**ACM (Association for Computing Machinery )[[1]](#footnote-2)**

ACM cung cấp một thư viện số mà tại đó người dùng có thể tìm kiếm các bài báo được công bố bởi ACM thông qua The ACM Digital Library, hoặc người dùng cũng thể tìm các bài báo,tạp trí ở các hội nghị khác thông qua The Guide.

Đối với The ACM Digital Library  chúng ta có thể tìm kiếm bằng một từ khóa bao gồm :

+ Thông tin về tác giả bài báo ([*Names*](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=4)*,* [*Institutions*](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=3208)*,* [*Authors*](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=2)*,* [*Editors*](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=3)*,* [*Reviewers*](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=3&dim=3209)*. )*

*+ Thông tin về nơi công bố (*[Publication Year](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3221), [Publication Names](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=1), [ACM Publications](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3224), [All publications](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3225), [Content Formats](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3210), [Publishers](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=6&dim=3213)*).*

*+ Thông tin về hội nghị (*[Sponsors](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=5&dim=3207), [Events](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=5&dim=3198), [Proceeding Series](http://portal.acm.org/results.cfm?coll=portal&CFID=100715260&CFTOKEN=11391884&query=datamining&dl=ACM&dimgroup=5&dim=3206). *)*

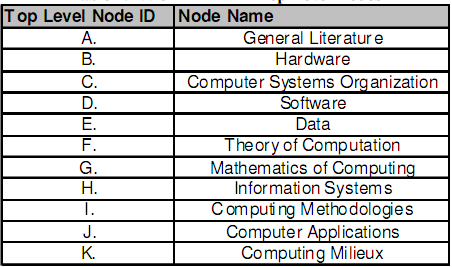
Thông tin lấy được từ bài báo bao gồm :

* abstracts
* citings (where the paper has been referenced by other papers)
* references (by the paper to other papers)
* index terms from ACM's [Computing Classification System (CCS)](http://www.acm.org/class/)
* reviews from ACM's Computing Reviews.

Người dùng không được download tài liệu về nếu không có tài khỏan.

Các bài báo được phân chia chuyên đề theo khung phân loại được định nghĩa bởi ACM, khung phân loại dạng cây và được định nghĩa và bổ xung bởi người dùng . Khi một bài báo đưa lên thư viện số người dùng sẽ dựa vào hướng dẫn của ACM để chọn đúng chủ đề cho bài được đăng lên. Khi bài báo được submit lene trên trang ACM sẽ có một đội ngũ review bài của người đăng.

CCL Cấu trúc của ACL gồm 1473 node được chia trên cây phân lớp thành 3 cấp, cấp đầu tiên bao gồm 11 node, cấp sau đó gồm 81 node và còn lại thuộc cấp số 3. Bảng dưới đây hiên thị các node trong cấp đầu tiên của bộ phân lớp.



**CCS classification tree** [[2]](#footnote-3).

**CiteSeer [[3]](#footnote-4)**

Là một thư viện số mà tài liệu được thư viện cung cấp chủ yếu là về lĩnh vực máy tính. Thư viện số này dùng hệ thống Autonomous Citation Indexing (ACI) để đánh chỉ muc và tím kiếm tài liệu từ đó tạo cơ sở để người dùng có thể tìm kiếm được các bài báo. Khi ngườ dùng tra cứu thông tin về một bài báo thì hệ thống sẽ trả về các thông tin sau: Link download, các thông tin metadata (abtract, title, year).

======================24/8=========================

Tìm hiểu về DBLP (Digital Bibliography & Library Project) :

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 [Germany](http://en.wikipedia.org/wiki/Germany). Ban đầu trang web chỉ tập trung vào cung cấp các thông tin về lĩnh vực DataBase systems và Logic Programming, sau đó hệ thống được mở rộng sang các mục khác trong khoa học máy tính.

Tính đến tháng 1/2010 DBLP chứa thông tin của 1,3 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.

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 dang CDF, XML, SQL người phát triển có thể download các file dữ liệu này từ trên web của chương trình.

# Hiện nay có một số ứng dụng được xây dựng cho phép tìm kiếm các thông tin về các bài báo được rút ra từ dữ liệu của DBLP trong đó có CompleteSearch DBLP , [Faceted search](http://dblp.l3s.de) và DBL – Brown phổ biến và nhiều người dùng.

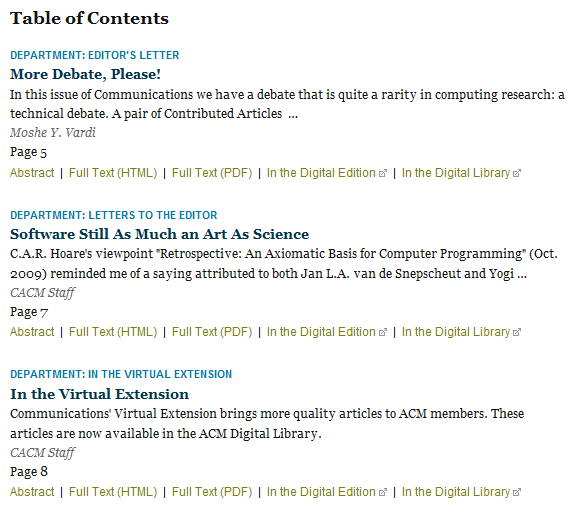
# CompleteSearch DBLP : cho phép tìm kiếm các thông tin từ người dùng đưa vào.

* [Faceted search](http://dblp.l3s.de) cho phép người dùng tìm kiếm thông tin theo những cách sau: theo metadata, theo tác giả và nơi công bố.
* DBL – Brown : Là chương trình sử dụng để tìm kiếm trên file dữ liệu DBLP offline Chương trình cho phép nhìn một cách trực quan về dữ liệu của một bài báo.

Ban dầu DBLP là một tập các danh sách các đề mục – mục lục (tables of contents –TOCs ) của các hội nghị và tạp trí trong lĩnh vực dữ liệu và lập trình logic . Các file TOCs được nhập bằng tay và theo định dạng HTML được dẫn link tới trang giới thiệu bằng thủ công .

Bước tiếp theo hệ thống sẽ phân tích file TOCs để xây dựng một "author pages" . Một Author Page của một tác giả chứa danh sách các bài viết của tác giả và các bài viết có tác giả tham gia (đồng tác giả ). Kết quả thu được từ sau bước này là file TOC\_OUT chứa thông tin chỉ mục của page hay tạp chí mà ta nhập vào.

Chương trình sử dụng kết hợp một bộ parser tên mkhtml và file chứa tất cả tên của các tác giả tiến hành tạo thành file của chương trình.



# TOCs ACM January 2010 (Vol. 53, No. 1)

# 

# TOC – Out

# 

# 

# Author Page

# -----------------------------------------------------------------------------------------------------------------------------------------------------------1/9 ------------------------------------------------------------------------------------------------------------------------------------------------------------------

# Các khái niệm trong rút trích cũng như xử lý thông tin:

# Inverted Index. An inverted index is an index structure that allows efficientretrieval of documents.

# 

# Tokenization:

# 

# Linguistic Preprocessing : stop words, stemming and lemmatization

# 

# -------------------------------------------7/9………………………………………………….

# ……………………………………………………………………………………………

# Khảo sát thư viện số : ScienceDriect :

# <http://www.sciencedirect.com>

# Sciencedriect là một thư viện số có thu phí, cung cấp cho người dùng sách báo và tài liệu thuộc rất nhiều lĩnh vực trong đó có lĩnh vực công nghệ thông tin. Hiện này trong cơ sở dữ liệu của nó chứa hơn 10 triệu bài báo, tạp chí và các cương sách. Trong đó có khoảng 2.500 tạp chí và 10 nghìn quyển sách, trung bình một năm có nửa triệu thông tin bài báo,tạp chí,sách được đưa thêm vào.

# Sciencedriect là một thư viện số có thu phí của người dùng,

# Khi người dùng tìm kiếm một bài báo trên thư viện số này, hệ thống cho phép người dùng tìm kiếm theo các trường sau:

# Tên tác giả (Author).

# Tựa đề sách báo, tạp chí (Journal/Book title).

# Tên bài báo.

# Các tác giả của bài báo.

# 

# Trong thư viện số cho phép người dùng duyệt theo subject của bài báo, tạp chí.

# Kết quả tìm kiếm trả về từ thư viện số của một bài báo bao gồm các thông tin sau :

# Tên tác giả bài báo.

# Tên tựa đề bài báo.

# Phần tóm tắt của bài báo (Abstract).

# Keywords của bài báo.

# Phác thảo về cấu trúc của bài báo.

# 

# -------------------------------------------17/9………………………………………………….

# ……………………………………………………………………………………………

# Doc về thuật toán rút trích thông tin từ ACM và bibtex file :

# Bibtex: là một định dạng của bài báo trong đó có cấu trúc chưa thông tin của bài báo

# Thuật tóan rút trích thông tin bài báo khoa học từ ACM :

# + Các Partem

# Pattern hitsPattern = ".\*Found <b>(\\d+,\*\\d\*)</b> of.\*"

# Pattern tìm kiếm số kết quả trả về từ thư viện số

# Pattern maxHitsPattern = ".\*Results \\d+ - \\d+ of (\\d+,\*\\d\*).\*"

# Pattern kết quả lớn nhất của Page trả về.

# Pattern bibPattern = popBibTex.cfm.\*)','BibTex'.\*"

# Pattern tìm kiếm file bibtex trong kết quả trả về của một link bài báo khoa học.

# Pattern absPattern = ABSTRACT</A></span>\\s+<p class=\"abstract\">\\s+(.\*)"

# Pattern tìm kiếm abstract của tài liệu.

# Pattern fullCitationPattern ="<A HREF=\"(citation.cfm.\*)\" class.\*"

# Pattern tìm kiếm các tài liệu tham chiếu .

# ------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

# Tìm kiếm sách báo :

**The use of titles for automatic document classification**

1. Karen A. Hamill,
2. Antonio Zamora

# Article first published online: 22 MAR 2007

## Journal of the American Society for Information Science<http://onlinelibrary.wiley.com/doi/10.1002/asi.4630310603/abstract>

### A Review of Machine Learning Algorithms for Text-Documents Classification

## Journal of Advances in Information Technology, Vol 1, No 1 (2010)

# <http://ojs.academypublisher.com/index.php/jait/article/viewArticle/01010420>

**Representation and Classification of Text Documents: A Brief Review**

IJCA Journal -

# Year of Publication: 2010

# <http://www.ijcaonline.org/specialissues/rtippr/number2/984-107>

Author = "Murugeshan, Meenakshi Sundaram",

author = "Lakshmi, K.",

author = "Mukherjee, Saswati",

title = "A negative category based approach for Wikipedia document classification",

journal = "International Journal of Knowledge Engineering and Data Mining",

volume = "1",

year = "8 April 2010",

pages = "84-97(14)",

url = {http://www.ingentaconnect.com/content/ind/ijkedm/2010/00000001/00000001/art00006"

doi = "doi:10.1504/IJKEDM.2010.032582"

# Gelernter, J, Cao, D., Lu, R., Fink, E., Carbonell, J. (2009). Creating and visualizing fuzzy document classification. IEEE Conference on Man, Systems and Cybernetics, October 11-14, San Antonio, Texas

# <http://www.cs.cmu.edu/~gelernter/>

# [Hierarchical Document Classification Using Automatically Generated Hierarchy](http://users.cis.fiu.edu/%7Etaoli/pub/Li-jiis.pdf)

# . [*Journal of Intelligent Information Systems*](https://www.editorialmanager.com/jiis/default.asp),29(2): 211-230, 2007.

# <http://users.cis.fiu.edu/~taoli/pub/home-new.html>

**A Hybrid Attribute Selection Approach for Text Classification**

# Journal of the Association for Information Systems: Vol. 11: Iss. 9, Article 1.

# <http://aisel.aisnet.org/jais/vol11/iss9/1/>

## Document Classification for Focused Topics

# Journal of the Association for Information Systems: Vol. 11: Iss. 9, Article 1.

2010

Bộ phân lớp tự động của IBM

<http://publib.boulder.ibm.com/infocenter/p8docs/v4r5m1/index.jsp?topic=/com.ibm.p8.doc/developer_help/content_engine_api/guide/autoclassify_concepts.htm>

1. <http://portal.acm.org/portal.cfm> [↑](#footnote-ref-2)
2. <http://www.acm.org/about/class/ccs98-html> [↑](#footnote-ref-3)
3. <http://citeseerx.ist.psu.edu/> [↑](#footnote-ref-4)