*Trước khi làm thì phải update trên server về, vào phần work này để biết đang làm gì !*

*Có 2 file : Tiến, Cường mỗi người làm trên file này sau đó tổng hợp lại sau*

--------------------------------------------------------------------------------------------------------------------------------------------------------12/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

**+ Công Việc đặt ra:**

- Hoàn thành Crawl.

- Tìm hiểu google scholar API.

- Thu thập tài liệu để làm tài liệu training của ngành CS <trước hết là theo cái khung >

- Sử dụng SVN lib.

- Đọc các tài liệu có trong Document, tài liệu đọc rồi thì cho ra phần đã đọc

+ Tóm tắt lại kiến thức có trong tài liệu.

**+ Một số thư viện số có trên mạng cần tìm hiểu**

ACM <http://portal.acm.org/dl.cfm>

IEEE Computer Society <http://www.ieeexplore.ieee.org/Xplore/guesthome.jsp>

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

Search arXiv.org <http://arxiv.org>

The DBLP Computer Science Bibliography : <http://www.informatik.uni-trier.de/~ley/db/>

**+ Ý kiến và câu hỏi của Thầy đóng góp cho nhóm :**

- Xác định chính xác khung phân loại (Tiến có đề xuất theo cái wiki đó đọc file khung phân loại cs ý ), tức là mình phải tìm hiểu một số thư viện số như IEE,ACM rằng những hệ thống đó có cho ta tìm kiếm theo chủ đề, topic không ? Nếu có thì những chủ đề đó nó phân chia như thế nào ,có nhiều cấp không ---> từ đó hoàn thiện cái khung của mình.

- Thu thập các bài báo (một số lượng đủ lớn ) để ta làm training và test thử sau đó .

- Tìm hiểu rõ ràng và cài đặt thử hoặc có thể là dùng lại socuercode của 1 phương pháp phân loại (SVM) phải chi tiết từng bước như là đánh chủ mục làm sao, rút đặc trưng, đánh trọng số làm sao.

- Xây dựng kiến trúc của chương trình ? Trên desktop or web ?

- Các thư viện số có phân loại không ? Nếu có thì phân loại như thế nào ? Lấy được thông tin gì trong web đó về tài liệu (abtract , tile , metadata ….? ) .

- Xác định lại mục tiêu của đề tài ? phân loại để làm gì ? tại sao phải crawl trong khi đó các thư viện có thể nó không cho download tài liệu về và các thư viện số đó đã phân loại rồi có cần thiết phải phân loại không ??????? <gợi ý là crawl các thông tin về bài báo như abtract để phân loại làm dữ liệu làm giàu ontology dùng để lưu vết các bài báo cung cấp các thông tin …… , >.

**Công việc làm :**

* Xây dựng kiến trúc của chương trình.
* Xem các thư viện điện tử được liệt kê ở trên và viết tóm tắt về các thông tin gồm :

+ Có cho tìm kiếm theo chủ đề,tác giả …

+ Khung phân loại của nó nếu có.

+ Thông tin của 1 bài báo lấy được gì ? metadata,abtract, …

Tiến : ACM,Citier,DBLP.

==🡺 Chỉnh sửa lại đề cương !

--------------------------------------------------------------------------------------------------------------------------------------------------------16/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Việc đang làm :**

* Khảo sát các thư viện số .
* Tìm một số bài báo để làm key trong đề tài.
* Chỉnh sửa đề cương.
* Xem crawl có thu được file HTML để rút ra abtract của một bài báo.
* Suy nghĩ kiến trúc của chương trình (các modul, Package, class ….).

**Góp ý của Thầy :**

* Phân lớp tài liệu trong lĩnh vực CS dựa trên phần Abtract của mỗi bài báo .

**=🡺 tìm hiểu các paser để mình lấy các phần abtract của bài báo trong file PDF,HTML,DOC,…**

* Suy nghĩ về việc dính cái tìm kiếm bài báo về (Lấy phần abtract) .
* Dùng Crawl

Tool trong đó cho phép lấy và search thông tin về bài báo từ các thư viện điện tử : <http://jabref.sourceforge.net/index.php>

Phân loại download và phần lấy abtract dựa trên cái gói này chương trình này làm .

--------------------------------------------------------------------------------------------------------------------------------------------------------17/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

* **Tiến :**

+hoàn thành khảo sát các thư viện số .

+Đọc các bài báo để tìm bài báo keyword cho đề tài.

+ Chỉnh sửa laị đề cương .

+ Suy nghĩ các module của chương trình .... đề xuất để thống nhất !

* **Cường :**

+ Download chương trình này chạy thử chức năng search trên các thư viện số của chương trình : <http://biznetnetworks.dl.sourceforge.net/project/jabref/jabref/2.6/JabRef-2.6-setup.exe>

+ Check out source code của chương trình trên về để chạy thử và xem cách họ kết nối lấy các thông tin từ trên thư viện số về : (ACM,IEEE , Citerrr ..) và lấy được thômg tin metadata nếu mà thư viện số cung cấp.

Địa chỉ check out SVN bản mới nhất : <https://jabref.svn.sourceforge.net/svnroot/jabref/branches/beta_2.6>

Địa chỉ check out code kết nối với ACM của chương trình trên :

<https://jabref.svn.sourceforge.net/svnroot/jabref/branches/svnsupport/plugins/net.sf.jabref.imports.acm1>

Gói thư viện của chương trình : <http://biznetnetworks.dl.sourceforge.net/project/jabref/jabref/2.6/JabRef-2.6.jar>

--------------------------------------------------------------------------------------------------------------------------------------------------------23/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Cường :**

* Hoàn thành survey các thư viện số .
* Khảo sát chương trình Jabref .
* Khảo sát http://papercube.peterbergstrom.com/ xem cơ sở dữ liệu nó là gì, lấy làm sao

**Tiến :**

* Viết đề cương .
* Tìm báo key.

--------------------------------------------------------------------------------------------------------------------------------------------------------24/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Cường :**

* Hoàn thành survey về bài báo điện tử để đưa vào đề cương phần các nghiên cứu liên quan.
* Cho phần khảo sát của jabref vào báo cáo.
* Đọc bài của cile biết cách index của thư viện ci, bổ xung vào báo cáo của phần thư viện số ci.
* Bổ xung : dẫn nhập, mục tiêu, ….. bảng kế hoạch.

**Tiến :**

* Hoàn thành khảo sát về DBLP.
* Vẽ xong kiến trúc của hệ thống đưa vào báo cáo.

--------------------------------------------------------------------------------------------------------------------------------------------------------25/8----------------------------------------------------------------------------------------------------------------------------------------------------------------------

* Hoàn thành đề cương in một bản để Thầy đọc và đóng góp ý kiến.
* Tiếp tục nghiên cứu soucecode của Jabref để lấy các thông tin thư viện số về.
* Tìm kiếm các bài báo đề tài về phân loại dữ liệu, đọc sourcecode bài báo liên quan.
* Làm silde báo cáo

-------------------------------------------------------------------------------------------------------**-------------------------------------------------31/8-------------------------------------------------**-------------------------------------------------------------------------------------------------------

<http://code.google.com/p/dblp-expert-finder/> -🡪 đây là 1 đề tài tương tự của nhóm bên nước ngoaì . Vẫn chưa tìm được mail để liên hệ với tác giả !

**Tiến :**

* Chỉnh sửa lại kiến trúc chương trình.
* Chỉnh sửa lại cách đặt vấn đề của đề tài.
* Chỉnh sửa lại khung thời gian làm việc.

**Cường :**

* Tìm hiểu rõ cái Jabref xem lại cái đó dùng để làm gì, nó nói là dùng để cho vào word như endnote thì phải tìm hiểu rõ dùng như thế nào và endNote nó có tác dụng gì ? -🡪 Viết vào slide và báo cáo.
* Viết cách Citeseer sử dụng ACI để index Citations của báo cáo.
* Bắt tay lại việc tìm hiểu cách lấy tài liệu từ thư viện số của jabref <Tuần này là ACM để có gì Thứ 6 trình bày nội bộ nhóm.

Mục tiêu: Chỉnh sửa xong báo cáo để ngày mai nộp cho Thầy !

-------------------------------------------------------------------------------------------------------**-------------------------------------------------1/9 -------------------------------------------------**-------------------------------------------------------------------------------------------------------

Nộp báo cáo và nghe đóng góp ý kiến của Thầy cho đề tài của nhóm.

🡺 Thứ 6 gửi báo cáo tiến độ công việc và kết quả đạt được trong tuần.

**Cường:**

* Bắt tay lại việc tìm hiểu cách lấy tài liệu từ thư viện số của jabref : ACM, Cicier,IEE Xplo

**Tiến:**

* Đọc bài báo cáo của **Philip Fisher-Ogden** 🡺 Viết báo cáo về hệ thống này
* Bắt tay lại việc tìm hiểu cách lấy tài liệu từ thư viện số của Jabref

-------------------------------------------------------------------------------------------------------**-------------------------------------------------7/9-------------------------------------------------**-------------------------------------------------------------------------------------------------------

**Cường :**

* Chỉnh sửa hoàn chỉnh đề cương.
* Tìm hiểu lại cách lấy tài liệu trong Jabref <cái này cần làm ngay ~~!>

**Tiến:**

* Viết lại dẫn nhập của đề cương đưa ra thảo luận.
* Khảo sát thư viện số ScienceDrirect bổ xung vào báo cáo.
* Tìm hiểu lại cách lấy tài liệu trong Jabref.
* Đọc tài liệu của chương trình tương tự,chạy thử đưa lên phần soucecode tham khảo.

-------------------------------------------------------------------------------------------------------**-------------------------------------------------9/9-------------------------------------------------**------------------------------------------------------------------------------------------------------

**Cường**

* Chỉnh sửa hoàn chỉnh đề cương.
* Tìm hiểu lại cách lấy tài liệu trong Jabref <cái này cần làm ngay ~~!>

**Tiến :**

* Đọc tài liệu của chương trình tương tự,chạy thử đưa lên phần soucecode tham khảo.
* Tìm hiểu lại cách lấy tài liệu trong Jabref.

-------------------------------------------------------------------------------------------------------**-------------------------------------------------13/9-------------------------------------------------**-------------------------------------------------------------------------------------------

**Cường**

* Tìm hiểu lại cách lấy tài liệu trong Jabref .

**Tiến :**

* Đọc tài liệu của chương trình tương tự,chạy thử đưa lên phần soucecode tham khảo.
* Tìm hiểu lại cách lấy tài liệu trong Jabref.
* Database.

-------------------------------------------------------------------------------------------------------**-------------------------------------------------17/9-------------------------------------------------**-------------------------------------------------------------------------------------------

**Cường**

* Lấy thông tin của IEEE và một số thư viện số còn lại.
* Viết doc.

**Tiến :**

* Lấy thông tin dựa vào ACM Fetcher.
* Viết lại thuật toán lấy thông tin từ ACM .
* Tìm hiểu về cấu trúc Bibtex của 1 bài báo .
* Đọc LIGHTWEIGHT FEDERATION OF NON-COOPERATING DIGITAL LIBRARIES một luận văn tương tự về lấy thông tin từ các thư viện số.
* Tạo Database và viết lại database của nhóm dựa vào ý kiến của Hoàng .

-------------------------------------------------------------------------------------------------------**-------------------------------------------------20/9-------------------------------------------------**-------------------------------------------------------------------------------------------

**Cường**

* Viết thiết kế chương trình : câu trúc , chức năng có thể nghĩ ra để thống nhất thiết kế chương trình.
* Tiếp tục làm phần của IEEE.

**Tiến :**

* Lấy thông tin dựa vào ACM Fetcher.
* Viết lại thuật toán lấy thông tin từ ACM .
* Đọc LIGHTWEIGHT FEDERATION OF NON-COOPERATING DIGITAL LIBRARIES một luận văn tương tự về lấy thông tin từ các thư viện số.
* **Tạo Database và viết lại database của nhóm dựa vào ý kiến của Hoàng .**

-------------------------------------------------------------------------------------------------------**-------------------------------------------------4/10 ------------------------------------------------------------**-------------------------------------------------------------------------------------------

Hệ thống tương tự của cả hệ thống.

**<http://cindi.encs.concordia.ca>**

-------------------------------------------------------------------------------------------------------**-------------------------------------------------15/10 ------------------------------------------------------------**------------------------------------------------------------------------------------------

**Cường :**

* Hoàn chỉnh thu thập của ACM
* Viết lại thuật toán lấy dữ liệu.

**Tiến :**

* Thống nhất database.
* Hoàn thiện phần hibernate cho kết nối cơ sở dữ liệu.

**Góp ý cho giai đoạn tiếp theo của Thầy :**

***Viết giới thiệu tổng quan về phân lớp :***

* ***Thuộc ngành gì lĩnh vực gì, phần gì.***
* ***Các phương pháp, các hướng nghiên cứu, các thuật toán sử dụng***

***Từ đó đưa ra nhận xét và dựa vào hướng tiếp cận nào để làm việc (chú ý ta không có đánh giá về hiệu quả của vấn đề )***

***==================================21/10==============================================================================================================***

***Tiến :***

* ***Không thể lấy thông tin từ ScienceDriect lý do : không thể thành lập câu query khi nhập từ khóa từ người dùng để lấy được dữ liệu từ thư viện số về, nguyên nhân là do khi thực hiện câu truy vấn hệ thống sẽ mã hóa từ khóa tìm kiếm, việc này là không thể***

***Ví dụ khi tìm kiếm với từ khóa database thì đối với***

***ACM :*** [***http://ieeexplore.ieee.org/search/freesearchresult.jsp?newsearch=true&queryText=database&x=0&y=0c***](http://ieeexplore.ieee.org/search/freesearchresult.jsp?newsearch=true&queryText=database&x=0&y=0c)

***ScientDriect:***

***http://www.sciencedirect.com/science?\_ob=ArticleListURL&\_method=list&\_ArticleListID=1507058552&\_sort=r&\_st=13&view=c&\_acct=C000050221&\_version=1&\_urlVersion=0&\_userid=10&md5=9c4d2376b7e583e10248a3d5b683fd6a&searchtype=a***

***Đây là tìm kiếm theo từ computer***

[***http://www.sciencedirect.com/science?\_ob=ArticleListURL&\_method=list&\_ArticleListID=1507056819&\_sort=r&\_st=13&view=c&\_acct=C000050221&\_version=1&\_urlVersion=0&\_userid=10&******md5=e7ed529e6e06fc719a8211a498b8d1fd&searchtype=a***](http://www.sciencedirect.com/science?_ob=ArticleListURL&_method=list&_ArticleListID=1507056819&_sort=r&_st=13&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=e7ed529e6e06fc719a8211a498b8d1fd&searchtype=a)

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------Công việc phải thực hiện trong 2 tuần--------------------------------------------------------------------------------------------------29/10 – 12/11 ---------------------------------------------------

* Thiết kế database. Tiến
* Hoàn thiện phần hiển thị : Cường
* Kiểm tra trùng lặp dữ liệu Tiến
* Viết thuật toán rút trích từ thư viện số (Cường ACM, Tiến Citeseer,IEEEP)

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------Công việc phải thực hiện trong 2 tuần--------------------------------------------------------------------------------------------------12/11 – 26/11 ---------------------------------------------------

* **Viết báo cáo thuật toán rút trích từ thư viện số (Cường - ACM ,IEEEXplore Tiến - Citeseer ).**
* **Viết khảo sát về rút phân lớp chủ đề (Tiến) :**

Viết giới thiệu tổng quan về phân lớp :

+ Thuộc ngành gì lĩnh vực gì, phần gì.

+ Các phương pháp, các hướng nghiên cứu, các thuật toán sử dụng

Từ đó đưa ra nhận xét và dựa vào hướng tiếp cận nào để làm việc (chú ý ta không có đánh giá về hiệu quả của vấn đề)

* **Hoàn thiện các phần còn lại của chương trình:**

+ Thay đổi partent của ACM Fetcher : Cường.

+ Lấy đúng thông tin của Citeseer (Tiến).

+ Xử lý phân cập nhật, insert,delete, update dataabase trong tab Database.(Cường) .

+ Chuyển các phần handcode các partent xuống file (Cường) . (thể thết kế một layour cho phép chỉnh sửa các partent đó )

+ Thêm vào core các phần về google, googlescholar để search lấy link có liên quan dựa vào title của bài báo (Tiến).

* **Nghiên cứu cách để lấy abstract của các bài báo trong dblp để cập nhật abstract cho các bài báo của dblp (Tiến).**

***[1 ] Sebastiani, F.: Machine learning in automated text categorization. In: ACM Computing Sur-***

***veys, 34(1), Kluwer Academic Publishers (2002) 1–47***

**--------------------------------------------------------------------------------------------------------------------------------------------------------------------------Công việc phải thực hiện trong 2 tuần--------------------------------------------------------------------------------------------------26/11 – 10/12 ------------------------------------------------**

***Tối ưu hóa code trong chương trình bao gồm:***

+ Các handcode: các code về regular expression , các hằng số phải đưa ra file.

+ Các câu thông báo lỗi 🡪 cho hiện ra màn hình không cho hiện ra cosole nữa : nếu có thể new một cửa sổ cho hiện những thông tin lỗi giống trong jabref . kể cả cách câu throw của try catch

+ Những giá trị get được từ file bitext phải làm lại, nếu không get được phải trả về khoảng trắng chứ không trả về null, hoặc khi get lên table cần kiểm tra nếu nó là null thì thay bằng khoảng trắng.

+ Thiết kế giao diện config cho các regular Expresssion của các thư viện , và các con số max khi lấy kết quả - nếu có thể cho người dùng config luân luồng xử lý như sau : get số giá trị trả về từ thư viện số 🡪 kiểm tra giá trị max người dùng nhập vào nếu không thỏa mãn bắt thực hiện lại và show ra màn hình là số kết quả lớn nhất từ page đó là ………… thực hiện với từng thư viện số.

+ Render để thấy khi bấm vào link có thê tới được trang web.

+ Xử lý thêm bớt dữ liệu trong dataabase manager.

***Gợi ý : render để hiển thị hyperlink :***

<http://java-swing-tips.blogspot.com/2009/02/hyperlink-in-jtable-cell.html>

hoặc

<http://stackoverflow.com/questions/527719/how-to-add-hyperlink-in-jlabel>

**Công việc lên kế hoạch trong 2 tuần:**

- Đọc và trình bày nội dung bài báo mà nhóm chọn làm chủ đạo trong việc phân lớp chủ đề!  
  
- Đọc và viết báo cáo về phần các nghiên cứu liên quan đến phân loại chủ đề.  
  
- Đọc và tìm hiểu sử dụng thuật toán SVM -- Bước đầu cài đặt thuật toán.  
  
- Bước đầu cài đặt các công việc liên quan đến phần xử lý dữ liệu (nội dung title,abstract ) bao gồm : tiền xử lý dữ liệu (remove stop word, Stemming ).

**Công việc phải thực hiện tuần này:**

Seminar:

* Những gì đã làm được.
* Công việc dự kiến tiếp theo.

<nếu làm rộng ra thì làm theo kiểu nào – nếu làm xâu cho vấn đề đã làm thì làm cái gì>

========================================================================================================================================27/12/2010==================================================================================================================================================

***Ý kiến của Thầy:***

***- Tập trung và xác định homepage của tác giả . Có thể dùng thuật toán phân lớp để xác định homepage. Tách ra thành bài toán khác.***

***- Cho thời gian 1 tuần để hoàn thiện chương trình dbsa đang có. <bao gồm cả các action của các button, bắt lỗi, chỉnh sửa parten, Giao diện , phần tiếng việt – anh, cắt bỏ các phần không sử dụng tới >.***

***- Tóm tắt – tổng hợp kết quả quá trình làm được để viết 1 bài báo về xây dựng làm giàu dữ liệu chỉ mục.***

* ***Nếu rảnh Cường viết phần hướng dẫn sử dụng cho chương trình đã thực hiện được !***
* ==========================================================================================================================================14/1//2010================================================================================================================================================
* **Trong phần tác giả thêm Thầy và thêm email vào.**
* **Khi không có subject trong database thì không thể bấm add để thêm.**
* **Khi đã xóa hết subject báo lỗi index báo lỗi -1.**
* **Action gọi từ tab database sang tab insert chưa có.**
* **Insert check publish tồn tại trong database- chưa check year là số**
* **Chưa disable nút delete publish khi trong database chưa có bài báo nào.**
* **Đối với các tab insert bài báo thì khi chưa input vào 1 trong các trường đưa ra thì nút insert disenable.**
* **Cho bắt đầu tìm kiếm từ 1 hoặc khi các số người dùng lựa chọn phải bắt đầu từ 1.**
* **Insert danh sách hết các chủ đề trong Wiki vào subject và tạo thành script để nhập sau cùng**

**------------------------------------------------------------------------------------------------------------------------------------------------------27/1 ------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

* ***Giải thích và đưa ra được có cần lưu toàn bộ hay lưu những bài không trùng giải quyết tại sao ?Khi tìm kiếm bài báo thì có thê tìm trên database của mình hay kết hợp với dblp ?? sao tìm những bài báo không có trong dữ liệu dblp nhưng có trong dbsa.***
* ***Tìm kiếm cơ chế tự cập nhật các bài báo theo định kỳ.***
* ***Thêm tab thống kê số liệu.***

***+ Số lượng reco trong dblp.***

***+ Số lượng reco trong dbsa.***

***+ Số lượng kêt qua tim kiem duoc tuong ung voi moi thu vien so. Trong do co ghi co bao nhieu ket qua trung trong dblp, bao nhieu truoc nam 2005 sau 2010.***

***Số kết quả recode của dblp số recode hiện có trong dbsa***

***Tên từ khóa tìm kiếm người dùng nhập vào: ……………***

|  |  |  |  |
| --- | --- | --- | --- |
| ***Thông số*** | ***ACM*** | ***IEEE*** | ***Cite*** |
| ***Số kết quả tìm kiếm đươck*** |  |  |  |
|  |  |  |  |
| ***Sốkết quả không trung lắp trong dblp*** |  |  |  |
| ***Số kết quả trước năm xx*** |  |  |  |

* ***Tách tác giả ra lưu riêng.***
* ***Thêm đoạn scrip cho phép delete cập nhật lại database của dblp.***
* ***Load tên tác giả, và concept chọn. Đã làm xong nhưng không thể load file tác giả 13M đc. Bị lỗi Out Of Memory***
* **FIX Bug :** 
  + **search nhiều lần liên tiếp không lưu được bài báo vào.**
  + **Đối với các tab insert bài báo thì khi chưa input vào 1 trong các trường đưa ra thì nút insert disenable.**
  + **Thêm các tooltip khi rê chuột vào các button.**
  + **Thêm checkall vào phần delete bài báo.**
  + **Lỗi khi set defau parten : Hướng giải quyết clear dữ liệu người dùng nhập vào.**
  + **Phần tiếng việt <hàm riêng khi set Text các frame>.**
  + **Compelte 🡪 Done.**
  + **Chỉnh lại tên chương trình.**
* **Đối với thêm tác giả:**

**+ Cân đối lại các ô trong các mục thêm vào.**

**+ Tạo nút open đến file để người ta copy thông tin vào.**

**+ Thêm tác giả như thế nào?**