Mục Lục

[1. Cách thức thu thập các bài báo từ thư viện số ACM 2](#_Toc278484207)

[2. Cách thức thu thập các bài báo từ thư viện số IEEExplore. 4](#_Toc278484208)

[3. Cách thức thu thập các bài báo khoa học từ thư viện số citeseerx. 8](#_Toc278484209)

# 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=](http://portal.acm.org/results.cfm?query=computer20%vision&dl=ACM&coll=Portal&short=0)****[computer20%vision](http://portal.acm.org/results.cfm?query=computer20%vision&dl=ACM&coll=Portal&short=0)****[&dl=ACM&coll=Portal&short=0](http://portal.acm.org/results.cfm?query=computer20%vision&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.*

# 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.*

# 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 [[1]](#footnote-2) 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 [[2]](#footnote-3) – 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. <http://www.xml.com/pub/a/2004/06/16/dive.html> [↑](#footnote-ref-2)
2. <http://en.wikipedia.org/wiki/Simple_API_for_XML> [↑](#footnote-ref-3)