệu DBLP, sau đó ứng dụng sẽ lưu các bài báo không có trong DBLP xuống cơ sở dữ liệu.

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

1. **Những 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. 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.Hiện nay có một số phương pháp và hướng tiếp cận đến các nguồn mà từ đó hệ thống có thể xây dựng dữ liệu chỉ mục bao gồm: xây dựng từ các file đề mục (tables of contents– TOCs) [1] và rút trích từ nội dung bài báo (postscript hoặc file PDF)[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 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 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 hệ thống và hình thành lên cơ sở dữ liệu chỉ mục DBLP. Dữ liệu của DBLP là dữ liệu mở, dẽ liệu được xuất ra các 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 sử dụng phiên bản mới nhất của dữ liệu DBLP(Tháng 12 năm 2010) để kiểm chứng.Chúng tôi đã thực bằng cách, hiện 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 thư viện số. Dựa vào bảng 1 ta thấy dữ liệu của DBLP không đảm bảo được tính đầy đủ và cập nhật dữ liệu.

Trông các bài báo [5],[6] , tác giả giới thiệu về Autonomous Citation Indexing - hệ thống đánh dấu chỉ mục sử dụng trong thư viện số CiteSeer. ACI sẽ sử dụng sử dụng một search engine để thu thập, tải các bài báo trên Internet về với các định dạng file PostScript hoặc PDF, sau đó hệ thống 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 cho phép người dùng tìm kiếm thông tin 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 thư viện số ACM, IEEEXplore thì hệ thống chưa thể thu thập được dữ liệu chỉ mục, trong khi đó các thư viện số này đã đánh dấu dữ liệu chỉ mục sẵn.

Jabref 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 chắc 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 [11], từ đó dễ dàng cho phép người dùng tìm kiếm. 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.

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 luật và các mẫu để rút trích các thông tin chỉ mục của bài báo. Sau khi rút trích được thông tin bài báo hệ thống sẽ kiểm tra tính tồn tại của bài báo trong dữ liệu có sẵn của hệ thống và dữ liệu chỉ mục DBLP và cuối cùng sẽ lưu thông tin bài báo vào cơ sở dữ liệu của ứng dụng.

* 1. Kiến trúc ứng dụng.



Hình 1: Kiến trúc hệ thống DBSA

Trong hình 1 chúng tôi giới thiệu kiến trúc của ứng dụng, từ khóa tìm kiếm các bài báo trong lĩnh vực khoa học máy tính 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, hệ thống sẽ sử dụng các phân tích và áp dụng các luật để nhận rút ra các thông tin chỉ mục bài báo. Từ thông tin chỉ mục của bài báo, hệ thống sẽ kiểm tra xem bài báo đã tồn tại trong DBLP nếu có trùng lặp Sau sau quá trình kiểm tra thì tùy vào quyết định của người dùng xem có lưu thông tin bài trùng lặp không mà hệ thống sẽ lưu kết quả vào dữ liệu chỉ mục.

* 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.

Ứng dụng sẽ sử dụng bộ phân tích như HTMLParser và các luật (Regular Expression), để xác định số các kết quả trả về từ thư viện số. Đối với thư viện số ACM và IEEEXplore, Citeseer thì tương ứng với mỗi bài báo hệ thống sẽ cung cấp file Bibtex, việc dùng các luật để rút ra các file Bibtex này giúp hệ thống xác định 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 bài báo được ứng dụng thu thập về không bị trùng lặp với những dữ liệu đã có trong hệ thống cũng như trong DBLP, ứng dụng sẽ kiểm tra bài báo dựa vào tiêu đề cũng như năm xuất bản của bài báo.

Đầu tiên hệ thống sẽ 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 của hệ thống xem có bài báo nào có tựa đề như vậy không (tất cả các tiêu đề của các bài báo trong DBLP cũng đã được làm sạch 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 thì hệ thống sẽ kiểm tra năm xuất bản của bài báo điều này đảm bảo được tính chính xác của kết quả kiểm tra.

* 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. Công cụ được xây dựng bằng Java do đó có thể chạy tốt trên bất cứ hệ điều hành nào như Windows, Linux. Công cụ có cách hiển thị dữ liệu trực quan giúp người dùng chỉnh sửa các thông tin của bài báo hoặc thêm bớt dữ liệu một cách trực tiếp.



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 được liệt kê trong bảng 1.

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 hệ thống DBLP không chứa  (%) |  |
| ACM |  |  |  |  |
| CiteSeer |  |  |  |  |
| IEEEXplore |  |  |  |  |

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 hệ thống DBLP không chứa  (%) |  |
| ACM |  |  |  |  |
| CiteSeer |  |  |  |  |
| IEEEXplore |  |  |  |  |

1. **Kết luận và hướng phát triển.**

Trong bài báo này, chúng tôi đã giới thiệu một ứng dụng cho việc xây dựng và cập nhập dữ liệu chỉ mục các bài báo khoa học. Ứng dụng sử dụng các luật và các mẫu 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 hệ thống 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 hệ thống thu thập dựa trên những thông tin về chỉ mục của bài báo.
3. Xây dựng hệ thống tìm kiếm các bài báo khoa học dựa trên dữ liệu mà hệ thống thu thập được.
4. **Tham khảo**

# [1][Michael Ley](http://www.informatik.uni-trier.de/%7Eley/addr.html) The DBLP Computer Science Bibliography <http://www.informatik.uni-trier.de/~ley/db/>

# [2] ACM Digital Library

# <http://portal.acm.org/>

# [3] IEEEXplore Digital Library

# <http://ieeexplore.ieee.org/Xplore/guesthome.jsp>

# [4] CiteSeerx Scientific Literature Digital Library and Search Engine

# <http://citeseerx.ist.psu.edu/>

# [5] C.L. Giles, K. Bollacker , and S. Lawrence, “CiteSeer: An Automatic Citation Indexing System,” Digital Libraries 98: Third ACM Conf. Digital Libraries, ACM Press,New York, 1998, pp. 89-98.

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

[7] J. Diederich and W.-T. Balke: [*FacetedDBLP - Navigational Access for Digital Libraries*](http://www.ieee-tcdl.org/Bulletin/v4n1/balke/balke.html), Bulletin of IEEE Technical Committee on Digital Libraries, Volume 4 Issue 1, Spring 2008, ISSN 1937-7266.

# [8] 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*

[9] The JabRef – an open source bibliography reference manager. <http://jabref.sourceforge.net/>

[10]Computer science in Wikipedia <http://en.wikipedia.org/wiki/Computer_science>

[11] BibTeX http://www.bibtex.org/