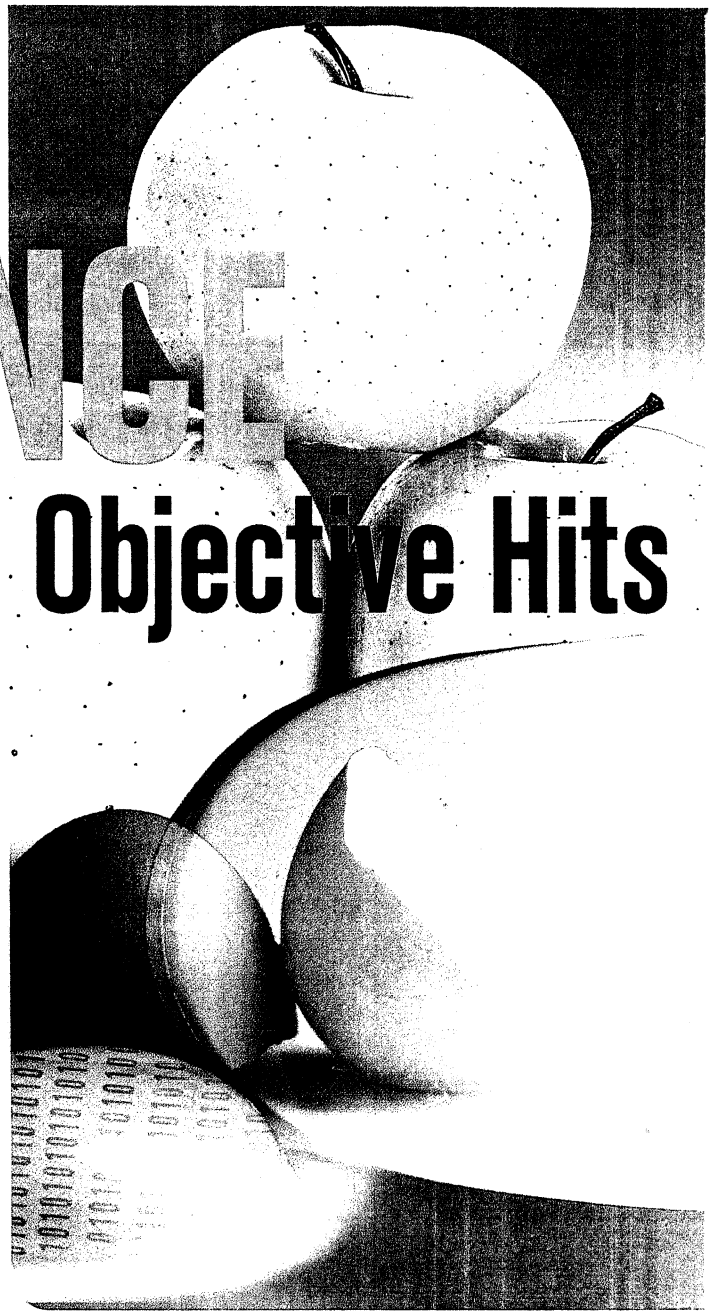
RELEVANCE and the End of Objective Hits



Information professionals expect search results to reflect their search query. This is what happens with traditional online search services.

By Stephen E. Arnold

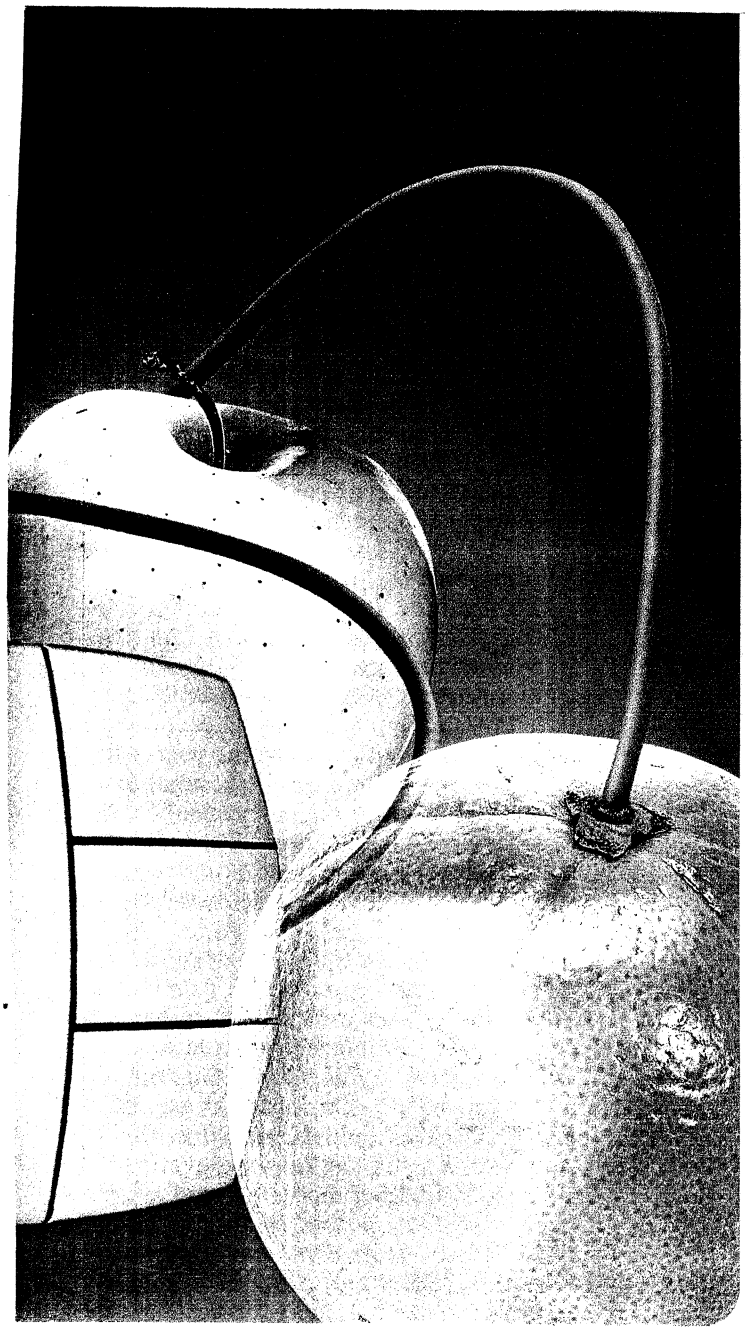
Ask LexisNexis, Factiva, Dialog, EBSCO*host*, or ProQuest to return information on, say, `Macedonian weapons`, and that's what you get. An exact match on the terms entered to describe the topic. It's up to the searcher to ensure that the terms entered adequately describe the topic. Searching for `Macedonia`, for example, will not retrieve `Macedonian` unless the researcher truncates the term. Traditional search services do not try to outguess the searcher. Nor do these services introduce advertising, aka sponsored links.

Web search engines, such as Google, Yahoo!, MSN, and Ask Jeeves, are different. Search results are not exact matches to the query terms. These results do not display in chronological or reverse chronological order. Relevance ranking replaces objectivity. Web search has spawned a cottage industry of search optimizers, who diligently work on Web sites to raise their relevance level and put them in the magic "top 10" of results—the ones that the searcher sees with no need to scroll down.

Given this new search environment, the big question for information professionals should be, "Does search engine optimization work?" The short answer is, "Yes." The next question is, "How does it affect the validity and relevance of my search results?" That is the more interesting question.

MAKING SENSE OF ADSENSE

Take a Google search on `AdSense`. Since AdSense is Google's product, you would expect Google's site to appear first. On May 25, 2005, the number-one most relevant site for Google's AdSense, according to Google, is a site called All in One Business. Google's PageRank algorithm seems to have gone astray—or has it? (See Figure 1 on page 17.)



with Web pages in order to return "false drops" or listings in a Google result list that send the user to an expected destination.

BELIEVING IN SEARCH

Search is an umbrella word like love, trust, and honor. One syllable embraces a mind-boggling range of meaning.

Today, search is a synonym for Google. In an interview on National Public Radio's *On the Media* program [www.onthemedia.org/transcripts/transcripts_052005_books.html], Jean-Noel Jeanneney, president of the French National Library, said in response to a question from Bob Garfield about his principal objections to Google's library book digitization project, "I think Google is quite a precious invention, and we all use Google." He then went forward with his concern that Google's Print initiative would crush French culture.

Relevance is another slippery fish. Those who have endured the rigors of professional training in information retrieval have a Pavlovian response to the word. Say "relevance," and the intellectual guard dog barks, "Precision. Recall."

The distance between a formal calculation of a query's result set in terms of recall and precision is the substance of master's and Ph.D. theses. Recall, in mathematical terms, measures how well a search system finds what a searcher wants. MSN, Ask Jeeves, and even Google identify a query expressed as a question and try to generate an "answer."

Precision, again in mathematical terms, measures how effectively information a searcher does not want is eliminated from the result set.

The American Society for Information Science and Technology (ASIST) has long produced conferences and publications that explore formulae, examples, and analyses of how to mathematically calculate recall and precision. However, we are now in the late Pleistocene

Clicking on the "all-in-one-business" link redirects the user to the Google AdSense page on Google. The owner of All-in-One-Business.com, Kevin Birdwell, writes and speaks about a wide range of business issues. His articles about search engine optimization mention Google and its AdSense service. All-in-One-Business's content includes links to Google's AdSense page as well as to other pertinent Web sites.

What appears to be happening is that Google indexes the content of All-in-One-Business.com, finds pertinent content, and determines a PageRank. The Google AdSense page does not have comparable substantive content. PageRank notes this and assigns a higher relevance score to the All-in-One-Business.com page. As this interesting PageRank result shows, solid content and proper Web page design influence the PageRank algorithm. There is a flip side as well. A person can tinker

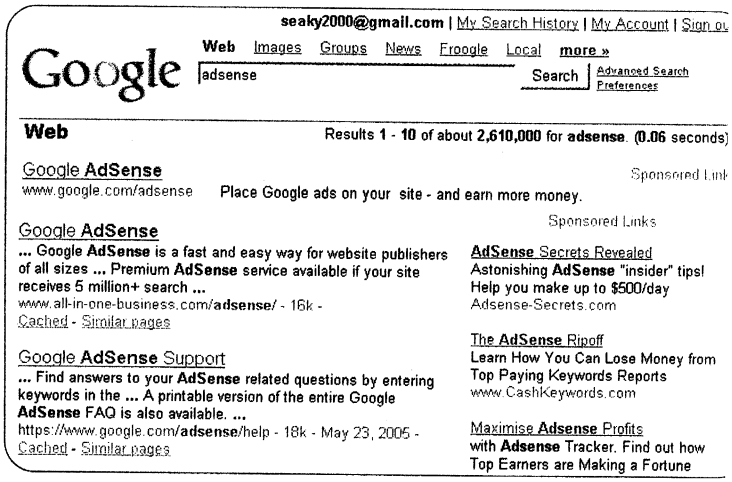


Figure 1

Age, with the extinction of certain notions about relevance and its children, precision and recall. Mathematics may have failed information scientists in determining, once and for all, the recall and precision scores for a given query within a given corpus indexed by a specific search system. However, mathematics delivered a spot-on way to handle the majority of queries submitted by users of Google.

VOTING VERSUS MATHEMATICS

The math embodied in the Google PageRank algorithm is now recognized as the painfully obvious way to figure out whether an average user keying `spears` in a Google search box wants `Macedonian weapons` or `Britney Spears`, the pop star. PageRank looks at the query, consults its index that includes a "score" indicating popularity, and delivers the highest-ranking match as the most relevant hit.

Figure 2 below shows the query run on May 25, 2005, for `spears` and its result set.

Contrast the spears query with the three-word query `Macedonian weapon spears` in Figure 3 (see below).

In both cases, Google delivers on-point results for the average user. An expert in matters Britney may be able to pinpoint specific omissions in the results list. The expert will cavil at the order in which Britney Spears' results appear. A search purist may note that the in-line advertisement for a company running a survey to determine who is hotter, Paris Hilton or Britney Spears, is not as relevant as BritneyZone.com and its pictures of the pregnant Britney. But those arguments are ones brushed aside by average Internet users looking for a link to a site that focuses on the star.

The hits for Macedonian weapons are not scholarly. The most relevant hit is to an encyclopedia entry. Google mingles weapons of Macedonia with those of ancient Egypt. A specialist in warfare is likely to grouse about the generalist nature of the results. A person owning shares of Google stock would point out that there is no in-line text advertisement.

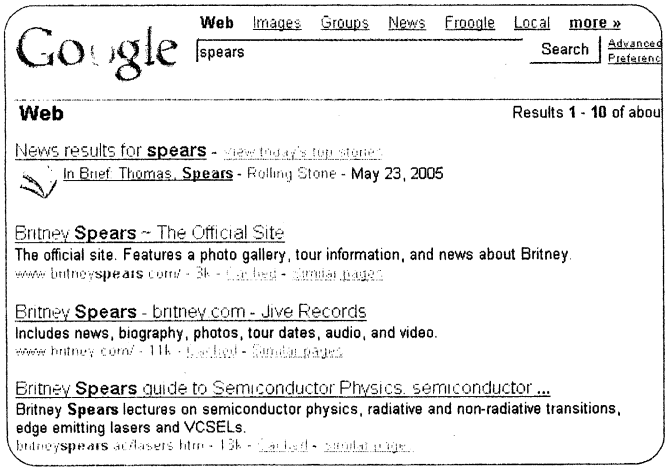


Figure 2

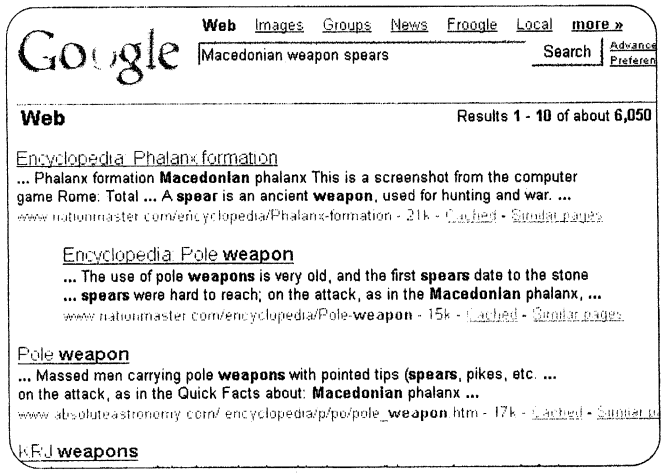


Figure 3

PAGERANK ALGORITHMS

Running queries and having experts comb through the results sets is tough work. It is important, particularly when a search system is making an effort to tune its algorithms to meet the needs of a particular user community. Remember, Google allows the PageRank algorithm and emergent behavior identified by such factors as clicks on a particular Web site, the number of times users enter a particular word or phrase, and the number of high-traffic in-bound links a particular Web site enjoys.

Google, in its March 2005 patent application 20050071741 [www.tinyurl.com/4o9vj], provides a glimpse at the hidden gears and wheels inside its PageRank algorithm. Google mavens speculate about the specific way Google's PageRank system uses as many as 60 to 80 "factors" to determine, for example, that the query `spears` should display the official Britney Spears Web site as the most-relevant hit. Keen-eyed searchers will see the possibility of a false drop in the news story from *Rolling Stone* containing the word spears. This is easily explained in terms of the Google PageRank algorithm. News is a separate cluster of content built from 4,500 sources. The top news "hit" on `spears` is, in that news corpus, number one, at least for May 25, 2005. For the larger mass of content with more data to inform the algorithm, BritneySpears.com is number one and probably will be until Ms. Spears' half-life is reached.

Google is a "popularity contest" algorithm. The larger the number of votes, the more "accurate" the PageRank's outputs. The mathematics of votes works brilliantly. Google's users know that Google overall at this point in time does a better job of converting a query into information that the user finds useful. Advocates of Yahoo! argue that Yahoo!'s search system, with its richly faceted interface, does a better job. Both sites enjoy traffic measured in the hundreds of millions of unique visitors per day, billion-dollar revenue flows, and services that are sufficiently magnetic to keep people coming back. Thirty years ago, in the good old days of LexisNexis, SDC Orbit, and Lockheed Dialog, information professionals were the important customers. Google and Yahoo! have figured out how to hook other

ers. They generate orders of magtude more cash than their predessors did in the past or do now.

ONSPIRACY THEORIES

Search engine optimization (SEO) the discipline of crafting publicly cessible Web content in order to ost a Web site's ranking in a oogle search results list. In the ok *The Google Legacy* (Infonortics, d., August 2005), a collection of out 75 of PageRank's "factors" apars. These factors do not come dictly from Google. Various SEO exrts ran tests, shared knowledge, d attempted to reverse engineer ith digital water witching techques to uncover the secret of a top nking on the first page of a Google sults page.

The reasons for this concentration search engine optimization are mewhat obvious. The language ed to describe these reasons nges from the pseudo-scholarly to e P.T. Barnum school of rhetoric. one strips away the linguistic mping that goes on in the SEO mmunity, the reasons include the llowing:

Traffic. Get a high ranking on a Google results list and the Web site gets visitors whether the content is good, bad, or indifferent. Most searchers click on the top results. A tiny percentage venture down the results list or explore "hits" on the second and third page of results.

Money. Google AdSense shares advertisers' payments for clicks on ads that run on a Web site. A high-traffic site with 250,000 unique visitors per month can earn anywhere from $4,000 to $6,000 per month, although participants in AdSense report widely varying results.

Happy clients. Web design firms, service providers, and marketing consultants point to usage reports and click-stream data to justify their fees or salaries.

The opportunity for a knowledgele person to influence a Web site's nking exists and will continue to a feature of the search landscape.

CHEATING TO GET HIGH RELEVANCY RANKING

How do the "black hats" of the SEO world manipulate the system to obtain high relevancy rankings? Here are a few examples that represent a small selection of "cheats." It provides a glimpse into the illicit techniques that Google, MSN, and Yahoo! are known to try to identify. A Web site using one of these cheats can be removed from the index.

Neither *ONLINE* magazine nor the author approves of these cheats. Ethical SEO professionals do not use these techniques.

1. *Hidden Text Cheat A*: Identify a high-ranking site. Copy chunks of content from this site. Paste them into your Web page. Color the text so that it is the same color as the original source's page background.

2. *Hidden Text Cheat B*: Follow the steps in cheat one, but put the high-value content in one or more metatags.

3. *Identify a vendor of link farms*: Pay the fee to list your site. You can locate a link farm vendor by doing a search on Google with this syntax: `"link farm" +vendor`. Note: Automated link farms can generate links in response to certain indexing robot behaviors.

4. *Blog seeding*: Identify Weblogs that are somewhat relevant to the topic of your Web site. Submit a favorable comment about your Web site and a link to it in the blog. Note: Due to the need to acquire a user name and password for posting to certain blogs, this approach will require both automated and manual processes.

5. *Metatag spamming*: Identify key words that are popular. Yahoo! Overture and Wordtracker provide useful lists. Include these popular terms in the appropriate metatags on your Web pages.

6. *Doorway cheat*: "Sniff" whether the visitor is an indexing robot or a normal visitor. If a robot, display Web pages optimized to ensure a high ranking. If a normal visitor, display the standard Web pages.

7. *Buy variant domain names*: Create Web pages that point to your core Web site. Use one or more of these techniques on each of these "shadow" Web sites.

Within what the SEO wizards call organic search results, there is considerable room to influence a Web page's ranking in a Google results list.

Google adjusts its PageRank algorithm to stay one step ahead of those who have figured out how to beat the Google system. Data about click-stream fraud is scarce. SEO experts pooh-pooh fraud, trusting that their degrees in fine arts will make this assertion accurate. Google does not talk. Yahoo! keeps its fraud data well away from other Yahoo! units as well as curious researchers. Serious researchers, however, should question their results' relevance in light of SEO "influencers," even when the activity is deemed by the search engines not to be fraudulent.

ORGANIC OPTIMIZATION

Within what the SEO wizards call organic search results, there is considerable room to influence a Web page's ranking in a Google results list. An *organic search result* refers to changes made to a Web site that influence that Web site's ranking in a Google results list. An *inorganic search result* describes traffic that comes from buying an advertisement. When the user clicks on the ad, the traffic is coming from this "Yellow Page" type message; *ergo*, inorganic, or not based on the natural content of a Web site. At least in the world of SEO, this type of lingo sounds scientific.

What can anyone with an HTML editor do to boost a site's ranking? Based on the research and interviews conducted for *The Google Legacy*, there are some surprisingly common-sense actions a person can take to influence PageRank. Here is what a short list of the Top 10 includes:

1. Dynamic URLs. Some content management systems such as BroadVision and others generate pages when users take an action. The adjustment may be painful; for example, generating a static HTML version of certain pages designed for the Google spider to process.

2. iFrames (invisible frames) and frames in general. Frames appear in a number of Google applications. Frames create issues with the Google spider. The optimization technique is to use tables.

3. Site map. Some Web sites have discarded Web site maps. A straightforward Web site map can point the indexing scripts to content that might otherwise be missed or ignored.

4. Indexing in metatags. A close reading of Google's suggestions for Webmasters leaves one with a sense that Google is aware of Dublin Core. Good indexing is not a negative, and librarians and indexing specialists can provide substantial value in selecting terms for metatags.

5. Valid code. Many Web sites create indexing issues because of flawed programming. A Web page is a software program. The correction is to identify flawed code and make appropriate changes.

6. Use Flash and similar objects wisely. Hollywood-inspired mini-motion pictures are OK as long as there is a way for the Google spider to see links to the textual content on a site.

7. Update frequently. Content that is not updated contributes to a site's steady downward drift in the ranking.

8. Solid, thematically related content. Informed, logical information makes a Web site appealing to the Google spider and researchers. Using content that violates Google's notion of quality, semantic tricks that try to fool Google, and copyright violations that Google detects may lower a Web site in a result list. In some severe instances, a Web site may be dropped from the Google index.

9. The "Can you say it to your daughter?" test. Remove any content that violates this "daughter" test. The probability is that certain content, if offensive, will trigger downward steps in Google rankings.

10. Links from reputable sites are, in general, positive. Getting links today requires effort. Inadvertent links to a Web site from a banned site can wreck havoc on a Web site's ranking. Finding links to a Web site requires time, but the investment can be an important form of "ranking insurance."

HIDING IN PLAIN SIGHT

What type of a boost can one expect? Consider a site from The Leasing Group, a large financial services firm. The company's URL is http://theleasinggroup.com. Its splash page shows a legitimate company greeting customers and inviting them to log on to the company's commercial financial services software. If the customer has forgotten The Leasing Group's URL, it would be logical to search Google. Unfortunately, this search would fail to locate the company's Web site (see Figure 4 on page 21).

The PageRank algorithm misses this firm entirely on the first page of results. The closest match is The Leasing

up, a dot-net domain with company headquarters in ıada, not Louisville, Ky. Investigation of the code used he top-ranked page, Premier Leasing Group, shows ımber of the optimization tactics being used.

ontent presented in tables. Content is on the same ubject, leasing, so the semantic vector score is high.

lean, although complex, code. The Google spider an figure out where the words and links are. No razy flash or other gimmicks much loved by Buffy he Web page designer armed with an MFA and Macromedia's DreamWeaver.

clear page label which has the tag <Title>.

n abbreviated site map that links to content elsewhere on the site.

lone of these tips skew results in terms of the aver- user. However, the owner of the domain name thesinggroup.com rightly wonders why his site is not in first page of results. The reasons are not far to seek. eview of his site reveals:

linimal content that is digestible by the Google spider.

lo explicit site map.

ynamic pages without opaque page-naming conventions and without useful indexing in the form of netatags.

lo inbound or outbound links, not even a listing in MOZ or the Yahoo! Directory.

or the owner of the domain theleasinggroup.com, orıic SEO would help the site move up in the Google ıkings. Google does know the site exists, but the site elf lacks the basic content and tags that PageRank efers. In this example, SEO is a positive factor and ll worth the investment in SEO expertise.

EO has a dark side as well, with not one Darth Vader t dozens, if not hundreds, using the SEO light saber slay "objective relevance." These experts work hard subvert the information profession's devotion to obtive, unbiased search results. Instead, they focus on esenting results favorable to their client, whether they relevant to a query or not.

LEVANCE IN THE REAL WORLD

oogle indexes sites without charge. The rankings of es move up and down as new sites are found, indexed, d get a boost over sites that are not updated or in some y fail to trigger a high PageRank "vote."

ankings can be influenced by following some nmon-sense rules. None of the optimization "tricks" much of a trick. Good content, accurate programming, d useful indexing are basic guidelines to follow.

ode, written by SEO Darth Vaders, that tries to spoof ogle indexing and page constructers specifically to

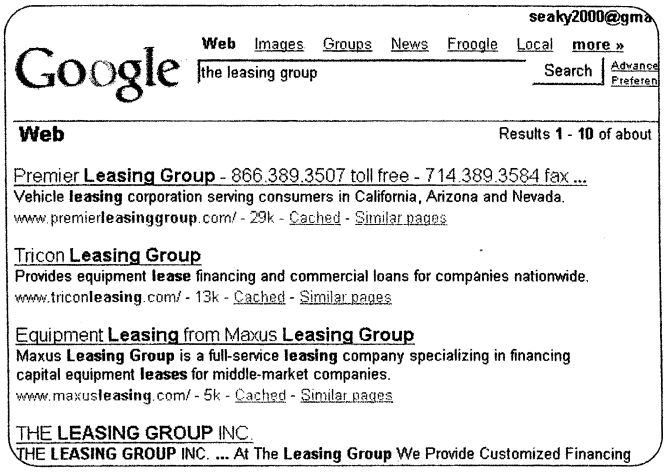


Figure 4

attract a user looking for one thing into ordering quite another, exists. Google engineers try to identify these pages and write code or adjust their PageRank factors to address the problem.

As more people become skilled in exploiting Google's blind spots, the relevance of the results erodes. One must look for chinks in Google's indexing armor. Explore long enough and gaps can be found. Some are interesting, as in the AdSense example. Others are more problematic and quite sophisticated, leading the unwary researcher to pages better left unseen.

The fix for a lack of relevance is to buy an advertisement that will appear on Web pages directly related to the content of the advertisement. If the relevancy-ranking algorithms of Google, or any other Web indexing service, are compromised, the usefulness of these services crumbles.

The best safeguard is knowledge. Google does a good job of delivering relevance. After all, more than 300 million unique queries per day are evidence that people get what they are looking for most of the time with Google. Yahoo! is not far behind, with 250 million unique queries per day. MSN is moving up fast.

Taking a broad view, one must accept that relevance of Web search results can be altered. Some relevance problems are inadvertent. Web page authors do not know what to do. Other Web page authors know exactly what to do and do it. Others work to exploit loopholes in complex algorithms. The reality is that a percentage of results will be skewed in some way.

The biggest threat, however, is not unscrupulous Web page authors. The risk is that the thirst for revenue may become so great that the only relevant results for a query may be results from paid listings. The future, then, may be the Web as a giant Yellow Pages with the content-rich days left behind like piles of old backup floppies.

Stephen E. Arnold *[sa@arnoldit.com] is president of Arnold Information Technology.*

Comments? E-mail letters to the editor to marydee@xmission.com.