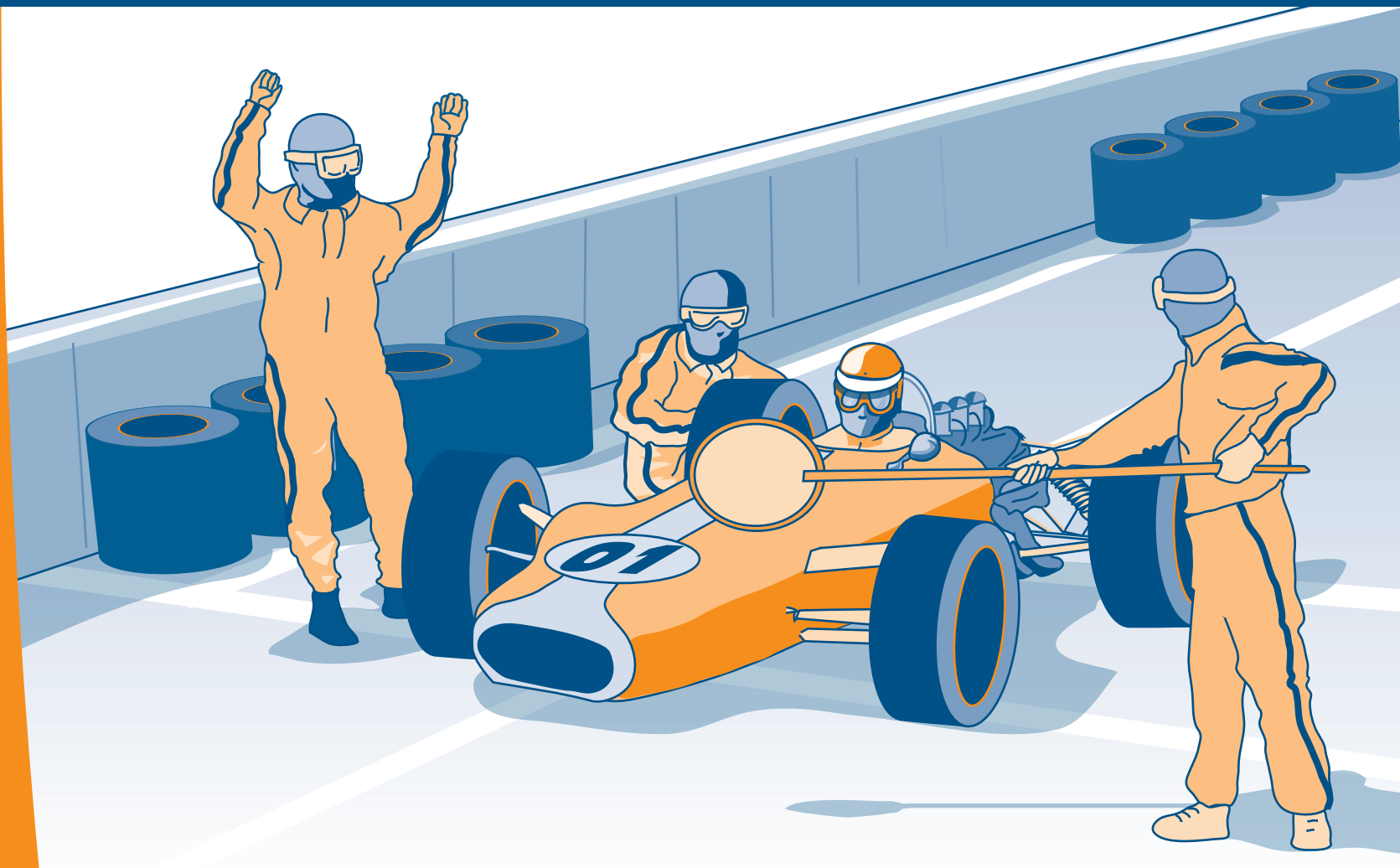


Covers SEO  
and PPC strategies,  
tips, and techniques!



# *The Search Engine Marketing Kit*



By Dan Thies

Manual and CD-ROM

## The Search Engine Marketing Kit (Chapter 1)

---

Thank you for downloading this excerpt of Dan Thies's *The Search Engine Marketing Kit*. This excerpt contains the Summary of Contents, Information about the Author, Expert Reviewers, and SitePoint, Table of Contents, Preface, and a chapter of the kit. We hope you find this information useful in evaluating *The Search Engine Marketing Kit*.

For more information on *The Search Engine Marketing Kit* and to order, [visit sitepoint.com](http://www.sitepoint.com)

## Summary of Contents of this Excerpt

About This Kit.....	viii
1. Understanding Search Engines .....	1
Index .....	254

## Summary of Additional Kit Contents

2. Search Engine Optimization Basics .....	41
3. Advanced SEO And Search Engine-Friendly Design.....	82
4. Paying To Play: Pay-Per-Click And Paid Inclusion.....	120
5. Running A Search Engine Marketing Business .....	155
6. Interviews .....	196
7. Tools .....	222
A. Resources.....	234

# **The Search Engine Marketing Kit**

**by Dan Thies**

---

# The Search Engine Marketing Kit

by Dan Thies

Copyright © 2005 SitePoint Pty. Ltd.

**Managing Editor:** Simon Mackie

**Editor:** Georgina Laidlaw

**Expert Reviewer:** Ed Kohler

**Expert Reviewer:** Jill Whalen

**Expert Reviewer:** Gord Collins

**Printing History:**

First Edition: March 2005

**Cover Designer:** Julian Carroll

**Cover Illustrator:** Lucas Licata

**CD-ROM Designer:** Alex Walker

## Notice of Rights

All rights reserved. No part of this kit may be reproduced, stored in a retrieval system or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical articles or reviews.

## Notice of Liability

The author and publisher have made every effort to ensure the accuracy of the information herein. However, the information contained in this kit is sold without warranty, either express or implied. Neither the authors and SitePoint Pty. Ltd., nor its dealers or distributors will be held liable for any damages to be caused either directly or indirectly by the instructions contained in this kit, or by the software or hardware products described herein.

## Trademark Notice

Rather than indicating every occurrence of a trademarked name as such, this kit uses the names only in an editorial fashion and to the benefit of the trademark owner with no intention of infringement of the trademark.



Published by SitePoint Pty. Ltd.

424 Smith Street Collingwood  
VIC Australia 3066.

Web: [www.sitepoint.com](http://www.sitepoint.com)

Email: [business@sitepoint.com](mailto:business@sitepoint.com)

ISBN 0-9752402-5-0

Printed and bound in the United States of America

---

---

## About The Author

Dan Thies lives in Frisco, Texas with his wife, two sons, two cats, and a very hyperactive miniature pinscher. He has been a student and practitioner of search engine marketing since the earliest days of the industry. Since 2001, he has been an active writer, speaker, and teacher for beginners and professionals alike.

He is a member of the Executive Committee of SeoPros, the Organization of Search Engine Optimization Professionals; a membership committee member and volunteer for SEMPO (the Search Engine Marketing Professional Organization); and a frequent speaker at Jupiter Media's Search Engine Strategies conferences.

## About The Expert Reviewers

Ed Kohler is president of Haystack In A Needle, a Web marketing firm based in Minneapolis, MN, offering pay-per-click advertising, search engine optimization, and email marketing consulting services.

Jill Whalen of High Rankings is an internationally recognized search engine optimization consultant and host of the free weekly High Rankings Advisor search engine marketing newsletter and forum. She is the author of the handbook *The Nitty-gritty of Writing for the Search Engines*.

Gord Collins owns Bay Street Search Engine Optimization, an SEO company in Toronto. He has been an SEO specialist since 1998 and has authored two books on the subject.

## About SitePoint

SitePoint specializes in publishing fun, practical, and easy-to-understand content for Web professionals. Visit <http://www.sitepoint.com/> to access our books, newsletters, articles and community forums.

---

---

*This kit is dedicated to my wife, Gina, and my sons, Jeremy and Jordan. Without their support and extreme patience I would never have been able to complete this work. In fact, without them, there isn't much point in getting out of bed!*

---

---

# Table of Contents

<b>About This Kit .....</b>	<b>viii</b>
Who Should Read This Kit? .....	ix
What's in This Kit? .....	ix
What's on the CD-ROM? .....	xi
Your Feedback .....	xiii
Acknowledgements .....	xiii
Getting Started .....	xiii
<b>1. Understanding Search Engines .....</b>	<b>1</b>
A Brief History of the Search Engine .....	1
The Early Days of Web Search .....	2
The Great Search Engine Explosion .....	2
Google Dominates, the Field Narrows .....	3
Anatomy of a Web Search Portal .....	5
Crawler-Based (Organic) Listings .....	6
Sponsored (Pay-Per-Click) Listings .....	7
Directory (Human-Edited) Listings .....	8
Other Listings .....	9
Search Engine Marketing Defined .....	10
The Crawling Search Engines .....	11
Major Tasks Handled by Search Engines .....	11
The Crawling Phase: How Spiders Work .....	12
Scheduling: How Search Engines Set Priorities .....	17
Parsing and Caching .....	18
Results of the Crawling Phase .....	19
Indexing: How Content is Analyzed .....	20
Link Analysis .....	22
How Queries Are Processed .....	28
Ranking and Retrieval Strategies .....	31
Other Considerations .....	32
What Search Engines Want .....	33
Snapshot of the Search Market .....	34
The Future of Search .....	37
Localization .....	37
Context and Personalization .....	39
Structure and the Semantic Web .....	39
Summary .....	40
<b>2. Search Engine Optimization Basics .....</b>	<b>41</b>
The Process and Craft of SEO .....	41
Phase I: Keyword Strategy .....	42
Understanding the Keyword Hierarchy .....	42
Step 1–Keyword Discovery .....	47

---

Step 2–Keyword Research and Metrics .....	48
Step 3–Keyword Selection .....	57
Phase 2: Site Design and Structure .....	58
Mapping Search Terms to Content .....	58
Crawlability and Site Navigation .....	61
Phase 3: Optimizing Web Pages .....	63
Key Page Elements .....	63
Page Layout .....	66
SEO Copywriting .....	66
Keyword Density and Overdoing It .....	67
HTML Issues .....	68
Phase 4: Link Building .....	69
Managing the External Profile .....	69
Directory Submissions .....	71
Other One-Way Links .....	72
Link Exchanges and Partnerships .....	73
Keeping it Relevant .....	73
Local Sites, Local Links .....	74
Linking Out .....	74
Phase 5: Getting Indexed .....	74
The Easy Way: Links and Crawlability .....	75
Submission and Submission “Services” .....	75
Paid Inclusion Options .....	75
For Indexing Problems, Look at the Site .....	76
Search Engine Spam .....	76
How Search Engines Define Spam .....	77
Cloaking and Variable Delivery .....	78
The Rules Have Never Changed .....	79
Best Practice SEO .....	79
Summary .....	81
<b>3. Advanced SEO And Search Engine-Friendly Design .....</b>	<b>82</b>
Harmonizing Design and SEO .....	82
Designing with Tables .....	83
Designing with CSS .....	84
The Blended Approach .....	85
Dynamic Text Replacement .....	86
Search Engines and Frames .....	86
Why Designers Use Frames .....	86
How Search Engines Handle Frames .....	87
Solution: a Self-Referencing Frameset .....	87
Site Navigation .....	87
Integrating Text Navigation .....	88
Developing a Site Map .....	88
“Crawlable” DHTML and JavaScript Menus .....	89

Pop-Up Windows .....	90
Forced Cookies and Form-Based Navigation .....	91
Working with Flash .....	91
Why Designers Like Flash .....	91
Search Engines and Flash .....	92
Solution: Mixing Flash with HTML .....	92
Solution: Using <noembed> .....	93
Warning: Heavy Content Ahead! .....	93
Duplicate Content: a Definition .....	94
HTTP Headers: a Peek Under the Hood .....	95
Dynamic Site Issues and Opportunities .....	98
Content Management Systems .....	99
Shopping Carts .....	100
Link Directories .....	101
Database and Server Error Handling .....	101
URL Rewriting .....	103
Duplicate Content .....	104
Spider Control and robots.txt .....	104
Diagnosing Duplication .....	108
Sessions and Cookies .....	109
www.example.com vs example.com .....	109
Checking and Fixing Scripts and Variables .....	110
Empty Pages .....	111
Server and Domain Issues .....	112
Redirection .....	112
Custom Error Pages .....	114
Managing Multiple Domain Names .....	114
Moving a Domain .....	117
Watching the Clock .....	118
The Importance of Reliable Hosting .....	119
Summary .....	119
<b>4. Paying To Play: Pay-Per-Click And Paid Inclusion .....</b>	<b>120</b>
Introduction to Pay-Per-Click .....	120
The Pay-Per-Click Marketplace .....	122
Major Players: AdWords and Overture .....	123
Minor Pay-Per-Click Services .....	128
The Pay-Per-Click Process .....	130
Triggering: Targeting Ad Displays .....	132
Click-Through: Qualifying and Motivating Visitors .....	138
Landing Pages and Landing Zones .....	144
Interaction: Improving Website Conversion .....	147
Measurement and Reporting .....	150
Other Pay-To-Play Programs .....	150
Paid Inclusion .....	151

Trusted Feed .....	151
Paid Directories .....	152
The Future of Paid Search .....	152
Supply and Demand Issues .....	153
Advertisers Demand Greater Control .....	153
Big Budgets .....	153
Summary .....	154
<b>5. Running A Search Engine Marketing Business .....</b>	<b>155</b>
Building an SEM Business .....	155
Essential Functions and Skills .....	156
Processes and Tools .....	160
People .....	163
Are You In? .....	165
Getting Business .....	165
Understanding the Selling Cycle .....	166
What Prospects Look For .....	169
Gaining Experience and References .....	171
Finding Prospects .....	173
Consultative Selling .....	175
Effective Proposals .....	176
Doing Business .....	179
Pricing .....	179
What to Sell .....	184
“Difficult” Clients .....	185
Developing SEM Strategy .....	186
Assessment .....	187
Goal Setting .....	188
Planning: Keyword Strategy .....	189
Planning: Linking Strategy .....	192
Being a Professional .....	193
Lifetime Value: Results Matter .....	193
Methods Matter .....	194
Lifelong Learning .....	194
What’s Coming Up? .....	194
<b>6. Interviews .....</b>	<b>196</b>
Andy Beal, Keyword Ranking .....	197
Greg Jarboe, SEO-PR .....	201
Jill Whalen, High Rankings .....	205
John Slade, Overture .....	209
Scottie Claiborne, Karcher Group .....	214
<b>7. Tools .....</b>	<b>222</b>
SEO Tools and Services .....	222
Keyword Discovery .....	222

Wordtracker .....	223
Overture Search Term Suggestion Tool .....	223
Position Technologies .....	224
Priority Submit .....	224
Areliis .....	224
SEO Elite .....	225
Mozilla Firefox .....	225
PPC Tools and Services .....	226
Atlas OnePoint .....	226
BidRank and BidRank Plus .....	227
Findwhat and eSpotting .....	227
Overture Keyword Selector Tool .....	227
Who's Clicking Who? .....	228
Analysis Tools .....	228
Clicktracks .....	229
NetTracker .....	229
Omniture SiteCatalyst .....	229
Webtrends .....	230
DigitalPoint .....	230
Advanced Web Ranking .....	231
Other Tools and Services .....	231
SEO-PR .....	231
Hitwise and Comscore .....	232
eLance.com .....	232
SEO Research Labs .....	233
Alliance Link .....	233
<b>A. Resources .....</b>	<b>234</b>
PageRank in Focus .....	234
PageRank Resources .....	235
White Paper: "The Classification of Search Engine Spam" .....	236
Abstract .....	236
Content Spam .....	238
Meta Spam .....	239
Links .....	241
Redirects .....	243
Agent-Based Delivery and Agent-Based Spam .....	243
IP Delivery and IP Cloaking .....	245
Conclusion .....	246
Resources .....	247
Recommended Reading .....	247
The Big Directory List .....	248
Websites .....	250
SEO/SEM Forums .....	250
Other Resources .....	251

SEM Organizations, Marketplaces, and Directories .....	253
Index .....	254

---

# About This Kit

Search engine marketing (SEM) is one of the hottest topics in the marketing world today. Even traditional “offline” marketing agencies are beginning to understand the powerful ways in which search engine marketing can help achieve their objectives. For those in the business of providing Web design, search engine marketing, and Website promotion services, this is great news, but it does come at a price.

As more individuals and organizations begin to use search engine optimization (SEO) and pay-per-click advertising (PPC) as part of their Website’s marketing strategy, competition for space in the search engine results will become increasingly fierce. In order to compete effectively, search engine marketers-be they individual site owners, or professional consultants-must increase their knowledge and skills.

This kit is intended to fill a large gap that exists between the many helpful but basic introductory texts for beginners, and the often expensive conferences, training programs, and workshops designed for full time search engine marketers. The bottom line is that nobody has taken the knowledge of the professional search engine marketer, and put it in writing. That’s what I’ve tried to do here.

While many things will change in the search engine marketing world, and search engines will continue to adapt their algorithms to deliver more useful information to searchers, some things remain static. In this kit, I hope to have captured those lasting truths, and provided a sound reference to an increasingly complex field.

*The Search Engine Marketing Kit* provides a fantastic road map for your successful journey into search engine marketing. It provides considerable detailed information that will enable you to affect your site’s position in search results. What you do with this knowledge is up to you, but I do hope you’ll pay attention to the strategic (and perhaps philosophical) aspects of the field as well.

Beyond the important how-to questions, and the technical information, I believe that it’s important for search engine marketers to better understand both the search engines and their users. There’s a great deal of conflict between today’s search engine marketers and the search engines on which they rely, but this doesn’t have to be the case.

Search engine marketing should not be carried out in a vacuum, and those practitioners who ignore other pertinent aspects of the Web and the user experience will ultimately fail. The reason is simple: a better Website will generate greater profits, and will ultimately have available more resources with which to compete for search rankings and traffic.

Those who embrace the concept that the entire Website must be optimized in every respect will be the real winners in the long run. A Website should be optimized not just

---

for the benefit of search engines, but for the interests of users and the business or organization behind the site as well.

The primary mission of search engines is to deliver relevant search results to users. Relevance is also the goal of the searcher. If you focus your efforts on enhancing the relevance of your Website, and dedicate your search engine marketing to reaching a well-targeted audience, then you will experience the success you deserve.

## Who Should Read This Kit?

This kit is intended for those who already have some knowledge of Website design and development. The information presented is, in some cases, very basic; in others, it's far more advanced. This is necessary—the kit aims to teach you the skills that professional search engine marketers use to work their magic. It therefore involves a natural graduation from the basics to more advanced material.

Although the primary audience is the Web professional (or skilled amateur), even those who do not participate in the actual design of Websites can learn a lot from this kit. Different chapters and sections will appeal to different interests—designers, IT folks, marketing people, and so on. Site owners and Web professionals who must fulfil all of these roles will find this kit especially useful, because it encompasses so many aspects of the industry.

Readers who find some aspects heavily technical or overwhelming should take heart, because the application of even the most basic elements of search engine optimization to a site can yield substantial results.

The advanced techniques are presented for those readers who require more than the basics, and those who want to develop their expertise over time. Don't be afraid to seek professional help or expert guidance if it's needed. By the time you finish this kit, you'll be an expert yourself—even if you don't fully grasp the technical details.

## What's in This Kit?

### Chapter 1: *Understanding Search Engines*

Do you think you understand how search engines work? So did I ... until I started doing a little in-depth research for this kit! In the first chapter, we'll take a revealing peek under the hood of modern search engines. We'll see where search results come from, how search engines crawl the Web, and how Web pages are ranked.

### Chapter 2: *Search Engine Optimization Basics*

Now that we understand how search engines work, it's time to look at how you can influence your site's position in search results. This chapter covers the basics of

search engine optimization including keyword strategy, optimizing page layout, and effective site structure.

### **Chapter 3: *Advanced SEO and Search Engine-Friendly Design***

It's time to move beyond the basics! This chapter is a little more technical, but necessarily so. The understanding you'll have developed from the first two chapters will serve you well as we explore such advanced topics as duplicate content, Web server issues, content management systems, and moving domains.

### **Chapter 4: *Paying To Play: Pay-Per-Click and Paid Inclusion***

In Chapter 4, we'll take an in-depth look at the world of pay-per-click (PPC) advertising and other pay-to-play options. If you feel that you can't afford to use PPC to promote your Website, think again! Here, you'll discover many new ways to optimize PPC campaigns to deliver a greater return on investment.

### **Chapter 5: *Running A Search Engine Marketing Business***

As I mentioned in the introduction to this kit, the current boom in search engine marketing represents a tremendous business opportunity for Web professionals. In this chapter, we'll look at the various elements involved in building your own search engine marketing business, or integrating search engine marketing services into your current offering.

### **Chapter 6: *Interviews***

In the course of writing this kit, I spoke with dozens of search engine marketing professionals. In this chapter, I've collected six interviews with a range of folks who provide expert perspectives on topics ranging from SEO strategy and pay-per-click, to running a successful search engine marketing business.

### **Chapter 7: *Tools***

The world of search engine marketing is simply filled with companies offering services, software, and other tools. No search engine marketer can do his or her job without a substantial number of these offerings. In this chapter, I'll review a variety of tools, focusing on the best that are currently available.

### **Appendix A: *Resources***

The Appendix provides references to a range of quality resources—some specifically related to search engine marketing, others dealing with broader questions of the Web and its users—that will allow budding search engine marketers to expand their perspectives and boost their knowledge.

# What's on the CD-ROM?

The CD-ROM included with this kit contains several useful tools for search engine marketers and professional SEM consultants.

## **Client Management Form (MS Word)**

This form is intended to help professional SEO/SEM consultants manage information about a client. It includes a contact form for new leads, an intake form for new clients, and a business assessment form.

## **SEM Sales Presentation (MS PowerPoint)**

Another tool for professionals, this PowerPoint presentation template will allow you to speak to the value of search engine marketing, the advantages and disadvantages of SEO and PPC, reasons to hire a professional, and the overall process involved in an SEM campaign.

## **SEM Process Flowchart (MS Visio/PDF)**

This flowchart provides a big-picture overview of the search engine marketing process as described in this kit. This can be used by professionals, in-house search engine marketers, or do-it-yourselfers—anyone who needs to communicate what's involved in search engine marketing.

## **Keyword Research Worksheet (MS Excel)**

This is the same keyword research worksheet that my own company delivers to its clients. The major advantages of this worksheet are that it allows you easily to make a weighted popularity calculation for search terms based on their actual relevance, and estimates monthly traffic for the top ten listings on major search engines.

## **Link Partnership Tracker (MS Excel)**

This worksheet represents a very simple and effective tool for tracking link exchanges, promotions, and partnerships. Keeping this information in Excel allows you to sort and filter the data quickly, and perform mail merges with Microsoft Outlook.

## **Directory Submission Tracker (MS Excel)**

Another simple Excel tool for tracking directory submissions: the directory, the title and description used, the date of submission, and any associated costs can be noted in this tracker. The spreadsheet includes my own seed list of general-purpose directories.

**Site Review Checklist (PDF)**

Intended mainly for the professionals, but useful for all search engine marketers, this site review checklist covers the main points that you'll want to address prior to beginning an SEO/SEM campaign.

**SEM Proposal Sample (MS Word)**

Professionals are often asked to deliver proposals to prospects and clients, and unfortunately, many such documents fall far short of what's required to sell SEM services. This sample proposal exemplifies several key points of effective proposal writing. It begins by addressing the client's business issues, maintains negotiating flexibility, and ties the proposed SEM activities back to business outcomes.

**SEM Service Agreement Sample (MS Word)**

When you start selling your services to clients, you'll need an agreement that sets out the work you'll be doing, how much you'll be paid, and the responsibilities of both parties. This is a basic "bare bones" agreement that you can use to gain ideas for your own contracts. Be sure to seek professional legal advice before entering into any agreement.

**Rates, Pricing, & ROI Calculator (MS Excel)**

This tool is intended for all search engine marketers, to help make realistic assessments of the true value of an SEM campaign. Set hourly rates for each activity, estimate the amount of work required for the campaign, and see how different outcomes affect the overall return on investment (ROI).

**SEM Project Planner (MS Excel)**

Another tool that all search engine marketers can use, this Excel spreadsheet contains a simple project planning tool. Identify the tasks involved in each phase of the campaign, assign responsibilities, and schedule the work. Project planning is especially important when multiple teams are involved, for instance, when an SEO consultant works with a site designer.

**Web CEO (Application)**

Web CEO is suite of software tools, including a keyword researcher, site optimization tool, and link checker, to help you to promote your site in search engines, analyze your visitors, and easily maintain your Website at optimal quality. We've included the free version of Web CEO on the CD-ROM so that you can take it for a test-drive.

## Your Feedback

If you have questions about any of the information presented in this kit, your best chance of a quick response is to post your query in the SitePoint Forums.<sup>[1]</sup> If you have any feedback, questions, or wish to alert us to mistakes, email [books@sitepoint.com](mailto:books@sitepoint.com). Suggestions for improvement, as well as notices of any mistakes you may find, are especially welcome.

## Acknowledgements

I would like to sincerely thank the folks at SitePoint (Georgina, Matt, and Simon) for all their help, and for giving me an opportunity to put my knowledge into writing. Thanks are also due to the kit's technical editors (Ed, Gord, and Jill) who made so many valuable contributions to the final product.

## Getting Started

I hope you enjoy using this kit! Please note that all the information presented here—from case studies to documentation, be it printed or in electronic format—is protected under international copyright laws.

SitePoint Pty. Ltd. reserves all rights to the content presented in *The Search Engine Marketing Kit*, which may not be copied, reproduced, or redistributed, in whole, or in part, under any circumstances, without their express written permission.

Also, while every effort has been made to ensure the accuracy of the information and documents herein, neither the authors, nor SitePoint Pty. Ltd. will be held liable for any damages caused by the instructions or documents contained in *The Search Engine Marketing Kit*.

What we're saying here is that it's up to you to decide what information and resources suit your business, and to seek professional advice if you're unsure about any of the topics covered in *The Search Engine Marketing Kit*.

That's the legals out of the way. Let's get started!

---

[1] <http://www.sitepoint.com/forums/>



# 1

## Understanding Search Engines

---

Every day, millions of people turn to their computers and look for information on the Web. And, more often than not, they use a search engine to find that information. It's estimated that more than 350 million English language Web searches are conducted every day!

In this chapter, I'll offer a brief history of search engines, explaining the different components of search portals, and how people use them. We'll dive into the inner workings of the major crawling search engines. Finally, we'll conclude with a review of today's search engine landscape, and some thoughts on the future of search engine technology.

You may be tempted to skip right past this chapter to the nitty gritty, but, trust me: this is required reading. Understanding where search results come from, how search engines work, and where the industry is headed is essential if you're to make successful search engine marketing decisions now and in the future.

*note*

In the search engine optimization business, one of the key distinctions between amateurs and professionals is that a professional truly understands how the system works, and why. An amateur might learn to tweak a page's content and call it "optimized," but a professional is capable of explaining the rationale behind their every action, and adapting to changing industry conditions without radically altering their methods.

## A Brief History of the Search Engine

The World Wide Web was born in November, 1990, with the launch of the first Web server (and Web page) hosted at the CERN research facility in Switzerland. Not surprisingly, the purpose of the first Web page was to describe the World Wide Web project. At the time, no search engine was needed—you could literally read the entire contents of the World Wide Web in less than an hour.

---

By early 1993, the stage was set for the Web explosion. In February of that year, the first (alpha) release of the NCSA Mosaic graphical browser provided a client application that, by the end of the year, was available on all major desktop computing platforms. The Netscape browser, based on Mosaic, was released in 1994. By this time, dial-up Internet access had become readily available and was cheap. The Web was taking off!

## The Early Days of Web Search

Even though the combination of cheap dial-up access and the Mosaic browser had made the Web semi-popular, there was still no way to search the growing collection of hypertext documents available online. Most Web pages were basically collections of links, and a popular pastime of Web users was to share their bookmark files.

This isn't to say that attempts weren't made to bring order to the swiftly growing chaos. The first automated Web crawler, or robot, was the World Wide Web Wanderer created by MIT student Mathew Gray. This crawler did little more than collect URLs, and was largely seen as a nuisance by the operators of Web servers. Martjin Koster created the first Web directory, ALIWeb, in late 1993, but it, like the Wanderer, met with limited success.

In February 1993, six Stanford graduate students began work on a research project called Architext, using word relationships to search collections of documents. By the middle of that year, their software was available for site search. More robots had appeared on the scene by late 1993, but it wasn't until early 1994 that searching really came into its own.

## The Great Search Engine Explosion

1994 was a big year in the history of Web search. The first hierarchical directory, Galaxy, was launched in January and, in April, Stanford students David Filo and Jerry Yang created Yet Another Hierarchical Official Oracle, better known as Yahoo!.

During that same month, Brian Pinkerton at the University of Washington released WebCrawler. This, the first true Web search engine, indexed the entire contents of Web pages, where previous crawlers had indexed little more than page titles, headings, and URLs. Lycos was launched a few months later.

By the end of 1995, nearly a dozen major search engines were online. Names like MetaCrawler (the first metasearch engine), Magellan, Infoseek, and Excite (born out of the Architext project) were released into cyberspace throughout the year. AltaVista arrived on the scene in December with a stunningly large database and many advanced features, and Inktomi debuted the following year.

Over the next few years, new search engines would appear every few months, but many of these differed only slightly from their competitors. Yet the occasional handy innovation

would find its way into practical use. Here are a few of the most successful ideas from that time:

- ❑ GoTo (now Overture) introduced the concept of pay-per-click (PPC) listings in 1997. Instead of ranking sites based on some arcane formula, GoTo allowed open bidding for keywords, with the top position going to the highest bidder. All major search portals now rely on PPC listings for the bulk of their revenues.
- ❑ Metasearch engines, which combine results from several other search engines, proliferated for a time, driven by the rise of pay-per-click systems and the inconsistency of results among the major search engines. Today, new metasearch engines are rarely if ever seen, but those that remain possess a loyal following. The current crop of metasearch engines display mostly pay-per-click listings.
- ❑ The Mining Company (now About) launched in February 1997, using human experts to create a more exclusive directory. Many topic-specific (vertical) directories and resource sites have been created since, but About remains a leading resource.
- ❑ DirectHit introduced the concept of user feedback in 1998, allocating a higher ranking to sites whose listings were clicked by users. DirectHit's data influenced the search results on many portals for a long time, but, because of the system's susceptibility to manipulation, none of today's search portals openly use this form of feedback. DirectHit was later acquired by Ask Jeeves (now Ask), and user behavior may well be factored into the Ask/Teoma search results we see today.
- ❑ Pay-to-play was introduced, as search engines and directories sought to capitalize on the value of their editorial listings. The LookSmart and Yahoo! directories began to charge fees for the review and inclusion of business Websites. Inktomi launched "paid inclusion" and "trusted feed," allowing site owners to ensure their inclusion (subject to editorial standards) in the Inktomi search engine.
- ❑ The examination of linking relationships between pages began in earnest, with AltaVista and other search engines adding "link popularity" to their ranking algorithms. At Stanford University, a research project created the Backrub search engine, which took a novel approach to ranking Web pages.

## Google Dominates, the Field Narrows

The Backrub search engine eventually found its way into the public consciousness as Google. By the time the search engine was officially launched as Google in September 1998, it had already become a very popular player.

The development of search engines since that time has been heavily influenced by Google's rise to dominance. More than any other search portal, Google has focused on the user experience and quality of search results. Even at the time of its launch, Google

offered users several major improvements, some of which had nothing to do with the search results offered.

One of the most appealing aspects of Google was its ultra-simple user interface. Advertising was conspicuously absent from Google's homepage—a great advantage in a market whose key players typically adorned their pages with multiple banners—and the portal took only a few seconds to load even on a slow dial-up connection. Users had the option to search normally, but a second option, called “I'm Feeling Lucky,” took users directly to the page that ranked at the top of the results for their search.

Like its homepage, Google's search results took little time to appear and carried no advertising. By the time Google began to show a few paid listings through the AdWords service in late 2000, users didn't mind: Google had successfully established itself as the leading search portal and, unlike many other search engines, it didn't attempt to hide paid advertising among regular Web search results.

Many other search portals recognized the superiority of Google's search results, and the loyalty that quality generated. AOL and Yahoo! made arrangements to display Google's results on their own pages, as did many minor search portals. By the end of 2003, it was estimated that three-quarters of all Web searches returned Google-powered results.

Within a few years, the near-monopoly that Google achieved in 2003 will be recognized as a high water mark, but the development of this search engine is by no means finished.

The years 2001–2003 saw a series of acquisitions that rapidly consolidated the search industry into a handful of major players. Yahoo! acquired the Inktomi search engine in March 2003; Overture acquired AltaVista and AllTheWeb a month later; Yahoo! announced the acquisition of Overture in August 2003.

In 2004, a new balance of power took shape:

- Yahoo! released its own search engine powered by a fusion of the AltaVista, Inktomi, and AllTheWeb technology they acquired in 2003. Yahoo! stopped returning Google search results in January 2004.
- Google's AdWords and AdSense systems, which deliver pay-per-click listings to search portals and Websites respectively, grew dramatically. Google filed for an initial public offering (IPO).
- The popularity of the Ask search portal, powered by the innovative Teoma search engine, steadily increased. Like most portals that Yahoo! doesn't own, Ask uses Google's AdWords for paid listings.
- The 800-lb gorilla of the computing world, Microsoft, announced plans for its own search engine, releasing beta versions for public use in January and June of 2004,

and formally launching the service in February 2005. Microsoft now offers MSN search results on the MSN portal.

That's enough history for now. We'll take a closer look at the current search engine landscape a little later in this chapter, when I'll introduce you to the major players, and explain how all this will affect your search engine strategy.

## Anatomy of a Web Search Portal

Today, what we call a search engine is usually a much more complex Web search portal. Search portals are designed as starting points for users who need to find information on the Web. On a search portal, a single site offers many different search options and services:

- AOL's user interface gives users access to a wide variety of services, including email, online shopping, chat rooms, and more. Searching the Web is just one of many choices available.
- MSN features Web search, but also shows news, weather, links to dozens of sites on the MSN network, and offers from affiliated sites like Expedia, ESPN, and others.
- Yahoo! still features Web search prominently on its homepage, but also offers a dazzling array of other services, from news and stock quotes to personal email and interactive games.
- Even Google, the most search-focused portal, offers links to breaking news, Usenet discussion groups, Froogle shopping search, a proprietary image search system, and many other options.

In this section, we'll examine the makeup of a typical search engine results page (SERP). Every portal delivers search results from different data sources. The ways in which these sources are combined and presented to the user is what gives each Web search portal its own unique flavor.

Changes to the way a major portal presents its search results can have a significant impact on the search engine strategy you craft for your Website. As we look at the different sources of search results, and the ways in which those results are handled by individual portals, I'll offer examples to illustrate this point.

A typical search engine results page has three major components: crawler-based listings, sponsored listings, and directory listings. Not all SERPs contain all three elements; some portals incorporate additional data sources depending on the search term used. Figure 1.1, from Yahoo!, shows a typical SERP:

Figure 1.1. A typical SERP.

The screenshot shows a Yahoo! search results page for the query "search engine optimization". At the top, there are navigation links for "Web", "Images", "Directory", "Yellow Pages", "News", and "Products". The search bar contains the query, and a "Search" button is visible. Below the search bar, there are links for "Shortcuts", "Advanced Search", and "Preferences". The results section shows "Results 1 - 20 of about 4,700,000 for search engine optimization. Search took 0.11 seconds. (About this page...)".

The page is divided into several sections:

- SPONSORED LISTINGS:** This section is highlighted with an orange arrow. It contains two sponsored results:
  - Professional Search Engine Placement:** We provide extensive optimization among the major search engines. Receive a free placement with your proposal. [www.submitawebsite.com](http://www.submitawebsite.com)
  - Search Engine Marketing:** Achieve maximum online visibility with search engine optimization. [www.refinery.com](http://www.refinery.com)
- ORGANIC LISTINGS:** This section is highlighted with an orange arrow. It contains two organic results:
  - Search Engine Optimization Inc.:** provides search engine optimization, search engine placement, and Internet marketing services. Category: B2B > Search Engine Optimization Services. [www.seoinc.com/](http://www.seoinc.com/) - 19k - Cached - More pages from this site
  - Search Engine Watch: Tips About Internet Search Engines & Search Engine Submission:** Search Engine Watch is the authoritative guide to searching at Internet search engines and search engine registration and ranking issues. Learn to submit URLs, use HTML meta tags and boost ... Free SEO Forum Search Engine Optimization. Cash out on your content with Google AdSense ... traffic to your web site. Search Engine Marketing. Search Engine Optimization Internet Marketing ... Category: Search Engine Optimization (SEO) Resources. RSS: View as XML - Add to My Yahoo! [Beta] [searchenginewatch.com/](http://searchenginewatch.com/) - 53k - Cached
- DIRECTORY LISTINGS:** This section is highlighted with an orange arrow. It contains three directory listings:
  - IGlobalMedia: Affordable SEO:** With several thousand marketing partners and several million dollars in payouts... [marketing.iglobalmedia.com](http://marketing.iglobalmedia.com)
  - Improved Search Engine Rankings:** Search engine evaluator improves search results and your business. Get started now. [www.strategic-e.co.nz](http://www.strategic-e.co.nz)
  - Search or Engine or Optimization:** Free white paper download. Learn the tips, tricks and practices in search engine... [track.did-it.com](http://track.did-it.com)

## Crawler-Based (Organic) Listings

Most search portals feature crawler-based search results as the primary element of their SERPs. These are also referred to as editorial, free, natural, or organic listings. Throughout the rest of this kit, we will refer to crawler-based listings as organic listings.

Crawler-based search engines depend upon special programs called robots or spiders. These spiders crawl the Web, following one link after another, to build up a large database of Web pages. We will use the words spider, crawler, or robot to refer to these programs throughout this kit.

Each crawler-based search engine uses its own unique algorithm, or formula, to determine the order of the search results. The databases that drive organic search results primarily contain pages that are found by Web-crawling spiders. Some search engines offer paid inclusion and trusted feed programs that guarantee the inclusion of certain pages in the database.

Paid inclusion is one of many ways in which search engines have begun to blur the line between organic and paid results. Trusted feed programs allow the site owner to feed the search engine an optimized summary of the pages in question; the pages may be ranked on the basis of their content summaries rather than their actual content.

Although all the search engines claim that paid inclusion does not give their customers a ranking benefit, the use of paid inclusion does offer SEO consultants an opportunity to tweak and test copy on Web pages more frequently. We will learn more about this in Chapter 2.

Organic search listings are certainly the primary focus for search engine marketers and consultants, but they're not the only concern. In many cases, the use of pay-per-click is essential to a well-rounded strategy.

Most of today's search portals do not operate their own crawler-based search engine; instead, they acquire results from one of the major organic search players. The major providers of organic search listings are Google and Yahoo! who, in addition to operating their own popular search portals, also provide search results to a variety of different portals.

Aside from Google and Yahoo!, only a few major players operate crawling search engines. Ask uses its own Teoma search engine, LookSmart owns Wisenut, Lycos, too, has its own crawler-based engine, and Microsoft's MSN search is also in the mix. That's a grand total of six crawler-based search engines accounting for nearly all of the organic search results available in the English language.

*note*

In order to have a meaningful chance to gain traffic from organic search listings, a Web page must appear on the first or second page of search results. Different search portals show varying numbers of results on the first page: Google displays ten, Yahoo! shows 15, and MSN's search presents eight. Any changes a major search portal might make to the listing layout will affect the amount of traffic your search engine listings attract.

## Sponsored (Pay-Per-Click) Listings

It costs a lot of money to run a search portal. Crawler-based search engines operate at tremendous expense—an expense that most portals can't afford. Portals that don't operate their own crawler-based search engines must pay to obtain crawler-based search results from someone who does.

Either way, the delivery of unbiased organic search results is expensive, and someone has to pay the bill. In the distant past, search portals lost money hand over fist, but today, even very small search portals can generate revenue through sponsored listings. Metasearch engines typically use sponsored listings as their primary search results.

In addition to helping search portals stay afloat, sponsored listings provide an excellent complement to organic search results by connecting searchers with advertisers whose sites might not otherwise appear in the search results.

Most portals do not operate their own pay-per-click (PPC) advertising service. Instead, they show sponsored results from one or more partners and earn a percentage of those advertisers' fees. The major PPC providers are Google AdWords and the Overture service

offered by Yahoo!. Other PPC providers with a significant presence include Findwhat and LookSmart.

The PPC advertising model is simple. Advertisers place bids against specific search terms. When users search on those terms, the advertiser's ads are returned with the search results. And, each time a searcher clicks on one of those ads, the advertiser is charged the per-click amount he or she bid for that term. PPC providers have added a few twists to this model over the years, as we'll see in Chapter 4.

Different PPC providers use different methods to rank their sponsored listings. All methods start with advertisers bidding against one another to have their ads appear alongside the results returned for various search terms, but each method has its own broad matching options to allow a single bid to cover multiple search terms.

*note*

The bidding for extremely popular search terms can be quite fierce: it's not unusual to see advertisers bidding \$10 per click—or more—for the privilege of appearing at the top of the sponsored listings. Reviewing the amounts that bidders are willing to pay for clicks to sponsored listings can give SEO practitioners a very good idea of the popularity of particular search terms—terms that may also be suitable for organic optimization.

In addition, PPC ranking systems are no longer as simple as allocating the highest position to the highest bidder. Google's methodology, for example, combines the click-through rate of an advertiser's listing (the number of clicks divided by the number of times it's displayed) with that advertiser's bid in assessing where the PPC advertisement will be located. Google's method tends to optimize the revenue generated per search, which is one of the reasons why its AdWords service has gained significantly on Overture.

*note*

In the example SERP shown above (Figure 1.1), Yahoo! displays the first two sponsored listings in a prominent position above the organic results. Understanding which sponsored results will be displayed most prominently will help you determine how much to bid for different search terms. For example, it may be worth bidding higher to get into the #1 or #2 position for the most targeted search terms, since those positions will gain the most traffic from Yahoo!.

## Directory (Human-Edited) Listings

Directory listings come from human-edited Web directories like LookSmart, The Open Directory[1], and the Yahoo! Directory. Most search portals offer directory results as an optional search, requiring the user to click a link to see them.

Because directories usually only list a single page (the homepage) of a Website, it can be difficult for searchers to find specific information through a directory search. As the quality of organic search results has improved, search portals have gradually reduced their emphasis on directory listings.

---

[1] <http://dmoz.org>

Currently, only Lycos displays a significant number of directory listings (from LookSmart), and that's likely to change as LookSmart transitions from its old business model (paid directory) into a standard PPC service.

The decline of directory listings within search results does not diminish the importance of directory listings in obtaining organic search engine rankings. All crawler-based search engines take links into account in their rankings, and links from directories are still extremely important.

*note*

The way that Yahoo! makes directory results available to users should be a significant factor in helping the site owner decide whether or not to pay for a listing in the directory. At \$299 per year, a paid listing in the Yahoo! Directory is a considerable expense for small businesses. Yet, while there is value in any link, the directory itself no longer generates significant traffic.

In addition, it is by no means clear whether the display of a directory category link below a site's organic search result listing may contribute to the click-through rate for that listing. In fact, it's possible that users might click this directory link and arrive at the directory category page, where the given listing could be buried at the bottom of a long list of competing sites.

Compared to other advertising options, paying \$299 for a link buried deep within the Yahoo! Website is not as appealing as it once was. In addition, sites listed in the Yahoo! Directory automatically have a title and description displayed alongside each of their listings in the organic search results. This style of listing can actually generate a lower click-through than an ordinary listing within the organic results.

Whether or not you currently have a Yahoo! Directory listing, you owe it to yourself to discuss other ways to make use of those funds. For example, at an average of 20 cents per click, you could bring in nearly 1500 visitors per year through PPC advertising.

## Other Listings

In addition to the three main types of search results, most search portals now offer additional types of search listings. The most common among these are:

- Multimedia searches, which help users find images, sounds, music etc.
- Shopping searches to help those searching for specific products and services.
- Local searches to find local business and information resources.
- People searches, including white pages, yellow pages, reverse phone number lookups.
- Specialized searches, covering government information, universities, scientific papers, maps, and more.

# Search Engine Marketing Defined

Throughout this kit, I'll use **search engine marketing (SEM)** to describe many different tasks. We'll talk about this concept a lot, so it will be helpful to have a working definition. For the purposes of these discussions, we'll define search engine marketing as follows:

Search engine marketing is any legal activity intended to bring traffic from a search portal to another Website.

The term search engine marketing, therefore, covers a lot of ground. Wherever people search the Web, whatever they search for, and wherever the search results come from—if you're trying to reach out to target visitors, you're undertaking search engine marketing. The goal of SEM is to increase the levels of high-quality, targeted traffic to a Website. In this kit, we'll focus on the two primary disciplines of SEM, which are:

## Search Engine Optimization (SEO)

The function of SEO is to improve a Website's position within the organic search results for specific search terms, and to increase the overall traffic the site garners from crawler-based search engines. This is accomplished through a combination of on-page content and off-page promotion (such as directory submissions).

## Pay-Per-Click Advertising (PPC)

PPC involves the management of keyword-targeted advertising campaigns through one or more PPC service providers, such as Google's AdWords, or Overture from Yahoo!. The advertiser's goal is to profitably increase the amount of targeted traffic that his or her Website receives from search portals.

In addition to these two major disciplines, there are other aspects of search engine marketing that we'll discuss to a lesser degree, including:

- Contextual advertising, which is offered by many PPC service providers. Contextual advertising delivers targeted advertising based on the content of each individual Web page that carries an ad. Advertisers who have used PPC to target people searching on the term fishing can also have their ads distributed across a great many Websites on which fishing is discussed. This is a fast-growing market, and one that's sure to become a very significant part of SEM over time.
- Directory submission, which involves the submission of Websites to general-purpose and vertical (topic-specific) directories, or vortals. We will discuss this mainly in the context of SEO, but many directories (both general-purpose and vertical) provide search-driven traffic to the Websites they list. Many operate on a paid advertising or PPC basis. As searchable business directories like Verizon's SuperPages and the

already established Business.com grow, so too will this area of search engine marketing.

Search engine marketing is a fast-growing and rapidly changing field. Before we get too far ahead of ourselves, though, let's take a close look at where organic search results come from: the crawling search engines.

## The Crawling Search Engines

In this discussion, we'll explore the major components of a crawler search engine, and understand how they work. The typical Web user assumes that when they search, the search engine actually goes out onto the Web to look around. In fact, the job of searching the Web is vastly more complex than that, requiring massive amounts of hardware, software, and bandwidth.

To give you an idea of just how much hardware it takes to run a large-scale, modern search engine, here's a staggering figure: Google runs what is believed to be the world's largest Linux server cluster, with over 10,000 servers at present, and more being added all the time (it was "only" 4,000 in June, 2000).

Searching a small collection of well-structured documents, such as scientific research papers, is difficult enough, but that task is relatively easy compared to searching the Web. The Web is massive and mobile, consisting of billions of documents in over 100 languages, many of which change or disappear on a daily basis. To make matters worse, there is very little consistency in terms of how information is organized and presented on the Web.

## Major Tasks Handled by Search Engines

There are five major tasks that each crawling search engine must handle, and significant computing resources are dedicated to each. These tasks are:

### Finding Web pages and downloading their contents.

The bulk of this task is handled by two components: the **crawler** and the **scheduler**. The crawler's job is to interact with Web servers to download Web pages and/or other content. The scheduler determines which URLs will be crawled, in what order, and by which crawler. Large crawling search engines are likely to have multiple types of crawlers and schedulers, each assigned to different tasks.

### Storing the contents of Web documents and extracting the textual content.

The primary components at this stage are the **database/repository** and **parser modules**. The database/repository receives the content of each URL from the crawlers, then stores it. The parser modules analyze the stored documents to extract

information about the text content and hyperlinks within. Depending on the search engine, there may be multiple parser modules to handle different types of files, including HTML, PDF, Flash, Microsoft Word, and so on.

### **Analyzing and indexing the content of documents.**

This is handled by the **document indexer**. The text content is analyzed by the indexer and stored in a set of databases called indexes. For simplicity's sake, I'll refer to these indexes as simply "the index." Included in the indexing process is the preliminary analysis of hyperlinks within the documents, feeding URLs back into the scheduler and building a separate index of links. The main focus of this phase is the on-page content of Web documents.

### **Link analysis, to uncover the relationships between Web pages.**

This is the work of the **link analyzer** component. All of the major crawling search engines analyze the linking relationships between documents to help them determine the most relevant results for a given search query. Each search engine handles this differently, but they all have the same basic goals in mind. There may be more than one type of link analyzer in use, depending on the search engine.

### **Query processing and the ranking of Web pages to deliver search results.**

The **query processor** and **ranking/retrieval module** are responsible for this important task. The query processor must determine what type of search the user is conducting, including any specialized operations that the user has invoked. The ranking/retrieval module determines the ranking order of the matching documents, retrieves information about those documents, and returns the results for presentation to the user.

## **The Crawling Phase: How Spiders Work**

As mentioned above, one of the largest jobs of a crawling search engine is to find Web documents, download them, and store them for further analysis. To simplify matters, we've combined the work of tasks 1 and 2 above into a single activity that we'll refer to as the crawling phase.

Every crawling search engine is assigned different priorities for this phase of the process, depending on their resources and business relationships, and what they're trying to deliver to their users. All search engines, however, must tackle the same set of problems.

## **How Search Engines Find Documents**

Every document on the Web is associated with a URL (Uniform Resource Locator). In this context, we will use the terms "document" and "URL" interchangeably. This is an oversimplification, as some URLs return different documents to the user depending on

such factors as their location, browser type, form input etc., but this terminology suits our purposes for now.

To find every document on the Web would mean more than finding every URL on the Web. For this reason, search engines do not currently attempt to locate every possible unique document, although research is always underway in this area. Instead, crawling search engines focus their attention on unique URLs; although some dynamic sites may display different content at the same URL (via form inputs or other dynamic variables), search engines will see that URL as a single page.

The typical crawling search engine uses three main resources to build a list of URLs to crawl. Not all search engines use all of these:

### **Hyperlinks on existing Web pages**

The bulk of the URLs found in the databases of most crawling search engines consists of links found on Web pages that the spider has already crawled. Finding a link to a document on one page implies that someone found that link important enough to add it to their page.

### **Submitted URLs**

All the crawling search engines have some sort of process that allows users or Website owners to submit URLs to be crawled. In the past, all search engines offered a free manual submission process, but now, many accept only paid submissions. Google is a notable exception, with no apparent plans to stop accepting free submissions, although there is great doubt as to whether submitting actually does anything.

### **XML data feeds**

Paid inclusion programs, such as the Yahoo! Site Match system, include trusted feed programs that allow sites to submit XML-based content summaries for crawling and inclusion. As the Semantic Web begins to emerge, and more sites begin to offer RSS (RDF Site Summary) news feed files, some search engines have begun to read these files in order to find fresh content.

Search engines run multiple crawler programs, and each crawler program (or spider) receives instructions from the scheduler about which URL (or set of URLs) to fetch next. We will see how search engines manage the scheduling process shortly, but first, let's take a look at how the search engine's crawler program works.

## **The Robot Exclusion Protocol**

The first search spiders developed a very bad reputation in a hurry. Web servers in 1993 and 1994 were not as powerful as they are today, and an aggressive spider could bring an underpowered Web server to a crashing halt, or use up the server's limited bandwidth, by fetching pages one after another.

Clearly, rules were needed to control this new type of automated user, and they have developed over time. Spiders are supposed to fetch no more than one document per minute (a rate that's probably much slower than necessary) from a given Web host, and they're expected to obey the Robot Exclusion Protocol[2].

In a nutshell, the Robot Exclusion Protocol allows Website operators to place into the root directory of their Web server a text file named `robots.txt` that identifies any URLs to which search spiders are denied access. We'll address the format of this file later; the important point here is that spiders will first attempt to read the `robots.txt` file from a Website before they access any other resources.

When a spider is assigned to fetch a URL from a Website, it reads the `robots.txt` file to determine whether it's permitted to fetch that URL. Assuming that it's permitted access by `robots.txt`, the crawler will make a request to the Web server for that URL. If no `robots.txt` file is present, the spider will behave as if it has been granted permission to fetch any URL on the site.

There are no specific rules about this, and each search engine will implement this differently, but it is considered poor behavior for a search engine to rely on a cached copy of the `robots.txt` file without confirming that it's still valid. In order to save resources, schedulers can assign the crawler program a set of URLs from the same site, to be fetched in sequence, before it moves on to another site. This allows the crawler to check `robots.txt` once and fetch multiple pages in a single session.

## What Happens in a Crawling Session?

For the sake of clarity, let's walk through a typical crawling session between a spider and a Website. In this particular scenario, we'll assume that everything works perfectly, so the spider doesn't have to deal with any unusual problems.

Let's say that the spider has a URL it would like to fetch from our Website, and that this URL has been fetched before. The scheduler will supply the spider with the URL, along with the date and time of the most recent version that has been fetched. It will also supply the date and time from the most recent version of `robots.txt` that has been fetched from this site.

The communication between a user agent (such as your Web browser or our hypothetical spider) and a Web server is conducted via the HTTP protocol. The user agent sends requests, the server sends responses, and this communication goes back and forth.

Once the document has been downloaded from the Web server, the crawler's job is nearly done. It hands the document off to the database/repository module, and informs the scheduler that it has finished its task. The scheduler will respond with another task, and it's back to work for the spider.

---

[2] <http://www.robotstxt.org/wc/exclusion.html>

## Practical Aspects of Crawling

If only things could always be as simple as our hypothetical session above! In reality, there are a tremendous number of practical problems that must be overcome in the day-to-day operations of a crawling search engine.

## Dealing with DNS

The first problem that crawlers have to overcome lies in the domain name system that maps domain names to numeric addresses on the Internet. The root name servers for each top level domain, or TLD (e.g. .com, .net etc.), keep records of the domain name server (DNS server) that handles the addressing for each second level domain name (e.g. example.com).

Thousands of secondary and tertiary name servers across the Internet synchronize their DNS records with these root name servers periodically. When the DNS server for a domain name changes, this change is recorded by the domain name registrar, and given to the root name server for the TLD.

Unfortunately, this change is not reflected immediately in all name servers all over the world. In fact, it can easily take 48–72 hours for the change to propagate from one name server to the next, until the entire Internet is able to recognize the change.

A search engine spider, like any other user, must rely on the DNS in order to find the resources that it's been sent to fetch. Although the major search engines all have reasonably fast updates to their DNS records, when DNS servers are changed, it's possible that a spider will be sent out to fetch a page using the wrong DNS server address. When this happens, there are three possibilities:

- The DNS server from which the spider requests the site's Web server address no longer has a record of the domain name supplied. In this case, the spider will probably hand the URL back to the scheduler, to be tried again later.
- The DNS server does have a record for the domain name, and dutifully gives the spider an address for the wrong Web server. In this case, the spider may end up fetching the wrong page, or no page at all. It may also receive an error status code.
- Even though it's no longer the authoritative name server for the supplied domain name, the DNS server still provides the spider the correct address for the Web server. In this case, the spider will probably fetch the right page.

It's also possible that a search engine could use a cached DNS record for the domain name, and go looking for the Web server without checking to ensure that the record is current. This used to be an occasional problem for Google, but probably will never be

seen again. It certainly hasn't appeared to be a problem for any of the major search engines in some time.

We will discuss exactly how to move a Website from one server to another, from one hosting provider to another, and from one DNS server to another, in Chapter 3. For now, the key point is that the mishandling of DNS can lead to problems for search engines, and this can, in turn, create major headaches for you.

## Dealing with Servers

The next challenge that spiders have to handle is HTTP error messages, servers that simply cannot be found, and servers that fail to respond to HTTP requests. There are also many other server responses that must be handled with particular care in order to avoid problems.

Rather than provide a comprehensive listing of every problem that could ever eventuate, I'll simply list a few broad categories and note how search engines are likely to deal with them. We'll dig more deeply into server issues in Chapter 3.

### Where's That Server?

If a server can't be found, or fails to respond, it's likely a temporary condition. The crawler will inform the scheduler of the error, and move on. If the condition persists, the search engine might remove the URL in question from the index, and may even stop trying to crawl it. It usually takes a long term problem, or a very unreliable server, to elicit such a drastic response, however. If a URL (or an entire domain) is removed because of server problems, a manual submission may be required in order to have the search engine crawl it again.

### Where's That Page?

If a page does not exist at the requested URL, the server will return a 404 Not Found error. Sometimes, this means that a page has been permanently removed; sometimes, the page never existed in the first place; occasionally, pages that go missing reappear later. Search engines are usually quick to remove URLs that return 404 errors, although most of them will try to fetch the URL a couple more times before giving it up for dead. As with server issues, it may be necessary to resubmit pages that have been dropped for returning 404 errors. In Chapter 3, we will discuss the right (and wrong) way to use custom 404 error pages.

### Whoops, There Goes The Database!

Database errors are the bane of dynamic sites everywhere. Unless the code driving the site has robust error handling capabilities, most database errors will cause the Web server to return a 200 OK status code while delivering a page that contains nothing but an error message from the database. When this occurs, the error message will be stored by the spider as if it were the page's content. Resubmission of the page

is not necessary, assuming the database issues have been corrected the next time the spider visits. Chapter 3 will include some recommendations on how best to manage database errors.

### **Sorry, We Moved It ... Or Did We?**

Redirection by the Web server can be a challenge for search engines. A server response of 301 Moved Permanently should cause the search engine to visit the new URL and adjust its database to reflect the change. Trickier for spiders is the 302 Found response code, which is used by many applications and scripts to redirect Web browsers. Search engines currently have varying responses to server-based redirects. In some cases, very bad things can happen if spiders are allowed to follow 302 redirects, as we'll see in Chapter 3.

## **Handling Dynamic Sites**

One of the most difficult challenges faced by today's crawlers is the proliferation of dynamic or database-driven Websites. Depending on the way the site is configured, it's possible for a spider to get caught in an endless loop of pages that generate more pages, with a never-ending sequence of unique URLs that deliver the same (or slightly varied) content.

In order to avoid becoming caught in such spider traps, today's crawlers carefully examine URLs and avoid crawling any link that includes a session ID, the referring URL, or other variables that have nothing to do with the delivery of content. They also look for hints of duplicate content, including identical page titles, empty pages, and substantially similar content. Any of these gotchas can stop a spider from fully crawling a dynamic site. We will review crawler-friendly SEO strategies for dynamic sites in Chapter 3.

## **Scheduling: How Search Engines Set Priorities**

In addition to the challenges that must be overcome in crawling the Web, there are a great number of issues with which search engines must grapple in order to properly manage their crawling resources. As mentioned previously, each search engine's priorities are different.

Five years ago, the major competition between the search engines was to build the largest index of documents. News networks like CNN played up each succeeding announcement of what was described as the new "biggest search engine," which, no doubt, pleased many dot-com investors, even if some of the search engines played it a little fast and loose when it came to the numbers.

Today, the total index size is no longer seen as a key indicator of a search engine's quality. Nonetheless, any search engine must index a substantial portion of the Web in order to deliver relevant search results. Google currently has by far the largest index,

which is especially evident to those searching for detailed technical information, as relevant pages may be buried deep within a site.

The scheduling of crawler activity must be guided by the search engine's individual priorities in four specific areas:

### **Freshness**

In order to deliver the best possible results, every search engine must index a great deal of new content. Without this, it would be impossible to return search results on current events. Most scheduling algorithms involve a list of important sites that should be checked regularly for new content. Indexing XML data feeds helps some search engines keep up with the growth of the Web.

### **Depth vs Breadth**

A key strategic decision for any search engine involves how many sites to crawl (breadth) and how deeply to crawl into each site (depth). For most search engines, making the depth vs. breadth decision for a given site will depend upon the site's linking relationships with the rest of the Web: more popular sites are more likely to be crawled in depth, especially if some inbound links point to internal pages. A single link to a site is usually enough to get that site's homepage crawled.

### **Submitted Pages**

Search engines such as Google, which allow the manual submission of pages, must decide how to deal with those manually submitted pages, and how to handle repeat submissions of the same URL. Such pages might be extremely fresh or important, or they may be nothing more than spam.

### **Paid Inclusion**

Search engines that offer paid inclusion programs generally guarantee that they will revisit paid URLs every 24–72 hours.

In terms of priority, a search engine that offers a paid inclusion program must visit those paid URLs first. After listings for paid inclusion, most search engines will probably focus resources on any important URLs that help them maintain a fresh index. Only after these two critical groups of URLs are crawled will they pursue additional URLs. URLs submitted via a free submission page are probably the last on the list, especially if they have been submitted repeatedly.

## **Parsing and Caching**

Once the contents of a URL have been fetched, they are handed off to the database/repository and stored. Each URL is associated with a unique ID, which will be used

throughout all the search engine's operations. Depending on the type of content, one of two things will happen next.

If the document is already in HTML format, it can be stored immediately, exactly as is. Additional **metadata**<sup>1</sup>, such as the Last-Modified date and page title, may be stored along with the document. This stored copy of the HTML code is used by some search engines to offer users a snapshot view of the page, or access to the cached version.

For documents that are presented in formats other than HTML, such as Adobe's popular Acrobat (PDF) or Microsoft Word, further processing is needed. Typically, search engines that attempt to index these types of documents first translate the document into HTML format with a specialized parser.

Converting non-HTML documents to an HTML representation allows search engines to offer users access to the document's contents in HTML format (as Google does), and to conduct all further processing on the HTML version. When the document contains structural information, such as a Microsoft Word file that makes use of heading styles, search engines can make use of these elements within the HTML translation. Adobe's PDF is notably lacking in structural elements, so search engines must rely on type styles and size to determine the most significant text elements.

At this point, all that has been accomplished is to store an HTML version of the document. Most search engines will perform further parsing at this stage, to extract the text content of the page, and catalog the various elements (headings, links etc.) for analysis by the indexing and link analysis components. Some of them may leave all of this processing to the indexer.

## Results of the Crawling Phase

By the end of the crawling phase, the search engine knows that there was valid content at the URL, and it has added that content (possibly translated to HTML) to its database.

Even before a search engine crawls a page, it must "know" something about that page. It knows that the URL exists and, if the URL was found via links, the search engine may also have found within those links some text that tells it something about the URL.

Once a search engine knows that a URL exists, it's possible that this URL could appear in search results. In Google, a page that has not yet been crawled can appear as a supplemental search result, based on the keywords contained in hyperlinks pointing to that page. At this point, the page's title is not known, so the listing will display the page's URL in place of the title.

---

<sup>1</sup>Metadata should not be confused with `<meta>` tags. Metadata is "data about data." For search engines, the primary unit of data is the Web page, so anything that describes that Web page (other than its content) is metadata. This might include the page's title, URL, and other information such as the Website's directory description, which Yahoo! uses within its search results.

After the crawling phase is complete, the search engine knows the document's title, last-modified date, and its size. Such pages can appear in Google's results as supplemental search results, based on keywords that appear in the page's title and incoming links. After the crawling phase, the page title can also appear in the search results.



The Google search engine provides an unusual amount of transparency around its process and results. It's possible, for example, to have Google return a list of all the URLs it has found within a particular site. The syntax for this search is `site:example.com`.

If some of the URLs listed for a `site:domain` search do not include page titles or page size information, this means that those URLs have not been yet been crawled. If this condition persists, as happens often with dynamic sites, there may be issues with duplicate content, session IDs, empty pages, or other problems that have caused the spider to stop crawling the site. We will cover these issues in Chapter 3.

## Indexing: How Content is Analyzed

After the content of a Web page (or HTML representation of a non-HTML document) has been stored in the database, the indexer takes over, breaking down the page piece by piece, and creating a mathematical representation of it in the search engine's index.

The complexity of this process, the extreme variations between different search engines, and the fact that this part of the process is a closely guarded secret<sup>2</sup>, makes a comprehensive explanation impossible. However, we can speak about the process in general terms that will apply to all crawling search engines.

### What Indexing Means in Practice

When a search engine's indexer analyzes a document, it stores each word that occurs in the document as a hit in one of the indexes. The indexes may be sorted alphabetically, or they may be designed in a way that allows more commonly used words to be accessed more quickly.

The format of the index is very much like a table. Each row in the table records the word, the ID of the URL at which it appeared, its position within the document, and other information which will vary from one search engine to the next. This additional information may include such things as the structural element in which the word appeared (page title, heading, hyperlink etc.) and the formatting applied (bold, italic etc.).

Table 1.1 shows a hypothetical (and simplified) search engine index entry for an imaginary (and very boring) document. The page's title is "Hello, World!" The document itself contains the same words in a large heading, followed by the words "Greetings, everyone!" as the first paragraph of text.

---

<sup>2</sup>A search engine's algorithm must be kept secret, in order to prevent optimizers from unfairly manipulating search results and, of course, to prevent competitors from "borrowing" useful ideas.

Each index contains hits for different groups of words. The hypothetical index entry for the document will therefore be spread across multiple indexes. The only place in which the entire document might remain intact is in the repository of a search engine that retains a full cached copy of each page, as Google does.

**Table 1.1. A Search Engine Index Entry**

Word	Document	Position	Type
hello	1	1	Title
world	1	2	Title
hello	1	3	Heading 1
world	1	4	Heading 1
greetings	1	5	Body
everyone	1	6	Body

The first thing that you will notice is that punctuation is not stored in the index. That's because search engines ignore punctuation—if any of them considered it at all, it would be news to me. When we talk about SEO copywriting later, you'll see the importance of this fact.

For now, you need to understand that search engines don't look at Web pages; they look at indexes that match words to documents. When we talk about the vector space model and the ranking/retrieval process later in this chapter, you may need to refer back to this table to refresh your memory.

I should note one thing before we move on. Although search engines do apply a different “weight” to words that appear in more prominent positions (such as headings), they do not necessarily attempt to store those values in the index records. When we talk about algorithm changes and term weights in just a moment, you'll understand why.

## Link (URL) Discovery and Indexing

URLs that are found within documents are fed back into the scheduler for crawling. Information about the source document may have an impact on the URL's priority within the crawling queue. For example, the URLs contained in links found on an important page (such as the Yahoo! homepage) may take precedence over links that have been found on lesser pages.

When a hyperlink is found within the document that's being indexed, the words in the hyperlink are recorded in the index as a hit, just like any other word, along with the fact that the words appeared in a hyperlink.

There are two ends to every link, though. The source document (the one being indexed) links to a target document, and some search engines also take this into account when indexing a page. The words contained within a hyperlink may also be indexed as hits for the target document.

In this manner, a URL that has never been crawled can still appear in search results, because the index still contains information about that URL. This definitely applies to Google, and may apply to other search engines as well. Only Google provides enough public information about its processes for us to be sure.

Even with Google, it's unclear whether such hits are stored in the main indexes, or in a separate index of link-related hits. There are indications, for example, that the ordering and proximity of words in anchor text is a factor in determining how much the link text affects a page's overall ranking for a given search query.

## Results of the Indexing Phase

At the end of the indexing phase, the search engine is capable of returning the indexed URL in search results. If the search engine makes heavy use of link analysis in its ranking algorithm, the URL may not rank well for competitive search terms, but it is at least visible in the search results.

*note*

We conclude our discussion of this phase with another example from Google. As mentioned in our examination of the crawling phase, Google's `site:domain` search shows all known URLs for a domain, including those which are not yet crawled and indexed. In order to find all of the indexed pages from a domain, a different approach is required. Most sites will have some signature text which appears on all pages, such as the copyright notice etc.

By combining the `site:domain` search with this signature text, it's possible to get a true measure of a site's index saturation, or the number of pages from the site which have been indexed. For example, the query `site:example.com copyright` will return all of the indexed pages from `example.com`, assuming that the work 'copyright' appears on all pages.

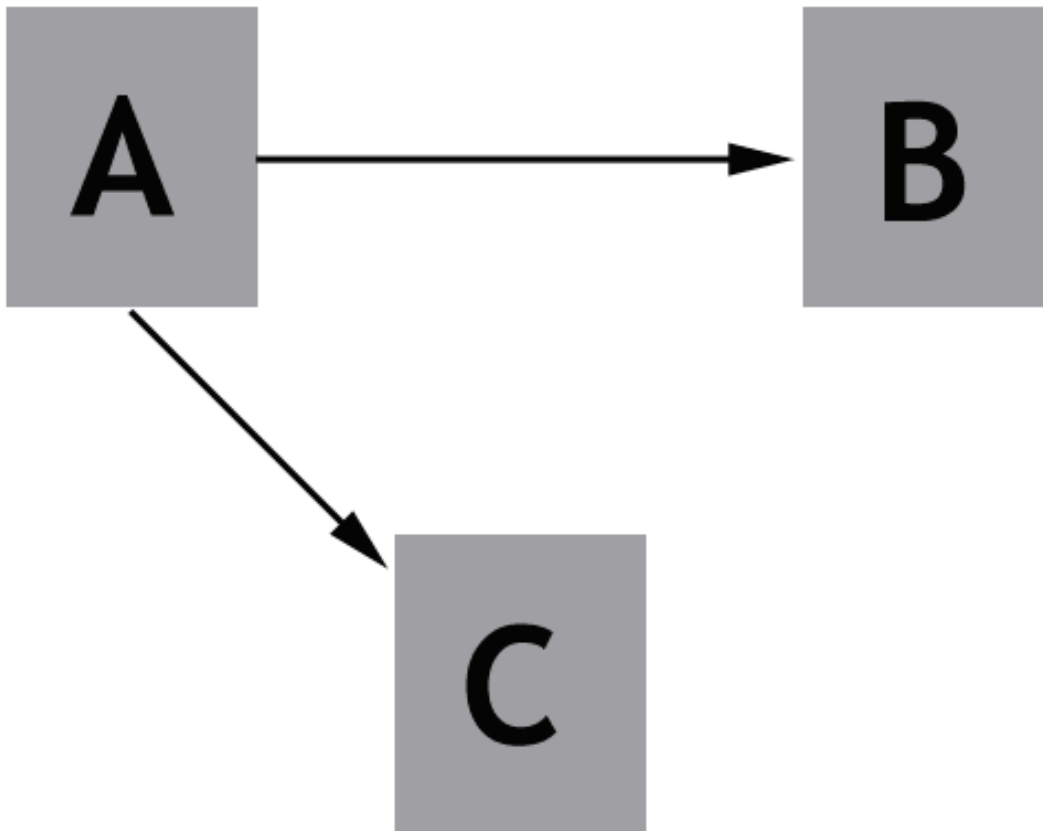
## Link Analysis

The content of documents is susceptible to manipulation or optimization, so there may be a large number of Web pages that appear, at first glance, to be relevant to given keywords. Indeed, there may be millions of pages that match the searcher's query to some degree. Conversely, some highly relevant pages may not be optimized. As a result, search engines can't simply rely on the content of documents as a means by which to assess them—to do so would prevent the engines from showing the best possible search results.

There are many ways in which search engines can derive information from the linking relationships between pages, and each search engine takes a different approach to doing so. In this brief summary, we'll see how links can imply topical relationships and help

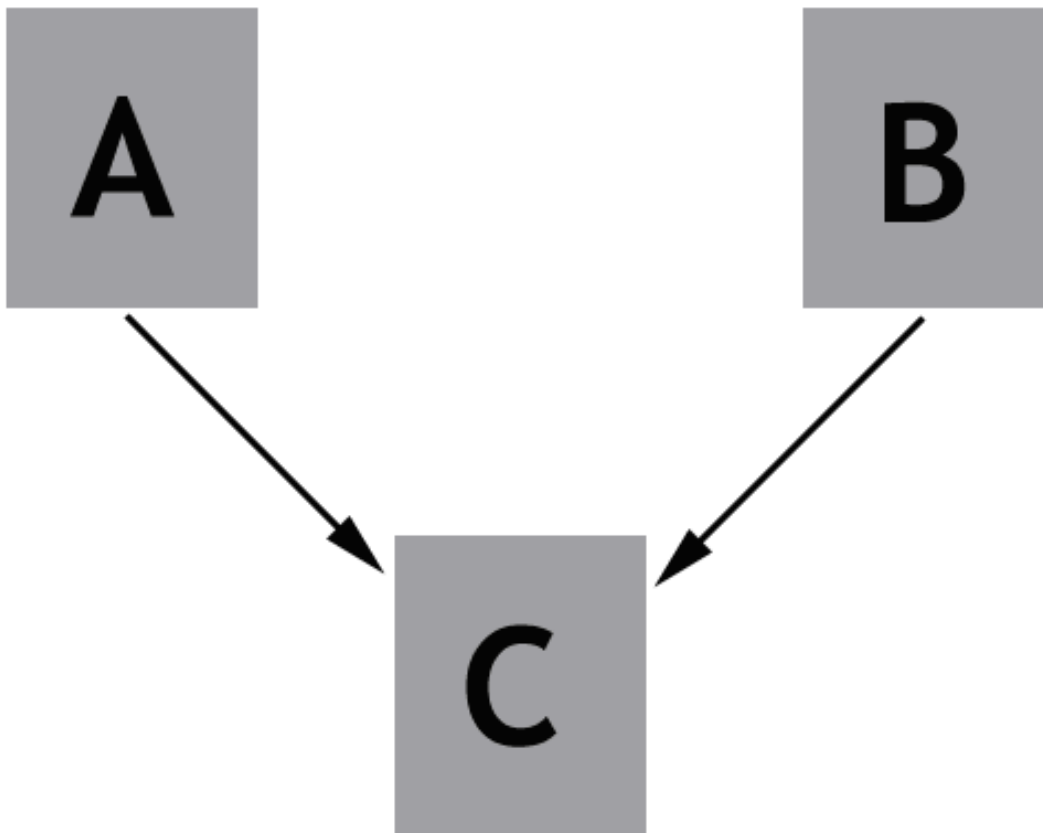
search engines find the Web pages that are most important to their index. To start off, let's look at a couple very small **Web graphs** to understand how a search engine might interpret the linking relationships between Web pages. A Web graph is simply a diagram of the linking relationships between a group of pages.

**Figure 1.2. Simple Web graph: one page “votes” for two others.**



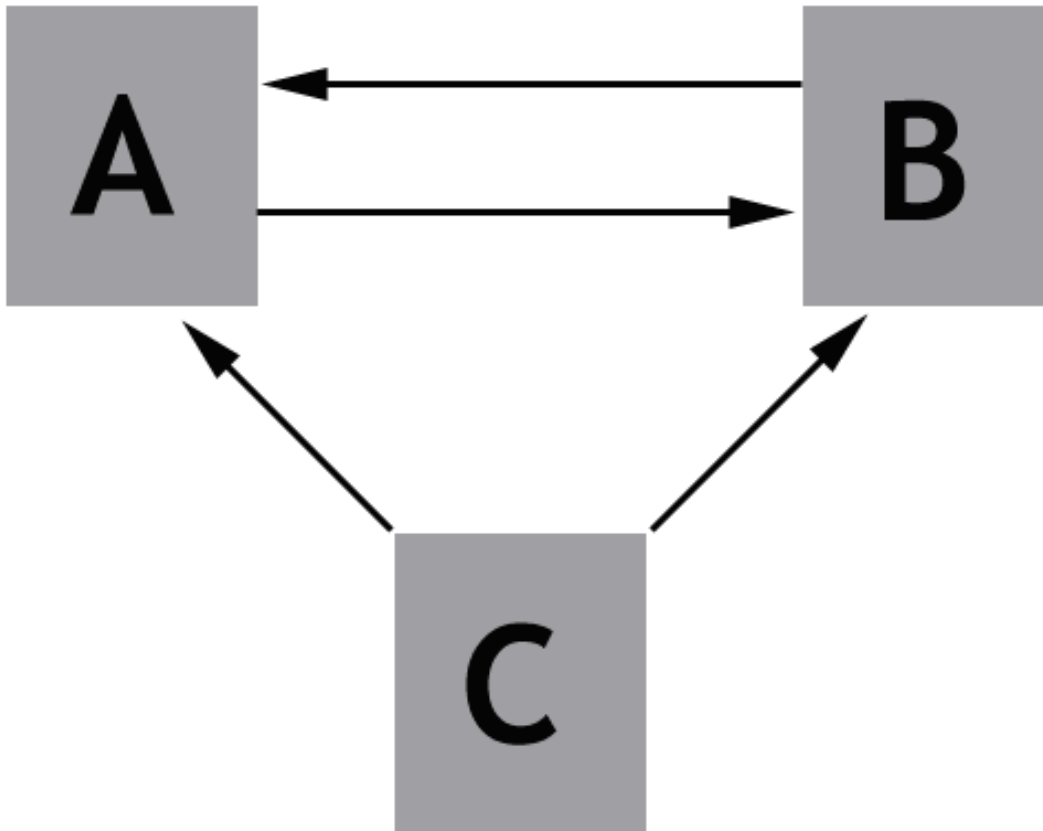
In Figure 1.2, we see three pages (A, B, and C). A links to both B and C. This implies that the content of B and C may relate to the topic discussed on page A. It also implies that the author of page A considers B and C useful—in effect, the author of A is “voting” for B and C.

**Figure 1.3. Simple Web graph: two pages “vote” for one page.**



In Figure 1.3, both A and B link to page C. This implies that pages A and B cover the same topic as page C, and that the authors of A and B are voting for page C. The implied topical relationship between A and B is as strong, or stronger, than in the previous example, because two pages that are found to link to the same resource are likely related to the same topic.

**Figure 1.4. Simple Web graph: two pages interlinked; a third links to both.**



In Figure 1.4, we see a more complex relationship. Pages A and B link to each other, implying that they may address the same topic. The existence of links from C to both A and B validates this interlinking and provides a strong indication that pages A and B mention the same topic.

## Hubs and Authorities

The first really interesting attempt to harness the linking relationships between pages was the HITS (Hypertext-Induced Topic Selection) algorithm developed by Jon Kleinberg at Cornell University. Kleinberg's great revelation was that communities on the Web tend to cluster around specific **hubs** and **authorities**.

A hub is a page that contains links to many other pages on the same topic. A good example of a hub would be a page from Yahoo! or the Open Directory which contained a list of pages about a single topic. An authority is a page to which many other pages

link. A page that's listed within the appropriate category on many directories, or is otherwise well-linked within the community of related pages, is considered an authority.

Though the concept is simple, its practical implementation involves many nuances. Many attempts have been made to improve on the basic idea of HITS and, no doubt, some of these ideas are in use within search engines today. One such idea advocates that more focused hub pages are those that link to specific pages, rather than the homepage of every site, thereby implying that more editorial effort went into creating the hub page.

## Google's PageRank

Because of the dominant role Google plays in today's search engine landscape, and because of the incredible amount of insight it allows into its inner workings, the PageRank algorithm that this search engine uses has taken on an almost mythical status among SEO practitioners.

Many *very* detailed explanations of PageRank are available on the Web. The original paper written by Google founders Larry Page (the "Page" in PageRank) and Sergey Brin is called "The PageRank Citation: Bringing Order to the Web[3]," and is available in multiple formats online.

In addition to the many papers published by Stanford and Google researchers, numerous competing (and occasionally conflicting) accounts have been prepared by SEO consultants. I list many authoritative papers and provide an explanation of PageRank in Appendix A; for now, let's briefly discuss how PageRank works without getting bogged in mathematical detail.

The concept of PageRank is very similar to the "wandering drunk" algorithm employed in many areas of computer science, and to the proverbial thousands of monkeys that eventually type long enough to reproduce Shakespeare's *Hamlet*.

To understand this concept, let's consider a random Web surfer. We'll make him male, and call him Bob. To get Bob started, we'll sit him down with the browser open at a Web page that's selected at random from our index. If there are a million pages in our index, there's a one-in-a-million chance that Bob starts at any particular one of them.

Bob's job is to pick a random link on every page he visits, and continue on to wherever that link sends him. On each page Bob visits, including the first, there's a chance that he'll get bored and ask for a different random Web page. So, when he gets bored, we select another page completely at random, and the process starts again. In Page and Brin's paper, there is a 15% chance that Bob will get bored on any page.

---

[3] <http://newdbpubs.stanford.edu:8090/pub/1999-66>

If we let Bob surf the pages in our index in this way for a decade or so, he will eventually visit every page. Once Bob has viewed every page at least once, we can count the number of times he's visited each page in the index. The pages that Bob has visited the most times will be allotted the highest PageRank. To put this another way, the PageRank score of a given Web page is an *estimate* of the probability that a random surfer would find that page if they followed the process that Bob followed.

Now, if Bob starts surfing from a page with a very high PageRank, we can assume that any page it links to will have a high probability of being found by Bob. As such, you might think that links from a page with a high PageRank would be the most valuable.

However, this is not the case. The more links there are on a page, the less likely it is that any particular link will be chosen by a random surfer. This leads us to PageRank Truth #1:

The value of any link from a Web page is decreased proportionally for every additional link on that page.

This is why links from a pure directory like the Open Directory may actually be more valuable than links from a Web portal that happens to include a directory. In a pure directory, nearly all the PageRank attributed to the homepage flows through to the category listings.

By comparison, of the vast number of links that appear on each page of the Yahoo! site, only a certain percentage link to directory pages. Over 200 links appear on the Yahoo! homepage, most of which lead away from the directory. Even the directory pages themselves display many listings and other links.

Likewise, links from a highly selective directory are likely to be worth more than a less selective directory of equal size, because there will be fewer links (or listings) on each of the category pages.

The same logic applies to all links. If you are interested in maximizing the PageRank of the pages on your site, simply looking for high PageRank pages may not be the best approach. Link placement (i.e. which page carries the link) matters much more than the average site owner realizes. We'll have much more to say on PageRank and linking strategies later in this kit. But for now, let's turn our attention to the final piece of the link analysis puzzle.

## Topics and Communities

Search engines know that many hyperlinks exist solely for the purpose of boosting the perceived popularity of a site in order to improve its ranking in the search results. PageRank is susceptible to this sort of manipulation, as are older link analysis schemes based on link **popularity** (a simple count of the number of links to a particular page or site).

The cutting edge of link analysis, therefore, goes into a deeper exploration of the topical relationships between Web pages. The Teoma search engine, for example, sees the Web as a set of topical **communities**, and looks for the most relevant and authoritative pages within a topic. This was actually the basis of the HITS research, but HITS involved the manual selection of pages. Any sort of practical implementation of a topically-driven link analysis scheme requires some sort of automated method of performing “topic distillation” on a given Web page.

Appendix A contains many references that deal with topic distillation and link analysis, including a topic-sensitive variation on PageRank, and a scheme referred to as LocalRank. LocalRank involves the calculation of an internal PageRank score within the set of pages returned by a Web search, so that the results can be rearranged on the basis of topical authority.

Link analysis and topic distillation are fascinating topics. What they mean to SEO consultants will be explained in greater detail in the next chapter.

## How Queries Are Processed

Today’s search engines handle many different types of search queries. Searches run the gamut from FedEx package tracking, phone numbers, and dictionary definition lookups, to the plain old text-based searches that started it all.

Before retrieving search results, the search engine needs to interpret the user’s query. Such an interpretation necessitates the extraction of any special syntax or search options that the user has invoked, such as a site-specific search, reverse phone directory lookup, or other options. Assuming that a text-based search is required (and it’s not something easier, like a FedEx number), the processes that take place will differ depending on the search engine being used.

All the major search engines make some attempt to interpret the searcher’s intent as indicated by the search terms he or she entered. Certain words, or types of words, may give clues that can help the search engine deliver the most satisfactory results. To give you a simple example, a person who searches for brake repair is probably looking for information. In this case, a mix of informational and how-to sites, along with a couple of nationwide chains, would be a good set of search results for most users. By comparison, someone who searches for brake repair Chicago obviously has one thing in mind: finding someone to fix their brakes. Once the searcher has made his or her intent clear like this, it’s much easier for the search engine to help them find what they’re looking for.

What should be clear to you, though, is that different search terms are interpreted in different ways by different search engines. Though semantic analysis (the art of trying to determine the meaning of the words used in a search) is just coming into its own, already it plays a significant role in the search process. It’s likely that at least some ele-

ments of the algorithms used by many search engines to determine search results and rankings are already modified to address the type of search query that's performed.

Search engines also have a set of so-called stop words—extremely common words like a, and, and the. Although these words are supposed to be ignored by the search engine, they do have an influence on search results. A search for watching and waiting on almost any search engine will return different results than a search for watching waiting, even though the search engines claim to ignore the word and. It's likely that search engines replace the stop word with a wildcard, so that any word positioned between watching and waiting could pass for a matching phrase.

We will discuss how people search, how different types of searches indicate different buying modes, and what it all means to your strategy, in the next chapter.

## Information Retrieval (IR) Theory

The godfather of text-based information retrieval was Gerald Salton (1927–1995), a professor at Cornell University. Salton's group at Cornell developed the SMART information retrieval system. This system pioneered the **vector space model** that's now used in some form or other by all crawling search engines.

The vector space model is conceptually simple: take the contents of a set of documents, create an index of every word occurrence in every document, and combine this information to create a mathematical representation of each document within a multi-dimensional vector space. Once you've done that, all you need to do is create a vector that represents your search query, and present the documents that are closest to it in the vector space.

Okay... maybe it's not so simple after all!

If you think the vector space model sounds complicated, you're not alone. And I didn't even explain how it takes the dot product of the magnitudes of the query and document vectors, calculates the cosine of the angle between them, and compares the cosines of the dot products of the query vector and the different document vectors, in order to find the most relevant documents for the query!

The good news is that an elaborate mathematical explanation of this concept is not necessarily important to search engine marketing. That said, there are a few things you need to understand about the field of information retrieval theory and the vector space model:

- ❑ Words that appear more frequently within the collection of documents being searched are seen as less important in the retrieval of documents. Conversely, words that appear less frequently within the collection of documents are deemed more important in the retrieval of documents. So, if you search for defenestration policy guidelines, the

less common word defenestration will have a greater influence on the retrieval process.

- ❑ The proximity and ordering of words within a document is significant. If you searched for red monkey shoes, a document containing that exact phrase would be considered more relevant than a document that contained the words in a different sequence or in close proximity, either of which would, in turn, be seen as more relevant than a document that merely contained all three words.
- ❑ Search engines make use of the structural and presentational elements of hypertext. Words that appear within key structural elements (page title, headings, hyperlinks etc.), to which significant formatting has been applied (bold, italic, large type), or which appear near the top of the document, are given more relevance than words that appear elsewhere within that document. In other words, occurrences of a given word can be weighted differently depending on where and how they appear in a document.

## The Vector Space Model in Action

Just in case you're interested in digging deeper into the vector space, I thought I'd take a moment to explore it in more detail here.

As we've already seen, the search engine is not looking at documents; it's using an inverted index that maps *words* to their specific *appearances* within indexed documents. This is important, so do whatever you have to do to lock this concept into your brain.

When you perform a search, the ranking/retrieval component of the search engine constructs a set of vectors for matching documents within the index, and a separate vector for the search query itself. Don't get hung up on the term vector—a vector is just a collection of variables that relate to a specific item.

Each occurrence of a word within a document is weighted differently depending on a number of factors, including where the word occurs (e.g. is it a heading or bold text?). If these considerations are factored into the indexing process, the search engine may save some time during the search, but it will also forfeit the flexibility to change elements' weights depending on the specific search query.

The weighted occurrences of terms within a document are combined with the overall frequency with which the word occurs in the total collection of documents, and within the document itself, to produce a set of “term weights” for the document. The collection of term weights for all the words in the query represents that document's vector.

The query vector is an idealized set of term weights for the words in the query. To find the most relevant documents in its index, the search engine applies a little math that identifies the closest document matches in vector space. I won't even attempt to explain the math this involves—see Dr Garcia's Website [4] for a detailed explanation.

[4] <http://www.miislita.com/term-vector/term-vector-1.html>

## Ranking and Retrieval Strategies

At some point, the search engine has to return results to the searcher. All the work that's lead up to this point has given the search engine a certain understanding of the user's query, and a great deal of information about the contents of each page.

There's one misconception about search result delivery that I should clear up now. A typical SERP will include some message like, "Results 1-10 of 843,000." What that means is that there were a total of approximately 843,000 pages that *might* have been relevant to the query. Search engines don't really examine all 843,000 in order to deliver search results.

In reality, none of the major search engines delivers more than 1,000 total matches, presumably because users would get tired before they actually clicked through to the hundredth page of search results. And, because search engines don't have to deliver more than 1,000 matches, they can use a pre-selection process to winnow the 843,000 candidates down to a smaller number of returned results. A page that isn't among the top 1,000 results for any specific factor in the selection process has no chance whatsoever of appearing in the search results.

The factors involved in that selection process are very dependent on the search query itself, and this is one of the instances in which less common words (such as *defenestration*) are likely to have a greater influence than commonly used words (like *free* or *cheap*).

Once the pre-selection is made, the search engine applies its ranking algorithm to the pages that made the cut, and presents a selection to the user as search results.

## Query-Dependent Ranking Strategies

The type of search query that's entered can affect the way in which the search engine approaches the problem of delivering results. Every search engine has its own unique algorithm, but the results will always come from some combination of on-page factors, including content, formatting and structure, and off-page factors, such as link analysis and topic distillation.

For shorter, more generic queries, the initial result set will be very large, and there's a very good chance that the vector space model based on page content will fail to deliver satisfactory results. With such queries, ranking factors based on link analysis must play a significant role in the pre-selection process, and in determining the final results.

For longer, more specific queries, the initial result set will be smaller, so content-based strategies such as the vector space model may be more appropriate. This doesn't mean that all search engines will treat these queries differently, but it's a definite possibility.

## Does the Topic Come into Play?

Although topical factors are definitely put to use by some search engines (notably Teoma), the extent of their impact is unknown. One of the difficulties search engines face in applying topic distillation and topical link analysis is that these algorithms need an idea of the query's topic in order to work.

For very long search queries, this is probably fairly easy to determine, but the vector space model already performs very well with long queries. In this case, a topical algorithm may be overkill, and may even lead to results that are less relevant to the specific query than they are to a related topic.

For very short queries, with which the vector space model needs help, it can be very difficult to determine an appropriate topic. My favorite example of this type of query is barber shop. This might be a place to get a haircut, but it's also the title of a popular comedy film series, and a form of a capella music performed by four men wearing striped shirts.

Because the importance of topical algorithms is uncertain, much of our discussion of topical factors may be less than completely relevant today. However, search engine optimization is very much a long-term game, and the trend towards topic distillation and topical link analysis is too strong to ignore.

Call it future-proofing if you like, but throughout this kit I will encourage you to adopt topical strategies in your content, copywriting, and link strategies.

## Other Considerations

In addition to the basic processes we've just described, which are responsible for creating search results and rankings, search engines have a number of parallel activities that are worth a quick look.

Nearly all search engines undertake some form of automated monitoring of search quality. Google and Yahoo! both use a tracking link that allows them to know which of the search listings has been clicked. If a highly popular search phrase does not generate an acceptable number of clicks for the top-ranked pages, the search engine may consider adjusting its algorithm to deliver more satisfactory search results. Note, though, that this is not the same as Direct Hit technology, in which user clicks directly influence search results.

The search quality team, in addition to seeking out areas for improvement in search results, is responsible for the review and removal of sites that attempt to "spam" or deceive the search engine. For the most part, search engines can't act on individual spam reports, but must instead identify the techniques being used by spammers. When a new

technique is identified, the software engineers attempt to find an automated means of detecting and filtering that form of spam from the search results.

Judging from the countless panic-stricken site owners who post every day to online discussion forums, many folks assume that a “spam penalty” is the most common reason for a site or page to drop out of the search results. In fact, such penalties are extremely uncommon and, in most cases, the explanation is far more mundane.

Any Website operator whose pages suddenly disappear from a search engine will be upset, but the problem almost always lies on their side of the equation. As mentioned above, not all Websites are ready to answer when the spider calls, and this can cause the search engines to stop crawling for a time. Server and DNS errors are the most common reasons for pages to be removed from a search engine’s index.

In addition to the automated processes that may remove pages, search engines must comply with copyright laws such as the Digital Millennium Copyright Act (DMCA), which applies in the US. DMCA notifications represent a significant burden for search engines and other online service providers.

Unlike Web hosting companies, search engines usually do not have the means to contact the site owner quickly in order to allow for an appeal, and pages may be removed on the grounds of copyright infringement without the site owner ever becoming aware of the action. When Google deletes a page from its index for this reason, a link may be displayed on some search results to indicate that the page has been removed.

All of the major search engines must deal with a high volume of search requests from users, as well as peak hour demands that can easily be five times higher than their quiet times. In order to deliver search results in a hurry, no matter what the traffic levels, today’s search engines use load balancing techniques across multiple data centers located strategically around the world.

It’s not easy to run a search engine, and that fact goes a long way toward explaining why there are so few of them in the market today.

## What Search Engines Want

One of the keys to developing a long term search engine strategy for any Website is an understanding of what search engines want. Repeat visitors drive revenue, so the major goal of any search engine is to keep its users happy. A few lessons from the past can help shed a little light on the future.

Freshness is critical. The once mighty AltaVista search engine (now a Yahoo! property) lost users to Google and the Inktomi-driven search portals for one major reason: it stopped crawling the Web aggressively, and failed to update the index on a regular basis. The lack of freshness in the AltaVista search results meant that many SERPs contained

mostly broken links. Don't expect any search engine operating today to make the same mistake.

## Snapshot of the Search Market

Search engine usage varies substantially from country to country. Here's a snapshot of the top ten search engines in various countries for a given week in December, 2004. The data was provided by Hitwise[5] and should prove valuable to an SEM practitioner who wishes to target campaigns to any of these areas.

The most popular search engines visited by US Internet users for the week ending 12/18/04 are shown in Table 1.2.

**Table 1.2. Popular Search Engines: USA**

Rank	Name	URL
1	Google	<a href="http://www.google.com">http://www.google.com</a>
2	Yahoo! Search	<a href="http://search.yahoo.com">http://search.yahoo.com</a>
3	MSN Search	<a href="http://search.msn.com">http://search.msn.com</a>
4	Google Image Search	<a href="http://images.google.com">http://images.google.com</a>
5	Ask.com	<a href="http://www.ask.com">http://www.ask.com</a>
6	iWon	<a href="http://www.iwon.com">http://www.iwon.com</a>
7	My Web Search	<a href="http://www.mywebsearch.com">http://www.mywebsearch.com</a>
8	Yahoo! Image Search	<a href="http://images.search.yahoo.com">http://images.search.yahoo.com</a>
9	Dogpile	<a href="http://www.dogpile.com">http://www.dogpile.com</a>
10	AltaVista	<a href="http://www.altavista.com">http://www.altavista.com</a>

The most popular search engines visited by UK Internet users during the week ending 12/18/04 are displayed in Table 1.3.

[5] <http://www.hitwise.com/>

**Table 1.3. Popular Search Engines: UK**

Rank	Name	URL
1	Google UK	<a href="http://www.google.co.uk">http://www.google.co.uk</a>
2	MSN UK Search	<a href="http://search.msn.co.uk">http://search.msn.co.uk</a>
3	Ask UK	<a href="http://www.ask.co.uk">http://www.ask.co.uk</a>
4	Google	<a href="http://www.google.com">http://www.google.com</a>
5	Google UK Image Search	<a href="http://images.google.co.uk">http://images.google.co.uk</a>
6	Yahoo! UK Search	<a href="http://uk.search.yahoo.com">http://uk.search.yahoo.com</a>
7	MSN Search	<a href="http://search.msn.com">http://search.msn.com</a>
8	Yahoo! Search	<a href="http://search.yahoo.com">http://search.yahoo.com</a>
9	Wanadoo UK Search	<a href="http://search.wanadoo.co.uk">http://search.wanadoo.co.uk</a>
10	Tiscali UK	<a href="http://www.tiscali.co.uk">http://www.tiscali.co.uk</a>

The most popular search engines visited by Australian Internet users in the week ending 12/18/04 are presented in Table 1.4.

**Table 1.4. Popular Search Engines: Australia**

Rank	Name	URL
1	Google Australia	<a href="http://www.google.com.au">http://www.google.com.au</a>
2	Google	<a href="http://www.google.com">http://www.google.com</a>
3	NineMSN Search	<a href="http://search.ninemsn.com.au">http://search.ninemsn.com.au</a>
4	Yahoo! Australia Search	<a href="http://au.search.yahoo.com">http://au.search.yahoo.com</a>
5	Yahoo! Search	<a href="http://search.yahoo.com">http://search.yahoo.com</a>
5	Google Australia Image Search	<a href="http://images.google.com.au">http://images.google.com.au</a>
6	AltaVista	<a href="http://www.altavista.com">http://www.altavista.com</a>
7	My Web Search	<a href="http://www.mywebsearch.com">http://www.mywebsearch.com</a>
8	AltaVista Australia	<a href="http://au.altavista.com">http://au.altavista.com</a>
9	Answers	<a href="http://www.answers.com.au">http://www.answers.com.au</a>
10	Google Image Search	<a href="http://images.google.com">http://images.google.com</a>

The most popular search engines visited by New Zealand Internet users in that same week are shown in Table 1.5.

**Table 1.5. Popular Search Engines: New Zealand**

Rank	Name	URL
1	Google NZ	<a href="http://www.google.co.nz">http://www.google.co.nz</a>
2	Google	<a href="http://www.google.com">http://www.google.com</a>
3	MSN Search	<a href="http://search.msn.com">http://search.msn.com</a>
4	XtraMSN Search	<a href="http://search.xtramsn.co.nz">http://search.xtramsn.co.nz</a>
5	Yahoo! Search	<a href="http://search.yahoo.com">http://search.yahoo.com</a>
6	Google NZ Image Search	<a href="http://images.google.co.nz">http://images.google.co.nz</a>
7	Sina China	<a href="http://www.sina.com.cn">http://www.sina.com.cn</a>
8	Daum Search	<a href="http://www.daum.net">http://www.daum.net</a>
9	Yahoo! Australia Search	<a href="http://au.search.yahoo.com">http://au.search.yahoo.com</a>
10	AltaVista	<a href="http://www.altavista.com">http://www.altavista.com</a>

The most popular search engines visited by Hong Kong Internet users for the week ending 12/18/04 are as show in Table 1.6.

**Table 1.6. Popular Search Engines: Hong Kong**

Rank	Name	URL
1	Yahoo! Hong Kong Search	<a href="http://search.hk.yahoo.com">http://search.hk.yahoo.com</a>
2	Google Hong Kong	<a href="http://www.google.com.hk">http://www.google.com.hk</a>
3	Google	<a href="http://www.google.com">http://www.google.com</a>
4	Yahoo! Australia Search	<a href="http://au.search.yahoo.com">http://au.search.yahoo.com</a>
5	Sina Hong Kong	<a href="http://www.sina.com.hk">http://www.sina.com.hk</a>
6	MSN Search	<a href="http://search.msn.com">http://search.msn.com</a>
7	Yahoo! Search	<a href="http://search.yahoo.com">http://search.yahoo.com</a>
8	Google HK Image Search	<a href="http://images.google.com.hk">http://images.google.com.hk</a>
9	Baidu	<a href="http://www.baidu.com">http://www.baidu.com</a>
10	MSN Search Hong Kong	<a href="http://search.msn.com.hk">http://search.msn.com.hk</a>

The most popular search engines visited by Internet users in Singapore in the week ending 12/18/04 were as shown in Table 1.7 below.

**Table 1.7. Popular Search Engines: Singapore**

Rank	Name	URL
1	Yahoo! Search	http://search.yahoo.com
2	Google Singapore	http://www.google.com.sg
3	Yahoo! Singapore Search	http://sg.search.yahoo.com
4	Google	http://www.google.com
5	MSN Search	http://search.msn.com
6	Seeq	http://www.seeq.com
7	Yahoo! Image Search	http://images.search.yahoo.com
8	MSN Search Singapore	http://search.msn.com.sg
9	Google Singapore Image Search	http://images.google.com.sg
10	Baidu	http://www.baidu.com

## The Future of Search

While search engines today have, without a doubt, reached the point at which they're exceptionally useful, there's still a lot of work to be done before users can trust them for every type of search. Additionally, in terms of search success, a great deal still depends on user knowledge of individual systems.

Anticipated improvements in search technology lie in three major areas: localization, context, and the so-called Semantic Web. The first two deal primarily with gaining a better understanding of users' needs; the third offers opportunities to improve the efficiency of indexing and the quality of search results.

## Localization

Many search engines offer local search options, whether they're explicit, such as Google's Local Search, currently in beta, or implicit. For instance, the addition of a zip/postal code to any search on Google or Yahoo! brings local search elements into action. Unfortunately, users aren't used to inputting a zip code when they search.

One of the most famous cities in the world is Springfield, the setting for the popular Simpsons cartoon series. It's sort of a running joke in the series that nobody knows which Springfield they're talking about—there are numerous towns of that name in the United States—although, in reality, the Simpsons live in a purely fictional city. This reflects a major challenge that's inherent in implicit local searches: many cities have the same name.

If an implicit local search isn't possible (as would be the case for a search on real estate Springfield), search engines can prompt users to clarify which city they're looking for—provided it can be identified as a city. But, what if someone searches for marketing wisdom? Do you know whether they want marketing advice, or marketing firms in Wisdom, Montana?

Part of the solution lies in the emerging geo-targeting technologies that attempt to identify users' locations on the basis of their IP addresses. This technology is already used by Google AdWords and Overture to let PPC advertisers target specific locations and exclude others.

Geo-targeting may be part of the overall localization solution, but the current methods suffer two problems that are not so easy to overcome. The first is the use of proxy servers by some Web users, especially those on corporate networks. The firewall through which these users access the Web may be housed a long way from their physical location. This becomes a significant issue with geo-targeting in PPC advertising campaigns, particularly in major business to business markets.

The second problem is that not all local searches are intended for the user's current location. If I'm sitting at my desk in Frisco, Texas, and I search for Cleveland hotels, a geo-targeting system might think that I'm looking for hotels in Cleveland, Texas instead of Cleveland, Ohio.

Given the popularity of the browser toolbar additions distributed by all the major search engines, it's possible that some user profile information may eventually be incorporated into the delivery of search results. That will take care of where the user is coming from. An even bigger problem is figuring out the geographic scope of a Website. The world headquarters of Exxon-Mobil, for example, are in Dallas, Texas, but someone searching for Exxon in Chicago just might be looking for a gas station closer to home.

To determine which country a Website belongs to can be tricky, too. For search engines like Google that offer country-specific search options, this is a very real problem. It's easy to include all sites that use the appropriate country's top level domain (such as .au for Australia), but many sites around the world use .com, .net, .biz, .info and other global TLDs. SitePoint.com, for example, is based in Australia, but serves a global audience. Google's solution is to use IP-based lookups to determine where the DNS servers for the domain are located, but this has caused more than a few problems, and it's basically a "better than nothing" solution.

Despite all these challenges, we can expect search engines to move forward with efforts to localize their search offerings. Localization holds the promise of more relevant results (which helps keep users loyal), and offers substantial profit gains for those using PPC advertising programs.

## Context and Personalization

If there's one thing that's destined to shatter the old illusions about search engine optimization, it's the coming rise of context and personalization. Google and Microsoft have already shown personalized search offerings, and the results are simply too compelling to ignore.

Google's Personalized Search[67] system invites the user to select one or more topics from a list, then returns search results that are slightly skewed toward those topics. Depending on the mix of topics selected, it's possible to produce very different results. This system shows how far Google has progressed in its implementation of topic distillation.

Microsoft's system isn't a product yet—it's still a research project—but it provides some very useful insights into the ways that the company's offerings may differ from those of other major players. Called "Stuff I've Seen," it's designed to confine a user's search to Websites or documents that they've already seen.

When you put these two ideas together—as someone certainly will—you get a search that's different for each user, depending on their browsing and searching history. The search toolbars that the major search engines offer as browser plug-ins are already very popular; all they need in order to extend these services is users' permission to gather and use their browse and search data.

Not everyone will climb aboard the personalized search bandwagon (privacy concerns will need to be allayed), but you can expect the search engines to make an effort to promote this kind of service in coming years. My bet is that the majority of users will be happy to share a little semi-anonymous information if it helps them find what they're looking for.

Search results for many queries already differ from one location to the next on some search engines. Given the compelling opportunities created by personalized search, will search engine rankings as we know them today even exist in 2006?

## Structure and the Semantic Web

Tim Berners-Lee, inventor of the World Wide Web, has a plan for the future of the Web, or at least, the next version of it. The Semantic Web is built on top of the existing Internet structure, in a sense, in that it doesn't really involve eradicating the old Web. The Semantic Web is all about creating structure by providing a great deal more data about the different resources available online.

---

[67] <http://labs.google.com/personalized>

As I mentioned earlier in this chapter, some search engines already accept XML-based trusted feeds from some Websites, and/or read RDF Site Summary (RSS) files when they're available. Yahoo! has been particularly active in this area.

RSS, and Google-endorsed alternative Atom, are designed to let site owners provide more detailed and structured data about the resources available on their sites. In current practice, this is used mainly to provide news feeds for blogs, news sites, and similar content portals. Other sites can process these news feeds and make headlines and other information available to their visitors.

The Semantic Web is just beginning to take hold, but as more sites offer summaries and structured data, it is inevitable that search engines will find ways to make use of this new information source. It's not going to change the world tomorrow, but it's definitely the shape of things to come.

## Summary

The main goal in this chapter was to provide a detailed picture of how search engines work, and where search results come from. To keep you from falling asleep, I've mixed in a few little tidbits and insights that you can use right away.

*note*

Note to search engine marketing consultants: as you begin to think of yourself as a professional SEO or SEM consultant (and maybe you already do), I'm sure you can see how this kind of knowledge will help you position yourself in the eyes of clients. People trust experts, and you will certainly be seen as one if you can answer questions in detail, and clear up your client's own misconceptions about search engines.

The next chapter will dig more deeply into the process of search engine optimization, and introduce you to many of the practical aspects of what you've just learned.

## What's Next?

If you've enjoyed this chapter from *The Search Engine Marketing Kit*, why not order yourself a copy?

In the rest of the Kit, you'll get much more search engine insight, strategies, and tips from author Dan Thies. The Kit also contains comprehensive information on using pay-per-click advertising effectively, combining your SEO and PPC strategies, and even running your own SEM business.

Buy the Kit today and

- Learn best-practice strategies to maximize traffic
- Discover keyword strategies you won't find anywhere else
- Find out the best ways to optimize pages and build links
- Avoid getting banned from search engines
- Work around obstacles such as Flash, Content Management Systems, and server issues
- Create, optimize, and manage advanced pay-per-click campaigns
- Learn to sell professional search engine marketing services

The Kit also includes a CD-ROM packed with tools and documents for use in your own search engine marketing efforts or as part of your search engine marketing business.

For a limited time only, buyers of *The Search Engine Marketing Kit* will receive \$150 worth of free PPC advertising credit with Google Adwords, Overture, and Findwhat.

[Order now and get it delivered to your doorstep—  
for a limited time with FREE shipping!](#)

---

# Index

## Symbols

<meta> tags, as spam location, 240

## A

A/B/C testing, 144

accessibility techniques and spam, 239

accountants, ethics of lawyers and, 80

acquisitions, 4

multiple domain names from, 115

Adobe Acrobat, 19

Advanced Web Ranking tool, 231

advertising

(*see also* contextual advertising; PPC campaigns; targeted advertising)

campaign tracking tool, 230

multi-lingual advertising, 128

offer-based advertising, 140

paid advertising on Google, 4

PPC targeting of agencies, 154

testing PPC ad copy, 142

Adware and PPC suppliers, 129

AdWordAccelerator tool, 50

AdWords (*see* Google AdWords)

affiliate programs, minor PPC providers, 129

agent-based delivery, 243

agent-based spam, 244

distinguished from IP cloaking, 246

Alliance Link, 233

AllTheWeb search engine, 4

alt text misuse, 68, 88

intended use and, 239

AltaVista search engine, 2

acquisition by Overture and Yahoo!, 4  
decline of, 33

PRISMA tool, 48

analysis of PPC measurements, 150

analysis tools and techniques, 131

anchor text

link building and, 69

page ranking and, 22

search terms within, 53, 65

AOL user interface, 5

Apple Pie Cart, 101

Architext project, 2

archiving and the noarchive directive, 107

Arelis link tracking tool, 224

Ask search engine, 3–4

(*see also* Teoma search engine)

assessments

business planning stage, 187

consultative selling preliminaries, 175

current situations, 156

as introductory services, 184

“selling the assessment”, 168

Atlas OnePoint tool, 226

Atom news feeds, 40

audience segments (*see* target audiences)

Australia, search engine popularity, 35

authorities and hubs

link analysis and, 25

link popularity and spam, 241

## B

Backrub, precursor of Google, 3

“bad” sites, 74

barber shops, 32

bCentral, Microsoft

Microsoft Small Business Directory, 71–72, 75

Submit It service, 52, 75

Beal, Andy

flat fee pricing policy, 181

interview with, 197–201

benchmarking, 143

Berners-Lee, Tim, and the Semantic Web, 39

best practice SEO, 79

professionalism and, 193

Bid For Position model, PPC, 124

bid management tool effects, 49

bidding for contracts, 173

BidRank and BidRank Plus tools, 227

---

- billing and payment terms, 183
  - blogs, 251
  - Bob, the random Web surfer, 26
  - body copy
    - optimizing for length, 204
    - search terms in, 65
  - <body> element, HTML
    - content spam location, 238
    - user-agent spam location, 245
  - bookmarking and frames, 87
  - books, recommended, 247
  - bounce rate measurement, 131, 147
  - brainstorming for keyword discovery, 47
  - brand names as keywords, 45, 59
  - Brin, Sergey, of Google and PageRank, 26, 235
  - broad matching, 133
    - campaign refinement using, 135
    - John Slade on, 212
    - large retailers and, 153
    - ranking of bids, 123
  - browsers
    - browser detection, 78
    - history, 2
    - Mozilla Firefox features, 225
    - toolbar extensions, 38–39, 54, 226, 234
  - budgets
    - changing, and scope of work, 182
    - establishing that sales prospects can pay, 168
    - PPC campaign options, 135
    - self service pricing model, 183
  - building a business, 155–165
  - business directories, 10, 71
  - business fit, 170
  - business knowledge in client areas, 169
  - business operation, in SEM, 155–195
  - business planning cycle, 186
  - business processes, 162
  - business requirements
    - functions and skills, 156
    - people, 163–165
    - processes and tools documentation, 160–163
  - business volumes, controlling, 135
  - Business.com, 11, 71
  - buying cycle
    - digital camera illustration, 210–211
    - landing pages and position within, 209
    - purchase stages, 58
    - search term targeting and, 212
- ## C
- caching content, 18
    - noarchive directive and, 107
  - Cascading Style Sheets (CSS)
    - integrating Flash with, 93
    - page optimization problems with, 66
    - search engine optimization and, 84
  - category pages, visitors to, 59
  - CD-ROM (accompanying the Kit)
    - client intake material, 161, 188
    - comprehensive tool set on, 160
    - custom quotation template, 180
    - Keyword Worksheet, 57
    - project management documentation, 161
    - SEM process documentation, 161
    - site review template, 174, 179
  - change orders, 180
  - Claiborne, Scottie
    - interview with, 214–221
    - vertical directory expertise, 73
  - Classification of Search Engine Spam, the (white paper), 236
  - click-through rates
    - Google rankings in sponsored listings, 123
    - improving, 138–139
    - measurement and reporting, 150
    - PPC ad testing, 142
  - click-through stage, PPC, 130
  - click-tracking scripts, 96
  - Clicktracks tool, 229
  - client intake process, 161
    - at High Rankings, 206
    - sample form, 188
  - clients
    - company size and business fit, 170
    - “difficult” clients, 185

- disengaging from, 186
- expanding relationships with, 169
- expectations of an SEM consultant, 169–171
- getting business, 165–179
- lifetime value of, 193
- relations at High Rankings, 206
- relations at KeywordRanking.com, 198
- relations with after mistakes, 186
- turning prospects into, 168, 248
- cloaking, 77–79
  - distinguished from IP delivery, 245
  - professionalism and, 194
- clock synchronization, 97, 118
- closing the deal, 176
- common sense and high rankings, 215
- common words
  - search results and, 46
  - stop words, 29, 46
  - use as modifiers, 44
- communication with clients, 185
- comparison shopping, 145
- competing sites
  - keyword discovery from, 47
  - locating links to, 74
- competition for search terms
  - balance between value and, 53
  - Google ranking policy and, 124
  - less for targeted terms, 56
  - off-page optimization and, 69
  - PPC keyword selection and, 134
- competitive intelligence tools, 232
- Comscore service, 232
- consultative selling, 175–176
  - closing the deal, 176
- consulting in search engine marketing, 155–195
  - building an SEM business, 155–165
  - consulting process, 156
  - directories of SEM consultants, 174, 252–253
  - doing business, 179–186
  - finding clients, 165–179
  - professionalism, 193–194
  - selling consultation services, 184
  - strategy development, 186–193
- contact information
  - credibility and, 149
  - ranking and, 44
- content
  - mapping search terms to, 58
  - page types of typical Websites, 59
  - shortcomings for relevance assessment, 22
  - written proposals, 177
- content analysis, 20
- content duplication (*see* duplicate content)
- content hijacking, 113
- content management systems, 99
- content spam, 77, 238–239
  - link content spam, 242
- content strategy, 191
- content type headers, 98
- content-based ranking, 31
- context and personalization developments, 39
- contextual advertising, 10
  - AdWords and Overture support, 127
  - future of paid search and, 153
  - topic and synonym matching, 133
- contractors
  - alternative to hiring employees, 164
  - eLance as source of, 233
- contracts
  - ensuring loyalty with, 164
  - need to secure written, 176
  - standard and negotiated, 162
- conversion rates
  - competitiveness and, 148
  - customer focus and trust, 148
  - difficulty of tracking, 142
  - further reading on, 247
  - importance of measuring, 131
  - improving, 147
  - setting realistic goals, 149
- cookies
  - diagnosing content duplication, 109
  - navigational use, 91
  - privacy policy and, 149

- shopping carts and, 100
  - copyright
    - infringement and page removal, 33
    - notices as signature text, 22
  - copywriting, 66
    - outsourcing common, 158
    - as SEM business skill, 158
    - virtue of brevity, 141
  - cost-effectiveness of PPC and SEO, 190
  - cost-per-action pricing, 180
  - costs and price negotiation, 181
  - countries
    - country-specific search options, 38
    - PPC targeting at, 128
    - search engine usage patterns, 34–37
  - crawlability, 61
    - crawable pop-ups, 90
    - dynamic sites and, 99
    - getting indexed and, 75
  - crawler-based listings (*see* organic listings)
  - crawlers (*see* spiders)
  - crawling search engines, 11–37
    - indexing, 20
    - link analysis, 22
    - parsing and caching content, 18
    - practical problems, 15
    - progress of a typical crawling session, 14
    - query processing, 28
    - resources used by, 13
    - results providers, 7
    - spider operations, 12
  - creative messages, targeting, 209, 211
  - credibility, 149
  - cross-linking Websites, 115
  - custom error pages, 97, 114
    - server errors and, 102
  - custom quote pricing, 180
  - customer focus and trust, 148, 169
    - (*see also* clients)
- D**
- daily spending limit feature, 136
  - database errors
    - crawler problems with, 16
    - SEO and, 101
  - database/repository
    - caching content, 18
    - search engine component, 11
  - databases of search terms, 223
  - dayparting, 126, 136
  - deliverables, payment based on, 183
  - demonstration projects, 171–172
  - depth and breadth of crawling, 18
  - descriptive copy in PPC, 140
  - design and structure (*see* Website design)
  - DHTML used for navigation, 89
  - digital camera illustration, 210–211
  - DigitalPoint tool, 230
  - DirectHit search engine, 3, 32
  - directories
    - of consultants, 174
    - creating, to attract links, 73
    - dynamic directories and redirect scripts, 101
    - Galaxy directory, 2
    - link building and, 70
    - linking strategy planning, 192
    - list of major directories, 71, 248–250
    - listing SEM consultants, 253
    - paid directory listings, 152
    - partial indexing, 95
    - paying up-front for, 216
    - types of directory, 71
    - use by Scottie Claiborne, 215
    - vertical and special-interest, 71, 152
    - vertical directory guides, 72
  - directory listings, 8
    - economics of, 9
    - example from Yahoo!, 5
    - origins of About, 3
    - value of links from, 27
  - directory submissions, 10, 71
    - Alliance Link service, 233
  - disallow: directive, robots.txt, 105
  - discussion forums
    - dedicated to SEO/SEM, 250
    - IHY forum, 215
    - keyword discovery from, 47

- networking and referrals from, 172, 174, 217
  - Scottie Claiborne on, 216
  - distribution of advertisements
    - AdWords and Overture, 127
    - PPC options, 137
  - distribution of press releases, 204
  - <div> sections
    - blending Flash and HTML, 93
    - hidden content, 84
    - repositioning links using, 85
  - DMCA (Digital Millennium Copyright Act), 33
  - DNS servers
    - crawler problems with, 15
    - localization problems and, 38
    - moving domains between, 117–118
  - document indexer component, 12
  - document location by search engines, 12
  - documentation and client relations, 185
  - doing business, pricing, services and client relations, 179–186
  - domain forwarding (*see* redirection)
  - domain names
    - (*see also* DNS servers)
    - managing multiple names, 114–117
    - www.example.com and example.com, 109
  - domain spam, 77
  - domains, moving between hosts or servers, 117
  - doorway pages, 65, 68, 77
  - drill-down searches, 44
  - Dublin Core tags and spam, 241
  - duplicate content, 94, 104
    - checking scripts and variables, 110
    - crawlability and, 62
    - diagnosing duplication, 108
    - dynamic sites and, 98
    - empty pages, 111
    - examples, 94
    - sessions and cookies, 109
    - shopping carts and, 100
    - www.example.com and example.com, 109
  - dynamic sites, 98
    - content duplication, 98, 110
    - database and server error handling, 101
    - empty pages on, 111
    - examples of, 99
    - problems with site:domain searches, 20
    - robots <meta> tag use, 108
    - spider problems with, 16–17
    - URL rewriting, 103
  - dynamic text replacement, 86
- E**
- editorial guidelines, PPC providers, 139
  - eLance.com online marketplace, 165, 173–174, 232
  - <embed> tag, 93
  - employee hiring, 164
  - empty pages on dynamic sites, 111
  - empty table cell workaround, 83
  - eSpotting PPC service, 227
  - ethics and spamming, 80
  - exact matching, 132
    - ranking of bids, 123
  - excluded words (*see* negative matching; stop words)
  - experience
    - gaining references and, 171–173
    - requirement to demonstrate, 170
  - external profile
    - combining with internal, 55
    - link building, 69, 192
    - measuring competition and, 53
    - measuring off-page factors, 53
  - external redirects, 113
- F**
- Fallon, Brad, demonstration project, 172
  - Findwhat service, 128, 227
  - Flash content
    - Flash accents, 220
    - mixing with HTML, 92
    - search engines and, 91

- flat-rate pricing, 179, 181
  - <font> tag optimization problems, 66
  - footers, for contact information, 44
  - formatting and relevance, 30
  - form-based navigation, 91
  - forums (*see* discussion forums)
  - frames, 86
  - fraud detection tools, 228
  - freebies, advertising, 141
  - freelance service providers, 232
  - freeloaders, avoiding, 134, 141, 168
  - freshness of content
    - search engine popularity, 33
    - search engine quality and, 18
    - trusted feed programs and, 13
  - functions essential to an SEM business, 156
  - funnel model of selling, 166
  - future of search engines, 37
  - future-proofing with topical ranking, 32
- G**
- Galaxy as the first directory, 2
  - generic query ranking, 31
  - generic search terms, 43
    - relevance, 56
  - geographical targeting, 38, 128
    - location modifiers, 44
    - Overture Local Match, 214
    - PPC options, 137
    - qualifying visitors, 142
    - search engine improvements, 37
  - goals, setting realistic
    - business planning stage, 188
    - conversion rate improvement, 149
    - landing pages, 144
    - SEM consulting, 157
    - written proposals, 178
  - GoGuides directory, 72
  - Google AdSense, 4, 153
  - Google AdWords, 4
    - Atlas OnePoint support, 226
    - geo-targeting technologies, 38
    - Overture compared to, 123–128
    - PPC market position, 122
    - professional certification program, 139
    - ranking methods, 8
    - reviewing submitted advertisements, 125
    - sponsored listings provider, 7
    - testing support, 142
  - Google Groups tool, 48
  - Google Keyword Suggestion Tool, 50
  - Google News, 201–204
  - Google search engine
    - (*see also* PageRank algorithm)
    - browser toolbar showing PageRank, 54, 226, 234
    - cached DNS records, 15
    - diagnosing duplication, 108
    - implementing a site search, 116
    - page title length limit, 64
    - PageRank algorithm, 26
    - Personalized Search system, 39
    - redirect bug, 114
    - rise to dominance, 3
      - as search results provider, 7
    - site:domain syntax, 20
    - supplemental search results, 19
    - transparency of processing, 20, 26
    - user interface, 3, 5
  - Google Web services API, 230
  - googlebot <meta> tag, 107
  - graphic design and credibility, 149
  - guidelines
    - defining spam, 246
    - PPC advertisers and, 139
- H**
- <head> element, HTML
    - as meta spam location, 239
  - headings, HTML
    - resizing using CSS, 85
    - search terms in, 64
  - headlines
    - news search engines display, 204
    - PPC advertisement example, 121
    - PPC advertisement with search terms, 140

- sales-grabbing headings and rankings, 67
  - text as images, 86
  - hidden content
    - Flash movies and, 92
    - problems with, 84
    - search forms as, 89
  - High Rankings, interview with Jill Whalen of, 205–208
  - hit counters and redirection, 113
  - HITS (Hypertext-Induced Topic Selection)
    - algorithm, 25, 28
  - Hitwise service, 232
  - homepages
    - acting as a landing page, 145
    - generic search terms and, 61
    - information gathering mode and, 59
  - Hong Kong, search engine popularity, 36
  - hosting
    - importance, 119
    - moving domains between, 117
  - hourly rate pricing, 179
  - HTML
    - mixing with Flash content, 92
    - problems with redefining tags, 85
    - storing retrieved documents as, 19
    - table-based layout code, 83
    - validation and optimization, 68
  - HTTP headers, 95
    - checking tools, 98
    - Mozilla Firefox view, 225
    - user-agent field, 243
  - HTTP status codes, 96–97
    - 200 OK, 96
    - 301 Moved Permanently, 96, 112, 114
    - 302 Found, 96, 98, 113
    - 304 Not Modified, 96, 102, 118
    - 401 Unauthorized, 97
    - 403 Forbidden, 97
    - 404 Not Found, 16, 97, 114
    - 500 Internal Server Error , 97, 102
    - 503 Service Unavailable, 97
    - crawler problems and, 16–17
    - custom error pages and, 97, 114
  - hubs
    - link analysis and, 25
    - link popularity and spam, 241
  - human edited listings (*see* directory listings)
  - human readability and spam, 242, 244
  - hyperlinks (*see* links)
- I**
- If-Modified-Since header, 96, 102
  - IHY discussion forum, 215
  - images
    - displaying text as, 86
    - including in news releases, 203
    - use for site navigation, 88
  - inanchor: operator, Google, 53
  - incoming links
    - count of, 54
    - popularity and placement, 53
  - indexing
    - duplicate content and, 95
    - ensuring, for a Website, 74–76
    - index size and search engine quality, 17
    - problem diagnosis, 76
  - indexing process
    - crawling search engines, 12, 20
    - hypothetical example entry, 21
    - results of indexing, 22
  - information pages, 60
  - information retrieval theory, 29
  - informational offers, 174
  - in-house resources, employing, 183
  - Inktomi search engine, 2–3
    - acquisition by Yahoo!, 4
  - interaction stage, PPC, 131, 147
  - internal profile
    - measuring competition and, 53
    - measuring on-page factors, 54
  - interviews with SEM practitioners
    - Andy Beal, Keyword Ranking, 197–201
    - Greg Jarboe, SEO-PR, 201–205
    - Jill Whalen, High Rankings, 205–208
    - John Slade, Overture, 209–214
    - Scottie Claiborne, Karcher Group, 214–221
  - intitle: operator, Google, 54

- invisible text (*see* hidden content)
- IP addresses
- geographical targeting and, 38, 78
  - IP delivery and cloaking, 245
  - moving domains, 117–118
- “I’m Feeling Lucky” feature, Google, 4
- J**
- Jarboe, Greg, interview with, 201–205
- JavaScript
- alternative to hidden text, 84
  - Google and, 62
  - navigational menus and pop-ups, 89–90
  - self-referencing framesets, 87
- K**
- Karcher Group, interview with Scottie Claiborne of, 214–221
- keyboard errors, 46
- keyword campaign organization, 137
- keyword density analyzers, 47, 51, 67
- keyword discovery, 47–48
- Keyword Discovery tool, 222
- Keyword Effectiveness Index (KEI), 56
- keyword matching options, PPC, 132, 134
- keyword metrics, 48
- competition for terms, 53
  - popularity of terms, 49
  - relevance and weighted popularity, 55
- keyword phrases, 43
- Keyword Ranking
- flat fee pricing policy, 181
  - interview with Andy Beal of, 197–201
- keyword research
- business skill, 157
  - SEO Research Labs, 233
- keyword research tools, 48
- Keyword Discovery, 223
  - Overture Search Term Suggestion Tool, 223
  - Wordtracker tool, 223
- keyword strategy, 42–58
- hierarchical arrangement of keywords, 42
  - at High Rankings, 206
  - PPC and SEO, 189
- Keyword Worksheet tool, 57
- keywords
- (*see also* search terms)
- brand names as, 45, 59
  - defined, 42
  - placement effect on landing pages, 58
  - placement, and rankings, 65
  - selecting for PPC campaigns, 134
  - selection, 57
  - singular and plural forms, 45
- Kleinberg, John, 25
- L**
- landing pages
- buying cycle position and, 209–210
  - calls to action, 145
  - example screenshot, 145
  - including search terms, 147
  - John Slade on targeting, 209
  - keyword placement and, 58
  - navigability and, 62
  - pop-ups as, 91
  - position in PPC sequence, 130
  - PPC campaigns, 144
  - setting realistic goals, 144
  - specific search terms and, 61
  - testing, 147
- landing zone concept, 145
- languages, PPC targeting, 128
- large companies
- Keyword Ranking as, 197
  - working with, 171
- lawyers
- contract preparation and negotiation, 163
  - ethics of accountants and, 80
- layout options, 83
- optimization and layout tables, 66
- lead generation services, 128
- “lead products”, 174
- leads, turning into prospects, 167

- legal advice (*see* lawyers)
  - link analysis, 22
    - hubs and authorities, 25
    - PageRank algorithm, 26
    - ranking generic queries, 31
    - SEO Elite tool, 225
    - topics and communities, 27
  - link analyzer component, 12
  - link brokers, 193
  - link building, 69–74
    - Alliance Link service, 233
    - attracting one-way links, 72
    - directory submission, 71
    - getting indexed and, 75
    - at High Rankings, 207
    - link placement and context, 70
    - linking out, 74
    - localized links, 74
    - managing the external profile, 69
    - relevance, 73
    - skills and responsibilities, 158
    - target page selection, 70
  - link content spam, 242
  - link directories and redirect scripts, 101
  - link exchanges, 70, 73, 192
  - link farms, 242
  - link meta spam, 242
  - link partnerships
    - local link partnerships, 74
    - research tools, 224–225
  - link popularity
    - (*see also* PageRank)
    - authority, hub and spam, 241
    - importance of placement and, 53
  - link spam, 77, 242
  - link tracking tool, 224
  - link: operator syntax, 54
  - linking strategy
    - directory submissions and, 152
    - planning, 192
  - links
    - (*see also* external profile; navigation)
    - cross-linking Websites, 115
    - incoming links, 53–54
    - internal links and ranking, 58
    - location of navigation, 66
    - placement of, and PageRank, 27
    - repositioning, 85
    - search terms in, 65
    - supplemental search results from, 19
    - URL discovery and indexing, 21
    - use by crawling search engines, 13
    - value of, and PageRank, 27
  - load balancing techniques, 33
  - localization (*see* geographical targeting)
  - LocalRank algorithm, 28
  - log file analysis tools
    - landing page analysis, 146
    - PPC keyword selection, 135
  - Lycos search engine, 2, 9, 151
- M**
- Macromedia Corporation (*see* Flash content)
  - major search terms (*see* generic search terms)
  - managed services, 180
  - manual submission (*see* submitted URLs)
  - margins and negotiating prices, 181
  - marketing, 173
    - special features of paid search, 211
    - Website on online marketing, 251
  - marketing messages in page titles, 64
  - Marketleap link popularity tool, 54, 74
  - markups on outsourced and managed services, 180
  - mass marketing, 173
  - measurement
    - establishing business measures, 187
    - Greg Jarboe on, 205
    - importance of in PPC, 131, 150
    - reporting and, business process, 162
  - mergers and acquisitions, 4
    - multiple domain names from, 115
  - meta spam, 238–241
    - examples, 240
    - link meta spam, 242
  - metadata storage, 19

- metasearch engines, 2–3, 7
  - Microsoft Corporation
    - (*see also* MSN search engine)
    - “Stuff I’ve Seen” project, 39
  - Microsoft Small Business Directory, 71–72, 75
  - Microsoft Word document storage, 19
  - Mining Company, the, 3
  - misspellings, 45–46
  - modifiers
    - locations as, 44
    - product/service pages, 60
    - supplementing keywords, 44
  - monetizing, 216
  - “money pages”, 60
  - moonwalk rental business, 215
  - Mozilla Firefox browser
    - PageRank display extension, 54
    - viewed as an SEO tool, 225
  - MSN search engine, 4
    - PPC prospects, 122
    - search results provider, 7
    - Submit It service, 52, 75
    - user interface, 5
  - multi-lingual advertising, 128
  - multimedia resources as trusted feeds, 152
  - multiple domain names and spam, 116
  - multiple Websites, cross-linking, 115
- N**
- national usage patterns, 34–37
  - navigability, 62
    - crawlability and, 61
    - framed pages in search results, 87
  - navigation
    - (*see also* links)
    - crawlability and navigability, 61
    - crawlable DHTML and JavaScript, 89
    - forced cookies and form-based navigation, 91
    - pop-up windows, 90
    - site maps and, 88
    - table-based layout problems, 83
    - use of images for, 88
  - negative matching, PPC, 133
    - descriptive copy and, 140
  - Netscape browser, 2
  - NetTracker tool, 229
  - networking and finding new business, 174
  - new content (*see* freshness)
  - New Zealand, search engine popularity, 36
  - news search engines
    - algorithm changes, 204
    - optimizing press releases for, 201, 231
  - newsgroups and keyword discovery, 48
  - newsletter, Successful Sites, 214, 220
  - niche services, 173
  - noarchive directive, 107
  - <noembed> tag, 93
  - nofollow directive, dangers, 107
  - non-profit organizations, 172
  - <noscript> tag, 89
- O**
- Occam’s Razor, 52
  - offer-based advertising, 140
  - off-page factors (*see* external profile)
  - Omniture SiteCatalyst tool, 229
  - online marketing, 174
  - online marketplaces
    - bidding for work on, 173
    - eLance, 232
    - finding new business, 174
    - source of contract staff, 165
  - online resources (*see* Websites)
  - on-page factors (*see* internal profile)
  - Open Directory, the, 72
  - optimization (*see* page optimization; search engine optimization)
  - opt-in email, 204
  - organic listings, 6
    - example from Yahoo!, 5
    - PPC metric and competition for, 55
    - providers of, 7
    - sites appearing in sponsored and, 190
  - OSCommerce system, 100
  - outsourcing, 180
    - SEO copywriting, 158

- services listed, 231
  - to technical experts, 165
  - Overture Bid Optimizer, 226
  - Overture Content Match, 127, 153, 213
  - Overture Good Keywords, 48, 50
  - Overture Local Match, 128, 214
  - Overture search engine
    - (*see also* Yahoo! search engine)
    - acquisition of AltaVista and by Yahoo!, 4
  - Overture Search Optimizer, 126, 138, 213
  - Overture Search Term Suggestion Tool, 49, 223
  - Overture service
    - Atlas OnePoint support, 226
    - dayparting support, 126
    - geo-targeting technologies, 38
    - Google AdWords compared to, 123–128
    - interview with John Slade of, 209–214
    - ranking methods, 8
    - reviewing submitted advertisements, 125, 139
    - sponsored listings provider, 7
    - support for testing, 142
- P**
- package pricing, 179
  - packaged services, 179, 184
  - page element loading, 92
  - Page Not Found errors, 16, 97, 114
  - page optimization, 63–68
    - copywriting, 66
    - HTML validity and, 68
    - keyword density analysis, 67
    - layout and, 66
    - page elements, 63
  - Page, Larry, of Google and PageRank, 26, 235
  - PageRank algorithm, 26
    - as an iterative process, 234
    - nofollow directive and, 107
    - online resources, 235
    - pop-ups and, 90
    - resources on, 234–235
    - topic-sensitive variation, 28
  - PageRank Calculator tool, 235
  - PageRank values
    - apparent flow of PageRank, 234
    - browser toolbar display, 54, 234
    - leaking PageRank, 74
    - not the most important thing!, 235
  - paid advertising on Google, 4
  - paid directory listings, 152
  - paid inclusion programs, 151
    - crawling search engine scheduling, 18
    - ensuring indexing through, 75
    - organic listings and, 6
    - Position Technologies service, 224
    - PPC and future of, 76
    - Priority Submit service, 224
    - trusted feeds compared to, 151
    - XML data feeds, 13
  - paid search performance triangle, 209
  - paid search, John Slade on, 212
  - parser modules, 11
  - parsing stored documents, 19
  - partially indexed pages, 95, 109
  - partnerships
    - link exchanges and, 73
    - linking strategy planning, 192
    - localized links, 74
    - Overture dependence on, 127
    - Overture with Claria, 128–129
    - professional partners and vendors, 165
  - payment terms, 183
  - Pay-Per-Click listings (*see* PPC listings)
  - pay-to-play programs
    - (*see also* paid inclusion programs; PPC listings; trusted feed programs)
    - other than PPC, 150
    - paid directory listings, 152
  - PDF document storage, 19
  - people, building a team, 163–165
  - performance
    - dynamic sites and, 99
    - reviewing, for PPC campaigns, 131
  - performance bonuses, 181
    - alternative to lowering prices, 182

- Perkins, Alan, on the Classification of Search Engine Spam, 236
- personality and client expectations, 170
- personalization developments, 39
- phasing alternative to lowering prices, 182
- phrase matching, 133
- pipe character, 64
- placeholder pages, 112
- planning stage  
(*see also* goals, setting realistic)  
business planning cycle, 189  
linking strategy, 192
- plural forms  
exact matching options and, 132  
search engine recognition problems, 45
- poorhouse tycoons, 185
- pop-over elements, 77
- popularity of search engines, 34–37
- pop-under windows, 129
- position (*see* rankings)
- Position Technologies service, 224
- postal codes, 37
- PPC campaigns, 120–150  
adding listings, 125  
advertisers' control over, 121, 130, 153  
budgets and positioning, 135  
campaign management skills, 159  
choice between SEO and, 190  
dayparting, 126  
descriptive copy, 140  
distribution and targeting options, 137  
fraud detection, 228  
keyword selection, 134  
landing pages and landing zones, 144  
management in SEM businesses, 159  
measurement and reporting, 150  
minor PPC service suppliers, 128  
offer-based advertising, 140  
organization, 125–126, 137  
paid search performance triangle, 209  
PPC marketplace, 122  
price trends and conversion rates, 148  
selling management, as opposed to SEO services, 184  
significance of large retailers, 153  
testing advertisements, 142  
third-party applications, 125–126  
tools and services, 226–227
- PPC keyword metric, 53, 55
- PPC listings, 7  
(*see also* sponsored listings)  
advertisement example, 121  
Findwhat service, 227  
future of paid search and, 152  
history, 3  
listings providers, 7  
major process stages, 130–150  
Overture reviewing of, 213  
PPC advertising model, 8, 121  
pricing of clicks, 124  
ranking methods, 8  
as an SEM technique, 10
- PR Web service, 205
- pre-selection of results, 31
- press release optimization, 201, 231
- pressure cookers, 185
- pricing  
alternatives to lowering prices, 182  
consultancy work, 179–183  
cost-per-action pricing, 180  
flat fee pricing, 181  
at High Rankings, 207  
at Keyword Ranking, 199  
negotiating price, 181  
outsourced and managed services, 180  
PPC pricing models, 124  
self service pricing model, 183
- priorities  
link sources and, 21  
scheduling and, 17
- Priority Submit service, 51, 224
- PRISMA tool, AltaVista, 48
- privacy  
credibility and privacy policy, 149  
personalization and, 39
- problem clients, 185
- problem identification in consultative selling, 175

problem sites as sales leads, 167  
 problem solving skills, 156  
     pinning value to solutions, 175  
 problems, taking ownership, 186  
 processes  
     client expectations and, 170  
     documenting work processes, 160–163  
 product category targeting, 133  
 product/service pages as targets, 60  
 professionalism in SEM businesses, 193–194  
 profitability  
     conversion rates and, 147  
     margins and price negotiation, 181  
 project management documentation, 161  
 proposals  
     further reading on, 248  
     writing effective proposals, 176  
 prospects  
     finding new business, 173–174  
     requirements of, 169–171  
     turning into clients, 168, 248  
     turning leads into, 167  
 protected content and robots.txt, 106  
 proxy servers and geo-targeting, 38, 79  
 punctuation  
     omitted from indexes, 21  
     splitting search terms, 67  
 purchasing decisions (*see* buying cycle)

## Q

quadrant testing, 143  
 qualifying visitors, 141, 211  
 quality issues  
     avoiding links to “bad” sites, 74  
     search engine monitoring, 32  
 queries  
     processing by search engines, 28  
     query processor component, 12  
     search engine correction of, 46  
 query vectors, 29–30  
 query-dependent ranking strategies, 31

## R

random Web surfing and PageRank, 26  
 rank checking tools  
     Advanced Web Ranking tool, 231  
     DigitalPoint tool, 230  
     skew in keyword data from, 49–50  
 ranking strategies  
     query-dependent strategies, 31  
     retrieval strategies and, 31  
     topical factors, 32  
 ranking systems  
     PPC alternative, 3  
     ranking by feedback, 3  
     spam penalties and algorithm changes, 79  
     sponsored listings, 8  
 ranking/retrieval module  
     search engine component, 12  
 rankings  
     (*see also* search engine optimization)  
     benefits of trusted feeds, 152  
     common sense and, 215  
     contact information and, 44  
     deficiencies as a business goal, 193  
     directory submissions and, 152  
     generic search terms and, 43  
     internal links and, 58  
     keyword placement and, 65  
     manipulation of, 27  
     personalization and the future of, 39  
     position in sponsored listings, 122–124, 136  
     sales-grabbing headings and, 67  
 RDF and search engine spam, 241  
 reciprocal linking, 70  
 recommended additional reading, 247  
 redirect scripts, 113  
     link directory use, 101  
 redirect spam, 77, 243  
 redirection  
     crawler problems with, 17  
     example.com to www.example.com, 110  
     HTTP status codes for, 96  
     multiple Website versions and, 114

- relocation and, 112
  - spam and, 77, 243
  - subdomains and, 116
  - types of, and problems with, 112
  - references and testimonials, 169, 172
    - including in written proposals, 178
  - referrals and new business, 174
  - refinement
    - PPC campaigns, 131, 134–135, 147
    - search terms, 56
  - regions (*see* geographical targeting)
  - related searches features, 48
  - relevance of search terms, 48, 55
  - relevancy, Alan Perkins' definition of, 237
  - relocation redirects, 112
  - reporting and competitive advantage, 162
  - repository (*see* database/repository)
  - request for proposal (RFP) process, 171
  - resource pages, linking strategy planning, 192
  - resource requirements for search engines, 11
  - resources (appendix)
    - the big directory list, 248–250
    - PageRank algorithm, 234–235
    - SEM organizations, marketplaces and directories, 253
    - spam white paper, 236
    - Websites relevant to SEM, 250–252
  - results pages
    - additional types of listing, 9
    - example from Yahoo!, 5
    - framed pages in, 87
    - homepage domination of, 61
    - number of results displayed, 7
    - numbers of results quoted, 31
    - pre-selection process, 31
  - results, as client expectations, 171
  - return on investment, PPC, 121, 129
  - Robot Exclusion Protocol, 13
    - duplicate content and, 104
    - hiding duplication, 111
    - robots <meta> tag, 106
    - robots.txt file, 104
  - robots (*see* spiders)
  - robots <meta> tag, 106–108
    - empty pages and, 112
    - Google extension, 107
    - hiding duplication, 111
  - robots.txt files, 14, 104–106
    - choice between robots <meta> tag and, 108
    - external redirects and, 113
    - hiding duplication, 111
    - multiple Website versions and, 114
  - rollovers, 84
  - Rosenberger, Stewart, on dynamic text replacement, 86
  - Roy, Sumantra, KEI tool, 56
  - RSS (RDF Site Summary), 13, 40
- S**
- sales (*see* selling)
  - Salton, Gerald, IR theoretician, 29
  - scheduling by search engines, 11, 17
  - scope of work
    - modifying, 182
    - scope creepers, 185
  - Scope of Work statements, 180
  - scumware, 129
  - search counts in keyword research, 49
  - Search Creative team, High Rankings, 206–207
  - search engine listing (*see* indexing)
  - search engine marketing, 10
    - Andy Beal's view on, 201
    - building an SEM business, 155–165
    - business challenges, 179–186
    - discussion forums, 250
    - finding clients, 165–179
    - further reading on, 247
    - process documentation, 161
    - professional organizations, 253
    - professionalism in, 193–194
    - running as a business, 155–195
    - seeing the big picture, 219
    - SEM strategy, 186–193
    - services to provide, 184

- tools, 222–233
- Weblogs on, 251
- search engine marketing consultants (*see* consulting in search engine marketing)
- search engine optimization, 41–81 (*see also* page optimization)
  - advanced, 94–95
  - best practice, 79
  - choice between internal and external profiles, 191
  - choice between PPC and, 190
  - crawlability and navigability, 61
  - CSS use, 84
  - defined, 10
  - duplicate content, 94, 104
  - dynamic sites, 98
  - evolution of, 208
  - further reading on, 247
  - getting indexed, 74
  - harmonizing design and, 82
  - HTTP headers and, 95
  - keyword strategy phase, 42
  - link building phase, 69
  - outsourcing copywriting, 158
  - press releases, 201, 231
  - search engine spam and, 76
  - selling, as opposed to PPC services, 184
  - sequence of activities, 41
  - servers and domain issues, 112
  - site design phase, 58
  - tools and services, 222–226
  - URL rewriting, 103
- search engine results pages (*see* results pages)
- search engine spam (*see* spam)
- search engine-friendly design, 83
  - Flash and, 92
  - limitations of images, 88
  - shopping cart systems, 100
  - site maps and, 88
  - URL rewriting inadequacies, 103
  - using various layout options, 84
- search engines (*see also* crawling search engines)
  - Alan Perkins' definition of, 237
  - anticipated improvements, 37
  - context and personalization, 39
  - history, 1–5
  - localization, 37
  - national usage patterns, 34–37
  - online guides to, 252
  - organic search listing providers, 7
  - parallel activities, 32
  - quality criteria, 17
  - search portals and, 5
  - search quality monitoring, 32
  - understanding requirements of, 33
  - volume of Web searches globally, 1
- search forms and site maps, 89
- search options and query processing, 28
- search portals
  - anatomy of, 5–9
  - B2B and B2C focus, 137
  - paid inclusion support, 151
- search results (*see also* results pages)
  - crawling phase results, 19
- search term analysis tool, 47
- search terms (*see also* keywords)
  - competition and keyword selection, 134
  - competition and off-page optimization, 69
  - competitive intelligence tools and, 232
  - databases of, 223
  - defined, 42
  - drill-down searches, 44
  - generic search terms, 43
  - Google ranking policy and, 124
  - identification from log files, 135
  - including in landing pages, 147
  - including in PPC copy, 140, 211
  - including in title tags, 54
  - locating in body copy, 65
  - locating in headings, 64
  - locating in links, 65
  - locating in subheadings, 65
  - location and page optimization, 63

- mapping to individual pages, 58
- mapping to page types, 59
- popularity revealed by PPC, 8
- purchasing decision stages, 58
- refinement, 56
- relevance, 55–56
- splitting phrases, 67
- targeting, 43, 191, 212
- title tag location, 63
- weighted popularity, 56
- seasonal trends, 223
- seed lists, 43, 47–48
- self service pricing model, 183
- self-referencing framesets, 87
- selling
  - additional opportunities, 169
  - the buyer’s perspective, 169–171
  - closing the deal, 176
  - conceptual selling, 248
  - consultative selling, 175–176
  - finding new business, 173–174
  - further reading on, 248
  - Keyword Ranking’s sales force, 200
  - sales cycle time, 167
  - sales generated and conversion, 150
  - turning leads into prospects, 167
  - turning prospects into clients, 168
  - understanding the selling cycle, 166
  - what to sell, 184
- selling cycle, 166
- “selling the assessment”, 168
- SEM (*see* search engine marketing)
- semantic analysis in query processing, 28
- Semantic Web, the, 39
- SEO (*see* search engine optimization)
- SEO Elite tool, 225
- SEO Research Labs, 233
- SEO-PR, 231
  - interview with Greg Jarboe of, 201–205
- separators in titles, 64
- SERP (search engine results page) (*see* results pages)
- server clock checking, 97, 118
- server error messages
  - possible workaround, 102
  - SEO and, 102
- server log data, 229
- server-based redirects, 17
- servers
  - crawler problems with, 16
  - crawling search engine requirements, 11
  - moving domains between, 117
  - SEO issues, 112
- services and what to sell, 184
- session maintenance
  - diagnosing content duplication, 109
  - shopping carts, 100
- Shared Vision Statement, Keyword Ranking, 198
- shopping carts and SEO, 100
- signature text, 22
- Singapore, search engine popularity, 37
- single point of contact ideal, 220
- singular and plural keywords, 45, 132
- site design (*see* Website design)
- site maps, 88
- site monitoring services, 119
- site navigation (*see* navigation)
- site reviews
  - offering as introductory services, 174, 184
  - pricing, 179
- site:domain syntax, 20, 108
  - finding indexed pages, 22
- skew, in keyword data, 49
- skills
  - essential to an SEM business, 156
  - lifelong learning, 194
  - “skip intro” links, 93
- Slade, John, interview with, 209–214
- slow boilers, 185
- small business Web presence benefits, 218
- SMART (specific, measurable, attainable, relevant, timed) goals, 188
- SMART information retrieval system, 29
- snippet descriptions, 83
- software reselling, 185

- solutions development, 156
  - source of comment on SEM tools, 222
  - spam, 76–80
    - actions not regarded as, 237
    - agent-based delivery and, 243
    - agent-based spam, 244
    - Alan Perkins' definition of, 237
    - content spam, 238–239
    - detection and filtering, 32
    - ethics of, 80
    - guidelines for search engines and marketers, 246
    - IP delivery and cloaking, 245
    - link popularity and, 241
    - meta spam, 238–241
    - multiple domain names and, 116
    - professionalism and, 194
    - redirects and, 77, 243
    - search engine rules and penalties, 79
    - standards, 236, 246
    - test for, 244
    - tests for, 239, 242
    - white paper on, 236
  - special-interest directories, 71
  - specialism
    - managing specialists, 220
    - niche service opportunities, 173
    - using specialist partners, 165
  - specific query ranking, 31
  - specific search term identification, 135
  - spiders, 12–17
    - control programs for Webmasters, 104
    - cookies and form-based navigation, 91
    - crawlability and, 61
    - database errors and, 16
    - DNS reliance, 15
    - dynamic site problems, 17
    - frames and, 86
    - history, 2
    - intolerance of HTML errors, 68
    - IP address identification and spam, 245
    - organic listings use of, 6
    - pop-up windows and, 90
    - practical problems encountered, 15
    - progress of interaction with a Website, 14
    - response to redirects, 17, 243
    - Robot Exclusion Protocol, 13
    - server related problems, 16
    - shopping carts and, 100
    - table-based layout problems, 83
    - user-agent identification and spam, 244
  - sponsored listings, 7, 120
    - (*see also* PPC listings)
    - bid price and position within, 122, 136
    - display positions, 8
    - example from Yahoo!, 5
    - ranking and charging methods, 123–124
    - sites appearing in organic and, 190
  - standards on spamming, 236, 246
  - STAT tool, 47
  - status codes (*see* HTTP status codes)
  - stemming, 45
  - stop words, 29, 46
  - strategy formulation in SEM, 157, 186–193
  - subdomains, 116
  - subheadings, search terms in, 65
  - Submit It service, Microsoft, 52, 75
  - submitted URLs, 13
    - crawling search engine scheduling, 18
    - submission services, 75
  - Successful Sites newsletter, 214, 220
  - summaries in proposals, 177
  - supplemental search results, 19
  - support pages, targeting, 60
  - syndication
    - (*see also* partnerships)
    - listing position and, 136
  - synonym matching, PPC, 133
- T**
- table cells, empty cell workaround, 83
  - table-based layouts, 83
    - page optimization problems with, 66
  - target audiences
    - designing content search terms for, 190
    - landing page subheadings, 145–146

- optimizing press releases, 202
- target="\_blank" attribute, 90
- targeted advertising, PPC, 10
  - geographical targeting, 128, 137
  - search engine facilities for, 126
  - targeting using copy, 142
- targeted search terms, 55
  - available methods, 190
  - competition for, 56
  - ease of targeting, 191
  - selecting targets, 189
- targeting creative messages, 209, 211
- team building, 163
- technical skills in SEM, 159
- Teoma search engine
  - Ask popularity and, 4
  - related search terms tool, 48
  - topical factor use, 28, 32
- term weights, 30
- "text ads", 192
- text-based navigation
  - alternative to images, 88
  - DHTML alternative, 89
- third-party applications
  - PPC, search engine support, 125–126
  - search engine marketing, 222–233
- three mouse clicks principle, 61, 88
- title and description targeting
  - John Slade on, 209, 211
- title and description targeting, John Slade on, 209, 211
- <title> elements
  - Google length limit, 64
  - influencing link text with page titles, 69
  - meta spam location, 240
  - value of search terms within, 54, 63
- titles
  - (*see also* headlines)
  - written proposals, 177
- TLD servers
  - DNS server changes and, 15
  - localization problems and, 38
- tools and service providers
  - (*see also* keyword research tools; log file analysis tools)
  - documenting SEM business tools, 163
  - other tools, 231–233
  - PPC tools, 226–227
  - SEM tools, 222–226
  - tools on accompanying CD-ROM, 160
  - traffic analysis, 150, 228–231
- top level domain (*see* TLD servers)
- topic and synonym matching, PPC, 133
- topic distillation, 28
- topical factors vs. the vector space model, 32
- track record, 171
- trademarks, 45
- traffic
  - attracting with phrases and modifiers, 43
  - display position in results and, 7
  - ensuring client sites can accommodate increases, 188
  - increasing, as SEM aim, 10
  - not being everything, 131, 218
  - reasons for limiting, 135
- traffic analysis tools, 228–231
  - PPC measurement and reporting, 150
- traffic exchange schemes, 132
- training
  - provision by SEM consultants, 185
  - provision from Google, 139
- Trellian Priority Submit service, 51, 224
- triggering stage, PPC, 130
  - targeting ad displays, 132
- troubleshooting partners, 160
- trust, 149
- trusted feed programs, 40, 151
  - acceptance by paid inclusion programs, 13
  - ensuring indexing using, 76
  - freshness and, 18
  - organic listings and, 6
  - Position Technologies service, 224
- tunnel model of selling, 167

- type resizing, 85
- typefaces using images, 86
- typos, 46
- U**
- UK, search engine popularity, 35
- unpaid work, 172
- updating though paid inclusion, 151
- URL rewriting, 103
- URLs
  - discovery and indexing, 21
  - discovery by search engines, 12
  - display URLs in PPC listings, 122
  - dynamic page detection, 17
  - embedded variables in, 110–111
  - listing retrieved URLs, 108
  - manual submission, 13
  - supplemental search result display, 19
- USA, search engine popularity, 34
- usability
  - credibility and, 149
  - further reading on, 247
  - underestimated importance of, 217–218, 220
- Usenet newsgroups and keyword discovery, 48
- user agent delivery, 78, 243
- user feedback, 3
- user interfaces
  - crawlability and, 62
  - Google focus on user experience, 3
  - offered by different portals, 5
- user profiles and geographical targeting, 38
- User-agent directive, robots.txt, 104
- V**
- validator for robots.txt, 106
- value
  - pinning to solutions, 175
  - selling, rather than price, 183
  - stressing in proposals, 178
- variable delivery
  - browser detection as, 78
  - cloaking and, 78
  - IP delivery as, 78
  - variables within URLs, 110
    - effect of position, 111
  - vector space model, 29
  - Verizon SuperPages, 10, 71
  - vertical directories, 71, 152
    - creating, to attract links, 73
    - guides to, 72
  - vertical market PPC providers, 128
  - vortals (*see* vertical directories)
  - voting, and Web graphs, 23
- W**
- Web CEO tools, 51
- Web crawlers (*see* spiders)
- Web graphs, 23–25
  - link farm detection, 242
- WebCrawler search engine, 2
- Weblogs on SEO/SEM, 251
- Website design
  - blending Flash and HTML, 92
  - blending tables and CSS, 85
  - graphic design and credibility, 149
  - layout tables, 83
  - page optimization and, 66
  - reviewing client Websites, 188
  - search engine optimization and, 58–63, 82
- Website designers
  - inherent sales advantage, 168
  - SEO/SEM skills, 220
- Website development skills, 158
- Websites
  - duplication of, 114
  - importance of reliable hosting, 119
  - moving between hosts or servers, 117
  - not needing search engine visits, 219
  - PageRank online resources, 235
  - problem sites as sales leads, 167
  - relevant to SEM, 250–252
  - search engine guides, 252
  - typical page types, 59
- Webtrends tool, 230
- weighted popularity, 56

keyword selection and, 57  
weighting word occurrences, 30  
Whalen, Jill  
    interview with, 205–208  
    reliance on partnerships, 165  
Who's Clicking Who? tool, 228  
wildcards, stop words as, 29  
word stemming, 45

#### words

(*see also* common words)  
frequency and IR theory, 29  
frequency and weighting, 30  
index storage, 20  
position, formatting and IR theory, 30  
proximity, order and IR theory, 30  
Wordtracker tool, 50, 223  
    KEI and, 56  
    seed lists and, 48  
    weighted popularity estimates and, 56  
work processes, describing, 178  
World Wide Web history, 1  
World Wide Web Wanderer, 2  
WorldMall.com, 197  
written proposals, 176, 248

#### X

XML data feeds (*see* trusted feed programs)  
XML/RDF tags, metadata and spam, 241

#### Y

Yahoo! Directory, 72, 113  
Yahoo! News, 201–202, 204  
Yahoo! Overture market position, 122  
Yahoo! search engine, 4  
    (*see also* Overture)  
    acquisitions, 4  
    directory listings economics, 9  
    example results page, 5  
    origins, 2  
    as search results provider, 7  
    user interface, 5  
Yahoo! Site Match system, 13, 76, 151

#### Z

zip codes, 37