

Web Usage Mining

What is Web Usage Mining?

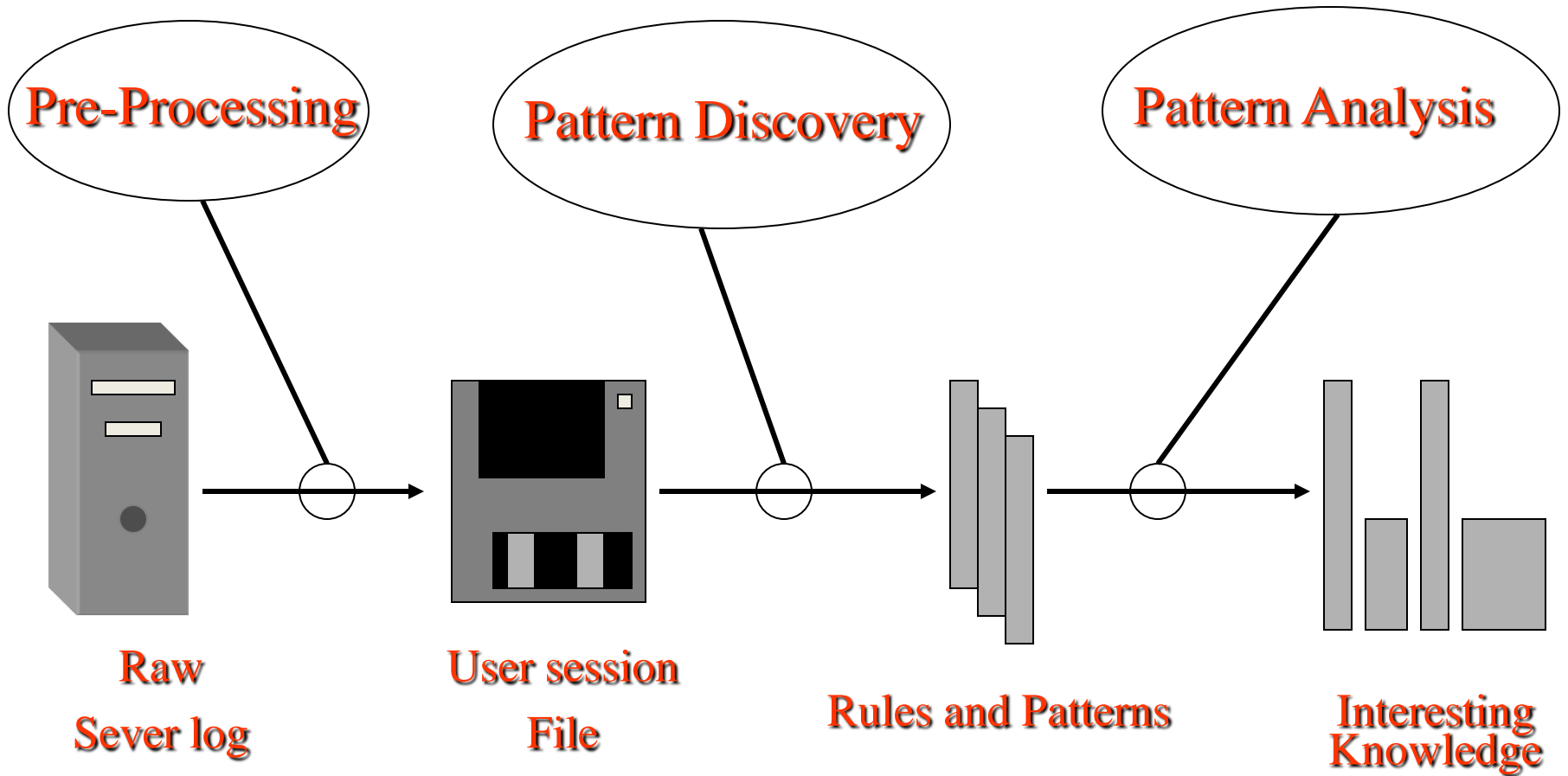
- A *Web* is a collection of inter-related files on one or more *Web servers*.
- *Web Usage Mining*.
 - ➔ Discovery of meaningful patterns from data generated by client-server transactions.
- Typical Sources of Data:
 - ➔ automatically generated data stored in server *access logs*, *referrer logs*, *agent logs*, and client-side *cookies*.
 - ➔ user profiles.
 - ➔ metadata: page attributes, content attributes, usage data.

Web Usage Mining (WUM)

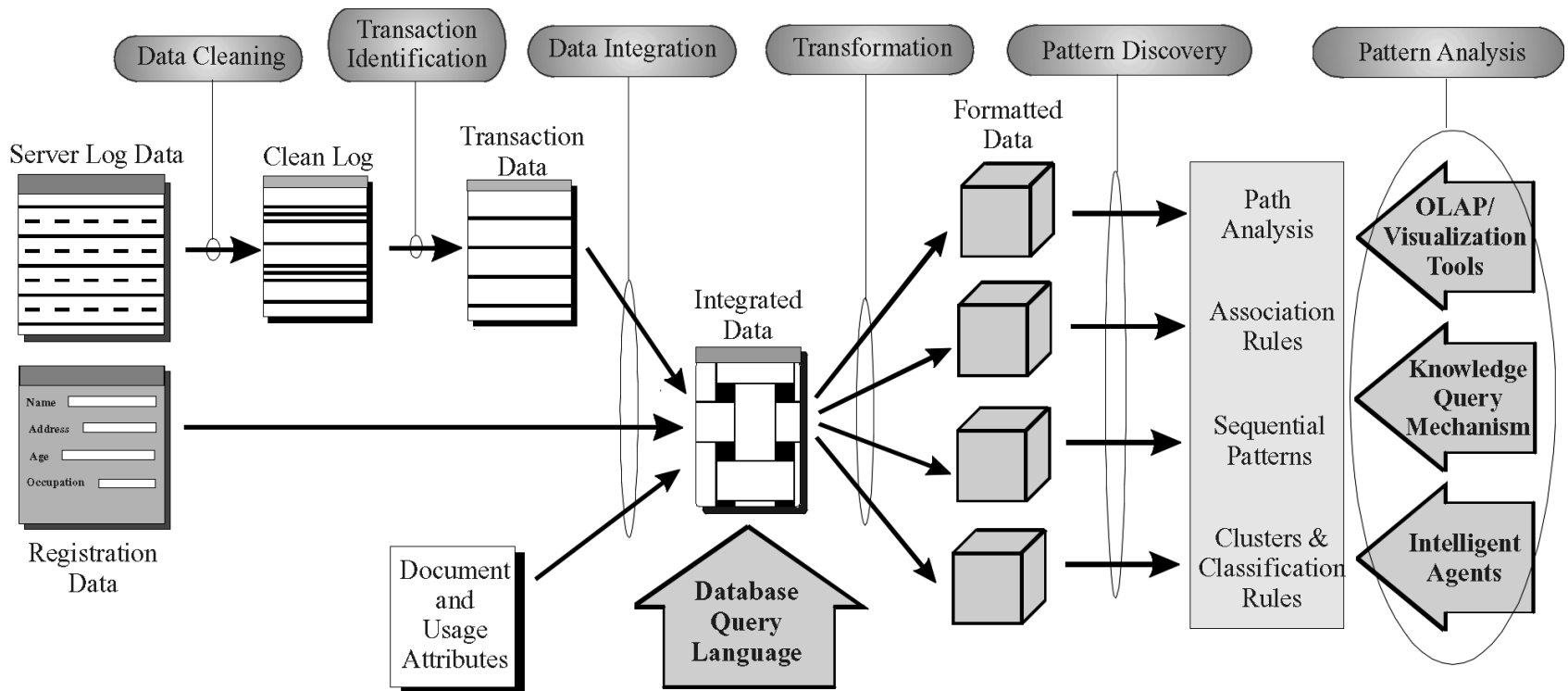
The discovery of interesting user access patterns from Web server logs

- **Generate simple statistical reports:**
 - A summary report of hits and bytes transferred
 - A list of top requested URLs
 - A list of top referrers
 - A list of most common browsers used
 - Hits per hour/day/week/month reports
 - Hits per domain reports
- **Learn:**
 - Who is visiting you site
 - The path visitors take through your pages
 - How much time visitors spend on each page
 - The most common starting page
 - Where visitors are leaving your site

Web Usage Mining – Three Phases



The Web Usage Mining Process



– General Architecture for the WEBMINER –

Web Server Access Logs

- Typical Data in a Server Access Log

```
looney.cs.umn.edu han - [09/Aug/1996:09:53:52 -0500] "GET mobasher/courses/cs5106/cs5106l1.html HTTP/1.0" 200
mega.cs.umn.edu njain - [09/Aug/1996:09:53:52 -0500] "GET / HTTP/1.0" 200 3291
mega.cs.umn.edu njain - [09/Aug/1996:09:53:53 -0500] "GET /images/backgnds/paper.gif HTTP/1.0" 200 3014
mega.cs.umn.edu njain - [09/Aug/1996:09:54:12 -0500] "GET /cgi-bin/Count.cgi?df=CS home.dat\&dd=C\&ft=1 HTTP
mega.cs.umn.edu njain - [09/Aug/1996:09:54:18 -0500] "GET advisor HTTP/1.0" 302
mega.cs.umn.edu njain - [09/Aug/1996:09:54:19 -0500] "GET advisor/ HTTP/1.0" 200 487
looney.cs.umn.edu han - [09/Aug/1996:09:54:28 -0500] "GET mobasher/courses/cs5106/cs5106l2.html HTTP/1.0" 200
...           ...           ...
```

◆ Access Log Format

IP address userid time method url protocol status size

Example: Session Inference with Referrer Log

	IP	Time	URL	Referrer	Agent
1	www.aol.com	08:30:00	A	#	Mozilla/2.0; AIX 4.1.4
2	www.aol.com	08:30:01	B	E	Mozilla/2.0; AIX 4.1.4
3	www.aol.com	08:30:02	C	B	Mozilla/2.0; AIX 4.1.4
4	www.aol.com	08:30:01	B	#	Mozilla/2.0; Win 95
5	www.aol.com	08:30:03	C	B	Mozilla/2.0; Win 95
6	www.aol.com	08:30:04	F	#	Mozilla/2.0; Win 95
7	www.aol.com	08:30:04	B	A	Mozilla/2.0; AIX 4.1.4
8	www.aol.com	08:30:05	G	B	Mozilla/2.0; AIX 4.1.4

Identified Sessions:

- S_1 : # ==> A ==> B ==> G from references 1, 7, 8
 S_2 : E ==> B ==> C from references 2, 3
 S_3 : # ==> B ==> C from references 4, 5
 S_4 : # ==> F from reference 6

Data Mining on Web Transactions

- Association Rules:

- ➔ discovers similarity among sets of items across transactions

$$X \xrightarrow{\alpha, \sigma} Y$$

where X, Y are sets of items, $\alpha = \textit{confidence}$ or $P(X \vee Y)$,
 $\sigma = \textit{support}$ or $P(X \wedge Y)$

- Examples:

- ➔ 60% of clients who accessed [/products/](#), also accessed [/products/software/webminer.htm](#).

- ➔ 30% of clients who accessed [/special-offer.html](#), placed an online order in [/products/software/](#).

- ➔ (Actual Example from IBM official Olympics Site)

- {Badminton, Diving} \implies {Table Tennis} ($\alpha = 69.7\%$, $\sigma = 0.35\%$)

Other Data Mining Techniques

- Sequential Patterns:
 - ➔ 30% of clients who visited [/products/software/](#), had done a search in **Yahoo** using the keyword “**software**” before their visit
 - ➔ 60% of clients who placed an online order for WEBMINER, placed another online order for software within 15 days
- Clustering and Classification
 - ➔ clients who often access [/products/software/webminer.html](#) tend to be from educational institutions.
 - ➔ clients who placed an online order for software tend to be students in the 20-25 age group and live in the United States.
 - ➔ 75% of clients who download software from [/products/software/demos/](#) visit between 7:00 and 11:00 pm on weekends.

Path and Usage Pattern Discovery

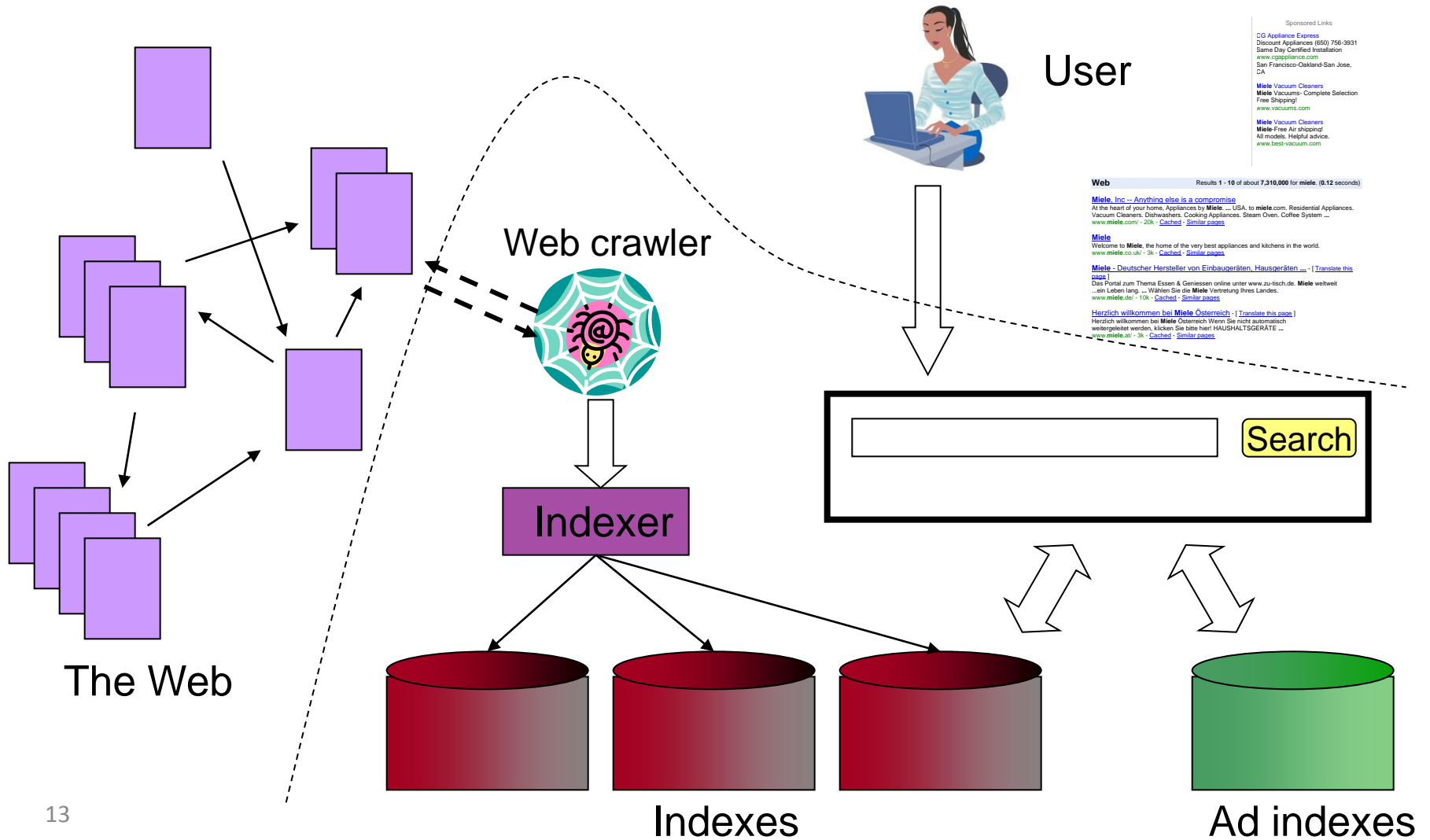
- Types of Path/Usage Information
 - ➔ Most Frequent paths traversed by users
 - ➔ Entry and Exit Points
 - ➔ Distribution of user session duration
- Examples:
 - ➔ 60% of clients who accessed `/home/products/file1.html`, followed the path `/home ==> /home/whatsnew ==> /home/products ==> /home/products/file1.html`
 - ➔ (Olympics Web site) 30% of clients who accessed [sport specific pages](#) started from the [Sneakpeek](#) page.
 - ➔ 65% of clients left the site after 4 or less references.

Search Engines for Web Mining

The number of Internet hosts exceeded...

- 1.000 in 1984
- 10.000 in 1987
- 100.000 in 1989
- 1.000.000 in 1992
- 10.000.000 in 1996
- 100.000.000 in 2000

Web search basics



Search engine components

- Spider (a.k.a. crawler/robot) – builds corpus
 - Collects web pages recursively
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - Additional pages from direct submissions & other sources
- The indexer – creates inverted indexes
 - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc.
- Query processor – serves query results
 - Front end – query reformulation, word stemming, capitalization, optimization of Booleans, etc.
 - Back end – finds matching documents and ranks them

Web Search Products and Services

- Alta Vista
- DB2 text extender
- Excite
- Fulcrum
- Glimpse (Academic)
- Google!
- Inforceek Internet
- Inforceek Intranet
- Inktomi (HotBot)
- Lycos
- PLS
- Smart (Academic)
- Oracle text extender
- Verity
- Yahoo!

Specifying field content in HotBot

<input type="text"/>		SEARCH
Look For	<input type="text" value="all the words"/>	? click to see explanations of how these options work.
Language	<input type="text" value="any language"/>	
Word Filter	<input type="text" value="must contain"/> <input type="text" value="the words"/>	
	<input type="text" value="must not contain"/> <input type="text" value="the phrase"/>	
	<input type="text" value="more terms"/> <input type="button" value="+"/>	
Date	<input checked="" type="radio"/> <input type="text" value="anytime"/>	
	<input type="radio"/> <input type="text" value="After"/> or on <input type="text" value="January"/> <input type="text" value="1"/> , <input type="text" value="2000"/>	
Pages Must Include	<input type="checkbox"/> image <input type="checkbox"/> audio <input type="checkbox"/> MP3 <input type="checkbox"/> video	
	<input type="checkbox"/> Shockwave <input type="checkbox"/> Java <input type="checkbox"/> JavaScript <input type="checkbox"/> ActiveX	
	<input type="checkbox"/> RealAudio/Video <input type="checkbox"/> extension: <input type="text"/> (.gif)	
Location/Domain	<input checked="" type="radio"/> Region <input type="radio"/> Domain	
	<input type="text" value="anywhere"/>	<input type="text"/> <small>(.com, .edu) website: (wired.com, etc.) country code: (.uk, .fr, .jp)</small>

Natural language interface in AskJeeves



The screenshot shows the Ask Jeeves website interface. At the top right, there are links for [About](#) and [Help](#). Below these are four navigation buttons: [Ask Jeeves Home](#), [Browse by Subject](#), [Ask Other People](#), and [Shopping Guide](#). On the left, a cartoon character of a man in a suit is pointing towards the search area. The main heading is "Ask Jeeves" with "Ask.com" underneath. Below the heading is the question "What can I help you find?" and a search input field. To the right of the input field is a red "Ask" logo. Below the input field is a tip: "Tip: Use a question, phrase, or word - Jeeves is flexible." At the bottom, there are links for international sites: [Ask Jeeves Español \(Pregunta.com\)](#), [Ask Jeeves UK \(United Kingdom\)](#), and [Ask Jeeves Australia](#). Below that are general information links: [About](#), [Business-to-Business Solutions](#), [Advertise](#), [Investor Relations](#), and [Become an Affiliate](#).

Three examples of search strategies

- Rank web pages based on popularity
- Rank web pages based on word frequency
- Match query to an expert database

All the major search engines use a mixed strategy in ranking web pages and responding to queries

Rank based on word frequency

- Library analogue: Keyword search
- Basic factors in HotBot ranking of pages:
 - words in the title
 - keyword meta tags
 - word frequency in the document
 - document length

Alternative word frequency measures

- Excite uses a thesaurus to search for what you want, rather than what you ask for
- AltaVista allows you to look for words that occur within a set distance of each other
- NorthernLight weighs results by search term sequence, from left to right

Rank based on popularity

- Library analogue: citation index
- The Google strategy for ranking pages:
 - Rank is based on the number of links to a page
 - Pages with a high rank have a lot of other web pages that link to it
 - The formula is on the Google help page 😊

More on popularity ranking

- The Google philosophy is also applied by others, such as NorthernLight
- HotBot measures the popularity of a page by how frequently users have clicked on it in past search results


Expert databases: Yahoo!

- An expert database contains predefined responses to common queries
- A simple approach is subject directory, e.g. in Yahoo!, which contains a selection of links for each topic
- The selection is small, but can be useful

Expert databases: AskJeeves







- AskJeeves has predefined responses to various types of common queries
- These prepared answers are augmented by a meta-search, which searches other SEs
- Library analogue: Reference desk

Best wines in France: AskJeeves

You asked: 

Tip: Try to keep your search questions short and to the point.

Click Ask for Answers! I have found the **answers** to the following questions:

-  Where can I see a list of the top 100 wines of 1999?
 -  Where can I buy from an online wine shop?
 -  Where can I find information on wine from ?
 -  Where can I find ?
 -  Where can I get a crash course in choosing ?
-
-  Post your question and get answers from other users in our Community Forum

People with **similar questions** have found these sites relevant:

 [All French wines, Bordeaux Burgundy and Champagne wine - Buy French Wine](#)

[Online](#)

NouvellePage4

From: <http://www.french-wine-shop.com/>

 [Slow Food Guide to the Wines of the World](#)

Web site by Intesys 2000 wineries 6500 wines described 174 top wines | Info | Catalogue | Top Wines | Search | Help | Copyright © 1996, Arcigola Slow Food and Veronafiere

From: <http://www.veronafiere.it/slowwines>

Best wines in France: HotBot

WEB RESULTS 53,600 Matches 1 - 10 [next](#) >>

1. [Jeroboam : Wine Cellar Management software.](#)

Software which manages your own wine cellar, helps you to serve and to pair your best vintages. Other Features include: Stock management, Pairing wines and dishes, help to serve and age wines. Available in English and French version.

2. [The best wines of South Africa](#)

The best wines of South Africa Unique dans la région Mouscronnoise et Tournaisienne Invitation découverte en seconde page ... Le soleil de l'Afrique du Sud Apporte au fruit de sa terre une saveur qui jalouse les vins que nous connaissons. Les viticu

3. [All French Wine Shop](#)

They are French and they sell wine. Customers can't ask for more when shopping for Bordeaux, Burgundy, Alsace and Cote du Rhone wines.

4. [French Wines](#)

Check out this guide to French wines. Includes a wine vocabulary section and an overview of French wineries.

5. [French Wines - Other](#)

Named as the best wine shop on the Internet by Money magazine and one of the top 10 wine retailers in the nation by the publishers of the Wine Spectator.

Best wines in France: Google

[Au Web du Vin, Cave à vin et vente en ligne](#)

... Accord Mets et Vins. Our cellar shop of finest **french wines** and alcohols, Boutique de vente en ligne et vins et alcools. ...
[www.webduvin.com/](#) - 10k - [Afrit](#) - [Svipaðar síður](#).

[Best Cellars: French Wines](#)

... for its quality and diversity of **wines**, and only Italy can be compared with it in terms of quantity. **Best** Cellars currently stocks these fine **French wines**. ...
[www.bestcellars.ag/france.htm](#) - 68k - [Afrit](#) - [Svipaðar síður](#).

[All French wines, Bordeaux Burgundy and Champagne wine - Buy ...](#)

... on line sales of **French wines**. Bordeaux and Burgundy **wines**, Champagne red and white wine. The **best** vineyards of the **french** soils, grown up with skill from the ...
Lýsing: Online sales of **French** great **wines**: Bordeaux and Burgundy red and white, Champagne, Alsace, Loire...
Flokkur: [Regional](#) > [Europe](#) > [France](#) > [Business and Economy](#) > [Shopping](#) > [Wine](#) > [Retailers](#)
[www.french-wine-shop.com/](#) - 6k - [Afrit](#) - [Svipaðar síður](#).

[Best links concerning shopping for french wines](#)

... \$15.50 only on Amazon store !! Why waste your time searching for shops on line ? Here is my Selection among **best** sites about **French wines** on the Internet. ...
[www.french-wines.com/Link.htm](#) - 7k - [Afrit](#) - [Svipaðar síður](#).

[Best Cellars - French Red Wines](#)

... Red **Wines** Taste Guide - A (light) to F (full ... bodied Fitou is made from the **best** traditional southern **French** grape varieties grown between Narbonne and ...
[www.devon.directory.co.uk/bestcellars/pages/html/wines/france/ red2.htm](#) - 22k - [Afrit](#) - [Svipaðar síður](#).

Some possible improvements

- Automatic translation of websites
- More natural language intelligence
- Use meta data on trusty web pages

Predicting the future...

- Association analysis of related documents (a popular data mining technique)
- Graphical display of web communities (both two- and three dimensional)
- Client-adjusted query responses