

MetaShopper, a preliminary study and implementation

Nguyen Bach
Department of Computer Science
Johns Hopkins University
nguyen@cs.jhu.edu

1. Introduction

MetaShopper is an application of using Web robots to bring and extract information from multiple sources and condense it into a useful format. An example is MetaCrawler, it is simply the easiest way to find better search results from more of the Web. MetaCrawler uses innovative metasearch technology to search the Internet's top search engines, including Google, Yahoo, Ask Jeeves, About, Teoma, FindWhat, LookSmart, and many more. With one single click, MetaCrawler searches the best results from the combined pool of the world's leading search engines -- instead of results from only one single search engine.

In this project, I implement a system called VeryNaiveBookCrawler (VNBC) which will try to find the most relevance and cheapest textbook from many resources.

2. Methods

VNBC basically visits all of the shopping sites that it supports, download their pages, and look at the HTML forms and structure. Then it can find the fields it will need to retrieve such as title and price. I decide the webrobot go to five shopping sites which are Half, Buy, Amazon, Barnes&Noble, and Wordsworth. The reason for choosing those sites simply is each site is strong in different facets. For instance, Half, a division of Ebay, tends to found the cheapest used books. While Amazon and Barnes&Noble are two of the biggest online bookshop, they have many different titles. Buy is good at with technical textbook while Wordsworth is good at with art & science. Therefore, when combining results of all resources, user's demand can be satisfied.

VNBC do the following tasks

- Get user's query and parse it to be suitable with each distinct website.
- Apply the query for each site and fetch the HTML files for analyzing.
- Analyze HTML files to extraction information.
- Sorted results and show them to users.
- Cache searched query in order to boosting search process.

Figure 1 outlines the operation of VNBC

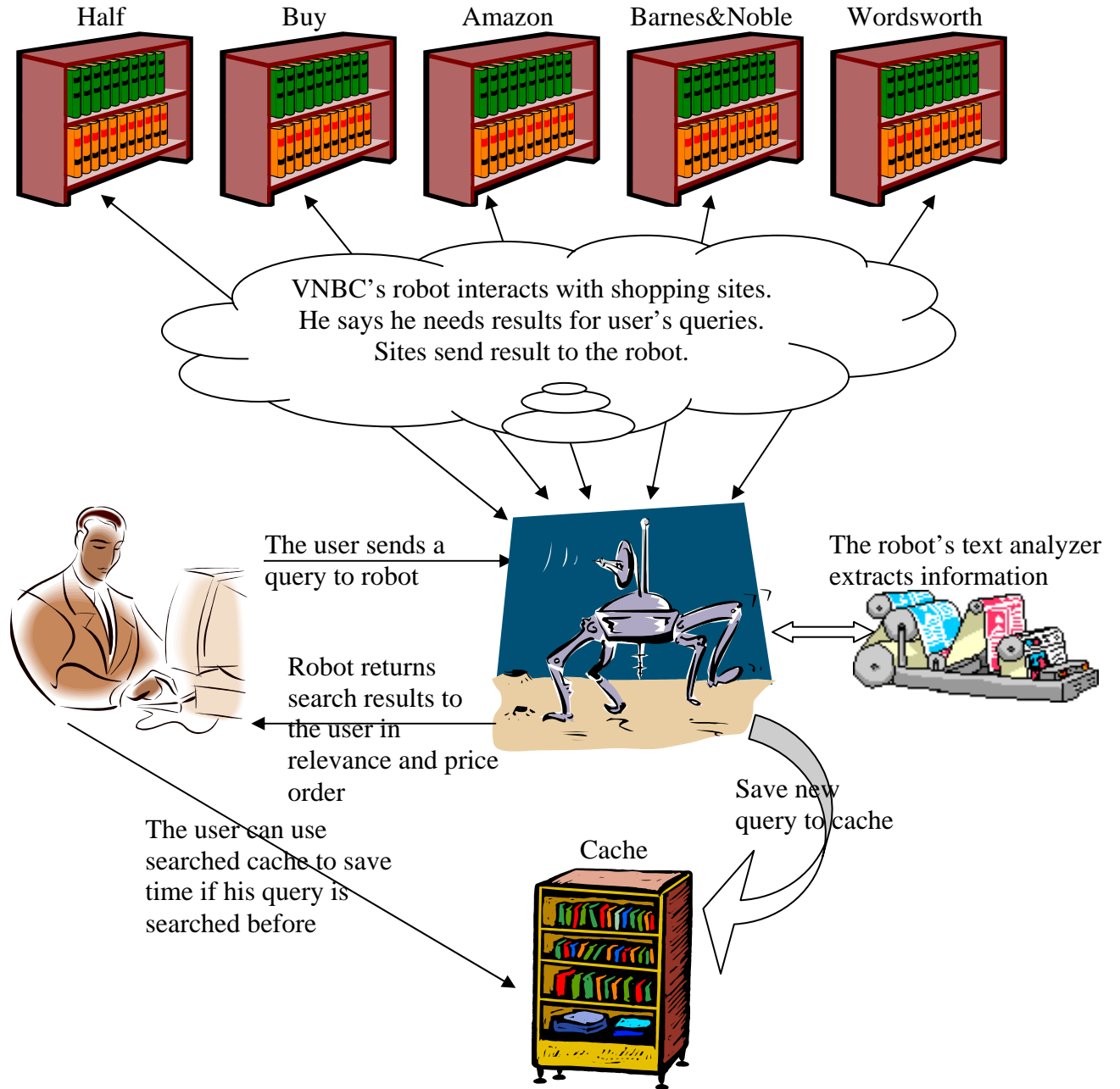


Figure 1: VNBC model

3. Implementation

VNBC uses CGI programming method to interact with server. VNBC contains the following files

search.cgi : the interface of the engine which allows the user to type queries they want. It also shows the searched queries and provides the user a faster way to access old results. The user can use delete command to refresh cache, see also Fig. 2.

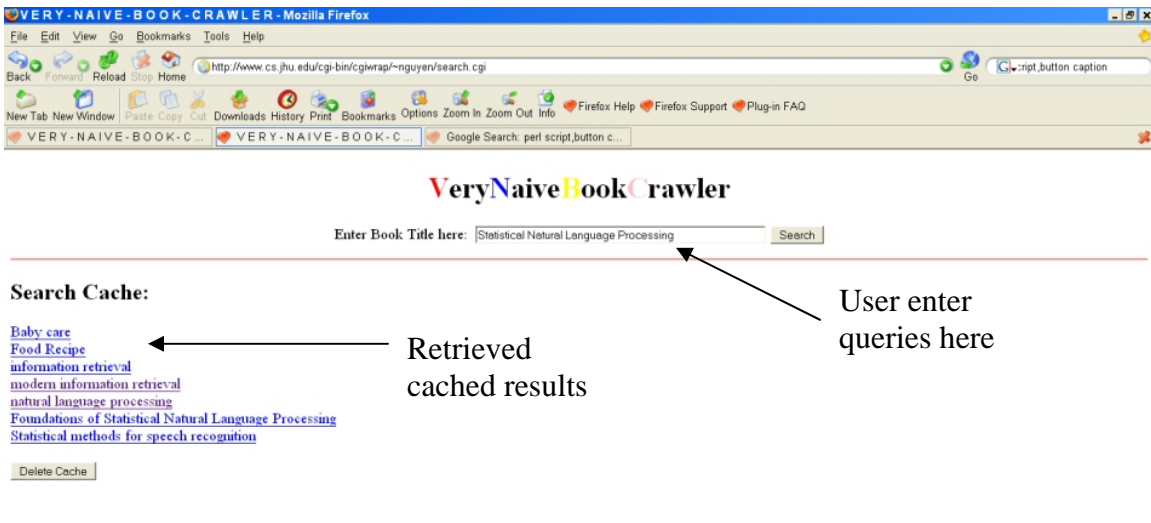


Figure 2: search.cgi

show.cgi: is the engine that receives queries from the user, send request and get response from shopping sites, text extraction, sorted by relevance and price for each group, cache and show results in HTML format, see also Fig.3 and Fig.4.

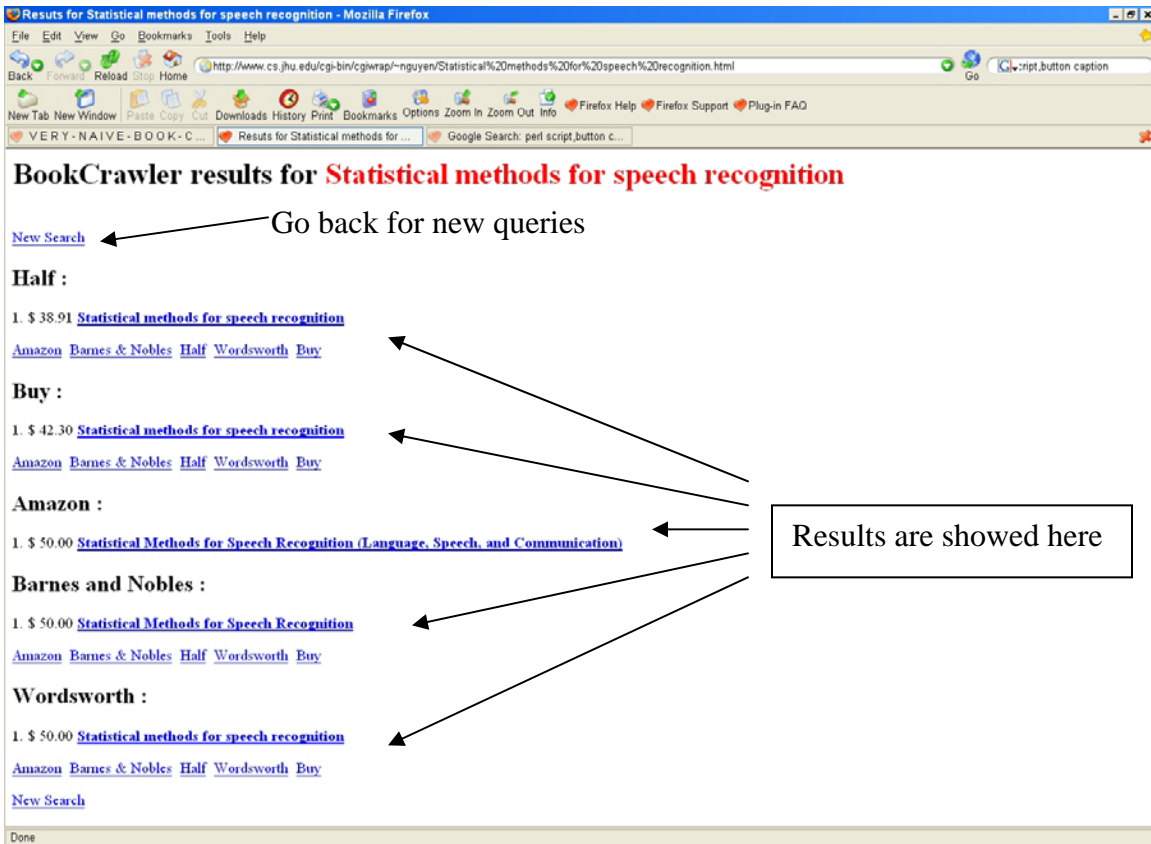


Figure 3: show.cgi's screen, multiple results for a specific query

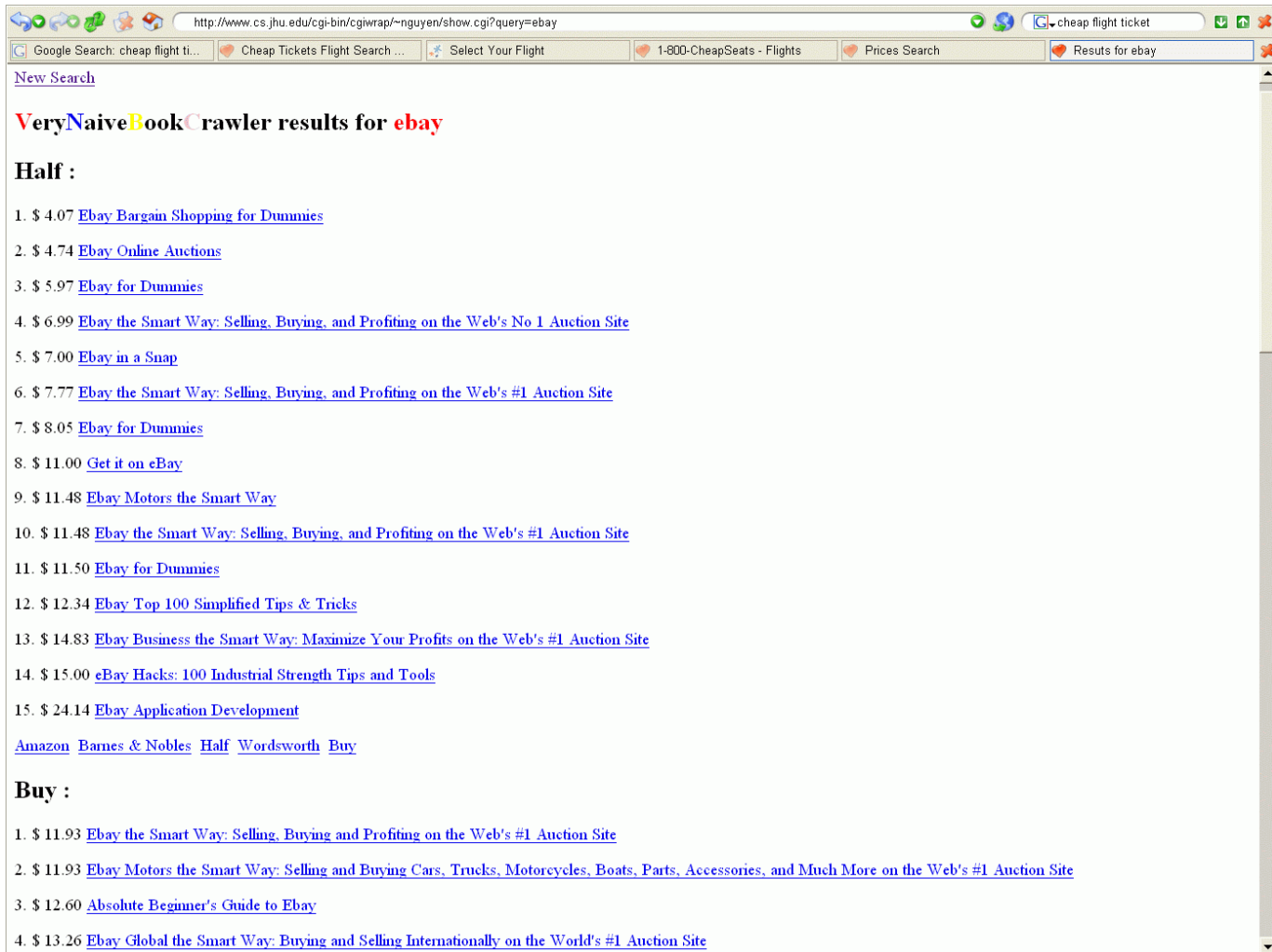


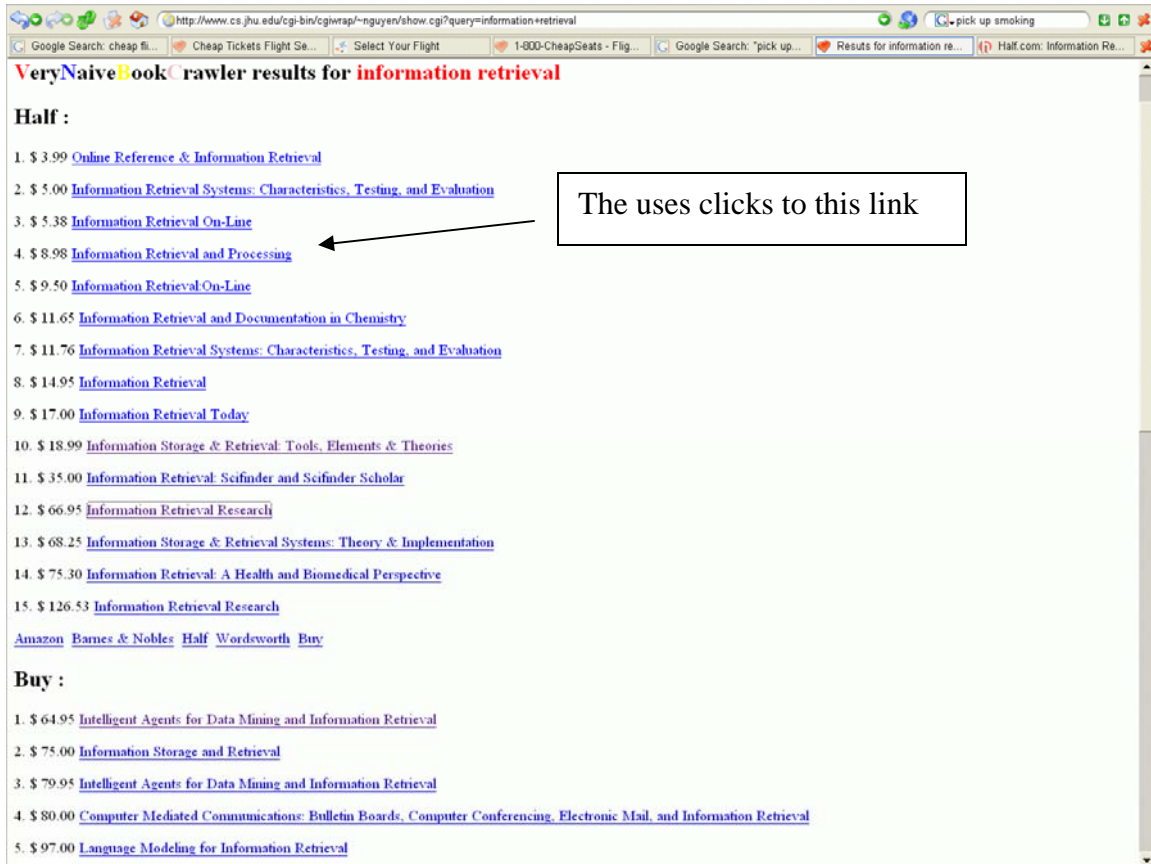
Figure 4: show.cgi's screen, multiple results are ranked by relevance and price for a vague query

config.txt: contain list of shopping sites, for each site there are two lines name and the link. The file will be

```
Barnes and Nobles
http://shop.barnesandnoble.com/BookSearch/results.asp?TTL=
Amazon
http://www.amazon.com/exec/obidos/search-handle-url/index=books&field-
keywords=&bq=1/ref=aps_more_b_1/107-1038526-9140510
Half
http://half.ebay.com/search/search.jsp?nthTime=1&srchType=adv&product=bo
oks&page=1&sort=BySearchRelevance&inStock=n&keyword=
Wordsworth
http://ishop.wordsworth.com/newsearch/redirect_search.asp?sessionID=ww15
41642346416&searchtype=T1&sort=title&order=asc&svalue=
Buy
http://www.buy.com/retail/searchresults.asp?search_store=3&querytype=boo
k&loc=106&dclksa=1&qu=
```

debug.log: for each query VNBC creates a file to store all status and data when it is processing the query.

We need the URL to call the retrieve prices function for the different online websites. We had to hack the individual websites and hunt out the url for their search engines. Except Half and Wordsworth which have a straightforward urls, three remain sites we chose to do had hidden their complete urls from the common user, (probably to prevent nosey students like us) but with a little perseverance we obtained these urls and edited them for the purpose of our own searches we chose to search and compile. I find urls of Half and Wordsworth quite easy, just go to the website, make a query, and copy the returned url. Moreover, we need to do something special for searching the Amazon site, because you need to get a cookie id number first. Therefore, I contact the site, get a redirect message that also sets a cookie on my computer. I get that id number and position it correctly into the search site variable to retrieve the webpage needed for our query. Remember that we get query form user we have to splits the title (or query search) of the book and adds '+' because this is the CGI standard for database lookups online.



The picture above shows results when the user queries “information retrieval”. This is a generic topic so the engine receives many suggestions form shopping site. Here the engine first retrieve top relevance items then sort them based on their prices. When the user click to any item, a new window will open any redirect the user concentration to the link. For example, the picture below demonstrate what happen when the userclick to Information Retrieval and Processing.

The screenshot shows an eBay product page for the book "Information Retrieval and Processing" (Binding Unknown, 1975) by Joseph Becker III and Lauren B. Doyle. The page is sorted by price, showing several listings. The top listing is from "krusenotes" for \$25.00. Below it is a listing from "alibris" for \$12.95. The bottom listing is from "rmphan" for \$8.98. The page also includes a "Save on Shipping!" banner, a "Rate this product" section, and a "Related Items on eBay" section.

Whenever we get a link, we need to investigate to see if it's relevant by seeing if the string online price is anywhere in the vicinity. If it is not, then we want to throw that link out, because this link is probably not a link to a book, and has nothing to do with a price on the current page. Therefore it is some other link with the 'title pieces' in it probably advertising to sell you something about that topic that is not related to a book. This picks up right where the big while loop is searching and gets the next 1500 characters, Amazon's site thins its content out with a lot of spaces so many times the characters we pick up for Amazon is spaces. Therefore the Price string can't be found in the 'looking_for_price' var and we end up throwing away good links. All of the other websites have no problem with the 1500 condition. Enlarging this number might include bad links on the other sites that we don't want.

There is a case here when user query exactly the name of textbook some websites as Half, Buy, and Wordsworth return a specific html page which has different structure with general html. Therefore we need a special treatment for this case. We notice that for Half's web pages has a string "search for signed", Buy has not got the string "browse titles" as usual, and Wordsworth has "subject areas this". So, based on these strings we know whether we are in a special webpage or not.

http://half.ebay.com/cat/buy/prod.cgi?cpid=1175673&domain_id=1856&meta_id=1

Reflections: Carly Simon's Greatest Hits - \$12.49 or Less!

half.com by ebay

Home Books Textbooks Music DVD/Video Video Games Computers Electronics Half Zone on eBay

Search: Books

Home > Books

Modern Information Retrieval (Paperback, 1999)
 Author: [Berthier Ribeiro-Neto](#), [R. Baeza-Yates](#)
 Best Price: **\$27.00**
 List Price: \$50.00 (Save \$23.00)
[Search for similar Berthier Ribeiro-Neto books on eBay!](#)
[Search for first editions of Modern Information Retrieval on eBay!](#)

Format: Paperback
 ISBN: 020139829X
 February 1999
 Publisher: Addison-Wesley
 513 pages
 Series: Acm Press Series D
 Language: English

Buy It Now on eBay

Price	Seller (Feedback)	Comments	Shipping	Ships From
\$30.00	blainerunner (13) ★		Media Mail	PA
\$35.99	tree_nine (108) ★	excellent condition...intl version with b...	Media Mail Upgrade	ML*
\$38.01	greatbookprices (1273) ★	Absolutely New Never Been Read Mint...	Media Mail	
\$38.60	yinqxuan169 (1)		Media Mail	IL

Buy It Now on eBay

Price	Title	Time Left
\$29.50	Buy It Now Modern Information Retrieval by Berthier Ribeiro-Net...	5d 22h 41m

Like New Items

Price	Seller (Feedback)	Comments	Shipping	Ships From
\$30.00	yvvincent (20) ★	New, never used. Intl edition, identical...	Media Mail	ML*
\$32.99	yuklee1 (21) ★	Very good condition. Ship quick.	Media Mail	CA
\$37.00	suziepoker-half (11) ★	No writings, no highlights inside. A li...	Media Mail	TN
\$38.80	bookbyte_com	Modern Information Retrieval by Baeza-Y...	Media Mail	OR

Buy.com

Books • Magazines • Toys • Music • Music Downloads • DVDs • Games • Sports • Bags • Today's Deals

Home > Books > Product Information

Search: Search Books

Modern Information Retrieval
 Our Price: \$36.25
 List Price: \$60.00
 You Save: \$23.75
 In Stock: Usually Ships in 1 to 2 business days.
 Qty: 1 BUY NOW Add to Wishlist Email a Friend

Item #: 87511W
 Larger image

Author: [Ricardo A. Baeza-Yates](#)
 Format: Paperback
 ISBN: 020139829X
 Publish Date: 2/1/1999
 Publisher: Addison-Wesley Publishing Company

Dimensions (in Inches) 9.5H x 6.75L x 1T
 Pages: 464
 Cover: Arm Dracc Karic

Product Description
 Modern Information Retrieval allows readers to familiarize themselves with the major IR techniques for document ranking, indexing, searching, visualizing data, operating on multimedia objects, and searching the Web.

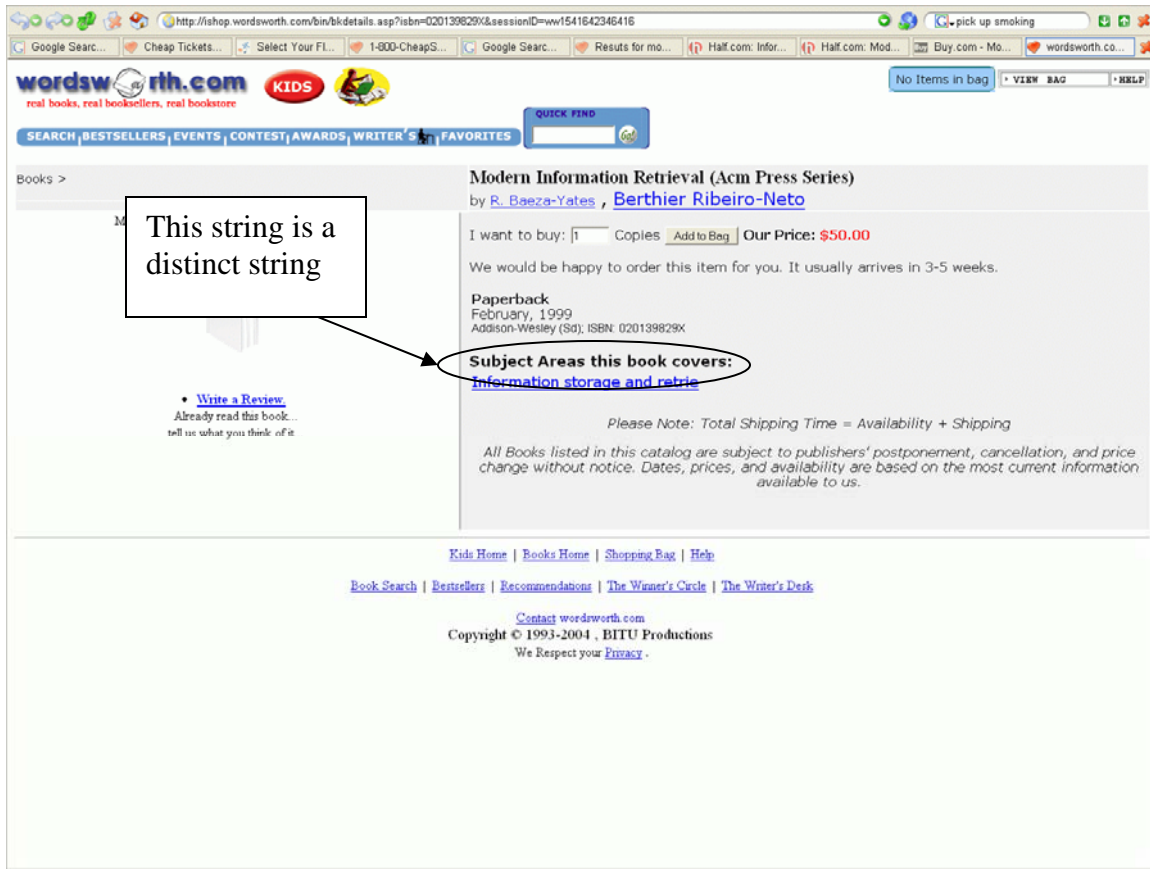
From the Publisher
 Modern Information Retrieval allows readers to familiarize themselves with the major IR techniques for document ranking, indexing, searching, visualizing data, operating on multimedia objects, and searching the Web.

By Bob Woodward
 The definitive account of a turning point in history as Bush, his war council, and allies launch a preemptive attack on Iraq. [Click here for more info.](#)

Oprah's Picks
 • The Heart Is a Lonely Hunter -- NEW!
 • One Hundred Years of Solitude
 • Cry, the Beloved Country
 • East of Eden
 • Sula

NETGEAR WIRELESS-G BUNDLE with \$35 Rebate
 FRIENDS SERIES FINALE

This string is a distinct string



The data structure contains three 2-dimension arrays, aPrices, aBuy_link, and aBook_Title. Each row represents a specific site, for example the second row is for Half. When we want to retrieve the jth item from Half, basically we call aPrices[2][j], aBuy_link[2][j], and aBook_Title[2][j]. To sort all items by price two hashes are implement that are %prices and %names with key is links.

It is much convenient to program, run, and test on ones own PC or laptop. My machine runs on top of XP. This appendix shows how to setup a Windows machine to run CGI scripts, in particular the web robot of the project.

First, Microsoft Internet Information Server (Ms IIS, or in short IIS) needs to be installed. For Windows 9x, IIS is also called Personal Web Server, while in Windows 2000 and XP IIS comes along with setup cd-rom. Do the following steps to setup IIS in XP:

Start; Settings; Control Panel; Add or Remove Programs; Add/Remove Windows components; Check IIS and click Next; then following the instructions.

It takes a couple of minutes to install IIS from a Windows XP Professional machine.

When we already have IIS, the next step need to do is install ActivePerl. ActivePerl is ActiveState's quality-assured distribution of Perl, available for Linux, Solaris, and Windows. Go to http://www.activestate.com/Products/ActivePerl/?_x=1 to get setup

package. After finishing install ActivePerl, try to test with a simple perl program such as C:\perl helloworld.pl .

Third, go to

Start; Settings; Control Panel; Administrative Tools; Computer Management; Services and Applications; Internet Information Services; -Your Web Site-

The job now is to associate file extensions for individual sites, or for the entire server. To do so right click on either the entire server, or the place where you want the association to be effective. Select [Properties] from the pop-up menu and press the [Configuration...] button. Select [Add] under the [App Mappings] tab. Next, browse for perl.exe and choose an associated file extension, here is .cgi. The path needs quotes if it has spaces, and you need to add %s %s to the end.

4. Future work

VNBC works smoothly but it is quite slow and the reason is VNBC only uses one robot to go to 5 shopping sites. So VNBC can boost up its speed by using several robots, and for each website one robot will fetch information to VNBC.

Another way to improve speed is upgrade the cache module. The cache module is pretty naïve because it simply stores the previous results. I can improve by driving VNBC to search in cache before it actually goes to search in the shopping sites.

VNBC works also better if it has a database of HTML structures of shopping sites. The structure of shopping sites is represented in XML format. VNBC will read the structure of these sites and base on its properties to analyze.