Relevance: The End of Objective "Hits"

The information in this chapter touches on several points that will be discussed in Stephen Arnold's Google seminar for Direct Marketing Magazine, Information Today, Inc., and VNU in the next several months. The bulk of the material in this section comes from the The Google Legacy: An Analysis of the Google Platform, to be published in August 2005 by Infonortics, Ltd. (www.infonortics.com), Tetbury, Glou. This material is provided for reference only and may not be redistributed or republished without the written permission of the copyright holder, Stephen E. Arnold, Postal Box 320, Harrod's Creek, Kentucky USA 40027.

The big question is, "Does search engine optimization work?" The answer, "Yes." Look at the results below. On May 25, 2005, the number one most relevant site for Google's AdSense is a site called All in One Business. Google's PageRank algorithm seems to have gone astray—or had it?

PageRank shows a third-party site as more relevant than Google's own AdSense Web page. Error, intentional, or complex interactions. PageRank's "Feel lucky" may not be so lucky for this query.

Click on the link and the Google system redirects the user to the Google AdSense page on Google. The owner of All-in-One-Business.com writes and speaks about a wide range of business issues. His articles about search engine optimization mention Google and its AdSense service. All-on-One-Business's content includes links to Google's AdSense page as well as 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 with Web pages in order to return "false drops" or listings in a Google result list that send the user to an expected destination.

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

Today, especially among Internet users, *search* is a synonym for Google. Even the directory of the French National Library admitted in an interview on National Public Radio said, "I, of course, use Google every day." 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 rigours 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. Recall, in mathematical terms, measures how well a search system finds what a searcher wants. MSN, Ask Jeeves, 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.

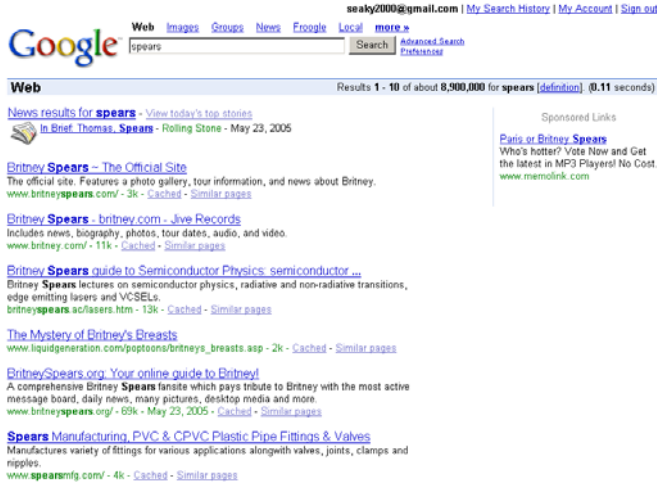
The American Society of Information Scientists exists to provide fora for presenting formulae, examples, and analyses of how to calculate *recall* and *precision*. We are now in the late Pleistocene Age with the extinction of certain notions about relevance and its children, precision and recall.

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 indexes by a specific search system. But mathematics delivered a spot-on way to handle the majority of queries submitted by users of Google.

PageRank: Voting Seems to Works

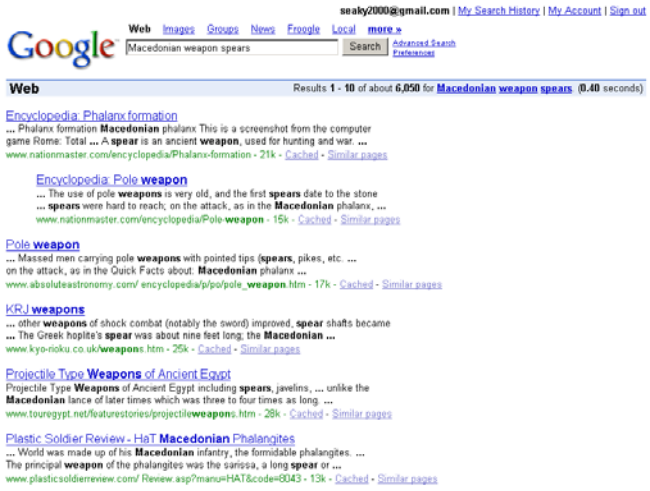
The math is embodied in the Google PageRank algorithm, now recognized as the painfully obvious way to figure out whether an average user keying *spears* in a Google search box wants *Macedonian weapons* or *Britney Spears*, the rock star. PageRank looks at the query, consults its index which includes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

Look at the query run on May 25, 2005, for *spears* and its result set:



Algorithms crunching user data return the correct results when Google uses its metrics to determine that most Google users keying *spears* want the celebrity, *Britney Spears*.

Contrast the spears query with the three-word query *Macedonia weapon spears*:



In order to retrieve, spears used in warfare, the user must craft a three-term query. Google's system correctly parses this query and returns results relevant for a user interested in Macedonia weapons. The results pointing to Egyptian weapons may be false drops.

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 result list. The expert will cavil at the order in which the Britney Spears's 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 Macedonia 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 with shares of Google would point out that there is no in-line text advertisment.

Running queries and having experts comb through the result 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 behaviour 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 will not reveal the gears and wheels inside their PageRank algorithm. Google mavens speculate that Google's PageRank system uses as many as 60 or 80 "factors" to determine that the query spears should display the official Britney Spears's Web site as the most relevant "hit." Keen-eyed searchers will see the false drop in the news story 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's half-life is reached.

Google is a voting 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 and 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. In the good old days of LexisNexis, SDC Orbit, and Lockheed Dialog, habitual users were the important customers. Google and Yahoo! have figured out how to hook users and generate orders of magnitude more cash than their predecessors did or do.

The Conspiracy Theory and Relevance

Search engine optimization is the discipline of crafting publicly-accessible Web content in order to boost that Web site's ranking in a Google search results list. In the book *The Google Legacy* (forthcoming from Infonortics, Ltd. in August 2005), a collection of about 75 of PageRank's "factors" appears. These factors do not come directly from Google. Various SEO "experts" run tests, share knowledge, and attempt to reverse en-

gineer with digital waterwitches the secret of a top ranking on the first page of a Google results page.

The reasons for this are somewhat obvious. The language used to describe these reasons ranges from the pseudo-scholarly to the P.T. Barnum school of rhetoric. If one strips away the linguistic pimping that goes into search engine optimization, the reasons include:

- 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 getting paid their fees or salaries. A SEO consultant in Louisville, Kentucky, working with a large ISP said, "First page! My client just expects me to move the site up in the Google rankings."

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

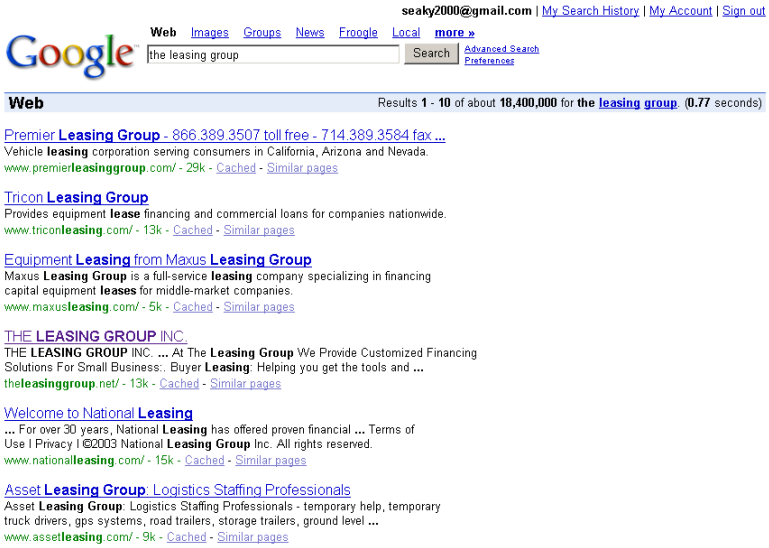
If we focus our attention on what the SEO wizards call organic search results, there is considerable room to influence a Web page's ranking in a Google results list. An *organic search result* refers to changes made to a Web site to influence that Web site's ranking in a Google results list. For the interested reader, 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. A short list of the Top Ten includes:

1. Dynamic urls. Some content management systems like BroadVision and others generate pages when users take an action. The adjustment may be painful; for example, generating a flat HTML version of certain pages specifically for the Google spider to process. Dynamic pages need to be considered for situations when Google indexing is not important.
2. iFrames or invisible frames. iFrames appear in a number of Google applications. iFrames seems to create issues with the Google spider. The optimization technique is to use table.
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 Web masters leaves one with a sense that Google is aware of the 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 okay 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 drift downwards 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. Content that violates this "daughter" test should be removed. The probabilities 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."

What type of a boost can one expect? Consider a site such as one operated by a large financial services firm. The company's URL is http://theleasinggroup.com. A query for

the leasing group returns this results list:



Certain content management systems generate pages difficult for indexing robots or spiders to parse in a meaningful way. Search engine optimization pays big dividends when pages are properly tagged and coded. Unfortunately, once an organization invests in a content management system, that CMS system is unlikely to be displaced easily. The dilemma is that using the CMS pages in their “native” form may make the pages hard to index. Fixing this means creating new types of pages, possibly by hand, so the cost-reductions of the CMS evaporate.

The PageRank algorithm misses this firm entirely on the first page of results. The closest match is The Leasing Group, a Dot Net domain with company headquarters in Canada, not Louisville, Kentucky. Investigation of the code used in the top ranked page Premier Leasing Group shows a number of optimization tactics being used. The reader can inspect the source for the page more closely, but here’s a checklist of the major factors:

1. Content presented in tables. Content is on the same subject leasing, so the semantic vector score is high.
2. Clean, although complex code. The Google spider can figure out where the words and links are. No crazy flash or other gimmicks much loved by Buffy the Web page designer armed with an MFA and Macromedia’s DreamWeaver.
3. A clear page label which has the tag <Title>
4. An abbreviated site map that links to content elsewhere on the site.

None of these tips skew results in terms of the average user. However, the owner of the domain name theleasinggroup.com rightly wonders why his site is not in the first page

of results. The reasons are not far to seek. A review of the site reveals:

1. Minimal content the is digestible by the Google spider
2. No explicit site map
3. Dynamic pages without opaque page naming conventions and without useful indexing in the form of metatags
4. No inbound or outbound links, not even a listing in DMOZ or the Yahoo! Directory.

For the owner of the domain theleasinggroup.com, organic SEO would help the site move up in the Google rankings. Google does know the site exists, but the site itself lacks the basic content and tags that PageRank favours. In this example, SEO is a positive factor and well worth the investment in SEO expertise.

SEO has a dark side as well, and SEO has not one Darth Vader but dozens, if not hundreds, using the SEO lightsabre to slay "objective relevance." The figure below shows a screenshot of a the result of a query for *Viagra* on www.infosonar.com.

This is one point of entry to a series of linked sites. Google sees each site as a distinct entity. Links point among the linked sites. Some spiders detect these; other spiders do not. Engineers try to stay one-step ahead of one another, leading to an "arms race" on the indexing side to find tricks and on the marketers side to figure out how to move up in the rankings.

The site is legitimate, but the principal content is sponsored listings. The site also has a sister site called Infosonarnet.com. This site also appears legitimate. Its content also seems to be a collection of pay-for-traffic listings, not substantive content. Both sites are owned and operated by DomainSpa. That company's principal business appears to

be another collection of services that include a type of clustering tool for general queries as well as more links to pay-for-click sites.

This type of site engineering is legitimate, and it is clear that the Google indexing system includes these domains in its crawl. However, many of the links in the Google index that include a snippet describing certain content link to one of these site's splash pages, not the content. A look in the Google cache reveals that some of the content consists of lists of links that point to sites sharing a theme such as travel.

The reader may want to explore http://domainspa.com, Infosonar.com, and Infosonar-net.com to get a better understanding of how links, different domains, and links can appear in the Google index at relatively high rankings without the type of substantive information usually perceived as important to a high ranking. PageRank is not infallible, and it can be outfoxed or at least shown content that sails through any filters Google has in place.

Now Relevance in the Real World

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

Rankings can be influenced by following some common sense rules. None of the optimization "tricks" is much of a trick. Good content, accurate programming, and useful indexing are basic guidelines to follow.

Code that tries to spoof Google indexing and page constructers specifically to attract a user looking for one thing and then delivering quite another exist. Google engineers try to identify these pages and write code or adjust their PageRank factors address the problem.

As more people become skilled in Google's blindspots, the relevance of the results may be eroded. One must look for chinks in Google's indexing armour. Explore long enough and gaps can be found. Some are interesting like the AdSense example used to begin this article. Others are more problematic and quite sophisticated like the DomainSpa.com example.

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. Obviously, if the relevancy ranking algorithms of Google or any other Web indexing service is 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 com-

promised. 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 back up floppies.