The Complete Reference

Chapter 9

Search

Many users will find browser-oriented navigation systems an inefficient way to find what they are looking for. Often, a user knows something exists and just needs to find it within a site. Search functions appeal to power users, frequent visitors, and the plain impatient, who are all looking to find a result quickly. A well-executed search facility is one major advantage a Web site has over printed media, as it gives users greater control over a site's content, allowing them to filter it to just what they want to see. Larger Web sites, especially those with complex data, must provide search facilities—and may consider making it the central navigation method. Searching facilities, however, must be designed with the user in mind. Before adding search to a site, give careful consideration to how users expect a search to work, the type of search required, the design of the search page, the help system, and the types of search-result listings.

How Users Search

Before getting into the theory of how search systems work and how to utilize both external and local search engines to improve site design, consider first how people actually use search facilities. People search for a variety of reasons. A big reason to search is to look for something known to exist. An example of known-item searching is when a user is looking for a particular part, like "RBA-4456." In this case, it is usually easy for the person to locate the item in question, assuming that the search facility has seen it before and particularly if the item is fairly uncommon.

Oftentimes, however, users may not know if the item they are searching for exists or not—in fact, they might just be searching to see *if* such an item exists. A query like "Robot shops" might be used for a general search that could have as its object the existence of a shop that repairs robots. Other times, a user may perform an exploratory search to get a sense of the extent of something. For example, a query for "Robot Butler" might be done not only for the existence of such a device, but to see the extent of sites offering information on a metallic servant. It would seem that known-item searching is what users would generally use search engines for, but, oddly, existence and exploratory searching are commonly employed.

Regardless of the reason for a search, users go through four basic steps.

Formulate a Query Depending on the search facility being used, the query formed by the user may vary greatly. A simple query might include only keywords, like "Robot Butler." More complex queries might include Boolean queries like "Robot AND Butler." Many search engines utilize queries filled with symbols, such as "+Robot +Butler –Jeeves." The search facility may even support a natural language interface where the user can ask something like "Where can I buy a robot butler?" The query formulation might not just include the selection of various search words, but also may offer refinements to search criteria, such as indicating the areas to search, a date range to query, data types to search, and so on. Users may also at this point specify how they would like their results returned— say, for example, ten at a time, sorted by last update. However,

further criteria beyond keywords are usually part of an advanced search and are usually performed only by more experienced users.

Execute the Search and Wait for the Result The second step of searching usually consists of a simple button click, followed by a short wait for network round-trip time plus time required for the search engine to run the query and list the result. While there isn't much going on interactively during this phase, don't ignore it. The user views this as a discrete step in the process and will not wait around forever for results to appear.

Review the Results Once the results have been listed onscreen, the user will peruse them to see if there is anything interesting in the list. During the review stage, the user will rely greatly on supplementary information, such as relevancy ranking and a description of the results including summaries, modification dates, and file sizes. During the review stage, the user may sort or filter the results in order to help them determine what to do. However, the actual decision concerning results will be influenced highly by what is actually returned by the query. Results will vary from the so-called negative result that contains no matches to the huge volumes of data when every document in a collection is returned. Most cases will be somewhere in between these extremes.

Decide What to Do with the Result On the basis of the results, the user decides what to do. For example, if there are no results, the user may search again with a new query or may simply give up. If the search didn't appear to provide the correct answer, the user may also search again. When the search provides too many results, the user may try to refine the search. Maybe the user selects a few of the choices in the search results to examine. While there may be numerous variations, basically the user decides to explore some of the results, redo or refine the search, or just quit.

This basic overview is important to keep in mind when designing a search facility. Later in the chapter we'll present theory and practical design suggestions that deal with each step the user takes during the search process. However, before doing this we'll present an overview of how search engines function.

How Search Engines Work

So how do search engines work? First, a large number of pages are gathered off a Web site (or the Web at large, in the case of public search engines) using a process often called *spidering*. Next, the collected pages are indexed to determine what they are about. Finally, a search page is built where users can enter queries in and get results related to their queries. The best analogy for the process is that the search engine builds as big a haystack as possible, then tries to organize the haystack somehow, and finally lets the user try to find the proverbial needle in the resulting haystack of information by entering a query on a search page.

Gathering Pages

Every day the Web is growing by leaps and bounds. The true size of the Web is unknown, and it will undoubtedly increase even as you read this sentence. At any given moment numerous documents are added and others are removed. Gathering all the pages and keeping things up-to-date is certainly a significant chore. Users always want to know which search engine covers the most of the Web, but the truth is that today even the largest search engines index maybe only a third of the documents online. Some index only a few percent. This may change in the future, but for now be happy that not everything is indexed. The resulting mess of information to wade through would be even worse. In the case of local site search engines, the index might also not cover the entire site nor be updated often.

Most search engines use programs called *spiders*, *robots*, or *gathers* to collect pages of the Web for indexing. We'll use the term "spider" to mean any program that is used to gather Web pages. Spiders start their gathering process with a certain number of starting point URLs and work from there by following links. In the case of a public search engine, starting URLs are either submitted by people looking to get listed or built by forming URLs from domain names listed in the domain name registry. Local search engines work in the same way, but may be given a very small number of starting points if the site is well connected.

As the spider visits the various addresses in the list, it saves the pages or portions of the pages for analysis and looks for links to follow. For example, if a spider were visiting the URL http://www.democompany.com, it might see links emanating from this page and then decide to follow them. Not all search engines necessarily index pages deeply into a site, but most tend to follow links—particularly from pages that are well linked themselves or contain a great deal of content.

Indexing Pages

The next step search engines take is attempting to determine what a page is about. This is usually called *indexing*. The method each search engine uses varies, but basically an indexer looks at various components of a page, including possibly its **<title>**, the contents of its **<meta>** tags, comment text, link titles, text in headings, and body text. From this information it will try to distill the meaning of the page. Each aspect of a page might have different relevance, and within the actual text, the position or frequency of different words will be taken into account as well. However, not all content within a page matters to a search engine. For example, *stop words* are words that a search engine ignores, normally because they are assumed to be so common as to carry little useful information. Examples of stop words might be "the," "a," "an," and so on. Most search engines have some stop words, but some engines like AltaVista claim to even index common stop words like "the."

While the use of stop words may improve a search engine by limiting the size of the index file and focusing it on more content words, it may not match how users think about queries. Novice users may feel "The Best Butler Robot" is a better query than

"Best Butler Robot." Sometimes the stop word may be important to the search. Consider searching for a song title like "Rock the Town." "The" is an integral part of the term and without it many other songs may come up. However, if the search were for "Rock the Casbah," it would be easier to throw out the noise word "the," given that "Rock" and "Casbah" rarely occur near each other. Deciding what stop words should be used can be very problematic given the broad topic domain of many Web sites.

Once a page has been analyzed for the various keywords, it is ranked in relation to other pages with similar keywords and stored in a database. Ranking is the very secret part of search engine operation. How a particular search engine decides one page should be ranked higher than another is what search engine promotion specialists are always trying to figure out. A very popular way to rank pages today is based upon determined site landmarks. Home pages and major section pages may be given higher weight than other pages in a site. Pages that have numerous incoming links will also be given extremely high ranking.

Providing a Search Mechanism

The final aspect of a search engine is the search page itself. A search page is the interface the user makes their query from, and it generally contains a primary query text box as well as other search fields for advanced users who may want to modify a query. The degree of complexity of the search page varies greatly in public search engines. Consider the difference in interfaces between basic and advanced search pages for various public search engines shown in Figure 9-1.

Users can enter queries as simple natural language questions—like, "Why is the sky blue?" (as encouraged by sites like www.ask.com)—or as complex Boolean expressions using advanced filters. Once queried, the search engine will retrieve the pages that meet the criteria and present them on a result page. Figure 9-2 shows a result page for the search engine Google (www.google.com).

From the result page, users can pick some results to explore, further refine the search with a new query, or just give up and try another method to locate what they were hunting for. The general function of search engines is illustrated in Figure 9-3.

Understanding what people expect Web-wide search engines to do is important, because users will bring their past experiences with searching to bear when using your local site search. Labeling, form layout, and result pages should somewhat mimic what users have come to expect from the public search engines. However, be careful not to directly imitate what public engines do. Public search sites aim primarily to get users to starting points for searching, while local search facilities on a site aim to provide a high degree of search accuracy. In fact, public search engines aren't always terribly accurate. They are often geared towards the needs of advertisers and the demands of dealing with the numerous tricks people employ to try to improve their site's ranking.

> **Rule: Utilize past user experience with public search engines by using similar layout and labeling in local search facility design, but avoid imitating aspects of public search engines that deal with the uncontrollable nature of public Web sites.**

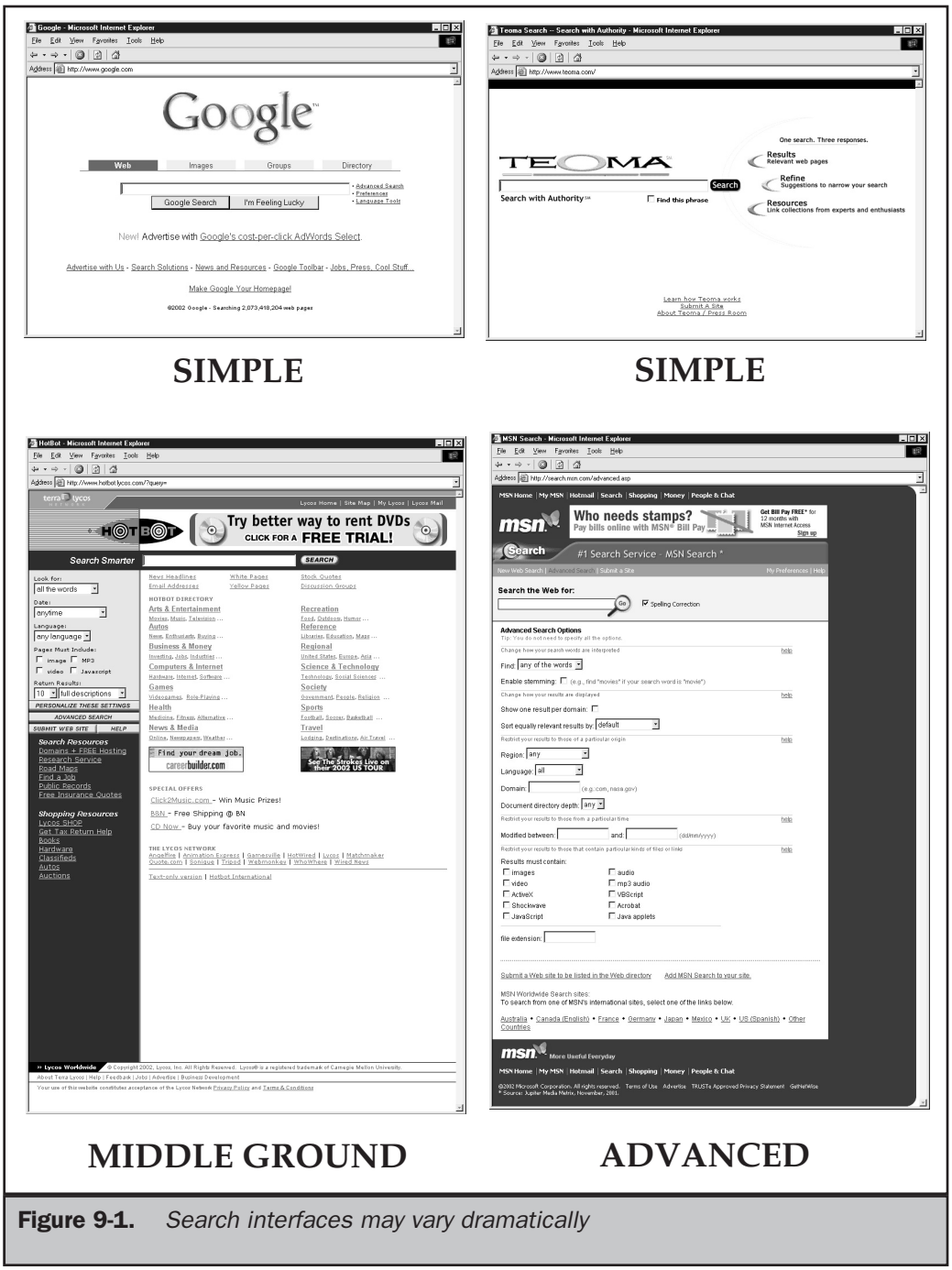


Figure 9-1. *Search interfaces may vary dramatically*

Figure 9-2. *Google's result page is clean and simple*

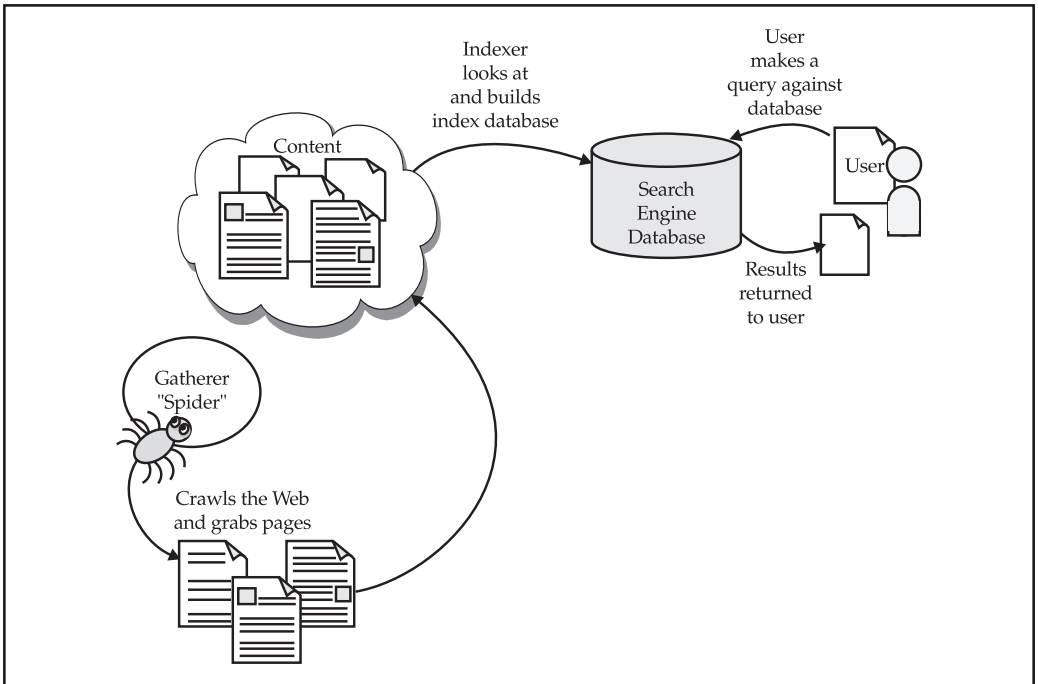


Figure 9-3. *Overview of search engines*

Adding a Search Facility

The following eight steps can summarize the process of adding a search facility to a site.

Step 1: Decide what to index

Do you want to index every document in a site or only certain documents? Often, it is only a parts catalog, technical support database, or other area that a user wants to search. Don't just index everything because you can.

Step 2: Decide how you want to index the information

Once you have determined what you should index, you will need to determine how it will be indexed. Should the search engine just create a free text index of the document set, where every non stop-word is recorded, or would it be better to create a special search term vocabulary and relate search terms to particular pages in the site?

Step 3: Select a search engine

It is very important not to select the search engine until you've figured out the volume and type of information you wish to search, as well as how it will be indexed. There are numerous search engines available, both free and commercial. Search engines can be installed locally on your system or outsourced to third parties, who will run the search facility for you. For pointers to some search engines and services, see http://www.searchtools.com.

Step 4: Design the search interface

Design the search screen to account for the types of searches the user may perform. Often, searches are separated into basic and advanced forms. The search interface should be integrated into the site, should meet the search needs of the users, and should fit the type of data being searched.

Step 5: Design the results pages

Make sure to consider building pages that deal with positive results when a query is successful, as well as negative results when nothing is returned.

Step 6: Index the data

During this step, the search engine is used to *crawl* all or part of the site and build an index. You may actually be forced to manipulate the index by hand to create optimal queries.

Step 7: Integrate the search engine with the search interface

This step involves making the search interface access the index. Generally, this is just a matter of setting the *action* attribute of the **<form>** tag used to implement the search form. Integrating the result page is a little more difficult, but is often a matter of taking the designed result page and making it into a special template the search engine can read.

Step 8: Test and monitor

A key aspect of implementing a search engine is making sure to test that it gives back the correct results for important queries. Search engines should also be monitored and common queries identified. Users also should be allowed to rate the value of the individual search results so that refinements can be made.

It is also very important for dynamic sites to re-index their search features on a regular basis. Such sites could be re-indexed manually by webmasters or editors when new content is added to the site or be automatically set up for regularly scheduled re-indexing.

The focus of the next few pages is not on how to actually create an index, which will vary greatly by the data being indexed, as well as the search engine being used, but to show how to design the various aspects of a search interface.

Designing the Search Interface

Assuming that a search facility is needed, a designer should first and foremost consider what the user wants to search for. Far too often, search engines are added to a site and set to index everything using a free text search. Similar to a Web-wide search, users pound their heads as they search for a particular part number like KF-456 only to be shown every single document the part number occurs in, ranging from press releases to technical notes. To the user, the ordering of the documents from this type of search may seem arbitrary, with the most important document not appearing first in the list. What's interesting is why this form of search was used. Designers assume that since public search engines work like this, so should their local search engine. This seems like a good idea—users are familiar with formulating search strings at public sites and bring this knowledge with them to your site. However, global search engines are not very accurate for a variety of reasons, including the fact that numerous sites try to fight their way to the top of returned results. Public search engine results don't always seem to make sense, and the ordering often seems more random than systematic.

Consider that in your own site, if you want a particular page to be shown when a user types in "Robot Butler," you can cause that page to be shown. Remember, when building a local search facility, to copy the style, syntax, and interface of public Web search engines, but don't imitate their imprecise functionality.

The main advantage of local searching is that you can utilize controlled vocabularies to deal with what users will probably want to search for. Besides relating keywords with certain pages in a more precise manner, you may even suggest common queries for users to run. Remember, local search engines provide designers with a much greater degree of control than public search engines.

Accessing Search

You should consider how your users will access the search facility. Some sites create a special button labeled "Search" that, when selected, takes the user to a special search page. Other sites utilize a search field within all pages. A visual comparison of the two approaches is shown in Figure 9-4.

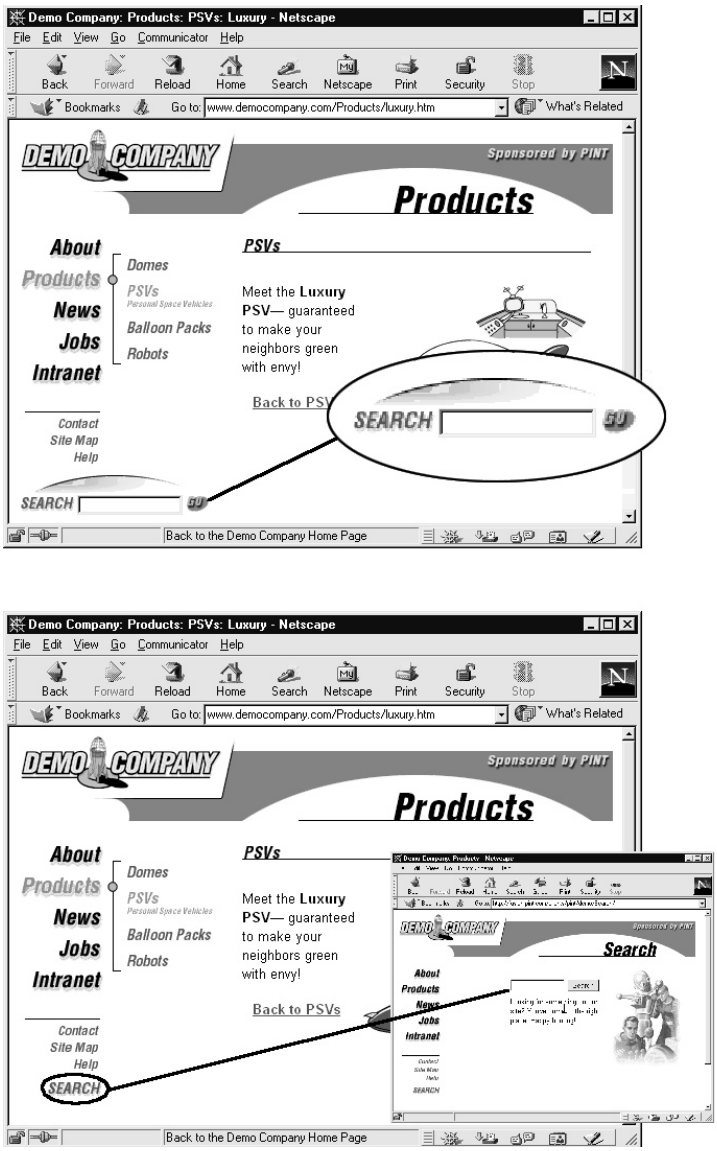


Figure 9-4. *An in-page search and a special search page*

While putting the search directly on the page eliminates one click for the user, a search field within a content page must be very basic. There still may be a need for

a special search page if more complex queries are to be formed. It really isn't possible to put advanced search mechanisms within every page, as it tends to make the search facility too prominent and takes away from the page's primary purpose of delivering content. So the question is really to expose a simple search facility on content pages or provide it on a special search page. Regardless of the choice, search should be easily found from every page in a site.

Suggestion: When search is available in a site, include a search button or field on all pages.

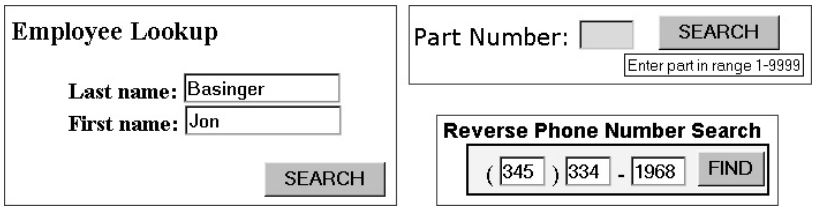
Designing a Basic Search Interface

The search facility of a site should look the same as the rest of the site. Often it is not the same because it is added by technical staff, who may not be concerned when setting up the search templates to match the site's look and feel. Users who utilize such search engines may feel they have left the site if the look changes greatly. Look at the two search facilities shown in Figure 9-5; the need for integration should be obvious.

Rule: Search forms and result pages must match the look and feel of a site.

Also, the search form should fit the type of data being searched. For example, if users are searching for objects that are colored, shouldn't the search form provide a way to specify by color? The example search interface in Figure 9-6 for searching for personal space vehicles shows how search forms should match the content that is being searched.

Consider the golden rule of designing a search facility for a site—the more we know about what users are looking for, the better able we'll be to help them find it. One way to do this is to analyze what people search for by looking at the queries they enter. No matter how we figure out what users search for, we need to help users narrow down their search properly. For example, if we are searching for names, try to help people enter in last names or first names into individual text boxes rather than just letting them type names into a single text box. If part numbers are being searched in a range from 1 to 10,000, then let people know that that is the range, limit them to the range, and alert them if they are out of range. A ToolTip set using the **title** attribute in HTML or a simple JavaScript is an easy way to let people know about ranges without explicitly printing them onscreen. A few search forms that fit the data being searched are shown here:



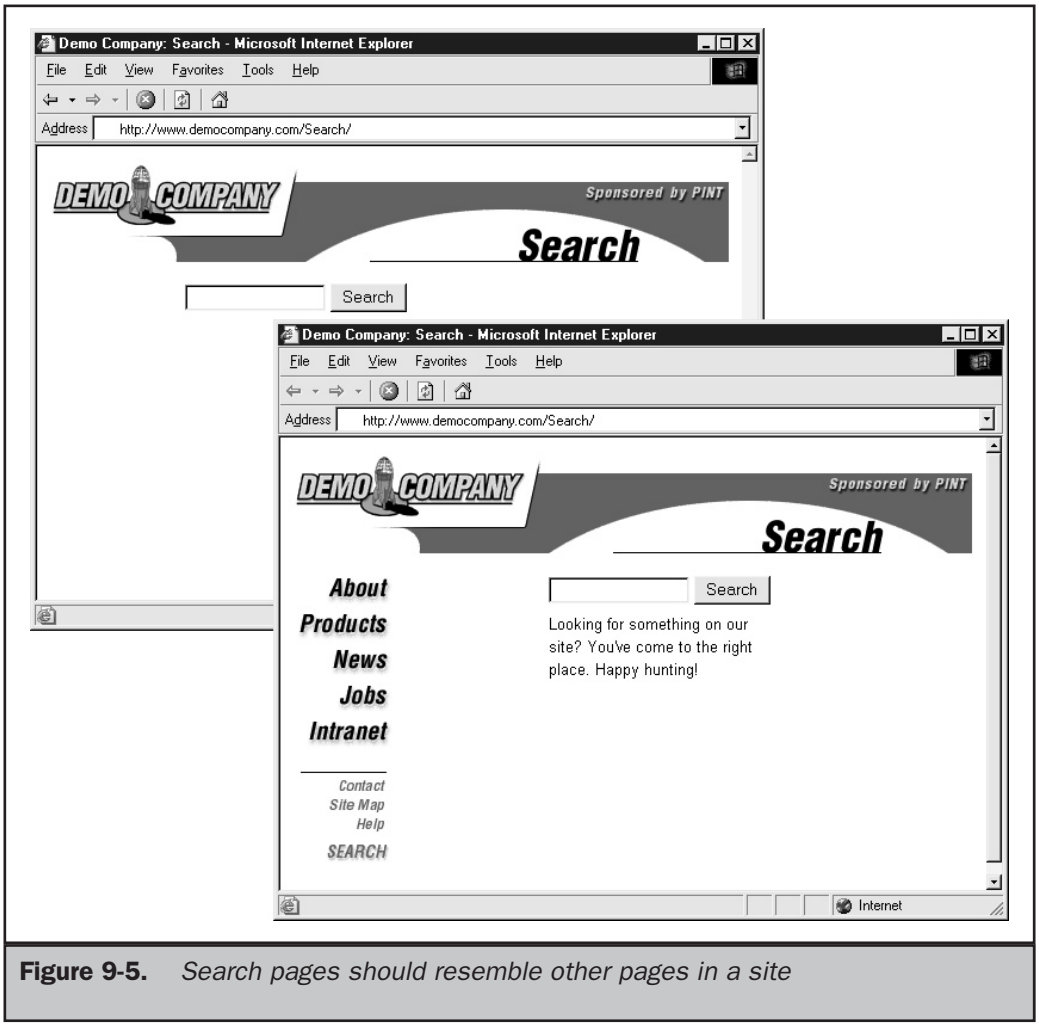


Figure 9-5. *Search pages should resemble other pages in a site*

Rule: A search form should match the content being searched.

The primary element of a search form is the actual search query field. A big question is how long should the search field be. The query text field should be large enough to hold at least a few search terms without scrolling. On average, users type two keywords in search fields.

The size of the search field also is related to the emphasis of the search task for the page. If search is the primary emphasis of the page and users are going to form complex searches, an input size in the range of 30 to 40 characters is common. A survey of the public search engines shows that most use a size of 30, 35, or 40 characters for their

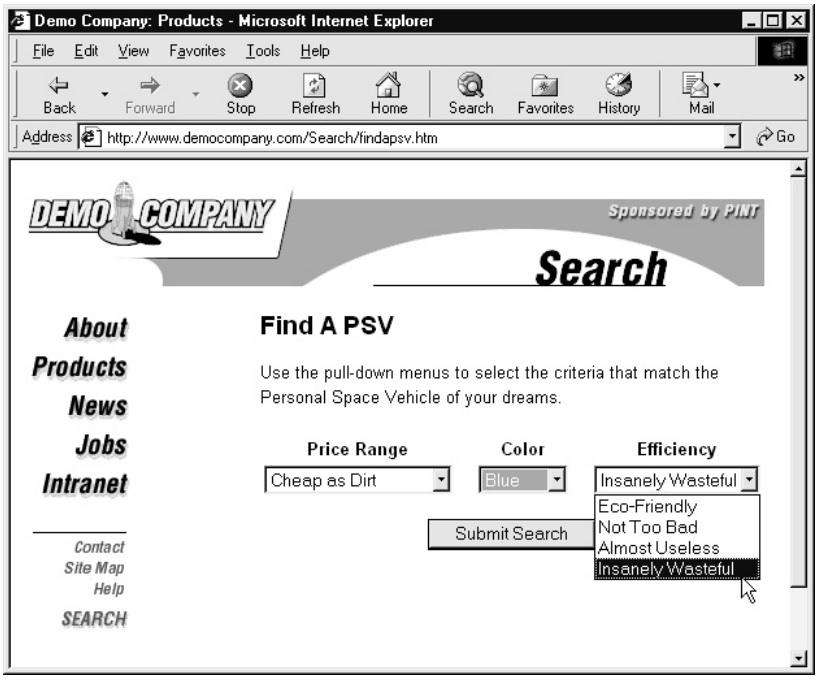


Figure 9-6. *Search forms vary based on content being searched*

primary search field, though Google is much larger at 55 characters. This size makes the search field a fairly large element, width-wise, on a typical page. When search is a secondary aspect of a page, the size should be about half the size—usually from 15 to 18 characters, which should allow a few keywords for a simple query. Of course, the size of the search box should always be designed with the search terms and the page layout in mind.

Suggestion: Primary search text boxes should be about twice as big as secondary search text boxes.

The second aspect of the search form is the button to execute the search. Sometimes a form button is used, while other times a custom button is used. The use of a form button is probably slightly more intuitive for users. The label of the button also varies. Some favor "Search," others "Find," and some even something as simple as "Go." A lot of this depends on the context of the search. If the word "search" is used to label the field, labeling the button "search" seems a little redundant.

The search form should fit the types of users the site is designed for. For example, a search facility for kids might be playful and have few instructions, while a search facility for engineers might contain a variety of fields for visitors to tune their searching. Simple search forms should be separated from advanced ones.

Advanced Search Form Design

Advanced search forms are more challenging to design, particularly if there are many ways for the user to tune the search. First, if the search is to allow Boolean searching using AND, OR, or NOT, the form must either be designed with pull-downs to separate search terms or provide explicit instructions for users on how to build Boolean queries, as shown in Figure 9-7. However, creating Boolean expressions can be a serious problem for many users. Try to avoid suggesting their use in basic searches where possible.

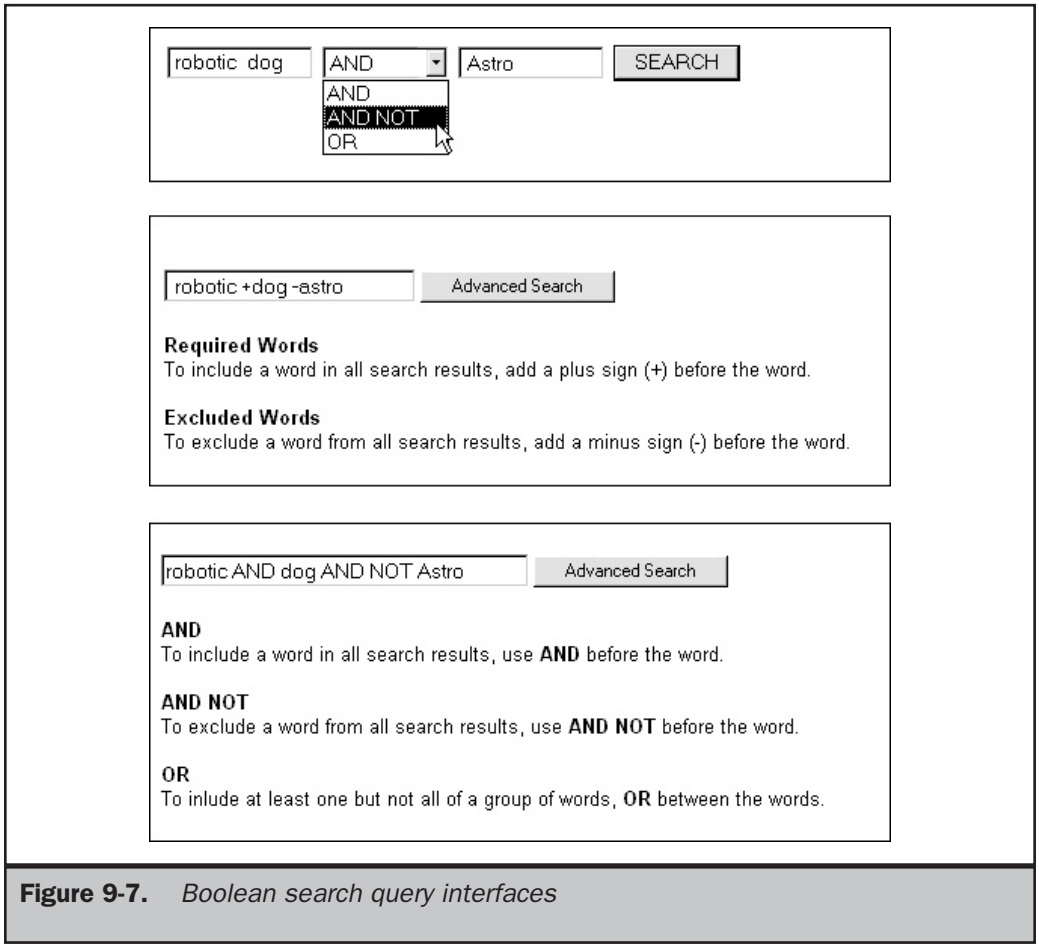


Figure 9-7. *Boolean search query interfaces*

Advanced search forms often include various fields to limit the time of search. For example, forms may allow users to specify a date range to search when looking for time-sensitive information. They may also be able to limit the type of data to search by format (image, PDF, sound, and so on), as well as by content type (for example, press releases or specifications). Some search facilities allow the user to search only certain parts of a site. This may be called a *scoped* search. Unfortunately, users may not understand a site's sectioning, so it may be better to allow limits on topics, categories, or ideas rather than on sections of a site. A common way to do either form of scoped search is using a pull-down as shown here:

Suggestion: It is generally better to limit a scoped search to a topic, category, or idea rather than a section of a site.

Other possibilities for an advanced search facility include allowing users to limit the number of results to be returned, to set the way results should be returned, and to search for meta information, such as document authors. Figure 9-8 shows an example of an advanced search form.

A very important part of advanced search forms is the instructions. Not all search engines work alike, and you should provide explicit instructions for the user, either directly on the search screen or using pop-up windows. Do not use a separate page for your search instructions, as it forces the user to either print out the instructions or quickly memorize the information. Besides instructions, example queries and field usage should also be provided in an advanced search page.

Rule: Advanced search facilities must provide instructions and examples.

Result Page Design

Designing result pages must take into account two extreme possibilities: no results and way too much information. Even when just about the right information is returned, a well-designed result page should help the user discern what is relevant. The rule of thumb for a result page: the more information the better—often people can't determine the value of one result over another. A well-designed result page should include the items shown in Table 9-1.

Not all types of search engines are able to provide all of these items—particularly advanced relevancy and matching indication. However, designers should strive to include all elements in a result page.

Rule: Result pages should provide as much information as possible so users can decide what items to peruse further.

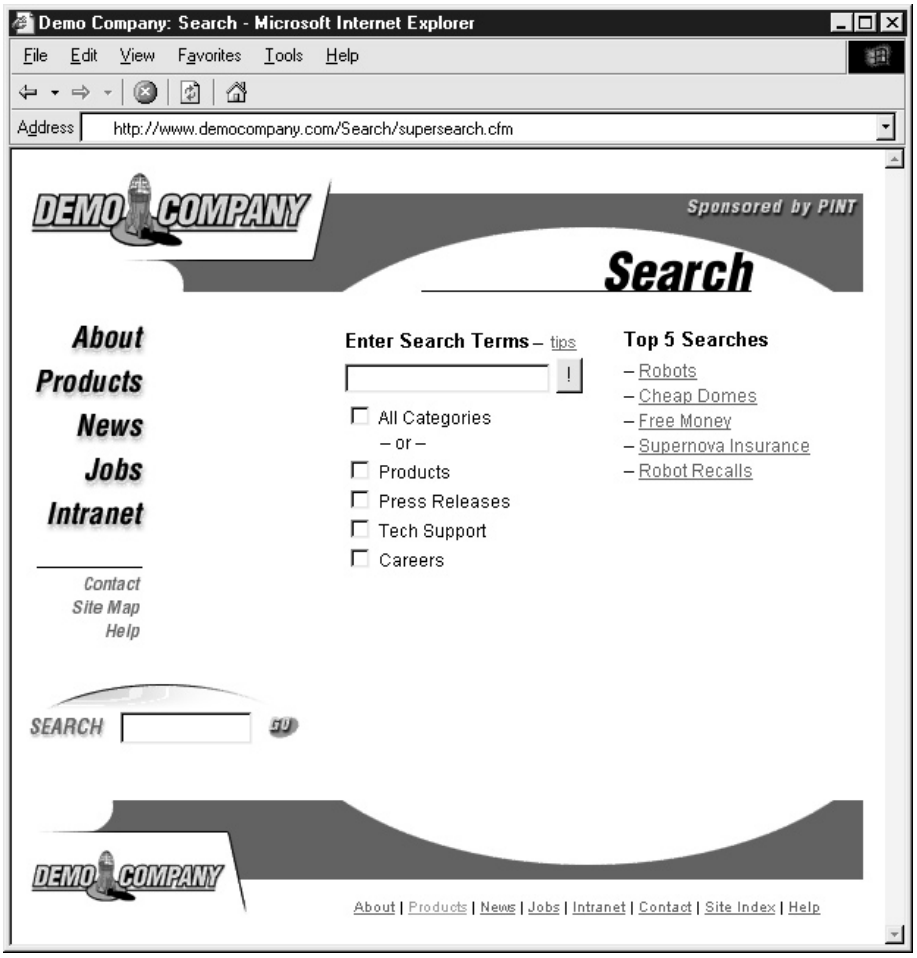


Figure 9-8. *Advanced search forms should be carefully designed*

Result Page Element	Description
Original query	The original query string used should be prominently displayed on all result pages so the users don't have to recall what search string they used.

Table 9-1. *Common Result Page Elements*

Result Page Element	Description
The scope of the search and the results found	The total number of documents searched and returned should be indicated (for example, 10,000 documents searched, 20 matches).
Context of current results	There should be some indication as to what part of the result list the user is looking at (for example, page 2 out of 10, or items 30–40 out of 200).
Page or document titles	Each item returned should be clearly titled.
URL of returned page	The actual URL of the individual documents should be shown, as it may provide useful information to the user.
Page summaries	A brief summary of a returned page's contents should be shown. This is often picked up either from the **<meta name="DESCRIPTION">** element or the first few lines of text in a document. A user may have the option to show or hide the page descriptions.
Date or time information of results	Minimally, the create date or date of last update of a returned document should be shown. Some search facilities also provide an indication of the time the index was last built, the time it took to search the index, and the time the query was performed.
Size of returned pages	The file size of the document returned should be indicated. This is especially important if the files being searched are large binaries.
Type of result	In some searches, other forms of data such as Adobe Acrobat, Microsoft Word, or multimedia data may be returned. Make sure to indicate the format of data with a label or icon.

Table 9-1. *Common Result Page Elements* (continued)

Result Page Element	Description
Relevancy of results	A relevancy ranking should be clearly indicated. Usually, search results are ranked from highest to lowest. A percentage score or bar should be used to show the difference between items.
Keyword matches	Since users are highly annoyed when they are unable to figure out why a particular page is returned for a query, show the keywords matched and, if possible, highlight these words in context in the summary. If possible, when the user selects a document, the query terms should also be clearly highlighted.
Navigation	Navigation to move through the result set should be provided. Common buttons include "Next 10 documents" or "Previous 10 documents," where the step value changes depending on the user's preference. Navigation to see the first or last page in a result set is also sometimes used.
Refinement options	The ability to refine the query should be present. Users may be able to search against the result set or even perform a brand-new query.
Help	Help information explaining the format of the results should be available.

Table 9-1. *Common Result Page Elements* (continued)

Search result pages often lack any provision for site navigation. When users access a results page, they are not just searching—they may also want to switch back to a browsing mode to investigate results. Remember, users are just looking for an answer, and they may move in and out of approaches in their hunt, so provide browsing facilities on search results when possible in case the user wants to leave the result page easily. Figure 9-9 presents a search results page that includes most of the elements listed in Table 9-1.

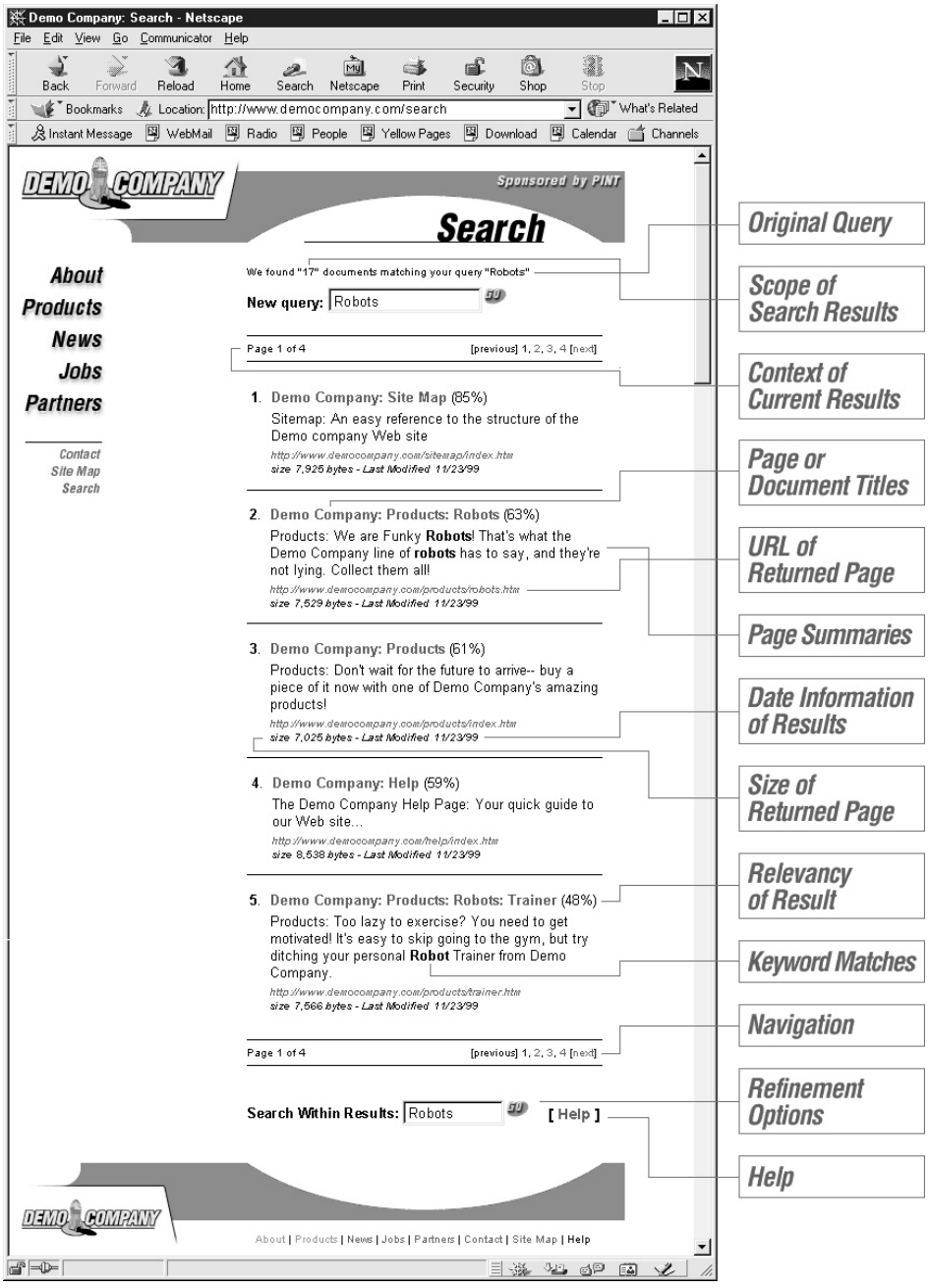


Figure 9-9. *Example search result page*

One aspect of search result pages that may appear obvious but is often overlooked is that the format of data returned should be carefully considered. For example, just listing a page title, URL, and description may not be enough for a user to make a decision about one choice over another. For example, if a user performs a search of products, it might be possible to output small thumbnails of the products that match the user's criteria, as shown in Figure 9-10.

Rule: The format of search results should fit the data that is being returned.

The key aspect of designing a positive search result page is helping the user find and make a decision about which returned items to pursue further. However, in view of the public Web search engines where far too much is often returned, designers should carefully consider the negative result when nothing has met a user's search criteria.

Negative Results Page

When a query results in no matches, the result page should try to help the user identify what went wrong. In some cases, it may be just that there is nothing that matches the search terms. In other cases, the user may have simply used the search facility incorrectly. A good negative result page should indicate which of these two conditions is applicable as well as perform the functions shown in Table 9-2.

Feature	Description
Clear failure message	Make sure the user knows that the query failed and perhaps why it failed. Indicate the number of documents searched and provide a clear message indicating that the search failed.
Search again mechanism	The query used should be shown, and the option to search again should be directly available from the result page (as with a positive results page).
Help information	Probably the most important aspect of a negative results page is to provide clear and useful help. First, provide tips that might explain why the search failed. For example, often search terms are misspelled. If the search engine doesn't provide spell checking, consider adding an option to spell check the query string. If possible, show terms that are similar to the term searched for. Consider showing the common search terms. Finally, make sure that help information on how to use the search facility is readily available.

Table 9-2. *Features of a Negative Result Page*

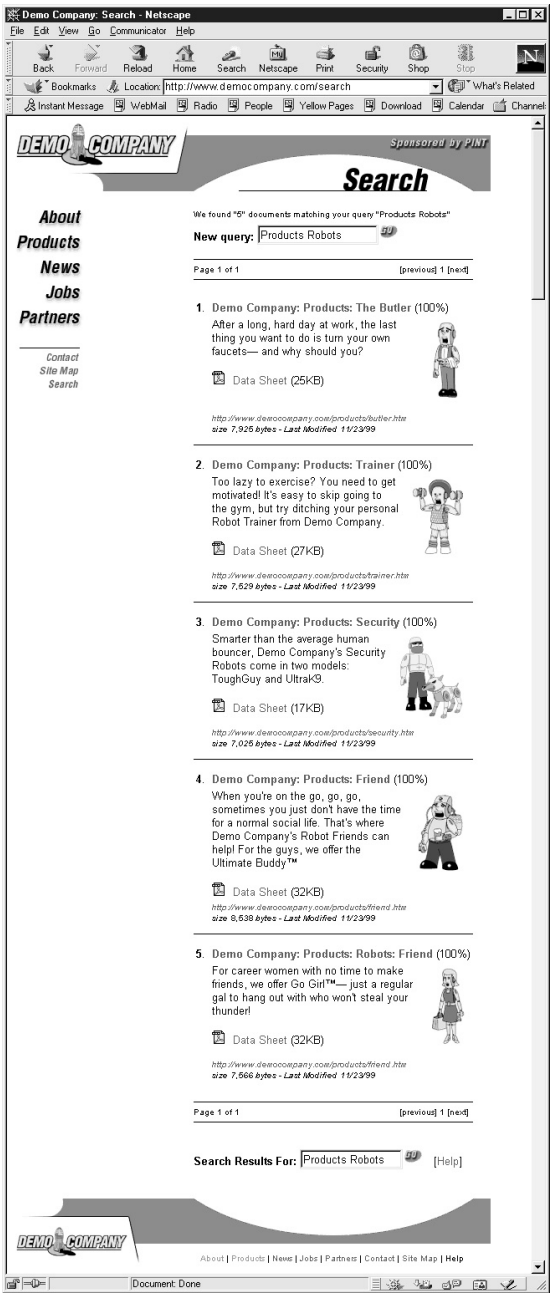


Figure 9-10. *Search results vary based on the data searched*

Figure 9-11 presents a negative result search page that provides all the features useful to help the user get back on track. Notice that the negative result page also fits with the design of the site.

Rule: Negative search result pages must include information on why a query failed and potentially how to fix the query.

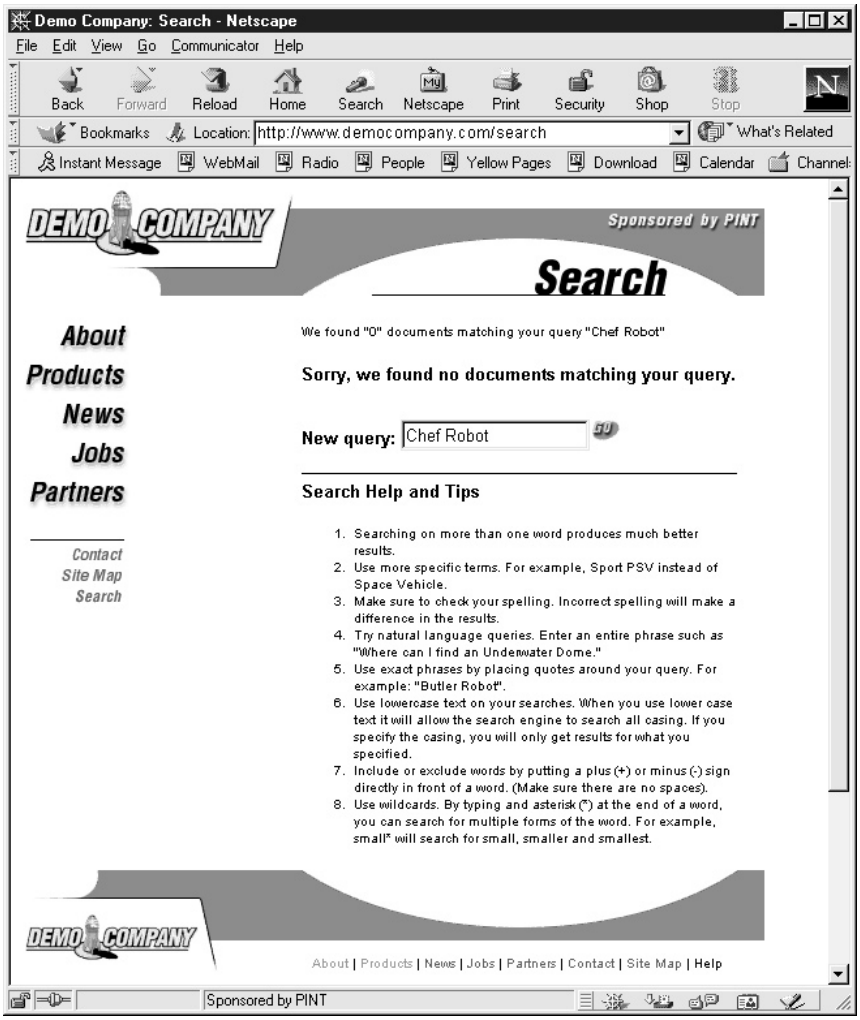


Figure 9-11. *Negative result pages should provide a wealth of help information*

Similar to broken link pages, negative result pages probably come up more than we would like. Make sure to monitor the negative queries to determine the usefulness of a search facility. Measure the percentage of negative queries and try to identify common bad queries. If your site is missing something, the negative queries may reveal the items that users are really looking for. Negative query monitoring is only one way to improve search facilities, so let's take a look at a few other strategies.

Improving Local Search

Despite our best efforts, local searching of Web site contents often fails. Why? Usually it is due to one of the following four problems:

- Search item not found on site
- Keyword mismatch
- Misspelled words and other near hits
- Search interface problem

The first problem really isn't solvable. If a user believes an item exists in the site and it doesn't, all we can really do is fail nicely. The other problems, however, can be addressed.

Addressing Near Hits

Keyword mismatch often has to do with the fact that how a user searches for something might not be exactly the way that the item is indexed. The basic problem has to do with vocabulary. For example, a user may enter "automobile" as a search term when the relevant pages were indexed under "car." Obviously, the two words are synonyms, so the search should not have failed. To solve this problem a site designer should come up with a controlled vocabulary of search terms, including related words. Generally termed a *thesaurus*, such a cross-reference of keywords can be generated uniquely, or for certain knowledge domains, a predefined set of words can be adopted.

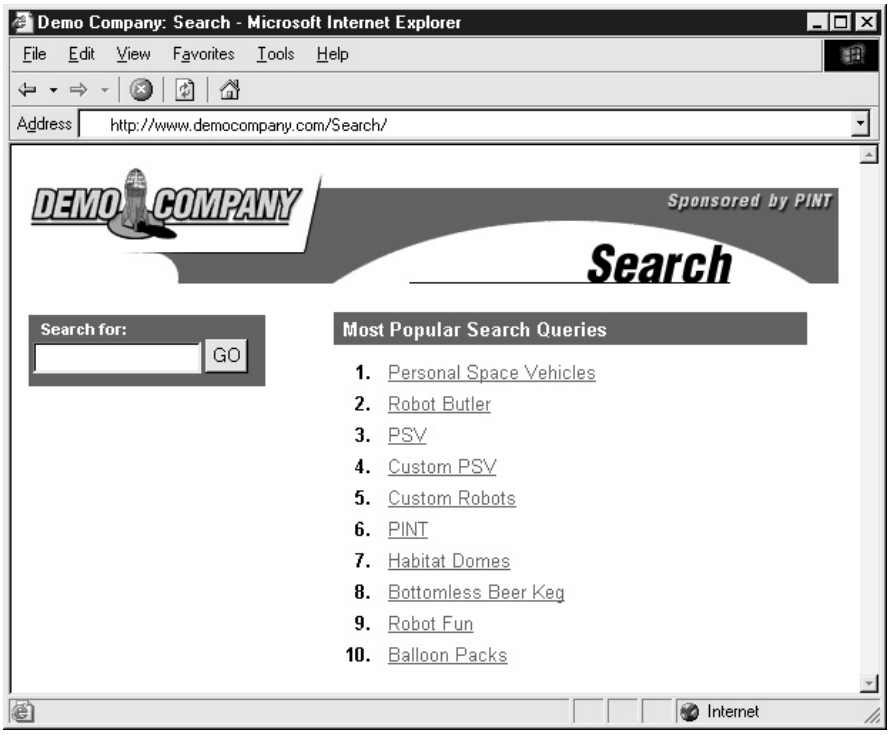
Similar to keyword problems are searches that include misspelled words or words that run together. Particular attention should be paid to alternative spellings of words related to regional language differences—for example, color and colour. If possible, the search system should provide a "Did you mean?" facility to get users to the items they were actually looking for.

Any search-fixing facility might also try to address run-together search terms, like "spaceship" and "space ship."

Show Common Queries

Local search engines obviously are not as user-focused as they could be, considering that few of them address the fact that most users will enter the same simple one- or two-word queries. Since this is the case, it is wise to ensure that such queries match up with their intended result. Understanding what these special keywords are does not require a detailed analysis of site contents; the search engine itself should be able to tell

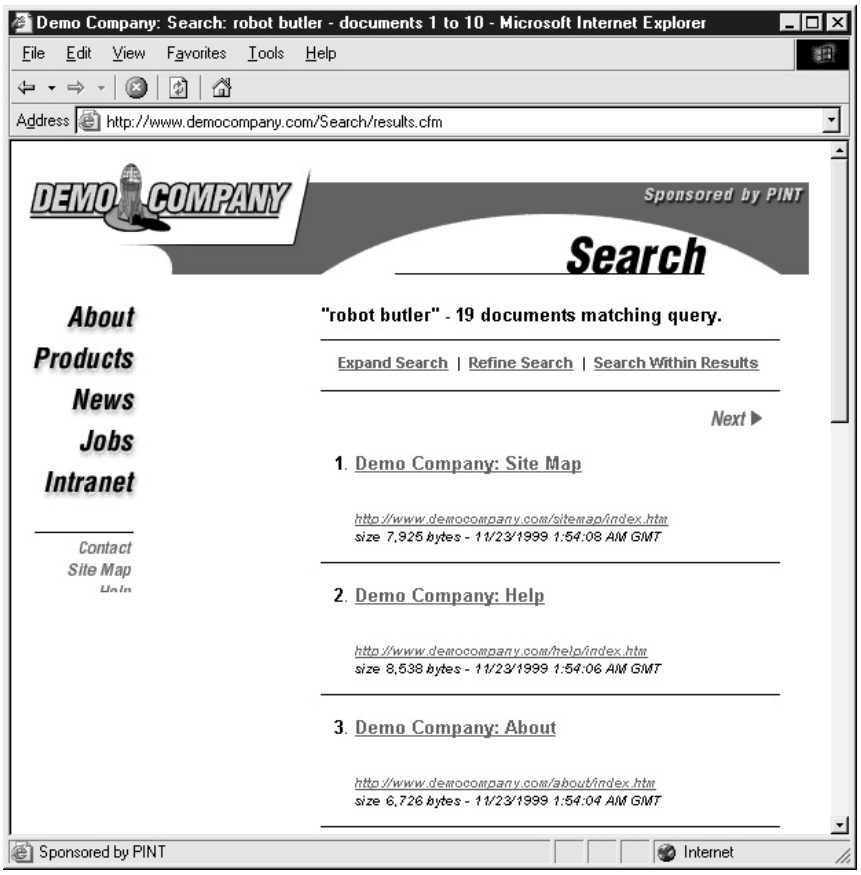
you what people are searching for. You may then consider showing the popular searches that have been fixed to return correct results right on the search interface, as shown here.



Scope Properly

When advanced searching facilities are provided, scoped searches are one of the best ways to improve the chances of search success. The first task is to limit the scoping. As mentioned earlier in the chapter, avoid scoping sites by hierarchical sections of site data but focus more on topics. Further scoping possibilities include limiting the search to a particular file type (such as PDF, GIF, HTML), date range, author, and so on.

Regardless of the scoping method used, make sure to allow the user to broaden or narrow the scope at will. For example, on the result page you might have a link that allows the search to be expanded to the whole site or to research within the set of results.



Add Polish

Search interfaces often suffer from a lack of usability and interface polish. For example, most local search sites will allow a user to enter in a blank query, only to have the query spit out an error message. If a blank search query is not allowed, try to address it right away using a simple JavaScript error message.

In some cases a blank query might return all pages in a given search space, but if it does not, allowing an obviously bad query is pointless.

Suggestion: Disallow blank search queries unless they return a complete page set.

Monitor and Maintain

Be as user-focused with search design and maintenance as with other areas of a site. Be sure to track your search logs so you can see why and how your site visitors are having problems with search. Watch searches that find zero matches and do your best to add new synonyms, terms and information that address these issues. For common searches, make sure the intended results come up.

Go Beyond Search

Finally, remember that users will move back and forth between browsing and searching. It isn't all-or-nothing when it comes to site navigation strategies. Try providing access to topic categories and browse facilities within a search interface, if possible, as shown in Figure 9-12.

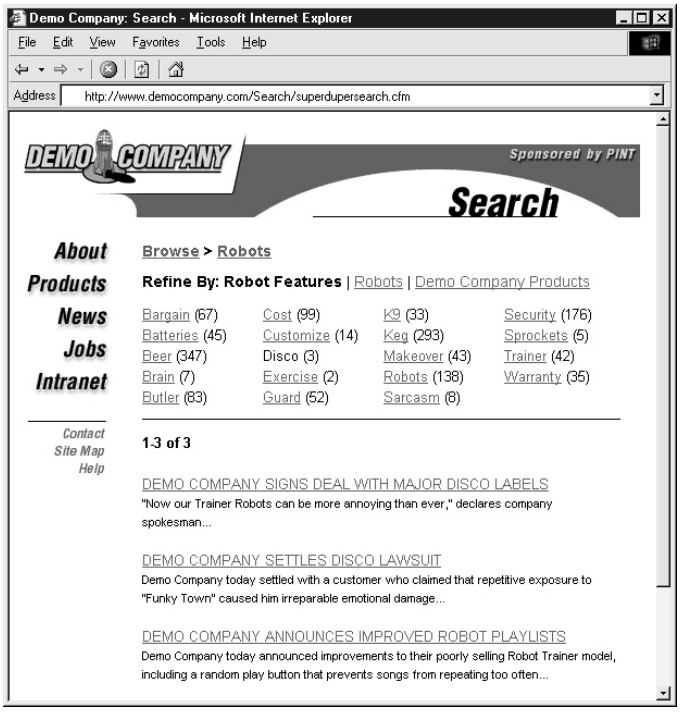


Figure 9-12. *Combine Search and Browse for Powerful Navigation*

The next chapter will explore other navigation facilities beyond search, but before we conclude let's turn our attention to how designers try to utilize search engines and other facilities to promote and drive traffic to their site.

Public Searching

As previously mentioned, site designers must be especially careful not to fall into the familiar trap of exactly imitating search on the Web at large. The needs of Web-wide searching are very different than those of a single site, or even a group of controlled sites. Unfortunately, users often expect site-search facilities to act similarly to public search engines like Google or Lycos. Public search engines have to deal with the extremely difficult task of gathering and indexing the enormous and ever-changing Web, which has numbers of documents that are purposely filled with misleading information. Then, from all this information, the user is supposed to quickly and easily retrieve a useful result using a simple query. In summary—finding a needle in a haystack is a much easier task than searching the Web. Regardless of the difficulty, users do rely on public search engines a great deal and their searches do work more often than not. Designers should consider a user's experience with Webwide search engines, since users will generally understand the functionality of these engines and apply that knowledge when they use a local search engine. Further, we need to understand Web-wide searching to see how it fits into the task of driving users to our Web site. The following sections will explore the components of Web searching and explore some of the problems encountered.

Full Web Searching Overview

The requirements for Webwide searching are daunting. Users expect to be able to quickly type in a simple search phrase at a global search engine like AltaVista or Google and end up with a realistic result. Consider the chances of walking into a public library and finding a particular passage in a book in a few seconds and you'll understand the near futility of the task. When searching, users are often overwhelmed with too much information, are shown irrelevant information, or do not get anything at all. Despite the resulting frustration, users keep pounding away at search engines, hoping to get a good result in a matter of minutes.

Many of the problems with search engine usage have to do with users not searching correctly. Searching really should be used only when looking for known items or for very specific topics. Consider searching for a general term like "hamburgers." Search engines may not necessarily pull up sites about hamburgers or even large hamburger restaurant chains. In fact, testing this query in some search engines resulted in numerous links to pages about Hamburg, Germany, as well as recipe sites for personal home pages and pages that appeared to have absolutely nothing to do with hamburgers. The problem is that the search term isn't specific enough. If you search for something like "White Castle Slyders"—a regionally famous hamburger in the United States—you may find a more useful list of results.

When looking for general information on a subject, users usually turn to a directory rather than a search engine. The main difference between a search engine and a directory is that a directory usually involves some human editing and usually contains a very limited number of links. Yahoo! is probably the most famous directory around, but it now provides search engine features as well. In fact, most of the search engines have begun to offer directory links as well as searching. Many popular search engines now focus more on delivering users to sites that focus on a particular topic rather than trying to get users to a very specific site. Directories like www.about.com or www.dmoz.org are organized by individuals who are responsible for a particular type of content. The benefit of a directory is that having a site organized by people can result in limiting content to just the "good sites." While automatic gathering and categorization of content can be a powerful tool, until Artificial Intelligence is vastly improved the value of human editing and categorization should not be underestimated.

Definition: A Web directory is a human-edited and organized collection of site links and associated information such as descriptions and reviews.

In comparison to a directory, a pure search engine is more like the phone book that you can only search. This is similar to calling your information service and asking for a phone number, except you ask for something related to a particular topic. Consider using a phone information service such as 411 in the United States and asking for the phone number of a "Chinese restaurant" rather than asking about a particular Chinese restaurant. If you ask for a particular restaurant, chances are you're going to get a good result. However, when asking for general information you'll be very lucky if the operator actually spends some time to give you a particular restaurant they know about, or even returns one that looks reputable based upon its ad in the print directory. In many cases, directory assistance might just give you the first one or even a random one from a list. Search engines tend to act the same way. They are good at returning specific answers, but results vary otherwise.

Search engines always attempt to be comprehensive and may list numerous sites without regard to content quality or freshness. Search engines are primarily automated in the collection and organization of links, though today some human editing as well as directory-oriented results are being used. This is due partly to the massive amount of search engine trickery going on, as well as a desire to improve the result sets for users. The reason for the trickery is a desire by Web site owners to use search sites to drive as much traffic to their sites as possible.

Search Engine Promotion

Site owners always want to be number one in search engines. Consider if you are a small travel agent. You probably would love it if people could go to a search engine, type **travel**, and have a link to your site show up as the first one. You'd get a large number of visits for sure. Unfortunately, there are probably a lot of other people who

would like to be number one, and being ranked 4,036th isn't going be worth much. In fact, if you are outside the first 20 sites or so returned, you probably aren't going to get many clicks at all. Because of this, page designers are always trying to determine how search engines categorize pages and then building their page with keywords in such a way to get a high ranking. In some ways, this idea is similar to how people name their company something like AAATravel in order to get listed first in the phone book. Unfortunately, consider how many travel agents in the world want their site to be in the top ten in search engines and you'll see a potential problem. The Web is not as geographically specific as the phone book. Imagine that there is only a single phone book for the United States. There would probably be dozen of pages filled with companies, all starting with AAA. The Web already has this problem, and that's one of the reasons you get so many results when you run a query for a competitive industry like discount travel.

The war to be first in the search engine has an obvious result—the rise of "pay for position." Consider that the tricks to be at the top of the search engine list spread rapidly. For common search phrases, it is nearly impossible to stay at the top of the list for long since other sites use the same search engine promotion techniques. Already search engines such as Overture (www.overture.com) are opting to push people to the top of the list that are willing to pay for position. Priority placement is also being made for banner ads triggered to correspond to particular search phrases. Just as with the phone book, naming your company AAATravel might put you at the top of the line listings, but readers may opt instead to look at the large display ads. Search engines will eventually adopt the same model. Further, as end users become more sophisticated, they will begin to rely more on directory listings for generic topics and use search engines only for very specific or complex lookups. The eventual outcome of the search engine war will almost certainly be a return to traditional models of information retrieval methods used in other advertising forms where you pay for audience relevancy and position. For now, designers should consider not taking advantage of search engine positioning methods, regardless of their long-term viability.

Adding to the Engines

Getting a site's pages gathered by a search engine is the first step in making a site findable on the Web. The easiest way to do this is simply to tell search engines that your site exists. Most search engines will allow you to add a URL to be indexed. For example, Google allows you to add a site for gathering by using a simple form (http://www.google.com/addurl.html). Of course, adding your site to every single search engine could be a tedious task, so many vendors (http://www.submitit.com) are eager to provide developers with a way to bulk-submit to numerous search engines. Most Web site promotion software, such as WebPosition Gold (http://www.webposition.com), also includes automated submission utilities. Today you may find that the simple guaranteed submission to a search engine costs money. Undoubtedly, this trend will continue.

You should consider how many search engines you'll want to submit your site to. Some people favor only adding few links to the important top ten engines, especially Yahoo! Numerous studies, as well as this author's experience, suggest that big search sites, particularly Yahoo, account for most search engine traffic. However, some site promotion experts feel this is not correct and believe it is best to create as many links to sites as possible. In fact, a whole class of link sites—"Free For All" links or FFA sites (not to be confused with anything related to the Future Farmers of America)—have sprung up to service people who believe that "all links should lead to me" works. The reality is that most of these link services are pretty much worthless and often generate worthless Traffic and spam messages. Further, consider that even if you do get back links and e-mail, it is mostly from people who are doing the same thing you're doing—trying to get links.

Robot Exclusion

Before getting too involved putting yourself in every search engine, remember that it isn't always a good idea to have a robot index your entire site, whether it is your own internal search engine or a public search engine. First, some pages such as programs in your cgi-bin directory don't need to be indexed. Second, many pages may be transitory, and having them indexed may result in users seeing *404* errors if they enter from a search engine. Finally, you may just not want people to enter on every single page—particularly those pages deep within a site. So-called "deep linking" can be confusing for users entering from public search engines. Because these users start out deep in a site, they are not exposed to the home or entry page information that is often used to orient site visitors.

Probably the most troublesome aspect of search engines and automated site gathering tools such as offline browsers is that they can be used to stage a denial of service attack on a site. The basic idea of most spiders is to read pages and follow pages as fast as they can. If you tell a spider to crawl a single site as fast as it possibly can, all the requests to the crawled server may very quickly overwhelm it, causing the site to be unable to fulfill requests—thus denying services to legitimate site visitors. Fortunately, most people are not malicious in spidering, but it does happen inadvertently when a spider keeps reindexing the same dynamically generated page.

Robots.txt

To deal with limiting robot access, the Robot Exclusion protocol was adopted. The basic idea is to use a special file called robots.txt that should be found in the root directory of a Web site. For example, if a spider was indexing http://www.democompany.com, it would first look for a file at http://www.democompany.com/robots.txt. If it finds a file, it would analyze the file first before proceeding to index the site.

Note *You will find that many spiders will ignore a robots.txt file with a URL like http://www.bigfakehostingvendor.com/~customer/robots.txt, where the robots.txt file is not located in the root directory. Unfortunately, you will have to ask the vendor to place an entry for you in their robots.txt file.*

The basic format of the robots.txt file is a listing of the particular spider or user agent you are looking to limit and statements including which directory paths to disallow. For example,

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /temp/
Disallow: /archive/
```

In this case, we have denied access for all robots to the cgi-bin directory, the temp directory, and an archive directory—possibly where we would move files that are very old but still need to be online. You should be very careful with what you put in your robots.txt. Consider this file:

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /subscribers-only/
Disallow: /resellers.html
```

In this file, a special subscribers-only and resellers file has been disallowed for indexing. However, you have just let people know this is sensitive. If you have content that is hidden unless someone pays to receive a URL via e-mail, you will certainly not want to list it in the robots.txt file. Just letting people know the file or directory exists is a problem. Consider that malicious visitors will actually look carefully at a robots.txt file to see just what it is you don't want people to see. That's very easy to do: just type in the URL like this: **http://www.*companytolookat*.com/robots.txt**.

Be aware that the robot exclusion standard assumes that spidering programs will abide by it. A malicious spider will, of course, simply ignore this file, and you may be forced to set up your server to block particular IP addresses or user agents if someone has decided to attack your site.

Robot Control with <meta>

An alternative method to the robots.txt file that is useful particularly for those users who have no access to the root directory of their domain is to use a **<meta>** tag to

control indexing. To disallow indexing of a particular page, use this **<meta>** tag in the **<head>** section of the HTML document:

```
<meta name="robots" content="noindex" />
```

You can also inform a spider not to follow any links coming out of the page:

```
<meta name="robots" content="noindex, nofollow" />
```

When using this type of exclusion, just make sure not to confuse the robot with contradictory information like

```
<meta name="robots" content="index, noindex" />
```

or

```
<meta name="robots" content="index, nofollow, follow " />
```

as the spider may either ignore the information entirely or maybe even index anyway. The other downside to the **<meta>** tag approach is that fewer search engines support it than do robots.txt.

Optimizing for Search Engines

Optimizing your site for a search engine is not difficult. The first thing to do is to start to think like a search engine—in other words, don't really think at all. Search engines literally look at pages and make educated guesses about what pages are about by following a set of rules to try to understand what the page is about. For example, search engines look for word frequency, **<meta>** tags, and a variety of other things. However, they really can't tell the difference between a page about the Miami Dolphins football team and a dolphin show in Miami. The reason is that search engines generally rely on keyword matching in conjunction with some criterion such as the placement of words in a page or the number of linking sites. So if a designer knows what a search engine is looking for, it is easy enough to optimize a page for the search engine to rank it highly. The next few sections provide a brief overview of some of the things search engines look for as well as some tricks people have employed to improve their search ranking.

<meta> Tags

Many search engines look at the **<meta>** tags for keywords and descriptions of a page's content. A **<meta>** tag like

```
<meta name="Keywords" content=" Butler-1000, Robot butler, Robot butler
specifications, where to buy a robot butler, Metallic Man Servant, Demo
Company, robot, butler" />
```

could be used in our example page about robot butlers. Notice how the content started with the most specific keywords and phrases and ended with generic keywords. This should play into how most users approach search engines.

Once a search engine looks at the **<meta>** tag, it may rate one site higher than another based upon the frequency of keywords in the **content** attribute. Because of this, some designers load their **<meta>** tags with redundant keywords:

```
<meta name="Keywords" content=" Robot butler, Robot butler, Robot
butler, Robot butler, Robot butler, Robot butler, Robot butler,
Robot butler, Robot butler, Robot butler" />
```

However, many search engines consider this to be keyword loading and may drop the page from their index. If the keyword loading is a little less obvious and combinations of words and phrases are repeated,

```
<meta name="Keywords" content=" Robot butler, Butler-1000, Metallic
Man Servant, Robot butler, Butler-1000, Metallic Man Servant, Robot
butler, Butler-1000, Metallic Man Servant, Robot butler,
Butler-1000, Metallic Man Servant " />
```

the search engine may not consider this improper. An even better approach is to make sure the pattern of repeating words isn't quite as obvious, as shown here:

```
<meta name="Keywords" content=" Butler-1000, Robot butler, Metallic
Man Servant, Robot butler, Butler-1000, robot, Robot butler,
Democompany, Metallic Man Servant, Butler-1000, robot, butler,
Robot butler, Butler-1000" />
```

However, be aware that search engines may still notice the heavy use of certain words or phrases and consider this spamming, potentially reducing the page's ranking or dropping it from the index completely.

Search engines also look at the description value for the **<meta>** tag. For example,

```
<meta name="Description" content="The DemoCompany Robot Butler is
the most outstanding metallic man servant on the market. The
Butler-1000 comes complete with multiple personalities and voice
modules including the ever-popular faux-British accent." />
```

would be included on the robot butler page and could be examined by the search engine, as well as returned by the search engine on the result page. Because the <**meta**> tag description may be output for the user to see, provide some valuable information in the description that will help users determine if they want to visit your site. Preferably, keep the description to a sentence or two and, at most, three or four sentences.

Titles and File Naming

One important aspect of search engine ranking is making sure your page has a very good title. For example,

```
<title>Robot Butler</title>
```

is a bad title as far as search engine ranking goes. A better title might be:

```
<title>Butler-1000: Specification of Demo Company's Robot Butler,
the leading metallic man servant on the market</title>
```

Remember that people also look at page titles, and they are used for bookmarking, so a really long title may be more appropriate for search engines than for users.

The name of a file can also be important for search engines. Rather than naming a file "butler.htm," use "butler1000robotbutler.htm." If you have a good domain name and directory structure, you may create a URL that almost makes sense. For example, if we named our server democompany.com, as well as www.democompany.com, we may have a URL like this:

```
http://democompany.com/products/robots/butler1000robotbutler.htm
```

Notice how this almost includes the same information as the title. This provides the secondary benefit of letting users know where they are, rather than resorting to cryptic URLs like this:

```
http://democompany.com/products.exe?prod=robots&mod=butler1000
```

Relevant Text Content

One of the best ways to get indexed is to have the keywords and phrases actually within the content of the page. Many search engines will look at text within a page, particularly if it is either towards the top of the page or within heading tags like **<h1>** or **<h2>**. Search engines may also look at the contents of link text. Thus,

```
<a href="specifications.htm">Specifications</a>
```

is not as search engine friendly as

```
<a href="specifications.htm">Robot Butler Specification</a>
```

One problem with search engines focusing on page text is that designers often create home pages that are primarily graphic. Search engines accessing such pages may have little to go on besides the **<meta>** tag and page title and thus rank the page lower. Consider using the **alt** attribute for the **<img>** tag to provide some extra information; for example,

```
<img src="robot.gif" alt="Butler-1000: Demo Company's industry
leading robot butler" />
```

Of course, putting the actual text in the page would be better. Some designers resort to either making text very small, or in a color similar to the background, or both, so that users won't see it but search engines may pick it up; for example,

```
<font size="1" color="white">The Demo Company Butler1000 is the
best robot butler. The Demo Company Butler1000 is the best robot
butler. The Demo Company Butler1000 is the best robot butler.</font>
```

Be careful with the small or invisible text trick. Many search engines will consider this to be spamming and may drop the page from the search engine.

Links and Entry Points

Another aspect of search engine ranking has to do with the number of links leaving a page, as well as the number of pages that link to a page. Landmark pages such as home pages tend to have a lot of outgoing and incoming links. Search engines would prefer to rank landmark pages highly, so it is important that key pages in your site have links to them from nearly every page. Some search engines also favor sites that have many sites pointing to them. Because of this, people are already starting to create sites solely for the purpose of pointing to other sites.

Another approach to improving search engine ranking is to submit many pages in a site, or even off a site, to a search engine. All of these entry pages, often called *doorway pages*, point to important content within your site. Unfortunately, doorway pages are more like decoy pages, as they can be loaded with false content to attract the visitor and eventually deposit the user at a page they didn't really want to see. The problem with search engine promotion is that the distance from simple logical keyword loading to various tricks is a short one—particularly if designers obsess with top-ten ranking.

Tricky Business

The tricks employed by search engine specialists are numerous and change all the time. Many ideas are simple add-ons to normal Web design techniques. For example, many designers rely on invisible pixel shims to force layout. Search engine promoters might say, "Why not put **alt** attributes on these images to improve things." Imagine, for instance, having the following all over your page:

```
<img src="pixel.gif" alt="robot butler robot butler robot butler" />
```

Pity the user who pauses on top of one of these invisible pixels only to have a ToolTip pop up screaming about whatever the page is promoting. Spamming pages with invisible text, small text, and multiple images, or just loading the **<meta>** or **<title>** tags, are not the most sophisticated tricks, but they often work.

Other tricks include the infamous "bait and switch," where a special search engine page is created and then posted to a search engine. Once the ranking is high, the bait page is replaced with a real page built for users. A more complicated version of this could be dubbed "feeding the dogs"(or page or site cloaking). In this scenario, you write a program that senses when a search engine hits the site, and then the program "feeds" the engine the page that it wants to see. Like a ravenous dog, it gobbles up the food with no idea it just ate the equivalent of informational pig snouts. As real users hit the site, they aren't served the dog food, but get the real site.

Distinguishing search engines from regular users isn't terribly difficult, since the engines identify themselves and come from consistent IP addresses. In reality, "feeding the dogs" is just a modified form of browser detection. Search engines can do little to combat this approach, since they would have to consider eliminating dynamically built pages—which is impossible given their growing importance—or not informing sites that they are search engines while indexing. A few search engines have already begun to provide a link to a page that shows what was indexed, so users can determine if they are being shown something different than what a search engine indexed. Others revisit the page in multiple guises and see if things are dramatically different; if they are, cloaking is considered to be in play and the page is dropped. Of course, this may just be because the page is dynamically created; thus, many search robots will tend to exclude pages with complex URLs, like www.democompany.com/products.cfm?robots=army&cost=expensive. In order to address this, some site owners will rewrite page URLs to make them more search engine friendly. We'll address that in Chapter 17, on site delivery.

The problem with all the search engine promotion ideas is that they tempt the designer to stop building pages for users and start building them for search engines. This is just another form of designing more for your own needs than for your users.

Rule: Do not design pages solely to attract search engines, as, ultimately, pages are for people.

One of the most interesting aspects about search engines is that many large organizations don't rely greatly on them for driving traffic. In fact, for many corporations, unless you type their name in directly, you'll be hard pressed to find them in a search engine. However, despite what appears to be a major oversight on their part, these sites continue to get huge amounts of traffic. According to studies such as the GVU Internet Survey, people type in URLs directly quite often.

How are they finding out about sites? Search engines aren't the only way to drive traffic. There are many ways to get users to visit your site. Banner ads, link exchanges, news group postings, mass e-mailings, and easily typed and remembered domain names all are well-known approaches to traffic generation. However, one increasingly popular way to attract visitors is to rely on things outside the Internet. Television, radio, print, billboard, direct mail, and a variety of other venues are being used to spread the address of the latest Web site.

Note *This is by no means a complete discussion of search engine promotion, as the topic literally changes on a weekly basis. Readers looking for more up-to-date information are directed to the numerous site promotion sites that exist on the Web, especially Search Engine Watch (www.searchenginewatch.com).*

Summary

If browsing is about following predefined trails in a Web site, then searching is going off-path, blazing your own direction through content. While it would seem that search facilities appeal primarily to power users and frequent visitors, the fact is that novice users are familiar with public search engines and rely more and more on sites like Google for searching. Understanding how public search engines work and are used is the first step in designing a local site search facility. Designers should also understand how users move from public search sites to local sites and attempt to guide users to what they are looking for. Search facilities must be designed with the user in mind. The best way to do this is to consider what users would actually want to search for in a site. Do not fall into the trap of blindly imitating the free text search qualities of global Web search engines. When providing local search, make sure to provide both basic and advanced search forms. Format the search form carefully and provide instructions. This will help users form good queries, but in case things go wrong, make sure the negative result page provides extra help to get users back on track. Once users do get a positive result from a search engine, make sure that enough information is provided so they can narrow down the potential choices. Having too much data is nearly as bad as having none at all. However, always consider that searching isn't everything. Like all forms of navigation, searching is a means to an end, not the end itself. There are many ways to help users find what they are looking for. The next chapter will present a variety of other navigational aids, such as site maps, site indices, and help systems.