International Journal of Computer and Communication Technology

Volume 1 | Issue 1 Article 4

January 2010

A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)

Pooja Gupta
Guru Gobind Singh IndraPrastha University, Maharaja Agrasen Institute of Technology, Sec-22, PSP Area,Rohini.,Delhi-85, poojaguptamait@gmail.com

Ashok Sharma
MDU Rohtak, YMCA Institute of Engineering, Faridabad, Sec-3, Faridabad,Haryana, ashokkale2@rediffmail.com

J. P. Gupta
Jaypee Institute of Information Technology University (JIIT), A-10, Sector-62, Noida-201307, jpgupta@jiit.ac.in

Komal Bhatia
YMCA Institute of Engineering, Faridabad, Sec-3, Faridabad, Haryana, kbhatia@rediffmail.com

Follow this and additional works at: https://www.interscience.in/ijcct

Recommended Citation

Gupta, Pooja; Sharma, Ashok; Gupta, J. P.; and Bhatia, Komal (2010) "A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)," *International Journal of Computer and Communication Technology*: Vol. 1 : Iss. 1 , Article 4.
DOI: 10.47893/IJCCT.2010.1003
Available at: https://www.interscience.in/ijcct/vol1/iss1/4

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer and Communication Technology by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)

Pooja Gupta*

Guru Gobind Singh IndraPrastha University,
Maharaja Agrasen Institute of Technology,
Sec-22, PSP Area,Rohini.,Delhi-85
Email:poojaguptamait@gmail.com
*Corresponding author

Ashok Sharma

MDU Rohtak,
YMCA Institute of Engineering, Faridabad,
Sec-3, Faridabad,Haryana
Email:ashokkale2@rediffmail.com

J. P. Gupta

Jaypee Institute of Information Technology University (JIIT), A-10, Sector-62, Noida-201307
Email: jpgupta@jiit.ac.in

Komal Bhatia

MDU Rohtak ,
YMCA Institute of Engineering, Faridabad,
Sec-3, Faridabad, Haryana.
Email: kbhatia@rediffmail.com

Abstract: Focused crawling aims to search only the relevant subset of the WWW for a specific topic of user interest; leading to the necessity to decide about the relevancy of a document to the topic of interest; especially when the user is not perfect in specifying the exact context of the topic. This paper provides a novel framework of a context based distributed focused crawler that maintains an index of web documents pertaining to the context of keywords resulting in storage of more related documents.

Keywords: Focused crawling; context; web documents; topic specific; crawler; word net; world wide web ; WWW; search engine; web index; query keywords; contextual information.

Reference to this paper should be as follows: Gupta,P, Sharma, A., Gupta, J. P. ,Bhatia, K. (2009) 'A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)', *Int. J.CCT,* Vol. 1 , No. 1 , pp.14-26.

Biographical notes: Pooja Gupta received the MCA degree with Gold Medal in 2002 and M.Tech degree with honours in Computer Science Engineering in 2006, both from Maharishi Dayanand University. Presently, she is working as a lecturer in Computer Science and Engineering Department in Maharaja Agrasen Institute of Technology (affiliated to I.P. University) Rohini, Delhi. She is also pursuing her Ph.D. in Computer Engineering and her areas of interests are Search Engines, Crawlers and Focused Crawling.

Prof. A. K. Sharma received his M.Tech. (Computer Sci. & Tech) with Hons. From University of Roorkee in the year 1989 and Ph.D (Fuzzy Expert Systems) from JMI, New Delhi in the year 2000. From July 1992 to April 2002, he served as Assistant Professor and became Professor in Computer Engg. at YMCA Institute of Engineering Faridabad in April 2002. He obtained his second Ph.D. in IT from IIIT & M, Gwalior in the year

Copyright © 2009 IPM (P) Ltd.

2004. His research interests include Fuzzy Systems, Object Oriented Programming, Knowledge Representation and Internet Technologies.

Prof. J.P. Gupta is the first Vice-Chancellor of JIIT, after its notification as a deemed University in November 2004. He assumed the charge of the Vice-Chancellor on July 09, 2005. He has a dynamic and extremely distinguished academic profile. He has the panache of achieving the best. An alumnus of Banaras Hindu University, Prof. Gupta obtained Master's Degree in Electronics & Communication Engineering with the Gold Medal in 1973 from the University of Roorkee. He obtained his Doctorate Degree in Computer Engineering from the University of Westminster, London under the Commonwealth Scholarship Award. In a meritorious academic career, he has held many coveted teaching and important administrative assignments. He held a position of Professor serving the University of Roorkee (now IIT, Roorkee) for over 25 years. Prof. Gupta has been first full time Member Secretary, All India Council for Technical Education (AICTE) (1994-98). He was founder Director, JIIT in 2000.

Komal Kumar Bhatia received the B.E. and M.Tech. degrees in Computer Science Engineering with Hons. from Maharishi Dayanand University in 2001 and 2004, respectively. He received his Ph.D. in Computer Engineering from Maharishi Dayanand University in 2009 and presently working as Assistant Professor in Information Technology department in YMCA Institute of Engineering, Faridabad. His research interests include Search Engines, Crawlers and Hidden Web.

1 Introduction

The WWW is a huge repository of hyperlinked documents accessible through Internet. It is an information-sharing model that is built on the top of the Internet. Owing to its growing source of data and information WWW has become an important resource for research that operates within the Internet's basic client-server architecture.

User searches for information related to a topic of interest by using a search engine. A traditional search engine comprises of three core components: download, index and search. Out of three the "download" is the most important activity wherein the web is traversed to download web pages by following links. A copy of all visited pages is stored for later use.

Currently there are more than some billion documents on the web and it continues to grow at an exponential rate. There may be tens of thousands of slightly relevant documents; from the user's point of view. A normal search engine returns thousands of matches in response to a user query. The size of the information is too large to go through, leading to the problem of information overkill. This problem aggravates in the case of inexperienced user trying to search the information on the web. In line with the famous "8 second rule", such users look at first few results and tend to turn away. This behaviour prompts the need to develop tools with very high precision that may provide links to relevant documents that too listed among the top ten of the total list of links, rendering search engines and related technologies a potential area of research.

Since the user enters only a combination of keywords to search information without giving a thought about the context, the crawler therefore search without concerning the user's context of search; providing results which may or may not fulfil the requirements.

Hence there is a need to design a crawler that is able to download topic specific documents from the web catering to the need of a user in search of highly relevant documents. In this paper a novel framework for focused crawling that utilizes topical locality on the web in order to perform resource discovery is being presented.

2 Related work

Most of the focused crawlers in the early age were based on the probability method; that is the relevancy of the document is determined based on some probability function as appears in [1] [2] [3] [4] [5].

Some recent literature results are on the basis of link semantics. LS Crawler [6] performs the searching on the semantic basis. It enhances the process of determining the relevancy of the documents before downloading. It generates a repository of the most relevant documents with better recall. It extracts the hypertext from the documents and then measures their relationship with the search keyword in the specific domain, by referring to the Ontology. Based on the relevancy of the terms, priority is assigned to the URL and the pair URL, its priority saved in a list for the extraction of the documents. The search is based on the matching keyword with the each term in the hyperlink.

Another research done is based on the content of the document and on link analysis as well. HAWK crawler [7] is implemented using user-defined relevant formula, shark-search and Page-Rank. It Matches the crawling page with the topic of search and computes the relevant score if it is higher than a specified value then get its entire child link and apply same to that too. Once the downloading of all relevant documents is completed; it computes the potential score of the page i.e. similarity to topic and anchor-text. If this is more than a specified value then compute anchor-text context. So it works over the similarity of topic with content and then with link.

Another literature CDFC [8] [9] has identified the need of contextual information to be sent along with the query terms and proposed method to represent the context in the form of augmented hypertext documents. It is an agent-based context driven focused crawler. It gets the user input and then presents the user different meaning of the input keyword for further selection of specific context. This is done by storing the volatile information and the context and the keywords in separate file say TVI (table of volatile information) and TOC (table of context) for every augmented hypertext document. These files are retrieved by agents in spite of the whole documents and are indexed by crawler. TOC information is presented to the user for their specific search.

A critical look at the available literature indicates that the existing focused crawler suffers from following drawbacks:

- Crawlers are not able to analyse the context of the keyword in the web page before they download it.
- The user submits his request for retrieval of information without explicitly mentioning the context in which he or she otherwise desire.
- Crawler treats user search requests in isolation
- Results returned are identical, independent of the interest of the user.
- There is a need to prepare separate files for each web document (TOC, TVI).
- Augmentation is required in HTML documents

In this paper a novel framework for Context based distributed focused crawler CBDFC) is being presented. It works on the different contextual interpretations of the keywords entered by the user in the form of query and the keywords present in the web documents. It presents these different contextual meaning of the keywords to the user to select the particular context of search, thus results in more specific web pages to accomplish the requirement.

3 Role of Word Net towards Extracting Contextual Meanings for Query Keywords

Since Google maintains only the word lexicon it returns synonyms words pair in alphabetical order in response to a given query keyword as discussed by Brin and Page [10]. For instance, in response to the query keyword “Student” the Google search engine returns a list of synonyms words pair as shown in Figure 1-3 and listed in Table I. The results for the query keyword ‘Spider’ are shown in fig 4 and listed in Table 2.

On the contrary when the same Keywords are tested on Word Net [11] it results with the list of distinct contextual pairs of the words that are represented as different senses of the word. The results for keyword 'student' are shown in Figure 5 & 6 and listed in Table 1. The results for keyword 'Spider' are shown in Figure 7 and listed in Table 2.

Table 1 Comparison for "Student"

Keyword	Meaning (Google)	Contextual meaning (Word Net)
Student	Student Loans Student universe Student portal Student flights Student loan consolidation Student doctor network ----------------	***(1st Sense)*** Pupil Educate Art student Medical student Scholar Non-reader Seminarian....
		(2nd Sense) Academician Graduate Scholiast Booklover, Reader....

Similar comparative analysis for a number of query keywords has been done and the results show that the Word Net produces in more meaningfully related pairs of words. Hence Word Net has its significance in the CBDFC to make the search proceed in a specific direction for specific or more related words.

Table 2 Comparison for "Spider"

Keyword	Meaning(Google)	Contextual meaning (Word Net)
Spider	Spiderman Spider solitaire Spiderman games Spider monkey Spider wick chronicles -------------------	***(1st Sense)*** Predatory arachnid Comb-footed spider Trap-door spider...
		(2nd Sense) Wanderer Program Computer programme...
		(3rd Sense) Skillet made of cast iron Frying pan Frypan...

4 Proposed Architecture

As we know that generally the end user is not a frequent surfer of the web to search the information and is not good in giving the search keywords related to its query as well. The same query keyword may lead to different interpretations and the crawler has no means to identify the user's interpretation and thus search the web in isolation to the user's interest.

It is concluded from the previous section that Word Net has its significance to produce the different contextual meanings and senses for a given query keyword, that can be used to index the web documents database that finally will help the end user to specify the topic of interest. In the proposed CBDFC indexing of database is done on these different senses produces by the Word Net.

For a given query, CBDFC (context based distributed focused crawler) consults its own database that is already indexed on different contextual meanings, to extract most suitable interpretations/senses corresponding to the keywords and presents to the user for selection. For examples some of the keywords entered by the user and its different interpretations are shown in Table 3:

Table 3 Keywords different interpretations /senses

Keyword	Different interpretations / senses
Student	Scholar Apprentice Undergraduate Learner
Spider	Computer Program Game of cards Insect

User selects his interest of search from the list of different contextual interpretations. CBDFC now uses this context to consult its database, which is indexed according to context of keywords embedded within documents. The search within database is carried out in search-insert fashion. The architecture of proposed Context Based Distributed Focused Crawler (CBDFC) is given in Figure 8. It consists of three layers. The bottom layer continuously searching the web servers and keep downloading the web pages, storing them in the local database.

Middle layer i.e. the indexing agent and the back-link extractor indexes the local database and enrich the database for different topics and contexts. This helps the top layer to response to the user query. It only searches the local database index and thus presents the quick response to the user query allowing him to choose /specify the context of the search topic. Hence, CBDFC results in a list of more specific web pages to accomplish the user requirement for a specific topic of search.

The detailed working of components of three layers is shown in table 4. All the components except the Crawl Worker works on the search engine side. The multiple instances of the Crawl Worker have been employed and works under the control of Crawl Manager.

The data flow within the CBDFC's components shown in Figure 9.

- Crawl manager gets the list of seed URLs. Distribute the list in multiple crawl workers those search the web servers and download the documents corresponding to the URLs assigned to them and return back to the crawl manager with the downloaded pages.
- Crawl manager stores the downloaded web documents/pages in the local database.
- Indexing agent extract the keywords from the web documents stored in the local database.
- Then it extracts the different contextual interpretations/senses of these keywords from the Word Net dictionary.
- It then prepares an index of the local database on the basis of extracted different contextual meaning and senses.
- Back-link extractor recursively finds the in-links of the URLs stored in the local database index.
- It forwards these in-link URLs to the crawl manager for further downloading and gets the downloaded pages.
- For all the in-link downloaded documents it extracts the keywords and gets the different contextual interpretations/senses from indexing agent.
- For the in-link documents, back-link extractor matched the keywords' contexts with the context of the first document and if it matched to some degree then it updates the index in search -insert fashion. This continues recursively till the downloaded page is out of context.
- User agent gets the end user query keywords from the console.

- It finds out a match for the keyword in the local database index in search-insert fashion and in case of a match gets the different contextual interpretations/senses of that keyword.
- Display these different contextual interpretations/senses to the user to enable him to select a specific context of the search.
- Get the user selection for a particular context.
- Searches the local database index for this particular context and gets the list of matched URLs.
- Display the web documents corresponding to the matched URLs to the end user.

The above technique of indexing database based on different contextual interpretations was tested over a number of keywords manually and it has been observed that the results are much relevant and specific to the topic of interest by using the context based indexed database as compared to the other existing techniques.

Table 4 Components and their Functionalities

Component	Functionality
Crawl Manager	Core component of the first layer of the framework. It works on the list of seed URLs. Distributes these URLs to multiple crawl workers for downloading. Receives the downloaded web documents from them and stores them in the local database.
Crawl Worker	This component of first layer works on the server-side and is under the control of crawl manager. It downloads the web documents for the list of URLs assigned by Crawl manager by visiting the different web servers under its range of search.
Indexing Agent	Second layer component that indexes the local database. It extracts the keywords from the web documents stored in the local database and gets the different contextual interpretations/senses of these keywords from the Word Net dictionary. Generates an index of local database based on the contexts of keywords. Maintain the index in the form that it stores the keyword, contextual interpretations and the corresponding URLs and keep updating it in search-insert fashion.
Back-link Extractor	Second layer component that recursively finds the in-links of the URLs stored in the local database index and sends them to Crawl manager for further downloading. Extract the keywords from these in-links documents and gets the different contextual interpretations with the help of indexing agent. If the context matches to that of the first page of recursive process to some extent or degree, add it to the index under the same context. This recursive process continues till an in-link page results in out of context.
User Agent	The top layer component of the framework, which act as an interface between user and the whole system. It accepts the user's query keywords from the console. Then search the index in search-insert fashion to get a match; if a match found, displays its different contextual interpretations/senses to user for further selection of specific context of the search and finally displays the web documents that matched with the context from the local database. Otherwise the context is added to the index for future search to update the database in that topic too.

Example

In the proposed architecture the documents are stored in a repository and index is maintained, that responses to the user query. So, for a keyword the match is first found from the index in alphabetical order. Each alphabet points to the hierarchy of keywords starting from the same alphabet e.g. keyword associated with alphabet ‘S’ will point to the hierarchy containing student, scholar, skillet, sneak, server etc.

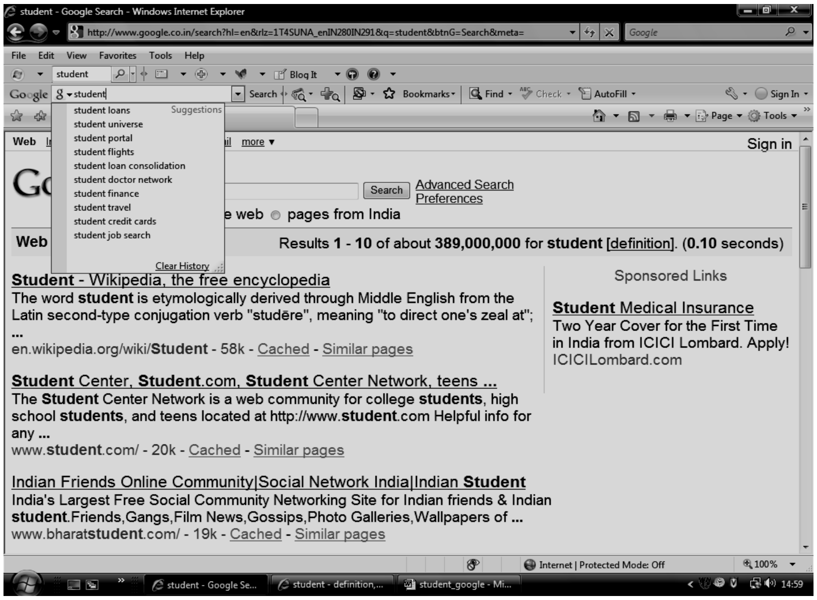
Each keyword in the hierarchy will point to a list containing their contextual meaning. E.g. Keyword “Server” will point to a list containing Waiter, Host, Server (Utensil), Server (Court Game) etc. Each of these contextual meanings points to the corresponding web documents stored in repository as shown in Figure 10.

So, whenever a query keyword comes, first the alphabetical index is searched for a match and then for the related match the corresponding list of keywords is searched to get a match. Then corresponding attached list of contextual meaning is displayed to the user to get the specific choice. Thereafter for that specific choice the pointed list of document is extracted from the repository to fulfill the user requirement. For instance, query for keyword “Server” is shown in Figure 10 where user has selected the meaning “Host” and corresponding list of documents i.e. D1, D5, D6, D11 will be displayed as the results of search.

5 Conclusion

In this paper a framework for CBDFC has been proposed that indexes the web pages with the help of Word Net [11]. Indexing the database is the core of the whole framework and as it is based on the different contextual senses of the query keyword, it focuses the search to the specific context of user interest. It also uses a novel technique to extract the back-links of a URL recursively until it finds the irrelevant links, enriching the database. It may be noted from the result analysis that the proposed framework results in more relevant and topic specific web documents.

Figure 1 Result of Google for keyword “Student”



Results **1** - **10** of about **389,000,000** for **student** [definition]. (**0.10** seconds)

Student - Wikipedia, the free encyclopedia
The word **student** is etymologically derived through Middle English from the Latin second-type conjugation verb "studēre", meaning "to direct one's zeal at"; ...
en.wikipedia.org/wiki/**Student** - 58k - Cached - Similar pages

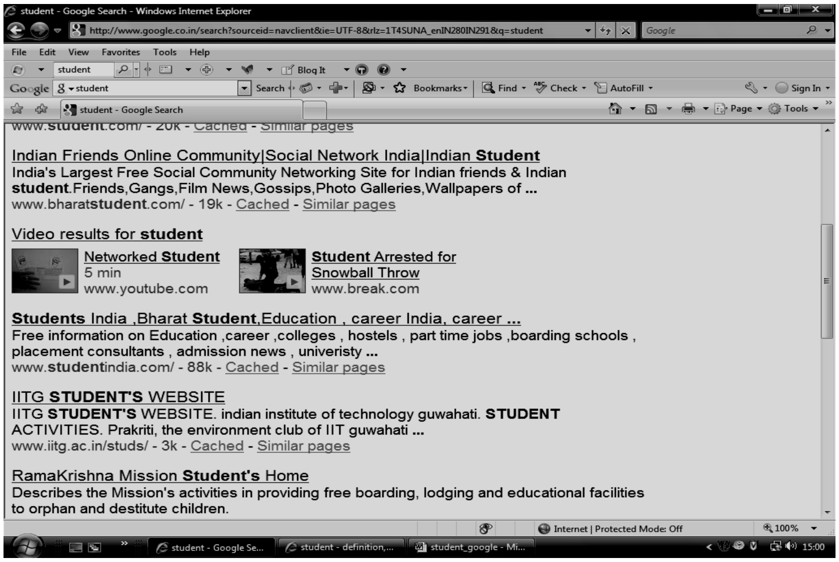
Student Center, **Student**.com, **Student** Center Network, teens ...
The **Student** Center Network is a web community for college **students**, high school **students**, and teens located at http://www.**student**.com Helpful info for any ...
www.**student**.com/ - 20k - Cached - Similar pages

Indian Friends Online Community|Social Network India|Indian **Student**
India's Largest Free Social Community Networking Site for Indian friends & Indian **student**.Friends,Gangs,Film News,Gossips,Photo Galleries,Wallpapers of ...
www.bharat**student**.com/ - 19k - Cached - Similar pages

Sponsored Links

Student Medical Insurance
Two Year Cover for the First Time in India from ICICI Lombard. Apply!
ICICILombard.com

Figure 2 Result of Google for keyword “Student” (Contd.)


student - Google Search - Windows Internet Explorer
http://www.google.co.in/search?sourceid=navclient&ie=UTF-8&rlz=1T4SUNA_enIN280IN291&q=student
File Edit View Favorites Tools Help
student
Blog It
Google student Search Bookmarks Find Check AutoFill Sign In
student - Google Search
Page Tools
Internet | Protected Mode: Off
100%
student - Google Se... student - definition,... student_google - Mi...
15:00

www.**student**.com/ - 20k - Cached - Similar pages

Indian Friends Online Community|Social Network India|Indian **Student**
India's Largest Free Social Community Networking Site for Indian friends & Indian **student**.Friends,Gangs,Film News,Gossips,Photo Galleries,Wallpapers of ...
www.bharat**student**.com/ - 19k - Cached - Similar pages

Video results for **student**

Networked **Student**
5 min
www.youtube.com

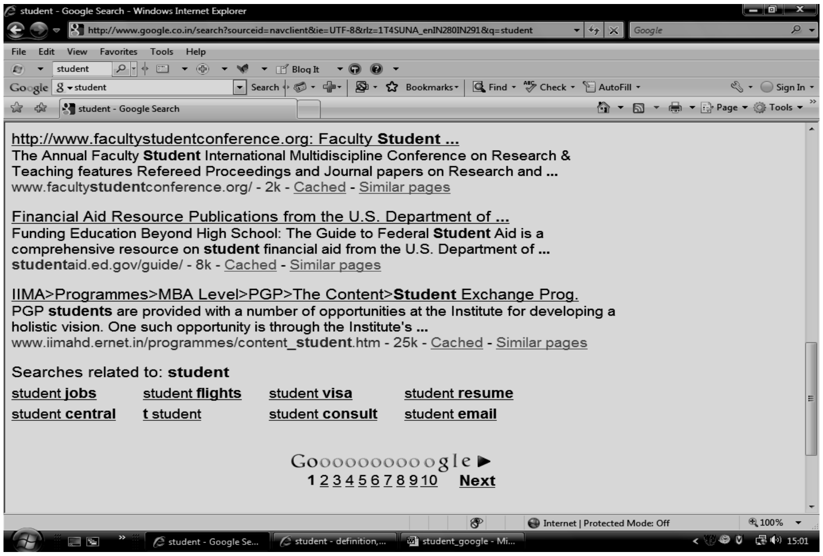
Student Arrested for Snowball Throw
www.break.com

Students India ,Bharat **Student**,Education , career India, career ...
Free information on Education ,career ,colleges , hostels , part time jobs ,boarding schools , placement consultants , admission news , univeristy ...
www.**student**india.com/ - 88k - Cached - Similar pages

IITG **STUDENT'S** WEBSITE
IITG **STUDENT'S** WEBSITE. indian institute of technology guwahati. **STUDENT** ACTIVITIES. Prakriti, the environment club of IIT guwahati ...
www.iitg.ac.in/studs/ - 3k - Cached - Similar pages

RamaKrishna Mission **Student's** Home
Describes the Mission's activities in providing free boarding, lodging and educational facilities to orphan and destitute children.

Figure 3 Result of Google for keyword "Student" (Contd.)


student - Google Search - Windows Internet Explorer
http://www.google.co.in/search?sourceid=navclient&ie=UTF-8&rlz=1T4SUNA_enIN280IN291&q=student
File Edit View Favorites Tools Help
student
Blog It
Google student Search Bookmarks Find Check AutoFill Sign In
student - Google Search
Page Tools
Internet | Protected Mode: Off
100%
student - Google Se... student - definition,... student_google - Mi...
15:01

http://www.facultystudentconference.org: Faculty **Student** ...
The Annual Faculty **Student** International Multidiscipline Conference on Research & Teaching features Refereed Proceedings and Journal papers on Research and ...
www.faculty**student**conference.org/ - 2k - Cached - Similar pages

Financial Aid Resource Publications from the U.S. Department of ...
Funding Education Beyond High School: The Guide to Federal **Student** Aid is a comprehensive resource on **student** financial aid from the U.S. Department of ...
studentaid.ed.gov/guide/ - 8k - Cached - Similar pages

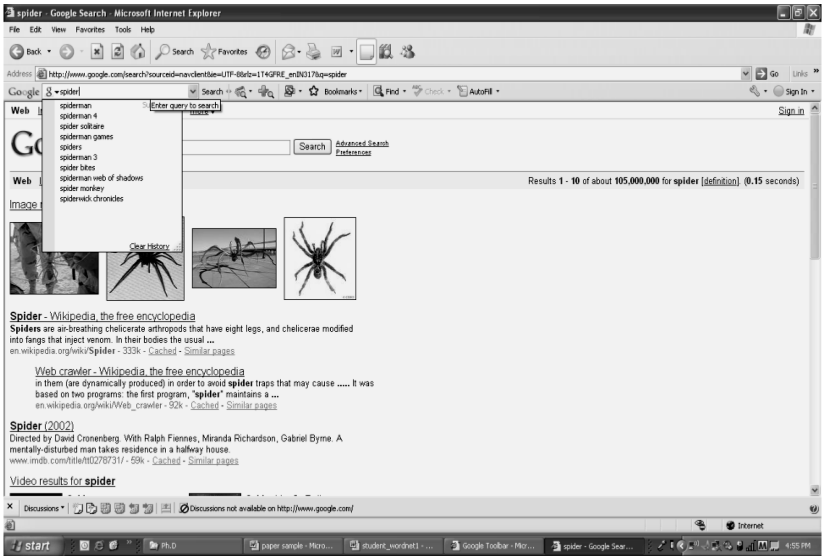
IIMA>Programmes>MBA Level>PGP>The Content>**Student** Exchange Prog.
PGP **students** are provided with a number of opportunities at the Institute for developing a holistic vision. One such opportunity is through the Institute's ...
www.iimahd.ernet.in/programmes/content_**student**.htm - 25k - Cached - Similar pages

Searches related to: **student**

student **jobs**	student **flights**	student **visa**	student **resume**
student **central**	**t** student	student **consult**	student **email**

Gooooooooooogle ►
1 2 3 4 5 6 7 8 9 10 **Next**

Figure 4 Result of Google for keyword "Spider"



spider - Google Search - Microsoft Internet Explorer

Address http://www.google.com/search?sourceid=navclient&ie=UTF-8&rlz=1T4GFRE_enIN317&q=spider

Google spider

spiderman
spiderman 4
spider solitaire
spiderman games
spiders
spiderman 3
spider bites
spiderman web of shadows
spider monkey
spiderwick chronicles

Results **1** - **10** of about **105,000,000** for **spider** [definition]. (**0.15** seconds)

Spider - Wikipedia, the free encyclopedia
Spiders are air-breathing chelicerate arthropods that have eight legs, and chelicerae modified into fangs that inject venom. In their bodies the usual **...**
en.wikipedia.org/wiki/**Spider** - 333k - Cached - Similar pages

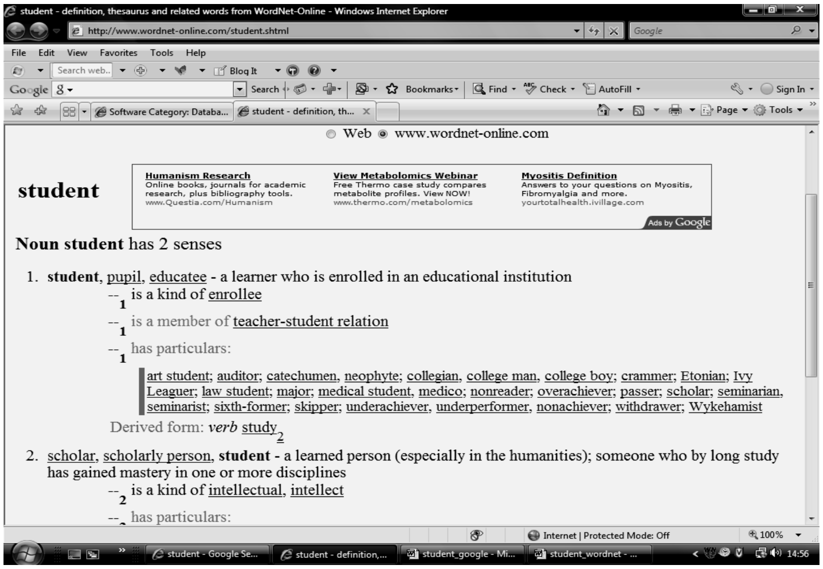
Web crawler - Wikipedia, the free encyclopedia
in them (are dynamically produced) in order to avoid **spider** traps that may cause **.....** It was based on two programs: the first program, "**spider**" maintains a **...**
en.wikipedia.org/wiki/Web_crawler - 92k - Cached - Similar pages

Spider (2002)
Directed by David Cronenberg. With Ralph Fiennes, Miranda Richardson, Gabriel Byrne. A mentally-disturbed man takes residence in a halfway house.
www.imdb.com/title/tt0278731/ - 59k - Cached - Similar pages

Video results for **spider**

Discussions not available on http://www.google.com/

Figure 5 Result of Word Net for Keyword "Student"



student - definition, thesaurus and related words from WordNet-Online - Windows Internet Explorer

http://www.wordnet-online.com/student.shtml

○ Web ● www.wordnet-online.com

student

Humanism Research
Online books, journals for academic research, plus bibliography tools.
www.Questia.com/Humanism

View Metabolomics Webinar
Free Thermo case study compares metabolite profiles. View NOW!
www.thermo.com/metabolomics

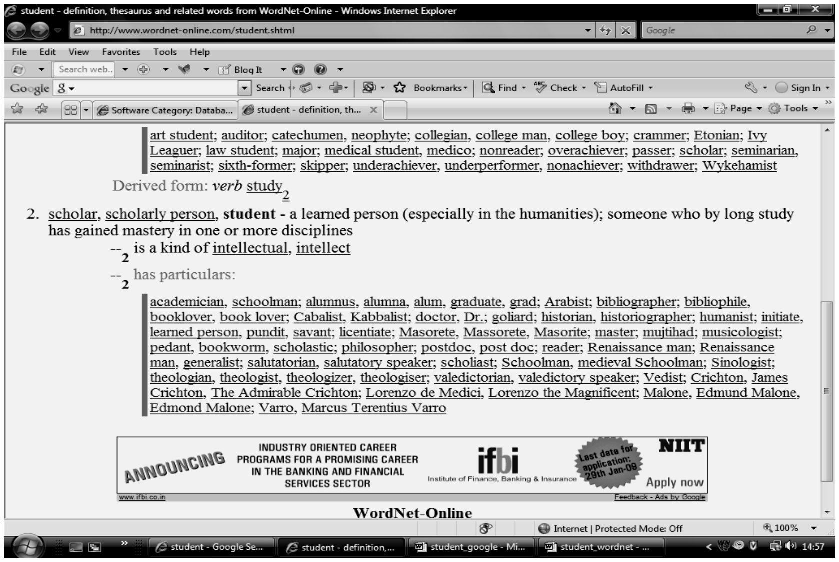
Myositis Definition
Answers to your questions on Myositis, Fibromyalgia and more.
yourtotalhealth.ivillage.com

Ads by Google

Noun student has 2 senses

1. **student**, pupil, educatee - a learner who is enrolled in an educational institution
   --$_1$ is a kind of enrollee
   --$_1$ is a member of teacher-student relation
   --$_1$ has particulars:
   art student; auditor; catechumen, neophyte; collegian, college man, college boy; crammer; Etonian; Ivy Leaguer; law student; major; medical student, medico; nonreader; overachiever; passer; scholar; seminarian, seminarist; sixth-former; skipper; underachiever, underperformer, nonachiever; withdrawer; Wykehamist
   Derived form: *verb* study$_2$
2. scholar, scholarly person, **student** - a learned person (especially in the humanities); someone who by long study has gained mastery in one or more disciplines
   --$_2$ is a kind of intellectual, intellect
   --$_2$ has particulars:

Figure 6 Result of Word Net for Keyword "Student" (Contd.)


student - definition, thesaurus and related words from WordNet-Online - Windows Internet Explorer
http://www.wordnet-online.com/student.shtml
Google
File Edit View Favorites Tools Help
Search web..
Blog It
Search
Bookmarks
Find
Check
AutoFill
Sign In
Software Category: Databa...
student - definition, th...
Page
Tools
Internet | Protected Mode: Off
100%
student - Google Se...
student - definition,...
student_google - Mi...
student_wordnet - ...
14:57

art student; auditor; catechumen, neophyte; collegian, college man, college boy; crammer; Etonian; Ivy Leaguer; law student; major; medical student, medico; nonreader; overachiever; passer; scholar; seminarian, seminarist; sixth-former; skipper; underachiever, underperformer, nonachiever; withdrawer; Wykehamist

Derived form: *verb* study$_2$

2. scholar, scholarly person, **student** - a learned person (especially in the humanities); someone who by long study has gained mastery in one or more disciplines

--$_2$ is a kind of intellectual, intellect

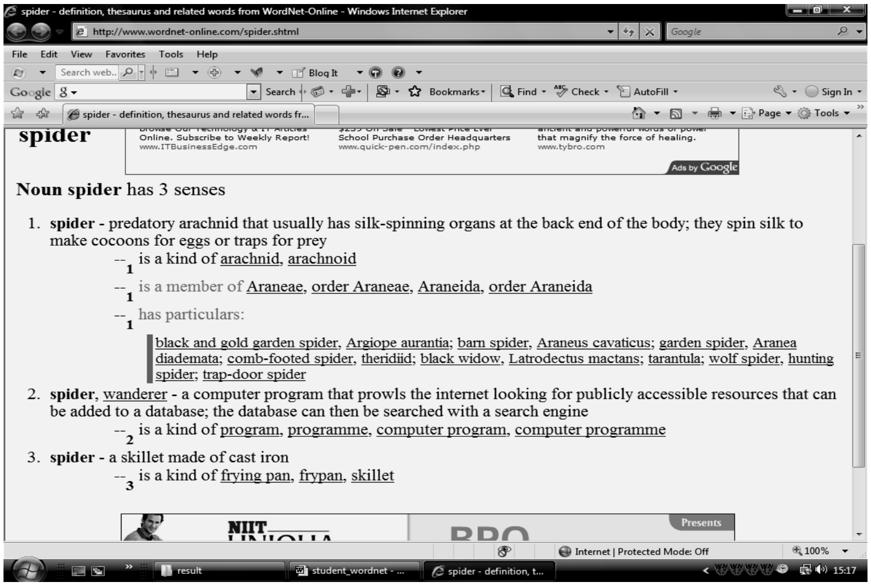
--$_2$ has particulars:

academician, schoolman; alumnus, alumna, alum, graduate, grad; Arabist; bibliographer; bibliophile, booklover, book lover; Cabalist, Kabbalist; doctor, Dr.; goliard; historian, historiographer; humanist; initiate, learned person, pundit, savant; licentiate; Masorete, Massorete, Masorite; master; mujtihad; musicologist; pedant, bookworm, scholastic; philosopher; postdoc, post doc; reader; Renaissance man; Renaissance man, generalist; salutatorian, salutatory speaker; scholiast; Schoolman, medieval Schoolman; Sinologist; theologian, theologist, theologizer, theologiser; valedictorian, valedictory speaker; Vedist; Crichton, James Crichton, The Admirable Crichton; Lorenzo de Medici, Lorenzo the Magnificent; Malone, Edmund Malone, Edmond Malone; Varro, Marcus Terentius Varro

ANNOUNCING INDUSTRY ORIENTED CAREER PROGRAMS FOR A PROMISING CAREER IN THE BANKING AND FINANCIAL SERVICES SECTOR ifbi Institute of Finance, Banking & Insurance Last date for application: 29th Jan-09 NIIT Apply now
www.ifbi.co.in Feedback - Ads by Google

WordNet-Online

Figure 7 Result of Word Net for keyword "Spider"


spider - definition, thesaurus and related words from WordNet-Online - Windows Internet Explorer
http://www.wordnet-online.com/spider.shtml
Google
File Edit View Favorites Tools Help
Search web..
Blog It
Search
Bookmarks
Find
Check
AutoFill
Sign In
spider - definition, thesaurus and related words fr...
Page
Tools
Internet | Protected Mode: Off
100%
result
student_wordnet - ...
spider - definition, t...
15:17

spider

Online. Subscribe to Weekly Report! www.ITBusinessEdge.com
School Purchase Order Headquarters www.quick-pen.com/index.php
that magnify the force of healing. www.tybro.com
Ads by Google

Noun spider has 3 senses

1. **spider** - predatory arachnid that usually has silk-spinning organs at the back end of the body; they spin silk to make cocoons for eggs or traps for prey

--$_1$ is a kind of arachnid, arachnoid

--$_1$ is a member of Araneae, order Araneae, Araneida, order Araneida

--$_1$ has particulars:

black and gold garden spider, Argiope aurantia; barn spider, Araneus cavaticus; garden spider, Aranea diademata; comb-footed spider, theridiid; black widow, Latrodectus mactans; tarantula; wolf spider, hunting spider; trap-door spider

2. **spider**, wanderer - a computer program that prowls the internet looking for publicly accessible resources that can be added to a database; the database can then be searched with a search engine

--$_2$ is a kind of program, programme, computer program, computer programme

3. **spider** - a skillet made of cast iron

--$_3$ is a kind of frying pan, frypan, skillet

NIIT Presents

Figure 8 High Level Architecture

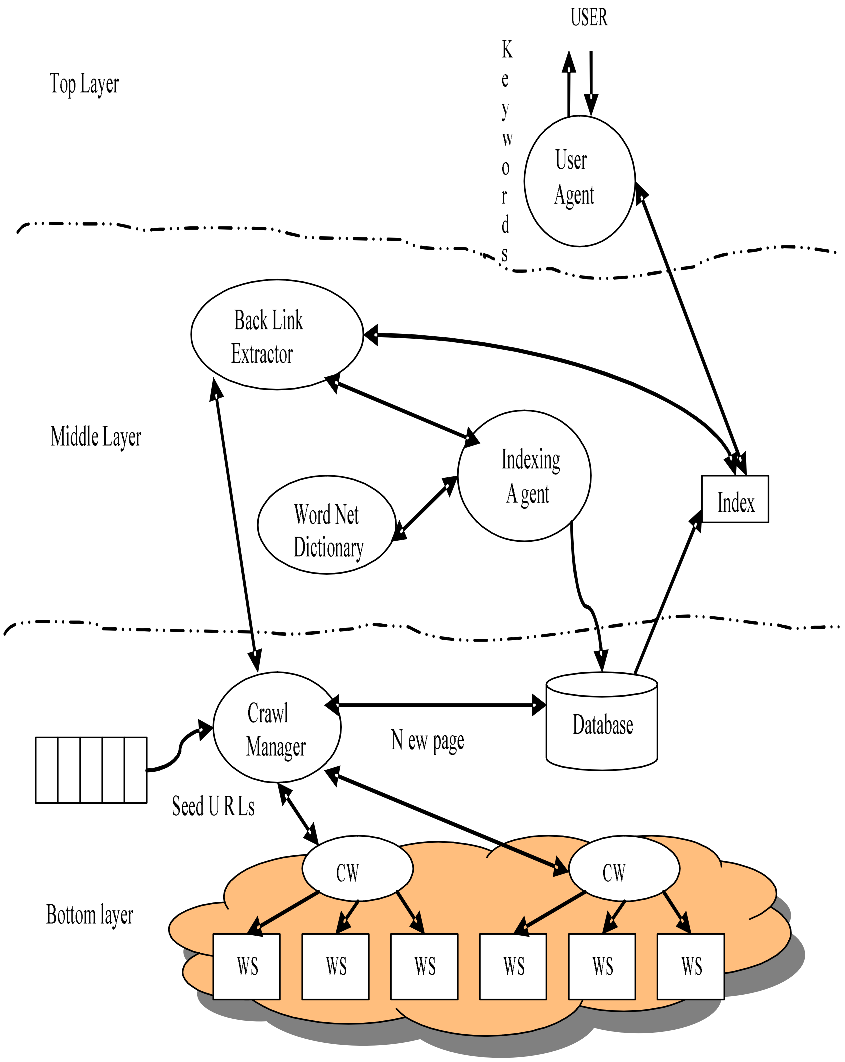


Figure 9 Data Flow within CBDFC

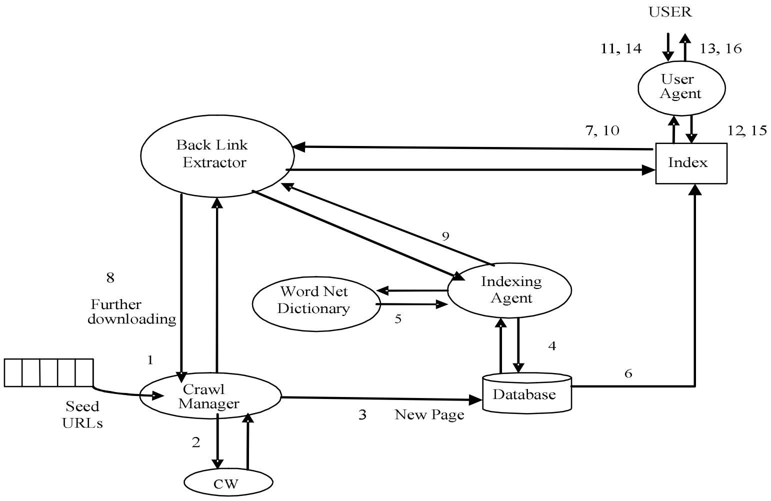
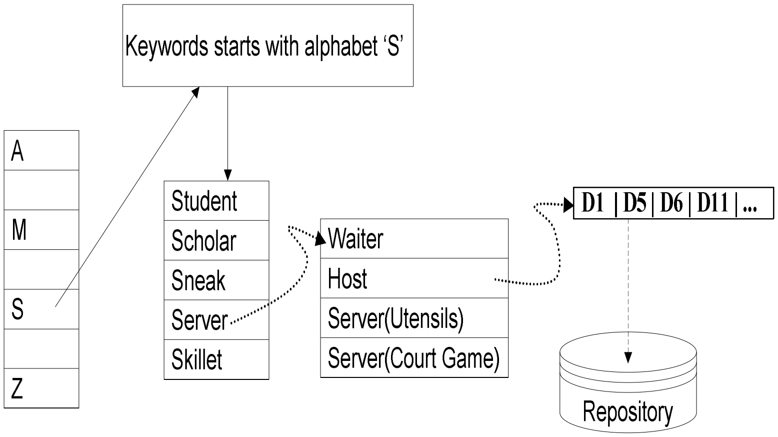


Figure 10 Example



References

[1] J. Cho, Hector Garcia Molina, Lawrence page, "Efficient Crawling through URL Ordering", paper presented at 7th *international WWW Conference*. April 1998. Brisbane, Australia.

[2] B. Novak, "A Survey of Focused Web Crawling Algorithms", paper presented at conferences on *Data Mining & Warehouses SIKDD*, 2004

[3] S. Chakrabarti, M. van den Berg and B. Dom, "Focused Crawling: A new Approach to Topic-Specific Web Resource Discovery", paper presented at 8th *International WWW Conference*. May 1999. Toronto, Canada.

[4] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles and M. Gori, “Focused Crawling using Context Graphs”, paper presented at *26th International Conference on Very large Databases (VLDB 2000)*. September 2000, Cairo, Egypt.

[5] S. Chakrabarti, K. Punera, Mallena Subramanyam, “Accelerated Focused Crawling through Online relevance feedback”, paper presented at *WWW conference* December 2002.

[6] M. Yuvarani, N.Ch.S.N.Iyengar, A.Kannan, “LSCrawler: A Framework for an Enhanced Focused Web Crawler based on Link Semantics”. Paper presented at *International Conference on Web Intelligence (IEEE/WIC/ACM)*, 2006 pp 794-800

[7] X. Chen, X. Zhang, ‘HAWK: A Focused Crawler with Content and Link Analysis’. Paper presented at *IEEE International Conference on e-Business Engineering*. pp- 677-680, 2008

[8] N. Chauhan, A. K. Sharma, “Context Driven Focused Crawling: A New Approach to Domain-Specific Web Retrieval”, paper presented at *International Conference on information & Communication Technology (IICT)*, July, 2007.Dehradun.

[9] N. Chauhan, A. K. Sharma, “A Framework to Derive Web Page Context from Hyperlink Structure”, *International Journal of Information and Communication Technology*, 2008 Vol. 1, No.3/4pp.329-346.

[10] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine” *Computer Networks and ISDN Systems* Vol.30, issue 1-7, Pages: 107-117, (April-1998).

[11] Thompson, B.B., Marks, R.J., El-Sharkawi, M.A., Fox, W.J. and Miyamoto, R.T. (2003) ‘Inversion of Neural Network Underwater Acoustic Model for Estimation of Bottom Parameters using Modified Particle Swarm Optimizer’, in *Proceedings of the International Joint Conference on Neural Networks*, pp. 1301-1306.

[12] WordNet-Online dictionary and hierarchical thesaurus Obtained through the Internet http://www.wordnet-online.com [accessed 28/12/2009]