SAMPLE CHAPTER

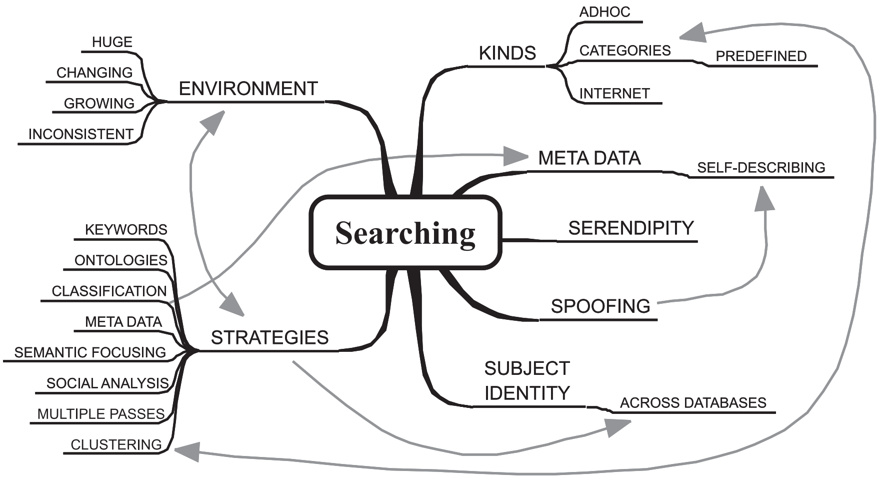
EXPLORER'S GUIDE TO THE SEMANTIC WEB

Thomas B. Passin

MANNING

5

Searching


Searching
ENVIRONMENT
HUGE
CHANGING
GROWING
INCONSISTENT
KINDS
ADHOC
CATEGORIES
PREDEFINED
INTERNET
META DATA
SELF-DESCRIBING
SERENDIPITY
SPOOFING
SUBJECT IDENTITY
ACROSS DATABASES
STRATEGIES
KEYWORDS
ONTOLOGIES
CLASSIFICATION
META DATA
SEMANTIC FOCUSING
SOCIAL ANALYSIS
MULTIPLE PASSES
CLUSTERING

> *...which gave the Pigeon the opportunity of adding, "You're looking for eggs, I know THAT well enough; and what does it matter to me whether you're a little girl or a serpent?"*
>
> —Lewis Carroll, *Alice's Adventures in Wonderland*

Anyone who searches for specific information on the Web learns to put up with frustration. Sometimes things go well, but other times irrelevant information buries those bits we seek. Here are two typical examples from my experience.

First, I want to find the home page for the Python XML project. I type "python xml" into the Google search page. In less than a second I get back 10 pages of hits, but soon I find that the third item is the one I want. If I remember that the project is also called pyxml, I still get 10 pages of hits, but now the first hit is the page I want.

Next, I need to get a replacement cooktop heating coil for my old stove, and I want to find a store in my area. Because I'm working on this chapter, I decide to try an online search. I enter the following search phrases in order, based on words I saw in otherwise useless hits:

- burner jennair
- burner jennair replace
- burner jennair new
- burner jennair distributor
- jennair distributor va
- jennair distributor virginia

The last phrase gives a few hits—only six—but they're useless too. They include a site that offers agricultural businesses for sale.

I know that I've saved a bookmark in my browser for a real store that carries the replacement parts I need. I find the bookmark and enter its URL into the search engine. I get back a page that contains the search engine's category for the URL and also a description. The category seems to make sense:

```
Shopping > Home and Garden > ... > Appliances > Parts
```

It's also a clickable hyperlink that gives me a page of 14 online appliance dealers. Adding the stove brand to the search finally gives me a list of seven dealers that apparently handle the brand. However, none of them are actual stores, let alone stores in Virginia, and the list doesn't even include the original store whose URL I tried. Later, I discover that, had I typed "jennair" in the first place, the return page would have displayed the category for the query. This would have helped, except that it's wrong—the site changed its categories and failed to update that part of the database.

These two experiences epitomize the strengths and weaknesses of searching the Web today. Search engines have indexed a staggering number of web pages, but a user has no good way to know which search terms have good chances of success. And each user has a different and often imprecise mental hierarchy of categories and terms, none of which may match what the search engines offer.

5.1 Searching the Web

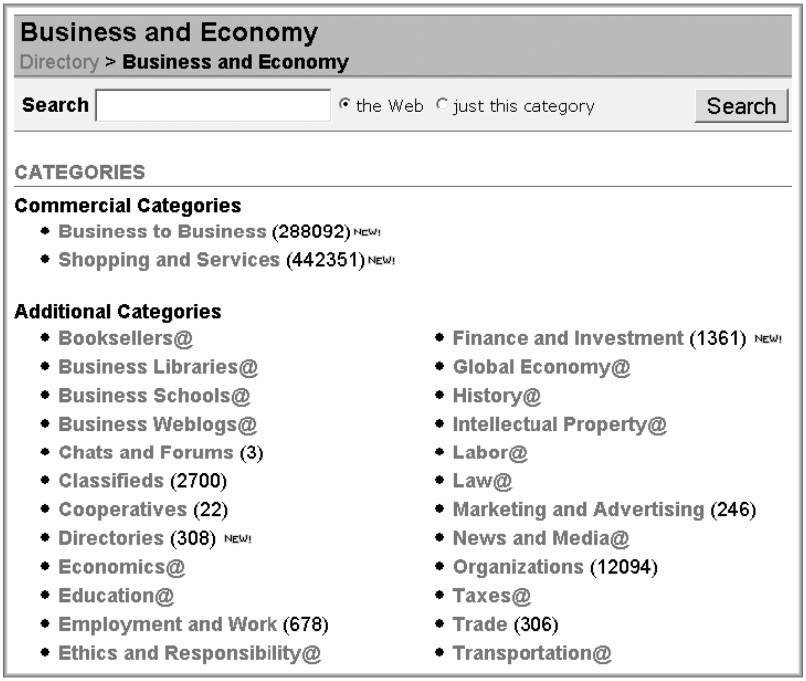
Surely every reader of this book has performed searches. On the Internet, there are search engines like Google and Altavista. Libraries have their card catalogs and electronic search terminals, and books have their indexes. These are also varieties of search aids. This section looks at current Internet searches, to suggest potential areas where the Semantic Web could bring substantial improvement.

5.1.1 Kinds of searches

The act of searching for information can be classified in many ways. One way is by the location—you can search your own computer, or your company's intranet, or the Internet. Another classification concerns the type of query. The distinction here is between an ad hoc search (which is what you do when, for example, you type "black Labrador kennel" into a search engine) and a more structured search in which you use predefined categories set up by the search site. Figure 5.1 depicts some of Yahoo's (www.yahoo.com) predefined categories.

A different kind of distinction can be made between a closed system, like a stable collection of reference documents that are similarly structured, and an open system containing a large number of different kinds of information. Library science knows many specialized distinctions between kinds of information to search for: inventory control, establishing the provenance of documents (their source or origin), locating similar items, finding similar sets (such as different editions of the same work), and many more.

In this book, we're most interested in information that can be found using the Internet and that can preferably be used by both people and computers. This covers quite a range.



Business and Economy

Directory > Business and Economy

CATEGORIES

Commercial Categories

- Business to Business (288092) NEW!
- Shopping and Services (442351) NEW!

Additional Categories

- Booksellers@
- Business Libraries@
- Business Schools@
- Business Weblogs@
- Chats and Forums (3)
- Classifieds (2700)
- Cooperatives (22)
- Directories (308) NEW!
- Economics@
- Education@
- Employment and Work (678)
- Ethics and Responsibility@
- Finance and Investment (1361) NEW!
- Global Economy@
- History@
- Intellectual Property@
- Labor@
- Law@
- Marketing and Advertising (246)
- News and Media@
- Organizations (12094)
- Taxes@
- Trade (306)
- Transportation@

Figure 5.1 An example of predefined categories, as presented by the Yahoo search site. Selecting any one category brings up a page presenting a more specialized list of categories. Categories like these work well when they match the words a person thinks of. When the categories don't match well, it can be hard to decide which ones to choose.

5.1.2 *So near and yet so far*

Today, Internet searches are usually done by people who type words into the input box of a search engine's web page. Some people also run search engine programs that index their hard drives and find files by name or text phrases. Either way, the results come back as text listings that a person can read, perhaps with hyperlinks to other web pages.

Search engines are also starting to provide interfaces so that programs can request and receive information; Google has such a service, for example. However, no standards exist yet for either the form of the request or the form of the response.

Web users everywhere have shared the experience of trying to guess the right words that will return a useful result, as the second vignette at the start of the chapter exemplifies. Some sites give no guidance; some, like Yahoo and Open

Directory (www.dmoz.org), have a hierarchy of terms you can navigate like an index; and some provide active help in focusing your search by showing additional terms to add to your search, like the wonderfully creative graphic interface at KartOO (www.kartoo.com/).[1]

Whatever search site you use, it can be difficult to get useful results. Either you get too many hits, or none of the results gives the information you want, or you get general information when you wanted specifics. The best search engines are much better at discovering useful results than lower grade sites, but the quality of results is erratic even at the best sites.

It can also be difficult to use search results in a program. A program that wants to look for a book in many different bookstores may receive data from each bookstore; this data can be read by a person, but it's difficult to combine and process automatically. The sites use different and incompatible data formats that don't lend themselves to machine processing. Fortunately, as search engines begin to return data in XML formats, this situation is improving.

Search engines today are sometimes so effective that it can be faster and easier to do a search than to find a site in your list of saved bookmarks. When I planned a vacation in Alaska, I found that I could type "alaska railroad schedule" into Google to get railroad schedules more easily than I could navigate to the bookmark saved in my browser.[2] To my way of thinking, for a large and distant search engine to find a reference faster that I can in my list of bookmarks is a stunning accomplishment, notwithstanding the limitations of current web searches.

5.1.3 Behind the scenes

The Internet must be the most difficult environment conceivable for search engines. It's huge, growing, always changing, and inconsistent, and it contains documents of all kinds and structures. Those documents may contradict each other, and no central registry lists them all. Of course, these facts are also precisely the strength and richness of the Internet, the reason it's so interesting and vital.

[1] Trying my search for the stove part on KartOO resulted in a small number of hits, essentially the top hits from the regular search engine. But the ingenious graphical display made it easy to see which hits might be useful. In fact, one of the hits led through an intermediate link to the manufacturer's page, where I got a listing of dealers near my home.

[2] Since I began to use the bookmark application discussed in the appendix, I can usually find my own bookmarks faster than using Google.

What Internet search sites do behind the scenes is amazing. They deal with so many information sources—millions or hundreds of millions of documents—and they usually give results within seconds:

1 A search engine must find web sites and documents for indexing (and there's a huge number of candidates).

2 The engine must analyze each page and each web site in many ways to be ready for the unknown queries that users will submit. The relationships between linked pages must be analyzed and a strategy developed to deal with those linked pages. Many sites *cache* (that is, save copies of) the pages they index in case the original page becomes unavailable. The search engine must analyze the contents of a wide range of document types, from the highly structured to the almost unstructured, from the informal to the highly technical.

3 Once the information has been stored and indexed for later retrieval, the search engine has to analyze queries, locate candidate results in its huge databases, select the best results from those candidates, possibly group them into sensible divisions, order them by their relevance, and present them to the user in a very short time.

Later in this chapter, we'll look at how Semantic Web technologies can help to improve this process, and how the results from a search could be made more useful for computer consumption.

SERENDIPITY: SURPRISE AS A GOAL

Search sites and users place a lot of emphasis on returning results that satisfy the original request as closely as possible, even when that request may have been vague. But it can be valuable for people to discover things they weren't looking for and didn't expect. In pre-Internet days, when we spent time in libraries (some of us still do), it was common to stumble across something tremendously interesting but previously unknown. (I myself acquired two new hobbies that way, thanks to the excellent library system of Shaker Heights, Ohio.)

We don't want to lose those opportunities in the name of efficiency and focused results. This is a subject I haven't seen discussed, and it shouldn't get lost. As I prepared to write this chapter, I accidentally stumbled across an unexpected and amusing site, using the AltaVista search engine (www.altavista.com): the Political Impersonator Speakers Bureau site (www.counterfeitbill.com/), which has links to comedians who impersonate presidents and other politicians. I was glad to have found it.

5.2 Search strategies

In this section, we look at some issues involved in searching. It isn't a comprehensive and detailed discussion, but it tries to frame the potential contributions of a future Semantic Web. At the time of this writing, no standards exist for indexing, cataloging, analyzing, or specifying the quality and performance of search engines, nor for specifying request and response message formats. Therefore we can't give specific technical examples of Semantic Web–oriented standard technologies—although many research papers delineate experimental approaches.[3] Of course, we expect the usual suspects—RDF, Topic Maps, and OWL (see chapters 2, 3, and 7)—to play a prominent role. Topic Maps could be used to organize and structure data, and OWL could define categories and relationships between them.

5.2.1 Keywords

The most straightforward approach to indexing the Web is to search web resources for a list of keywords and store their locations. When the user types in one or more search terms, the engine attempts to find them in its index; if it can't, it tries to find equivalent terms that it does contain. Pages known to contain the keywords are returned.

Aside from the fact that pages not containing any of the keywords may not be returned, there is the *natural language* issue. In language, as people use it every day, words can have many different *senses*, or classes of meaning. For example, according to the WordNet lexical dictionary (WordNet), the word *tool* has four senses when used as a noun and four more when used as a verb (for example, to "work with a tool" and to "joyride"). Some of these senses are closely related, and some aren't. To make effective use of keywords, the search engine must figure out the right sense when it analyzes a page and also when it analyzes a query. This gets into the difficult area of natural language processing (NLP). In addition, the search engine normally receives no context for analyzing the query; but in natural languages words often have different meanings according to the context (polysemy), which can lead to ambiguity.

[3] Many research papers discuss searching—far more than could be reported in this book. This chapter mentions only a few efforts that seem likely to be useful now or in the relatively near future, and that clearly involve Semantic Web technologies.

Another problem is that users often don't think of the same keywords that the search system uses. Keywords, although they can be useful, don't provide enough utility to support the kinds of search activity that goes on over the Web.

5.2.2 Ontologies

The classic AI-oriented (Artificial Intelligence) view would be that by classifying the words and concepts in web resources into categories defined by a suitable ontology, the essential contents of the resources could be captured and matched to the concepts behind the categories. Failing that, the terms used in a query could at least be automatically related to known terms from the ontology. Alternatively, terms from the ontology could be presented to the user, who would pick one or more for the query. This approach would be made more feasible if web pages were marked up with the right information.

There are two issues here: the analysis of documents on the Web and the creation of queries. For the creation of queries, the use of terms from ontologies hasn't been demonstrated to be nearly as effective as you might think. It can be useful in relatively small, closed systems where the vocabulary is small and well controlled, but a query for information on the Internet is very different.

A person who wants to find information is unlikely to have the same working ontology as the system uses for its classification. In other words, people have different associations to different words and concepts, and one person's hierarchy of concepts may not suit another's. In addition, the use of words is dependent on the context of the query and the expectations of the searcher. A real, useful ontology will probably be too large for a user to easily look through to find the right terms to use. Finding the proper terms can become a search exercise in itself, as I found when I tried to locate a heating coil for my stove. I started with *hob* because that's the term used in the restaurant industry, switched to *burner,* and ended up with *appliance* and *jennair distributor.* It's also hard to devise a good user interface that allows a user to choose terms from an ontology without disrupting their thoughts.

Whatever the reason, to date, the use of ontologies hasn't been particularly effective for helping users construct queries that produce good results. Note that this isn't the same as saying that that classification of terms is useless during analysis of either documents or queries, nor that classification systems can't help a person to craft a query. Still, ontology-based queries don't seem to be the largest part of the answer. This amounts to observing that a good index in a book can be invaluable for finding specific material; but for a large book written about many different subjects by many different authors, it's much harder to find what you want, index or no.

On the other hand, an approach in which the computer silently augments a query by using an ontology to relate the query terms to a knowledge base will probably be much more successful (see section 5.2.7).

5.2.3 Meta data

There are several kinds of meta data about any document or resource. There is meta data that is explicitly contained in a work, such as its author, keywords about its content, and its publisher. Other metadata is published separately from the work, perhaps in reviews or in annotations (see chapter 4 for more on annotations). In some cases, it may be possible to infer meta data that isn't explicitly stated.

Today, most documents and web pages contain little meta data. It has often been assumed that one of the most important consequences of the Semantic Web will be that most resources will be marked up with meta data in a few standard formats. It's undoubtedly true that more pages will be marked up with meta data, but it seems certain that most resources on the Web will continue to lack labeled meta data. Most web pages won't contain markup that says, "This is an important concept," or "This page uses terms drawn from the XYZ ontology."

Therefore, search engines need to analyze resources to deduce or extract the implicit and explicit meta data they may contain, and to do so even without marked-up sections in the documents. Some of this already takes place, and quite a lot of research focuses on such analysis. Two approaches, which may also be combined, appear frequently in the research: analysis of the language of the document (more in the next section) and analysis of common structural patterns to identify parts of particular cogency. Typically, such parts are consecutive sentences, or paragraphs with links, or runs of sequential paragraphs with certain kinds of links and terms. Even the visual layout of web pages has been shown to provide helpful information for identifying important concepts and links.

Identification of sections that have strategic hyperlinks is sometimes combined with what the discussion in section 5.2.6 calls *social analysis*. Social analysis is a key feature of Google's approach.

5.2.4 Semantic analysis

All web resources are about one or more themes and contain some number of concepts. Most of the time, the themes and concepts are expressed in natural language. If these concepts can be discovered and matched with the concepts contained in a query, it ought to make for better retrieval of the desired information. The discovery process will probably entail classification according to an ontology (as covered in section 5.2.2).

In the future, some resources' major concepts will be marked up with a standard language like Resource Description Framework (RDF). When this happens, it will make the analysis much easier. But a vast number of pages and documents aren't so tagged now and probably never will be. A great deal of research has gone into the subject of automatic analysis of natural language; it's a large and difficult subject. To be practical for search engines, the analysis must not only be reasonably accurate but also very fast. As another benefit, natural language analysis can also help a search engine to understand queries and to let a user type queries that are more like everyday questions.

However, there's more to it than just inferring concepts. Suppose you type in the name of a well-known person. Some research papers use Bill Clinton, the former American president, as an example. What kind of information should a system return for the query "Bill Clinton"? A biography? His current health? A history of the Clinton presidency? His email address? Let's try it: Table 5.1 lists the top hits for several major search engines.[4]

Table 5.1 Top search engine results for the query "Bill Clinton"

Search engine	Top results
AltaVista	Counterfeit Bill Productions—George W. Bush, Laura Bush, Bill Clinton, Hillary Clinton, Jesse Ventura, and Al Gore The "Unofficial" Bill Clinton New Book of Knowledge: Bill Clinton
Google	Welcome to the White House Clinton Presidential Center The "Unofficial" Bill Clinton
Netscape Search	Welcome to the White House Clinton Presidential Center www.whitehouse.gov/WH/Mail/html/Mail_President.html
Teoma	Clinton Presidential Materials White House Search Engine Clinton Library Book—satirizing the Clinton Library aka Clinton Presidential Center LindaTripp.com The Journal of Linda Tripp's fight for justice against the Bill Clinton Whi…
Ask Jeeves	Clinton Presidential Materials White House Search Engine Clinton Library Book—satirizing the Clinton Library aka Clinton Presidential Center Town Hall: Conservative News and Information—The Conservative Movement Starts Here

[4] Of course, these results will probably be different by the time this book is published.

The top response of both Google and Netscape is completely inappropriate, since at the time of writing Clinton hasn't been president for some time. All five search sites have satirical or comic sites in the top three, which probably isn't what most people expect (although it's a good example of the serendipitous response mentioned earlier). Ask Jeeves also produced a sidebar of suggested topics, each a hyperlink of its own, to narrow the search:

- Bill Clinton, Biography
- President Clinton
- Impeachment of Bill Clinton
- Monica Lewinsky
- Bill Clinton Jokes
- Bill Clinton Photos
- Bill Clinton Schedule
- Clinton Lewinsky
- Bill Hillary Clinton
- Bill Clinton Picture

Although helpful, the first return under "Bill Clinton, Biography" was for a biography of Hillary Clinton, not Bill.

All in all, the hits have some cogency but aren't particularly responsive. But what should the response have been, with such a plain question? That isn't easy to know, and the better search engines try to come up with something that makes sense, such as biographic material or primary employment.

The methods that commercial search engines use at any particular time are generally not known in detail. The principles behind Google were published while it was still an academic project, but no doubt it has extended its methods since then. The results of our test search do indicate two things. First, a number of sites evidently use some kind of semantic analysis, because the results tend to be about Clinton rather than, say, collections of his speeches or lists of old addresses. Second, the elevated position of satirical pages seems to reflect a measure of their popularity. How else would such pages rate so highly? (Section 5.2.6 goes further into this question.)

Clearly, the use of semantic analysis to discover concepts and word senses in resources can help find the right information during a search. Just as clearly, an understanding of the intended meaning of the search request is needed for searches to become substantially more effective. The next section discusses how a system could improve its understanding of a query.

5.2.5 Semantic focusing

What can be done at the query end to improve a search engine's understanding of a query? In some way, the query needs to be put into a context that will help a search engine do a productive search.

Context from the user

When a user makes a query because they're reading part of a document, the system could analyze that part of the document. Suppose you're reading a news story. You highlight a phrase and ask for more information. The system would try to analyze that part of the story, including any hyperlinks. The analysis would yield concepts and relationships between them, and the system would use them to augment the query.

This process would be useful even if the story contained no semantic markup. If it did, the system would be able to do its job with more precision. If the user's system could enhance the query using terms and concepts known to the search engine, the whole process of query-search-results could be much more effective.

Context from databases

Another strategy is to try to classify the user's query and then search a knowledge base using that classification. Information from the knowledge base would then be used to enhance the query sent to the search engine. Even better would be to combine information about the subject from several knowledge bases. For this to work, the system must be able to identify the subject of a query, even if it was called different things in different places. The project known as *TAP* (http://tap.stanford.edu/), being developed at Stanford University, has evolved a scheme for doing this that dovetails well with the design of RDF.

TAP has developed a way to ask a server to supply a graph containing information about a subject. A server that receives a request for information on "a person whose name is 'Tom Passin', who is writing a book about the Semantic Web, who has an email address of tpassin@example.com, who lives in such-and-such a town, and who is interested in flying" might find that it has data on a person named Thomas Passin, who lives in the same state as the town, who has a silver car, who has an email address of tpassin@example.com, and who likes flying and music. Putting the two together, the system can discover that there is a person who has the two names Tom Passin and Thomas Passin, who has an email address of tpassin@example.com, who lives in such-and-such a town in such-and-such a state, and who likes both flying and music.

TAP has built a knowledge base that can identify many well-known people and concepts. One example from TAP documents is a query for information about Yo Yo Ma, the famous cellist. The query is likely to be about music the cellist is involved with, and TAP can discover this from its knowledge base. So, TAP tries to get information about recordings, concerts, musical venues, and so on that relate to Yo Yo, as well as his current concert schedule. All this information is assembled and presented to the user. The result seems like an ordinary search result, except that it's much more focused on music-related information than usual. Most of the typical scattered, irrelevant results are gone. (The TAP site doesn't explain how to get non-music-related information about Yo Yo. Presumably this can be done as well.)

Subject identity—again

Recall from chapter 2 that RDF saves its information as a series of simple statements called triples, because they have three parts—for example, (Mycar, hasColor, Silver). Also recall from chapters 2 and 3 that a name may or may not identify a resource. RDF uses unique URIs as identifiers. But it's also possible to describe something without using its identifier: "my car is the small silver convertible parked near the end of the block."

Some statements are good for identifying their subject, and some aren't. "The person who has Social Security number xxx-yy-zzzz" is good for identifying a person in the U.S., because only one person is supposed to have a given Social Security number. By contrast, "The person who has brown hair" isn't good for identification, because many people have brown hair. Now, suppose we have some statements about a person in one database and some different statements in another database. If they're about the same person, and they're sufficiently good for identification, then certain parts of the two graphs—the two graphs that represent the data in the two databases—will match. The match can indicate that the two are really the same, as illustrated is figure 5.2.

In this way, the information in one database can sometimes be related to that in another, even though neither one knows the other's identifiers. (TAP) has more on this subject.

5.2.6 Social analysis

The apparent influence of popularity on the results found in section 5.2.4 is no accident. In the study of the influence of scientific research, one of the most effective measures turned out to be how often a given work is cited in other research papers. In a like manner, some search engines (notably Google) analyze

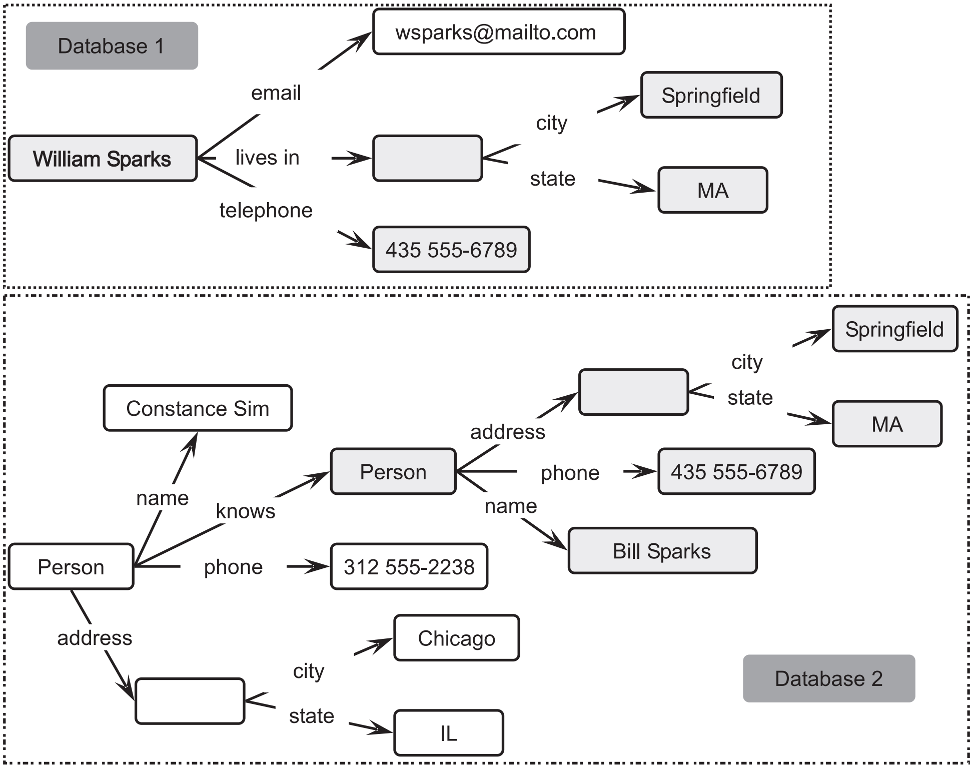


Figure 5.2 Data in two databases, depicted as RDF graphs. The data is structured differently in the two databases, but there are some similarities. A suitable processor could match "lives in" from database 1 with "address" from database 2. It could then discover that parts of the two graphs have essentially the same shape and type of connections. From this, it could infer that database 2 may contain the same person, William Sparks, as database 1, because the address has the same city and state, and the telephone numbers match. The fact that "Bill" is a common alternative for "William" strengthens the conclusion. The matching parts of the two graphs are shaded.

the pattern and number of links to a page, and the results play an important role in ranking a page's importance.[5] The analysis can become complicated, with certain pages being weighted more heavily depending on the patterns of links coming to them. The more a page or web site is judged to be of high quality, the more influence it has on the ranking of a page.

[5] Google calls its system PageRank. You can read more about PageRank on the Google web site at www.google.com/technology/ (they also have a delightful spoof of this at www.google.com/technology/pigeonrank.html) and in the research paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine" by Sergey Brin and Lawrence Page at www-db.stanford.edu/pub/papers/google.pdf.

What could be more web-like than using the very hyperlinks that characterize the Web to guide the discovery of information? Of course, popularity alone doesn't make something accurate (think of urban legends that are repeated over and over), but social ranking has a long history. Polls are a form of social ranking. Amazon.com, the bookselling site (which has branched out and now sells all kinds of other merchandise), makes it easy to see the opinions of other people who have bought the same item. Amazon includes the entire text of people's reviews, so these ratings affect buying decisions.

The ability of technologies like RDF to annotate resources (see chapter 4) could contribute powerfully to social analysis and thereby play a role in search strategies.

5.2.7 *Multiple passes*

It's clear that no one approach (or even two) to indexing Web resources and analyzing queries is enough. An approach that is frequently used in research papers and is now starting to show up on some search sites is to make a preliminary query using a web site like Google (or more than one site) and then to perform additional analysis. In (Amitay 2000), the work starts by searching for "Albert Einstein" on Google and other search sites. Each hit is analyzed by computer to find the best description of its contents, which are then presented to the user or filtered out. The hits are analyzed by looking for certain structural features in the pages; other research uses different strategies.

Some of these experiments have apparently delivered much better results than the methods normally used at present, with many fewer irrelevant hits and fewer cases of omitted pages that should have been returned.

5.2.8 *Peer to peer*

Another search strategy is to use peer-to-peer networks to allow a query to be answered by other computers that are likely to know useful information. The NeuroGrid project (www.neurogrid.net) takes this approach. A person who participates stores bookmark references into a database, which can be shared with other participants. A query is directed by the system to those peers most likely to answer it, based on past experience. The NeuroGrid software on a person's computer is also supposed to improve its performance over time by noticing the user's behavior in making queries and deciding which proposed sources to select.

NeuroGrid doesn't use RDF or any other standard language, but it does store its data in the form of triples. Thus it seems reasonable that it could be modified to use RDF if desired. Then the system could be adapted to use some of the other strategies mentioned earlier. For example, it could participate in a TAP-like system.

NeuroGrid is interesting, although it's still in an early stage of development. With all peer-to-peer systems, there is a question about how well they can scale up to large sizes. Still, NeuroGrid's use of peer information and its attempts to adapt its performance according to the actions of its users may prove useful for other systems.

5.2.9 Clustering

Many search engines return their results in a plain list, usually with the (presumably) most relevant results near the top. But an undifferentiated list can be cumbersome to deal with, especially when the search term has many possible meanings. It's desirable to group the search results into sensible categories. The term *swing* could bring results for the Java Swing user interface code library, swing music of the big band era, improving one's golf swing, a child's swing, and so on. Grouping the results is often called *clustering*.

Table 5.2 compares Google's results for a search on "child's swing" with those of today's apparent leader in result clustering, Vivisimo (www.vivisimo.com).[6] For brevity, the Vivisimo results show only the clusters and not the actual hits; the numbers after the cluster headings indicate the number of results for each head-

Table 5.2 Results for a search on the term "child's swing" from Google and Vivisimo

Google	Vivisimo
CHILD'S SWING JACKET Playland Adventures Inc. is a child's swing world of fun Little Tikes Toys, Beds and Playhouses at Just Little Tykes Child's Glider Swing Colerain Township Classified Ads: Wanted, Child's Swing Set National Safety Council Fact Sheet Library … Install the swing set legs in concrete below ground level to avoid a tripping hazard. ... more than five inches but less than 10 inches, since a child's head may … Seller's Past Listings … has ended. 71187, Antique Wooden Child's Swing, $88.00, 0, 2/15/2001 4:37:14 PM, Auction has ended. 71189, Antique Wooden Child's Swing, …	child's swing (185) Garden, Verses (27) Sets (24) Plans (12) Seat (11) Furniture (11) Slide, Residential (9) Safety, Dangerous (10) Toys, Little Tykes (8) Tree (9) Child's Play (9) Parent (2) Art (6) Swing Jacket (3) [… other minor headings omitted …]

[6] Some other sites that use a form of on-the-fly clustering are iBoogie (http://iboogie.tv/), Kartoo (www.kartoo.com), and Mooter (www.mooter.com).

ing. Many of the headings in the Vivisimo results also have their own subheadings, but they aren't shown for simplicity.

Which set of results would you rather work with? Most people would probably say "the clustered results on the right"—at least, provided the results were good hits. This isn't intended to take anything away from the accomplishments of the mainstream search sites, but to indicate one fruitful area for them to move into next.

Until recently, automatic clustering has generally been less than satisfactory, so the capabilities illustrated here are especially striking. The increase of computer capability has combined with progress in academic research (for example, Zamir and Etzioni [1998]) with surprisingly good results. But more is involved than first meets the eye.

Where should the clustering of results take place? The obvious thing to do would be to devise a set of categories—an ontology—based on the entire collection of indexed documents, and to tag each one with its appropriate classification. This approach has some drawbacks, but two stand out: Documents are rarely about one thing, and they should frequently be put into more than one category.

What's more, a set of categories that represents the entire collection known to the search engine may not suit the particular set of results for your current query. It can be more useful to cluster results on the fly as they come back from the search. That's what Vivisimo does—it performs its clustering using the short quotes that a search engine returns with each document. It's also possible to combine this ad hoc, on-the-fly clustering with predefined categories, although the Vivisimo site doesn't do so. Vivisimo is happy to put a given document into many categories if that's how the results turn out. The bookmark case study in the appendix also discusses the usefulness of ad hoc categories and multiple classifications.

The effectiveness of ad hoc clustering leads to a second point, a conundrum of sorts. For clustering like this to be usable on the Semantic Web, it should be possible to make the clustering available in a way that can be readily acted on by other computers. That's easy enough. But how can other computers know what these ad hoc categories mean? They wouldn't by definition be standard terms, since they would be created on the fly.

Chapter 7 discusses the design, construction, and interchange of ontologies. Usually it's taken for granted that this means the definition of a predefined set of categories, which again is at odds with the ad hoc nature of these clustered results. The OWL ontology language can state that one category or term is equivalent to another. Perhaps in the future a clustering engine will be able to link its ad hoc clusters to other vocabularies using such methods. Since clustering can also be done against terms from an ontology, it should be possible to relate those clusters to the ad hoc clusters.

Clearly, some basic work needs to be done in this area. Clustering results is an important way to present a large number of results so that people can work with them effectively, so it's important that the work be done.

5.3 Distorting results and spoofing search engines

HTML pages can include simple meta data by using the META element. A page designer can include keywords describing that page, and a search engine can use them for cataloging the page. It sounds ideal and well suited for extension with Semantic Web capabilities. The trouble is, people began packing their pages with multiple copies of the same keywords, and they added keywords that seemed to be popular with search engines even though they didn't suit the page. This (mal)practice can cause bad results and damage the utility of a search engine. As a result, the better search engines no longer make much use of self-describing META tags in a page.

There are other ways to fool search engines or users. A site can be *hijacked*—its home page can be hacked so that visitors are sent to a different web site that may look like the original one but belongs to someone else. Or, a URL can mislead a user into going to the wrong site. Here's an example. The Open Directory Project is found at www.dmoz.org. Open Directory aims to produce a directory in a noncommercial way, using the services of many volunteer editors. It uses a home-grown, changeable taxonomy—for example, Science/Biology/Agriculture/Forestry/Urban Forestry. But if you (by a perfectly natural mistake) visit www.opendirectory.org, you'll see a site that superficially looks like the www.dmoz.org site but is actually an advertising directory. Links to www.opendirectory.org might be misanalyzed by a page-ranking system.

Another source of bias is paid ranking. Some sites allow advertisers to buy the top places in the returned hits. Nowadays the better sites that do this keep the paid returns, sometimes called *sponsored links*, in a separate and well-labeled area, which is an improvement. Still, advertisers can be favored by having the site fail to return certain pages (possibly those of competitors).

The upshot is that there are many ways to distort or bias search results, and searches on the Semantic Web will have to take these into account (more in the next section). The issue of ensuring the reliability of Semantic Web data and operations is a difficult subject that is discussed further in chapter 10, "Distributed Trust and Belief."

5.4 Searching and the Semantic Web

There has been a tendency for the potential contribution of the Semantic Web to be presented largely in terms of adding classification and standard ontologies to resources on the Web. Search engines would add software that could extract that data, and thereby searching would become far more reliable. This seems logical, but we've seen that it will fall short in several areas.

5.4.1 Self-describing meta data: no panacea

As we've discussed in this chapter, the use of self-describing meta data for web searches has some limitations:

- Much of the material on the Web is likely to remain untouched by semantic markup.
- The value of classification beyond simple keywords hasn't been demonstrated, and there is some evidence suggesting that it doesn't bring materially better results.
- Nonstandard meta data may be ineffective (see [Utah]).
- Incorrect meta data inevitably provides worse results than no meta data (see [Sokvitne 2000]).
- Meta data can be expensive to create (see [St. Laurent 2003] and [Bray 2003]).
- The potential for spoofing and distortion is inherent in any self-describing markup.

The second and third points suggest that semantic markup and classification contained in each page could be useful within a company's intranet; where access is controlled, self-promotion would (presumably) be minimized, and the domains of interest would be relatively restricted. But for the fully connected Web, the situation is very different. Semantic markup should still be useful, but less for pure search purposes than for processing once a resource has been found.

5.4.2 Semantic Web possibilities for improving searching

Since self-describing meta data can be unreliable for general Web searches and social analysis is turning out to be increasingly valuable, it would seem that Semantic Web technologies could be the most useful in several general ways:

- *Make self-description more reliable*—This involves issues of trustworthiness, the subject of chapter 10. Meta data within a page could be digitally signed to attest to its source, and search engines would be able to take into account the source and its potential motives in using that information to rate the page. Social analysis might be applied to the trustworthiness of different sites or authors that added semantic data to their own documents, so that appropriate weights could be applied in deriving page rankings. So far, I haven't seen this approach discussed in literature or online technical discussion groups, but it seems logical and potentially powerful.
- *Make more information available for social analysis*—This could involve third-party annotations. That is, people would publish comments on particular web pages in a knowledge-representation language like RDF. Of course, people comment on web pages all the time now, but they do so in natural language terms. One way this approach could evolve would be for advanced natural and link analysis results to be transformed into RDF and published on the Web. Any pages that contained semantic markup would be that much easier to analyze and add to the data store. Such a movement—grass-roots social analysis—is starting to take place in the web logging (or *blogging*) communities. Bloggers link to each others' blogs, and rankings showing the blogs with the most links pointing to them are becoming available. Some people are experimenting with tools to do further analysis. The results aren't yet created in something like RDF, but the point is that there is a low-key spread of social analysis.
- *Use data integration techniques, such as those used by TAP*—This approach would require servers to present the data stored in their databases in a standard way, such as with RDF. A server would have to be able to retrieve data by matching graph fragments as well as by its identifier.

There are obviously other ways in which technologies developed for the Semantic Web could potentially improve search capabilities; many of them have been mentioned previously. Other search improvements are being developed that seem, at least on the surface, to not really involve the Semantic Web per se, and so they haven't been covered here. But the line is not always easy to draw.

5.4.3 Searching and web services

Chapter 8 discusses services. The search for services has a lot in common with the search for other kinds of information on the Internet. There is a difference in motivation for the information provider, especially for business and commercial services, because the service provider has a financial stake in having the service found and chosen.

If businesses are to cooperate in making services findable and usable on the Internet, they will have to adopt a means that will minimize any advantage one company has over the others. Is it better for a travel service to be able to find flights on any airline or just on those airlines that pay enough money to be privileged members of the service? Increasingly, businesses are coming to see that common access is good for everyone in the long run. If this trend prevails, the Semantic Web will have much to offer in the way of discovery and description of services.

5.5 Summary

This chapter has ended up in a surprising place. At first glance, it would seem that the future of searching, from a Semantic Web point of view, would be the large-scale introduction of marked-up semantic information into web resources, augmented by carefully developed ontologies. Classical logical reasoning techniques would be applied to the classification of the pages, and consequently searches would be far more effective than they are now.

Instead, you've seen that self-description in pages is unreliable and that social analysis of links and opinions together with the semantic analysis of the natural language of pages will probably continue to be of the greatest importance. You've also seen that intelligent analysis of the context of the queries themselves is likely to add significant effectiveness to searches. These realizations lead to questions about the trustworthiness of self-supplied information even if it's impeccably marked up with the latest technologies, and to the possibility that the stores of annotations and opinions the Semantic Web will make available may be an important contribution to searching.

WORLD WIDE WEB

EXPLORER'S GUIDE TO THE SEMANTIC WEB

Thomas B. Passin

A complex set of extensions to the World Wide Web, the Semantic Web will make data and services more accessible to computers and useful to people. Some of these extensions are being deployed, and many are coming in the next years. This is the only book to explore the territory of the Semantic Web in a broad and conceptual manner.

This Guide acquaints you with the basic ideas and technologies of the Semantic Web, their roles and inter-relationships. The key areas covered include knowledge modeling (RDF, Topic Maps), ontology (OWL), agents (intelligent and otherwise), distributed trust and belief, "semantically-focused" search, and much more.

The book's basic, conceptual approach is accessible to readers with a wide range of backgrounds and interests. Important points are illustrated with diagrams and occasional markup fragments. As it explores the landscape it encounters an ever-surprising variety of novel ideas and unexpected links. The book is easy and fun to read—you may find it hard to put down.

The Semantic Web is coming. This is a guide to the basic concepts and technologies that will come with it.

Threads Explored

- Scenarios of use
- Semantic Web layering
- A society of agents
- Trust and belief
- The life of annotations
- Collective knowledge
- The challenge of information:
  - incomplete
  - erroneous
  - conflicting

Since 1998, **Thomas Passin** has studied and published about flexible modeling of knowledge on the World Wide Web, especially on the subjects of XML, topic maps, and RDF. He is Principal Systems Engineer at the non-profit engineering firm, Mitretek Systems. Tom lives in Reston, Virginia.

AUTHOR ONLINE
Ask the Author

Ebook edition

www.manning.com/passin

53995

9 781932 394207

ISBN 978-1-932394-20-6

MANNING $39.95 US/$55.95 Canada