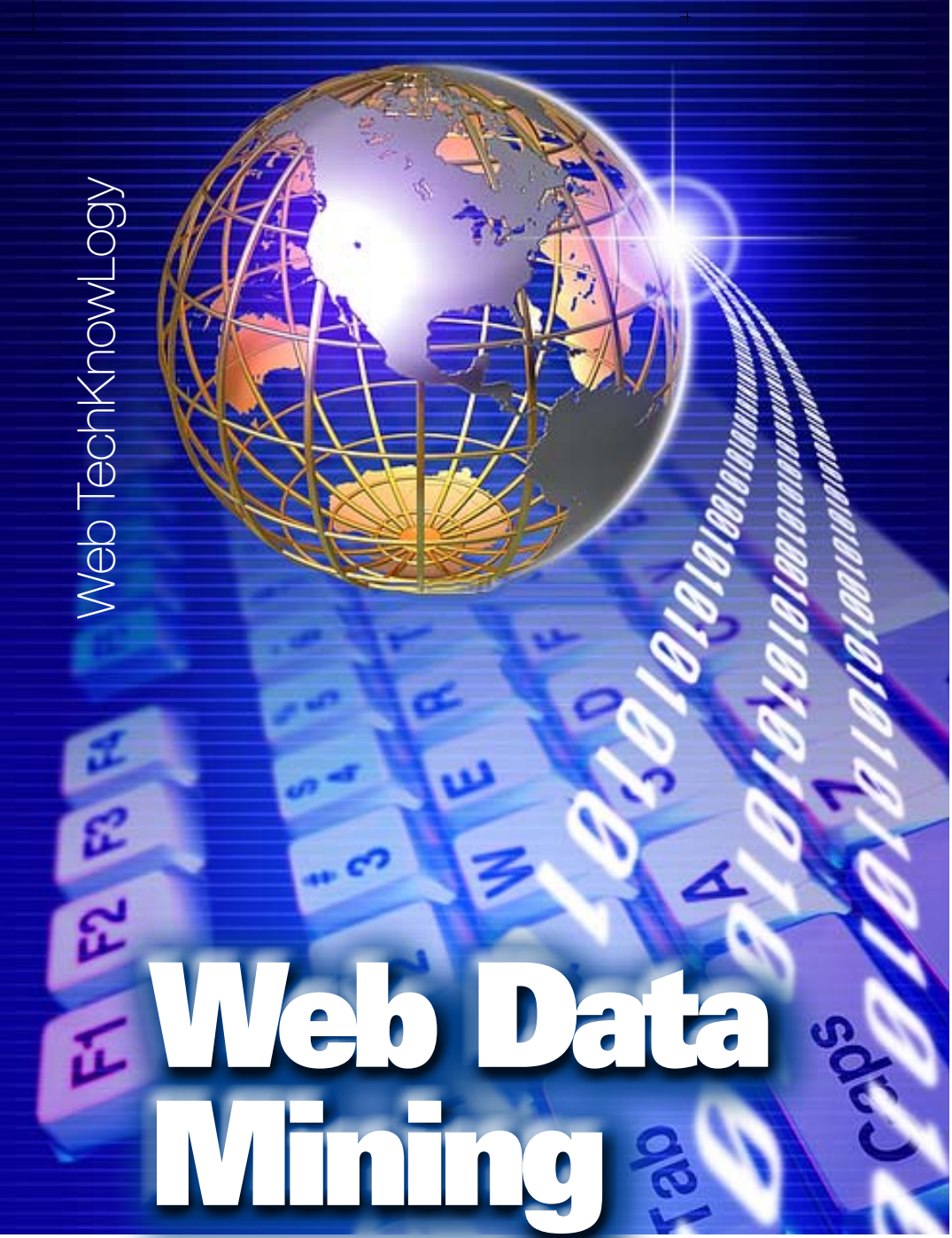


Web Data Mining

Here's a ready reckoner on how to mine the riches that the Internet throws up. As the Net grows, mining that vast repository of information is getting more daunting. But relax and read on, as there's enough help at hand.

Dorai Thodla
The author is a serial entrepreneur and technology consultant. Dorai Thodla has been involved in Web Data Mining for several years. His team built a product called InfoMinder (www.infominder.com) which helps track web pages and is currently building an RSS aggregator called NewsMinder. Dorai can be reached at dorai@thodla.com and he blogs at www.thodla.com

Information on the Web is increasing at a rapid rate. How can you leverage this information for your personal and professional use is what we will explore in this article. In follow-up articles, we will look at how to track various information sources, filter the right information, organise it and share it. While a lot of these tools and techniques apply to other industries as well, our focus in this article is the software industry.

Who needs it?

Almost all of us need information. A lot of information is freely available on the Web. Learning a few techniques on how to mine information on the Web is a useful skill. Here are some sample usage scenarios:

- You are an entrepreneur who is planning to start a new software business. You hear that Web 2.0 and social applications are hot. You want to do some research to understand the marketplace, and want to prototype a few product ideas.
- You are part of the CTO office of a software company, and are interested in short-, medium-, and long-term technology and business trends in your industry. You need this information to build skills in your organisation, and to build a few concept prototypes.
- You are part of the CIO office of an organisation. You need to balance early adoption of technologies with providing a stable environment for your business; you don't want to jump at *every* new technology. In addition to finding new tools and techniques, you also want to understand the risks and the maturity level of these technologies, which ones are being used for building applications, and you also want to track many non-technical factors.
- You are an outsourcing company and want to find customers for your business and track trends in outsourcing. Being a jump ahead of your competition and carving a niche are important differentiators.
- You are part of HR, or a Learning Officer, and need to plan for the skill development of your

employees. You want to keep your software team happy and so need to know the latest technologies, tools and resources to plan training and skill development.

- You are a development lead, and need to provide the team with the latest information on product releases, and access to product/technology knowledge bases. You need to know of any problems, including security issues, in the tools or software that you are currently using for your projects.

Broadly, there are several components to finding, using and sharing information.

- Identifying and discovering information sources
- Tracking information from various sources and filtering them for their relevance to your needs
- Organising collected information and sharing it with others

Finding Information

Information sources can be categorised as:

- News sources
- Company websites
- Blogs
- Search engines
- Wikis
- Discussion groups
- Social bookmarking sites
- Social networks

All these sources are complementary to each other. Each delivers a slightly differing type of information. We will discuss each one of them in more detail in the rest of this article.

Blogs are an interesting source of news—they contain content created by people like you and me, known as bloggers. In many ways, you could consider blogs to be 'citizen news'.

News sources

News has traditionally been our major source of information, but news sources are gradually becoming Web-based and are delivered in several ways.

News websites are the closest to the print newspapers that we have been accustomed to. Few layout changes are made to accommodate the smaller area for displaying news. But if you look at the *New York Times*, the *Wall Street Journal* or *The Hindu*, you will notice a lot of similarities to traditional newspapers.

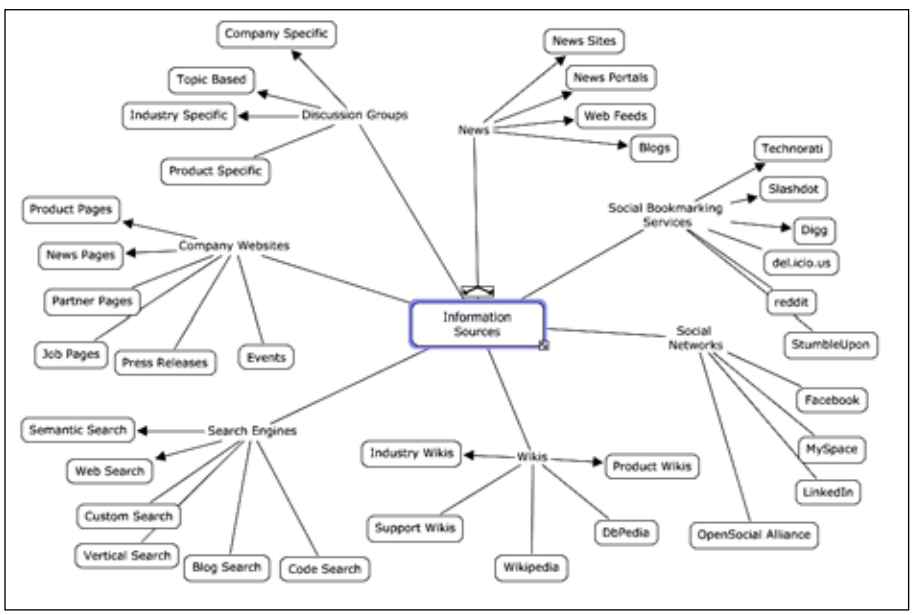
News portals are similar to news websites, with some difference—you are allowed to customise the news to suit your tastes. Some of the more popular news delivery is done using news portals. News portals are also popular for delivering custom news specific to an industry or subject area.

News feeds are lists of news items. These are also known as Web feeds or RSS feeds. Each feed has a title and description, and contains one or more news items. Each item, in turn, contains a title, description and a link to the original story. RSS feeds are a very popular means of delivering multiple news items.

Where do you find these feeds? Several publications (like *Business Week*, the *Wall Street Journal* and *Yahoo News* provide news feeds. You can subscribe to these feeds, and read them with news reader software. News readers aggregate several feeds of your choice, and present a list of current news items.

One of the benefits of RSS feeds is that they are based on a standard format—you can easily write an application that filters RSS feeds based on your own specific criteria, and have it present you with only selected items.

Blogs are an interesting source of news—they contain content created by people like you and me, known as bloggers. In many ways, you could consider blogs to be 'citizen news'. Bloggers pick an item of interest



The mind map above shows categories of information sources. (This is not an exhaustive list; we show a representative sample of actual sources in each category.)

and write a post. A post is like a mini news column. Blogs most often carry reference to the original news and commentary by the blogger. A more interesting aspect of blogs is that readers can comment on the news item or the commentary. On popular or controversial blog posts, the commentary often provides more information than the original posts themselves.

Discussion groups

We use the term 'discussion groups' to broadly cover mailing lists, company/product-based discussion groups, and collaborative Web products like Yahoo Groups and Google Groups. A lot of information is exchanged and debated in a discussion group. If you are considering a new programming language, Web framework, or a new development tool, these forums provide valuable information.

Company-based groups are normally run by a company's support group, or run independently by others outside the company. If you are looking at investing in a certain company's products, this may be a good place to check them out.

Topic-based groups focus on a certain topic of interest like wireless mobile devices. A typical group of this kind may focus on Web frameworks or learning software. They are not specific to a single company, but several companies in the arena may be discussed.

Product-based groups focus on a product or a product family. For example, Django is a Python-based framework for building Web applications. Members of the Django group on Google spend a lot of time discussing various approaches and best practices on using Django. They also discuss various issues and workarounds.

Company websites

Company websites are one of the best means of getting information about the company—directly from the source. You can typically get information about the philosophy, mission, team, products, or services—and a lot more. Company sites range from a few pages to a few thousand pages.

A lot of information is exchanged and debated in a discussion group. If you are considering a new programming language, Web framework, or a new development tool, these forums provide valuable information.

You can learn a lot about a company by visiting its website. Frequent updates to the site indicate a lot of activity. You can look at customer wins, business partnerships, product releases, and job postings. These core activities are reflected in frequent press releases and news coverage.

A few sites like Alexa, ZoomInfo and Hoovers not only provide information about a company, but also provide information on competing companies and products.

Search engines

Search engines from Google, Yahoo and Microsoft are probably the most frequently used resources to find information. Search engines index news sites, company sites, feeds and blogs, to provide answers to your search.

Web search engines in general locate information using a keyword search and some kind of a ranking scheme. Many of these search engines also provide an API (application programming interface) that you can use to create automated searches—for example, a search for your own company. Meta search engines submit searches to multiple search engines, and consolidate and group information.

Custom search engines allow you to customise one or more aspects of search. For example, Google Custom Search allows you to specify the sources (websites) to search, and some pre-defined keywords to include in your search.

Vertical search engines are focused on a specific industry or subject area. By focusing on a specific subject area, they can provide more effective search and more accurate results.

Blog search engines allow you to look for blogs that cover a certain topic or search area. While you can use regular search engines to find blogs, blog searches focus on indexing and searching only blog posts. Popular blog search engines include Google Blog Search and Technorati.

Code search engines let you search for code (programs) of a specific type. This is a great resource for developers. The search is typically done for open source, or other publicly available source code. You can specify certain keywords, and select a language or operating environment. Google Code Search enables you to search based on file paths or names, the licence under which the code is released, the programming language used, the name of the package, and more. Each

of these accepts regular expressions as the search expression, allowing you to construct powerful and very specific search expressions. Krugle is another code search engine, which was recently purchased by Yahoo, and also sports an open source code version.

Semantic search is a new type of search engine that is in its infancy. Instead of just indexing keywords and searching for them, semantic search engines allow you search a bit deeper. Some of the semantic search engines also allow humans to validate the results of the search, so that they can be improved.

Social bookmarks

Social bookmark services allow people to share their bookmarks. This is done by saving bookmarks on the Internet and making them accessible to every one. To make the access simpler, social bookmarking systems include descriptions and tags with bookmarks. Popular social bookmarking systems include del.icio.us, furl, Digg, Slashdot, Reddit and StumbleUpon. Some of them allow users to rank bookmarks, so the more popular ones are listed at the top, further increasing their visibility.

Other facilities provided by social book marking services include an API and RSS feeds to programmatically access bookmarks. They also provide tag clouds (where tags are displayed with the more frequently-occurring tags appearing larger).

Wikis

Wikis are a special type of website where content is created by multiple people—collaborative content development, editing, approval and publication make wikis a very powerful platform for creating content. Probably the most definitive example of a wiki is *Wikipedia*, which contains the collective knowledge of hundreds of thousands of people. A collaborative encyclopedia, Wikipedia has become the first stop in many people's search for encyclopaedic knowledge on the Net.

There are a wide variety of information sources, and they seem to increase rapidly. Most of them offer programmatic access through a language-independent API. All you need is a way to track them and be updated when these rich sources of information produce new information.

DbPedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries addressed at Wikipedia and to link other datasets on the Web to Wikipedia data.

Product wikis are specialised wiki sites that are set up to provide a collaborative community for product documentation, support and issues. Many open source projects/products have a product wiki. Even commercial products from major vendors have their own wikis.

Industry wikis cover information about specific industries. You can think of them as the aggregation of knowledge about a particular industry—for example, Web 2.0 or AJAX techniques. Industry wikis are normally associated with industry portals, or managed by independent groups. If you are trying to find customers or start a company in a specific industry, you may want to first check whether your industry has its own wiki. A good starting point is Wikipedia, which may contain links to other wikis.

Support wikis are knowledge repositories for products and services. These are maintained either by the vendor or the community. If you are in IT or any kind of software development, these can be gold mines of information.

Social networks

A discussion about information sources is not complete without some mention of social networks. Social networks allow people/groups to share information in a wide variety of ways. The most popular social network today is Facebook, closely followed by MySpace. LinkedIn is a business social network. Recently, Google introduced OpenSocial, which provides a common set of APIs for social applications across multiple websites. With standard JavaScript and HTML, developers can create applications that access a social network's friends and update feeds.

There are a wide variety of information sources, and they seem to increase rapidly. Most of them offer programmatic access through a language-independent API. All you need is a way to track them and be updated when these rich sources of information produce new information. That, however, is the focus of the next article in this series, and we hope you look forward to it. IT