**XÂY DỰNG KHO CHỈ MỤC BÀI BÁO KHOA HỌC DÙNG WEB CRAWLER**

*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:***

Trong bài báo này chúng tôi đề cập đến việc phát triển một hệ thống dùng Webcrawler[[1]](#footnote-1) để xây dựng kho dữ liệu chỉ mục bài báo khoa học trong lĩnh vực khoa học máy tính. Đầu vào của hệ thống là từ khóa được người dùng nhập vào hoặc được chọn tự động từ danh sách tên các tác giả, các chủ đề trong lĩnh vực khoa học máy tính, dựa vào từ khóa này hệ thống sẽ tìm kiếm các bài báo khoa học được công bố trên các thư viện số (ACM[[2]](#footnote-2), IEEEXplore[[3]](#footnote-3), CiteSeer[[4]](#footnote-4)). Kết quả trả về từ thư viện số là danh sách các bài báo tương ứng với từ khóa tìm kiếm, hệ thống sẽ sử dụng các luật để rút ra thông tin metadata của bài báo. Từ những thông tin của bài báo hệ thống sẽ kết hợp với dữ liệu có sẵn trong DBLP[[5]](#footnote-5) để xây dựng lên cơ sở dữ liệu chỉ mục các bài báo khoa học một cách đầy đủ và cập nhật nhất.

***Từ khóa***: Thư viện số, DBLP, Web mining, Web crawler, Database.

1. **Giới thiệu**

Hiện nay, phần lớn các bài báo khoa học được công bố trên các thư viện số vá các tiện ích chia sẻ cũng như trên các trang cá nhân của các tác giả. Các thư viện số chứa một số lượng lớn thông tin các bài báo của các hội nghị cũng như tạp chí, số lượng bài báo tăng lên rất nhiều sau mỗi năm. Đã có rất nhiều nghiên cứu cũng như các hệ thống dùng để đánh dấu, lưu trữ dữ liệu chỉ mục các bài báo [1], [2], [3], [7], [8] nhưng việc đảm bảo dữ liệu thu thập được đầy đủ và cập nhật vẫn là một thách thức lớn.

Trong bài báo này chúng tôi phát triển một hệ thống thu thập và đánh dấu dữ liệu chỉ mục bằng cách sử dụng Web Crawler trên các thư viện số ACM, IEEEXplore, CiteSeer dựa trên từ khóa là tên các tác giả hoặc tên các chủ đề thuộc lĩnh vực khoa học máy tính (Wikipedia[[6]](#footnote-6)), sau đó từ kết quả trả về chúng tôi sẽ sử dụng các luật để rút trích thông tin metadata của mỗi bài báo bao gồm tên tựa đề bài báo, tác giả bài báo, tóm tắt bài báo, nơi công bố và năm xuất bản bài báo. Từ dữ liệu thu thập được chúng tôi kết hợp với dữ liệu chỉ mục có sẵn là DBLP – một dữ liệu chứa thông tin 1,5 triệu bài báo (tính đến tháng 1 năm 2011). Hệ thống còn có chức năng tự động cập nhật những bài báo mới nhất từ các thư viện số để thêm vào dữ liệu chỉ mục. Như vậy với việc sử dụng từ khóa để crawler trên các thư viện số hệ thống đảm bảo tính cập nhật những bài báo mới nhất đồng thời kết hợp được dữ liệu có sẵn của DBLP việc này đảm bảo được tính đầy đủ của dữ liệu hệ thống thu thập được.

Trong phần 2, chúng tôi sẽ tiến hành khảo sát một số hệ thống đánh dấu, thu thập dữ liệu chỉ mục và dữ liệu chỉ mục DBLP. Trong phần 3 chúng tôi sẽ giới thiệu sẽ giới thiệu cách tiếp cận của chúng tôi. Một số đánh giá và kết quả thực nghiệm hệ thống giới thiệu trong phần 4. Phần cuối chúng tôi sẽ đưa ra kết luận và hướng phát triển của hệ thống.

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

Theo khảo sát [11],[12], có một số phương pháp và hướng tiếp cận từ 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 (sử dụng các file postscript hoặc file PDF của các bài báo được công bố trên mạng) [2],[3],[10] hay các hệ thống sử dụng crawler để thu thập thông tin các bài báo trên web [7],[8],[9].

Theo [1] DBLP (Digital Bibliography & Library Project) là một cơ sở dữ liệu chỉ mục các bài báo khoa học trong lĩnh vực khoa học máy tính.Tính đến tháng 1/2011 DBLP chứa thông tin của 1, 5 triệu bài báo được đánh dấu chỉ mục thông qua việc phân tích danh sách các đề mục (tables of contents– TOCs) của các hội nghị cũng như các tạp chí. 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[4],[5].

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

Để đánh giá tính đầy đủ và cập nhật dữ liệu trong DBLP chúng tôi tiến hành khảo sát 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 tính tồn tại của thông tin bài báo trong DBLP. Dựa vào bảng 1 ta thấy dữ liệu của DBLP chưa đảm bảo được tính đầy đủ và cập nhật dữ liệu.

Trong các bài báo [2], [3] tác giả giới thiệu về ACI(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. Hiện nay đối với các thư viện số như ACM, IEEEXplore, Spinger … khi người dùng muốn download một bài báo thì sẽ mất phí như vậy ACI chưa thể thu thập được dữ liệu chỉ mục các bài báo này trong khi đó thông tin của các bài báo đã được cung cấp trên thư viện số.

Trong các bài báo [7], [8], [9] giới thiệu một số hệ thống sử dụng Crawler kết hợp với các thuật toán để thu thập thông tin các bài báo trên Internet. Những kết quả trả về từ hệ thống trên chưa đảm bảo được tính đầy đủ khi sử dụng crawl thu thập thông trên các web không cho phép crawl cũng như độ chính xác của các thuật toán rút trích thông tin bài báo sử dụng trong các hệ thống trung bình 60%.

1. **Phương pháp thực hiện.**
   1. Kiến trúc hệ thống.

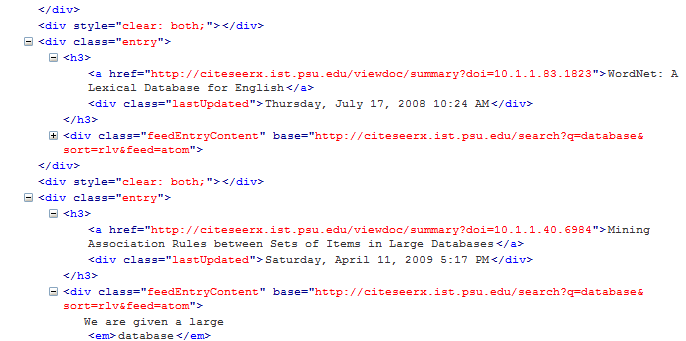


*Hình 1: Kiến trúc DBSA*

Trong hình 1, chúng tôi giới thiệu kiến trúc của ứng dụng, 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 với dữ liệu của DBLP và cập nhật dữ liệu của khi DBLP.

3.1 Crawl bài báo trên thư viện số.

Hệ thống sử dụng từ khóa được nhập vào từ người dùng để thu thập các đường dẫn chứa thông tin các bài báo có nội dung tương ứng với từ khóa nhập vào. Ví dụ đối với từ khóa Database trên thư viện số Citeseer:



Hình 2: Danh sách các đường dẫn sau khi crawl

Sau khi phân tích nội dung trả về thì kết quả quá trình thu được là danh sách các đường dẫn tới nội dung của các bài báo trên thư viện số giống như bài báo của Sheth [[7]](#footnote-7).

* 1. Rút trích thông tin chỉ mục.

Từ đường dẫn đến nội dung bài báo hệ thống sẽ sử dụng bộ phân tích HTMLParser[[8]](#footnote-8), SAXParser[[9]](#footnote-9) và các luật (Regular Expression) để rút ra các thông tin chỉ mục của các bài báo hay các file Bibex[[10]](#footnote-10).



Hình 3. Nội dung chi tiết thông tin bài báo.

Hình 3 hiển thị nội sùn HTML của bài báo của ví dụ trên. Từ nội dung HTML này hệ thống sẽ sử dụng các trình phân tích cũng như các luật để rút ra thông tin của bài báo bao gồm: tiêu đề bài báo, tên tác giả, tóm tắt, năm xuất bản, nhà xuất bản, đường dẫn tới bài báo.

Đối với thư viện có file Bibtex thì hệ thống sử dụng các luật để phân tích file bibtex để lấy thông tin bài báo. Sau đây là ví dụ của các luật để rút thông tin bibtex sử dụng trong thư viện số ACM

ACMStartGetBibtex###http://portal.acm.org/exportformats.cfm?id=%%%The first part of the link to get bibtex file base on bibtex id.

ACMEndGetBibtex###&expformat=bibtex%%%The last part of the link to get bibtex file.

ACMStartGetAbstract###http://portal.acm.org/tab\_abstract.cfm?id=%%%The first part of the link to get abstract part in the article base on article ID.

ACMEndGetAbstract###&usebody=tabbody%%%The last part of the link to get abstract part base on article ID.

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

Để tránh dữ liệu trùng lắp, hệ thống có một module để kiểm tra dữ liệu bài báo thu thập về với dữ liệu của DBLP cũng như các bài báo mà hệ thống thu thập được trước đó.

* 1. Hệ thống xây dựng dữ liệu chỉ mục các bài báo khoa học sử dụng Webcrawler.

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.*

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 sử dụng WebCrawler trên các 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 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).

# [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://en.wikipedia.org/wiki/Webcrawler> [↑](#footnote-ref-1)
2. <http://www.acm.org/> [↑](#footnote-ref-2)
3. <http://ieeexplore.ieee.org/search/> [↑](#footnote-ref-3)
4. <http://citeseerx.ist.psu.edu/> [↑](#footnote-ref-4)
5. <http://www.informatik.uni-trier.de/~ley/db/> [↑](#footnote-ref-5)
6. <http://en.wikipedia.org/wiki/Computer_science> [↑](#footnote-ref-6)
7. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.118.6768> [↑](#footnote-ref-7)
8. <http://htmlparser.sourceforge.net/> [↑](#footnote-ref-8)
9. <http://download.oracle.com/javase/1.4.2/docs/api/javax/xml/parsers/SAXParser.html> [↑](#footnote-ref-9)
10. <http://en.wikipedia.org/wiki/BibTeX> [↑](#footnote-ref-10)