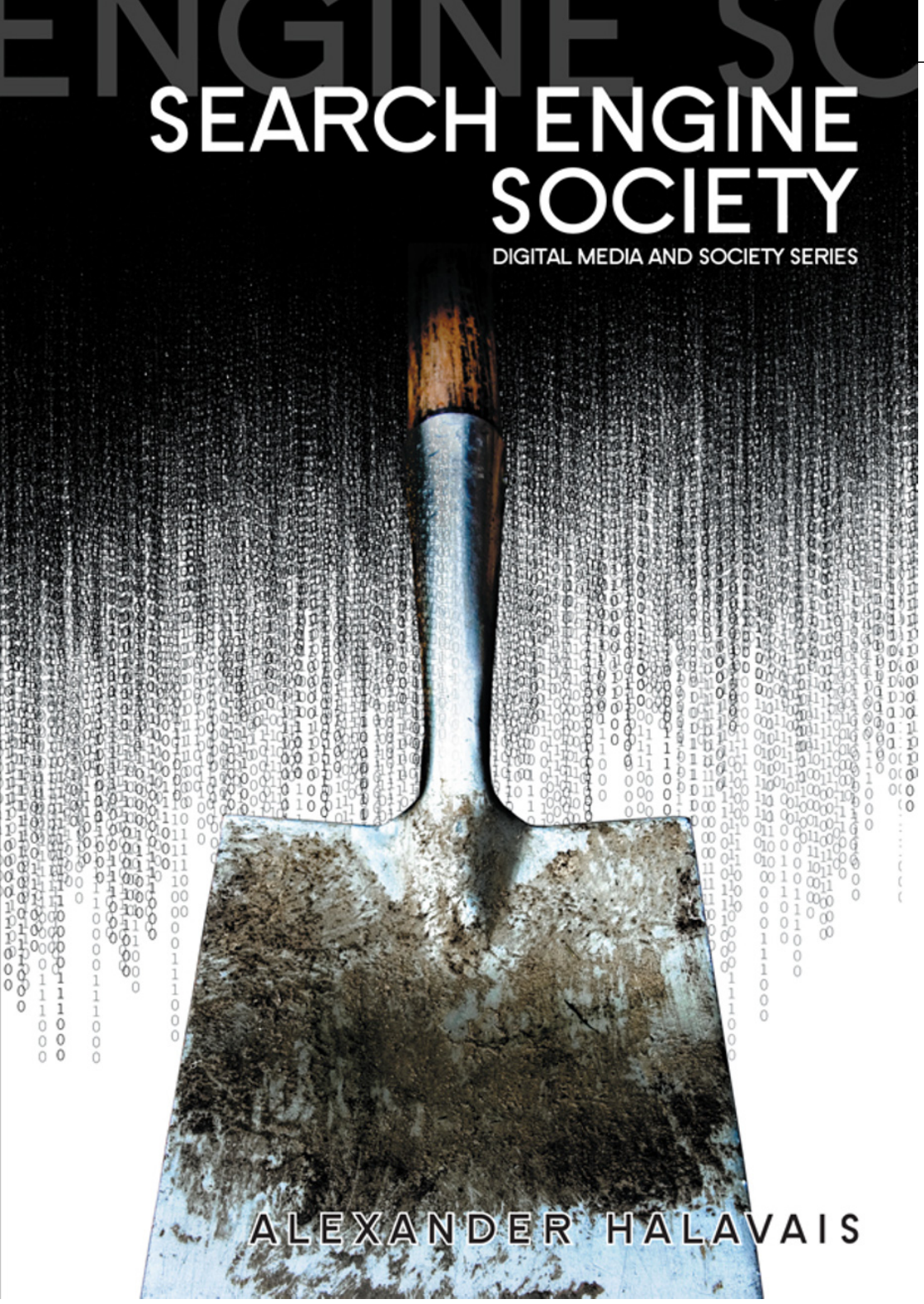
SEARCH ENGINE SOCIETY
DIGITAL MEDIA AND SOCIETY SERIES
ALEXANDER HALAVAIS

Search Engine Society

**Digital Media and Society Series**

Mark Deuze, *Media Work*
Alexander Halavais, *Search Engine Society*
Robert Hassan, *The Information Society*
Tim Jordan, *Hacking*
Jill Walker Rettberg, *Blogging*

# Search Engine Society

Alexander Halavais

polity

First published in 2009 by Polity Press

Polity Press
65 Bridge Street
Cambridge CB2 1UR, UK.

Polity Press
350 Main Street
Malden, MA 02148, USA



ISBN-13: 978-0-7456-4214-7
ISBN-13: 978-0-7456-4215-4 (paperback)

A catalogue record for this book is available from the British Library.

Typeset in 10.25 on 13 pt FF Scala
by Servis Filmsetting Ltd, Stockport, Cheshire
Printed and bound in Great Britain by MPG Books Ltd, Bodmin, Cornwall.

# Contents


*Introduction* 1

1 The Engines 5

2 Searching 32

3 Attention 56

4 Knowledge and Democracy 85

5 Censorship 118

6 Privacy 139

7 Sociable Search 160

8 Future Finding 181

*Notes* 192
*Glossary* 195
*Bibliography* 200
*Index* 226

# Introduction

Take a moment and type the following search query into your favorite search engine: "Google is your friend." Today, the number of hits on Google stands at "about 474,000." The company is successful, but who knew it was so friendly? Even the abbreviation of the phrase – GIYF – receives about 19,300 hits. If you have picked up this book, you can probably guess the context in which this phrase is used. If not, a description may be found at http://fuckinggoogleit.com/, which reads, in part:

Google Is Your Friend

All Smart People Use Google

It Appears That You Are Not One Of Them

The search engine has become so much a part of our culture that there is a common assumption that we have found a cure for stupid questions. Folded into that assumption we find a host of others: that even the unintelligent have access to and can use a search engine, that a search engine will lead someone to a page that contains accurate information, and that questions are best directed first to a machine, and only after that to other people.

This book suggests that those assumptions are dangerously flawed; that unpacking the black box of the search engine is something of interest not only to technologists and marketers, but to anyone who wants to understand how we make sense of a newly networked world. Search engines have come to play a central role in corraling and controlling the ever-growing sea

of information that is available to us, and yet they are trusted more readily than they ought to be. They freely provide, it seems, a sorting of the wheat from the chaff, and answer our most profound and most trivial questions. They have become an object of faith.

We ask many things of search engines, what do they ask in return? Search engines are at once the most and the least visible part of the digital, networked revolution. The modern search engine has taken on the mantle of what the ancients of many cultures thought of as an oracle: a source of knowledge about our world and who we are. Children growing up in the twenty-first century have only ever known a world in which search engines could be queried, and almost always provide some kind of an answer, even if it may not be the best one.

Search engines appear to be merely a functional tool, aimed at making the real work of the web easier, but they have the potential to reveal to us not only their internal structures, but the internal structures of the societies that build them. In *Troilus and Cressida*, Shakespeare (1912, p. 35) hints at why an examination of search engines is so enticing:

> *And in such indexes, although small pricks*
> *To their subsequent volumes, there is seen*
> *The baby figure of the giant mass*
> *Of things to come at large.*

Search engines represent the screens through which we view the content of the web, screens that allow us to inflict our own desires on the "giant mass" of the web, taming it and making it useful. At the same time, the view it presents is likely to shape future social values.

In his book *Information please*, Mark Poster (2006) reminds us of the change that has occurred: we once asked people for information and now we ask machines. He is interested in how this change to machine-mediated information affects our interactions. But it is all too easy to forget that the machines

we are asking are constructed in ways that reflect human conceptions and values.

As with any kind of filter, this book comes with a bias. That bias should be clear throughout, but to make it more findable, I state it plainly here. People who use search engines – and that is slowly approaching "everyone" – should know how they work, and what they mean to society. Once they know this, they will recognize the need to take collective action and participate in the management of these technologies. The appropriate search engine does not promote authoritarian dominion over knowledge, but invites communal finding and search sociability.

The first and second chapters introduce the basic mechanics of search engines and how users employ them. At a very basic level, understanding how to effectively search using a search engine requires that we know how the system collects information and how it is best accessed. But understanding this raises further questions. In the third chapter, we visit the trenches of search engine optimization and agonistic attempts to rise in search engine rankings. While the average searcher may not recognize what goes into the collection and ranking of pages on the web, industries that thrive on attention are far more aware of the process and how it might be manipulated. In chapter 4, it becomes clear that the structure of the web, and its representation through search engines, challenges visions of a platform that encourages democratic discussion.

During the early years of the search engine, governments were nearly as blind to their role as most users. Particularly in the last few years, search engines have become embroiled in substantial policy dilemmas. In particular, interactions with national governments have resulted in search engine censorship of various forms (as discussed in chapter 5), and intrusions on personal privacy (in chapter 6). Each of these hints at deeper questions of interactions between the traditional authority of governments and the new power of search engines. In the latter case, they may also mark a shift in how we identify as individuals within our own communities.

Given these challenges, what hope is there for more accountability in search? Chapter 7 addresses the new forms of "sociable search," search systems that incorporate not only collective searching, but searching that leads to community. The oft-remarked shift on the web from static publishing to user-created media and collaborative sites, including collaborative filters and collaborative tagging, provides new, more inclusive avenues for searching and finding. The final chapter ventures to suggest some areas in which search technology is expanding, and some of the ways in which those changes may relate to social changes in the next decade and beyond.

This is an important moment in the history of search engines, the internet, and a networked global society. Search engines emerged as the epicenter of the early web, and represent a nexus of control for the future. Those who recognize their power, and are able to exercise control over it, will help to shape our collective future.

CHAPTER ONE

# The Engines

It is tempting to treat the search engine as a free-standing technology, an invention that has made it easier to find things located on another independent technology, the World Wide Web. But even a cursory investigation suggests that the search engine, like most other technologies, is not something that can be treated without reference to a larger social context, and to evolutionary social and cultural changes. The search engine, far from being an isolated modern artifact, represents a touchstone of digital culture, and a reflection of the culture in which it exists.

This chapter provides a brief overview of what a search engine is and where it comes from, and a sketch of the industry that supports it, before outlining a few of the social changes it might represent. By understanding the historical, social, technological, and cognitive contexts of the search engine, we are better able to triangulate toward an understanding of the technology, toward an indication of the place of the search engine in our world and what it portends. The permanent loss of search engines is now almost unfathomable, but were it to occur, we would find the way we communicate, learn about the world, and conduct our everyday lives would be changed. And so we must look beyond the familiar "search box" and understand what it reveals and what it conceals.

## Search engines today

A basic definition of the search engine might refer to an information retrieval system that allows for keyword searches

of distributed digital text. While a search engine is usually a system that indexes webpages, the term has been extended more broadly to include a range of information environments and media forms, including multimedia and other content found on restricted intranets and individual computers. If you ask someone what a search engine is, however, they are less likely to provide a definition than they are to indicate one of the handful of top search engines that represent some of the most popular sites on the web: Google, Yahoo, Microsoft Live, or Ask.com, for instance.

And these sites are popular. Google is easily the most popular search engine today, and the various Google sites, including its search engine, are among the most popular sites on the web. According to one measure, Google properties are the most visited sites of any kind in the world, ranking first in the United Kingdom, France, and Germany (comScore July 1, 2007). They rank third in the United States, closely trailing sites owned by Yahoo and Time-Warner, and rank third behind Yahoo and Microsoft in Asia (comScore July 9, 2007). A listing of the most popular search engines globally appears in table 1.1. A Pew Internet and American Life study carried out in 2005 found that search engine visits were growing rapidly among American users, approaching the frequency of the most popular use of the internet, email (Rainie 2005). There can be little doubt that visits to search engines make up a large

Table 1.1 Global search engine use as of July 2007

| *Search engine* | *Global share* |
|---|---|
| Google | 53.3% |
| Yahoo | 20.1% |
| MSN/Live | 13.6% |
| AOL | 5.2% |
| Ask | 1.8% |
| Others | 6.0% |

*Source:* Nielsen//NetRatings, as cited in Sullivan (2007)

part of internet use, though it can be difficult to discover just how frequent that use is, and for what reasons.

One reason for this difficulty is that people often encounter the large search engines through the façade of another site; that is, without intending to. So a search on a particular website may rely on Google to do the actual searching, or it may draw on an internal search engine. Both of these are a form of search, but may be measured differently by different research firms (Hargittai 2004). Many portal sites are also search engines, so just measuring the visitors, for example, to Yahoo properties does not provide a useful metric of actual searches. As hard as measuring the use of public search engines is, it is nearly impossible to measure search more generally: people searching their company intranet or their hard drive, for example.

Particularly in recent years, there has been a rise in specialized search engines that seek to index not the entire web, but some constrained portion. This is often referred to as "vertical search," as opposed to the "horizontal search" of the general purpose search engines. Topically constrained search engines seek out only pages within a particular knowledge domain, or of a particular type of content. Some of these sites are efforts to move databases that have traditionally been found in libraries onto the web. ScienceDirect, for example, provides reference to scientific literature for web users, and Google Scholar provides the utility of a large article and citation index provided by scholarly journals combined with scholarly sources from the web. Some of these vertical search engines are focused on a particular industry. For example, an attorney in the United States might turn to open sources like FindLaw to provide news and information about their practice, to Lawyers.com to find an attorney within a particular practice area, or to THOMAS, a search engine maintained by the federal government to track legislation, in addition to a number of subscription-based search engines for legal information like Westlaw and Lexis-Nexis.

Some niche search engines focus on a specific geographical area. "Local search" has taken over the function of many local

telephone directories, providing information about local services. Rather than competing with local search, many of the largest business directories ("yellow pages") have created their own local search engines, as have local newspapers and television stations. Local search is often combined with mapping services, and there is an opportunity to create what is often called "locative media," providing information that depends on the geographical context of the individual, determined through GPS and other mobile devices. Others cater to particular modes of delivery; Google, for example, offers information about local business via voice for telephone users (Cheng 2007).

Sometimes it is not the content area that determines the search, but the type of media. Although large search engines, beginning with AltaVista, generally have had some ability to search for multimedia, it continues to present some of the greatest challenges, and some of the greatest opportunities for those creating vertical search engines. Sweden's Polar Rose, for example, seeks to use face-recognition technology to identify individual people in photographs on the web (Schenker 2006). Right now, searchers have to rely on accurate surrounding text, which may or may not identify individuals in the photographs, but this opens up photographs on the web in a whole new way.

Some of these efforts toward vertical search are created by groups with a particular interest in a certain narrow topical or geographical area; a good search engine may be seen as a way of promoting a particular topic, language, or region. Alternatively, for those who wish to create new kinds of search engines, it may make sense to try to capture a smaller audience, rather than go into head-to-head competition with the giants, at least in the early stages (Regan 2005). In many cases, however, the giants are now creating their own niche sites, or acquiring sites that show potential, in order to dominate the entire search market and integrate their services.

Many of the new approaches to cataloging the web attempt to leverage search technologies by combining them with tacit

and explicit coding of content by individual web surfers. While the taxonomic structures of some web directories, or the tagging structures of "folksonomic" sites like Flickr and del.icio.us, may seem to represent alternatives to search engines, in fact they both depend on the idea of search, and enhance search functionality. Many search engines trace the search behavior and even the surfing behavior of their users to better anticipate effective search results, or to better organize their indexes. Though the initial search engines marked a move away from human coding, clever engines now extract patterns from their users to exploit their social sense of what is an appropriate search result. It is likely that search will continue to become more closely tied to social relationships, moving to provide information not only about text, but about people who may have the expertise a searcher is seeking.

Although they may not be called "search engines," these technologies extend into even narrower domains. Early websites gradually adopted the practice of tree-like organizational structures, and eventually other ways of indexing content on the site. Especially after about 2000, though, the search box became a fixture on most sites of any size. Not only do people expect to see a search box, they expect it to behave in standard ways; anything outside of the expected will frustrate the average web user (Nielsen 2005). The search engine is now ingrained in our web experience, appearing as a box in the corner of e-commerce sites, personal blogs, dating sites, hospital websites, and nearly everything in-between.

Search engines have recently reached beyond the web. Of course, there have always been large collections of data that needed to be indexed in some way, particularly in libraries, but the specific technologies developed for search engines are now frequently to be found as a way of seizing hold of otherwise unmanageable unstructured and heterogeneous stores of data: email, documents, and the entire contents of home computers and corporate networks. As our use of digital media converges, mixing and combining computing applications

with more traditional media, we also find search engines becoming a part of our entire media ecosystem. It might once have been considered odd to search for a specific piece of information among a friend's collected email correspondence, a week's worth of your own television viewing, a novel, or the address book on your mobile telephone, but search is now becoming an expected feature of many previously unindexed collections of data.

Because the web is becoming an ever-expanding database of human knowledge, it represents the greatest challenge for those wishing to create systems to collect, summarize, organize, and retrieve information. Naturally, these tasks have existed before, but the size, extent, and diversity of the content of the web make it the ultimate target for such efforts. As a result, those who would have studied other topics in artificial intelligence, information design, library science, and a host of other fields have set their sights instead on developing a better search engine.

## Before the search engine

Some consider the greatest modern threat to be too much information, a glut of data that obscures what is really valuable. In his book *Data smog*, David Shenk (1997, p. 43) argues that computers are the "most powerful engines driving the information glut" by constantly drawing more data to our attention. While it is undoubtedly the case that the internet allows for the rapid delivery of ever growing amounts of information, it is also true that new computing devices were often created in order to manage and control increasingly complex environments. What once could be handled by a human, or a collection of individuals, became too time-consuming to result in effective control. So in 1823, when the British government recognized the need for an effective replacement for human "calculators" to come up with tide tables at their ports, they funded an effort by Charles Babbage to design the first mechanical

computer (Campbell-Kelley & Aspray 1996). Likewise, when the United States government found that it would take more than ten years to tabulate the decennial national census in 1890, they turned to Herman Hollerith, who founded the company that later became IBM, to create an automatic tabulating system (Aul 1972). That pattern of turning to information technology when faced with an overwhelming amount of data has occurred over and over: in libraries, in large businesses, and, eventually, on the World Wide Web.

It is natural to think of information technology as digital computing, since so much of contemporary information processing is relegated to networked computers. Computers are only the most recent in a long line of technologies that were created to allow for better control of complex collections and flows of information. The obvious example is the library: once a collection of books and papers grows to a significant size, finding the appropriate piece of information in a timely manner becomes the subject of its own techniques, records, and machinery. Collections of documents can be traced back nearly as far as history itself has been recorded; were cave drawings the first private libraries? As Kaser (1962) explains, many spiritual traditions conceive of the library as eternal, and the librarian as all-powerful. As early private collections grew larger, librarians emerged to organize and manage these collections. Because libraries were so important to many classical civilizations, the librarian was in a revered and politically powerful position which required special skills in collecting and manipulating information. Large libraries have always been a nexus of potential information overload, and so techniques and technologies evolved to help us filter and find information.

Sorting and finding items within these collections required the creation and maintenance of information about the collection: metadata. The Babylonian library at Nippur had such records of the collection as early as the twentieth century BCE. The nature of the need was simple enough: the librarian needed to be able to discover which books or documents addressed a

given topic, and then find where that book was physically located so that it could be retrieved for the person requesting information. Given that the subject of a work was often the issue most closely indexed to an informational need, the most popular indexes in the English-speaking world – the Dewey Decimal system and the Library of Congress System – provide a classification that is based on the subject matter of a book, so that books on similar topics are likely to be found in close proximity.

The use of computing systems in libraries has formed an important basis for how search engines now work. There is a long history of ideas about how to organize knowledge in the library, but the rise of computing in a library setting brought mathematics and linguistics to bear in new ways, and many of the techniques now used by search engines were first used by library indexes. The field of Information Retrieval (IR) now bridges the closed library index and the wider collection of documents on the web (Salton 1975), and draws from many areas of computing and information science to better understand the information available over computer networks.

Public and private libraries were not the only form of data collections. The industrial revolution led to new forms of social organization, particularly the rise of bureaucracy, which required a flood of new paper files. Records and copies of correspondence were generally kept on paper, and guides emerged for suggesting the best ways to organize these materials, including the best ways to stack papers on a desk. Paper stacking gave way to pigeonholes, and the business titans of the early twentieth century made use of a fabulously expensive piece of office furniture called the "Wooton desk," which contained hundreds of pigeonholes and could be closed and locked, allowing for the secure storage and access of personal work documents. The gradual development and innovation that led to vertical filing – a technology, perhaps unsurprisingly, developed by the inventor of the Dewey Decimal System – was a result of a data glut that began a century before anyone uttered the word "internet" (Yates 1982).

While subject-oriented classification made sense for the broad and relatively slowly changing materials of a library, it would have been useless when applied to the office of the last century. First, time was very much of the essence: *when* a document or file was created, changed, moved, or destroyed was often as important as the document's subject matter. Likewise, such records were often closely related to the people involved. Clearly this was true of customer records, and large insurance companies – whose very survival rested on increasing the size of their customer base – often drove innovations in business filing, right through to adopting the earliest electronic computers.

The earliest computer systems drew on the ideas of librarians and filing clerks, but were also constrained by the technology itself. While these earlier approaches provided metaphors for digital storage, they failed to consider the hardware constraints posed by the new computing devices and placed limits on the new capabilities of these machines. Computer programmers made use of queues and stacks of data, created new forms of encoding data digitally, and new imaginary structures for holding that data. Not housed in drawers or on shelves, these collections could be rearranged and cross-indexed much more quickly than their physical counterparts. Over time, this evolved into its own art, and database design continues to be a rapidly advancing subfield of computer science.

As the internet began its exponential increase in size during the 1990s, driven by the emergence of the World Wide Web, it became apparent that there was more information than could easily be browsed. What began as the equivalent of a personal office, with a small private library and a couple of filing cabinets, grew to rival and exceed the size of the largest libraries in the world. The change was not immediate, and, in the early stages, individuals were able to create guides that listed individual collections at various institutions, generally consisting of freely available software and a handful of large documents. Especially with the advent of the web, the physical machine

where the documents were stored began to matter less and less, and the number of people contributing documents grew quickly. No longer could a person browse the web as if it were a small book shop, relatively confident that they had visited each and every shelf. Competing metaphors from librarians, organizational communicators, and computer programmers sought out ways of bringing order, but the search engine, in many ways, was a novel solution for this new information environment.

## How a search engine works

Although the search engine has evolved considerably over time, all search engines share a common overall structure and function. Before outlining their development and commercialization over time, it is useful to understand how a basic search engine works. Our interaction with search engines, as users, is fairly uncomplicated. A website presents a box in which we type a few words we presume are relevant, and the engine produces a list of pages that contain that combination of words. In practice, this interface with the person, while important, is only one of three parts of what makes up a search engine. The production of the database queried by the web form requires, first, that information about webpages be gathered from around the web, and, second, that this collection of data be processed in such a way that a page's relevance to a particular set of keywords may be determined. By understanding the basic operation of each of these steps and the challenges they pose, an overall understanding of the technology may be reached. Figure 1.1 provides an overview of the process common to most search engines.

The process begins with a system that automatically calls up pages on the web and records them, usually called a *crawler*, *spider*, *web robot*, or *bot*. Imagine a person sitting at a computer browsing the web in a methodological way. She begins her process with a list of webpages she plans to visit. She types the

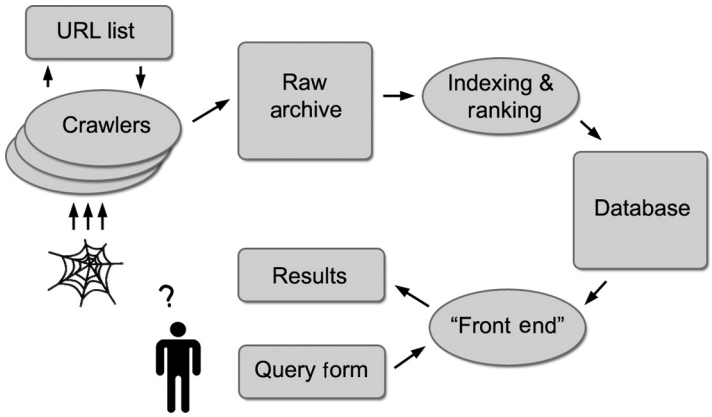


Figure 1.1. Conceptual organization of the typical search engine

URL for the first of these pages into the browser. Once it loads, she saves a copy of the page on her hard drive, noting the time and the date. She then looks through the page for any hyperlinks to other pages. If she finds hyperlinks that are not already on her list, she adds them to the bottom of the list. Following this pattern, she is likely to record a large part of the entire web. Once complete, she would begin again from the top of her list, as there are likely new pages that have been created and linked to since she began.

If the search engines really relied on individual humans to do this, it would take thousands of years to complete even a single crawl of the web. However, the operation described is not particularly complex, and creating a computer program that can duplicate this behavior is not difficult. Because the crawler is a relatively simple piece of technology, it has not evolved as much as other parts of the search engine. Even the smallest-scale crawlers are usually multi-threaded, making many requests at the same time rather than waiting for each page to be produced before moving on. They generally run not on a single computer, but on a large number of computers working in tandem. Most are careful to distribute their requests across the web, rather than ask for all of the pages

from one server at once, since the crush of requests could easily overwhelm a single server, and most are "polite," taking into account webpage authors' requests for certain pages to be ignored.

That does not mean that crawlers are all the same. There is an entire menagerie of crawlers out looking for new content on the web. On many pages, visits by web robots outnumber visits by real people. Some of these – going by exotic names like Slurp and Teoma – are gathering information for the largest general-purpose search engines, but others may be run just once by an individual. Small crawlers are built into a number of applications, including plug-ins for browsers and a robot used by Adobe Acrobat to create a PDF from a website. Because of small differences in how they are programmed, they behave slightly differently, following some links and not others, or coming back to re-check more or less frequently. There are a number of people who are trying to figure out just how these robots work so that they can ensure their message is presented to as many search engines as possible (Valentine 2005).

However, following hyperlinks may not be enough. Large portions of the web are now generated dynamically, according to various requests from website visitors. Think, for example, of an online site that provides theater tickets. The calendar, the pages describing available tickets, or even the seating maps may change depending on the show, the location of the person accessing the site, the current date, previous sales, and other variables. Because these are not static, hyperlinked pages, they are not easily accessed by most crawlers, and are part of what is sometimes called the "deep web" beyond the reach of most search engines (Sherman & Price 2001). There are substantial technological and potential legal barriers to accessing this hidden web, but there continue to be efforts to create smarter crawlers able to index dynamic pages (Ntoulas, Zerfos, & Cho 2005). This is particularly true for search engines that are focused on ferreting out the best price among competing