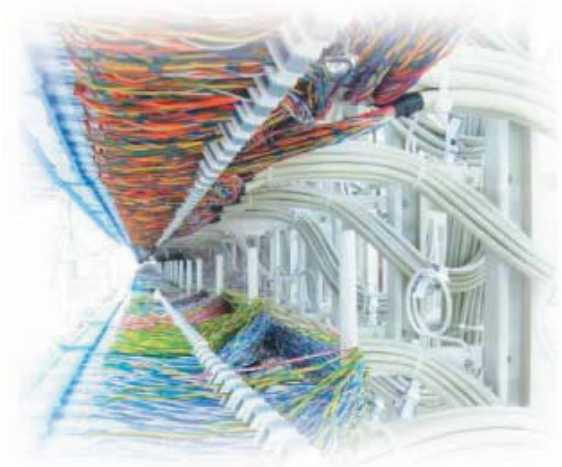
Is It Really Time for Real-Time Search?



David Geer

Due largely to the rise of social networking, the Internet constantly receives large quantities of new material that traditional search technology won't find in real-time. Real-time search addresses this shortcoming.

Internet search engines typically look in occasionally updated collections of documents, images, videos, and other content to produce responses to user queries. Periodically, search providers add new content to their indexes to give users access to fresh information.

However, the Internet now changes by the second. Not only do content providers add material regularly, but social media sites such as Twitter and Facebook constantly present large volumes of new information.

Typical search engines must first aggregate content via time-consuming Web crawlers and thus can't keep up with this flow.

But users want immediate responses to many search queries.

These factors have increased the demand for and opportunity offered by real-time search, in which users can obtain the most recent content and, sometimes, updated feeds relating to past queries.

Users get results immediately, from new posts to web pages, blogs, and particularly social-networking services such as Facebook, Flickr, and Twitter.

And during newsworthy events, such as the recent Haiti earthquake, real-time searchers can stay up-to-date on posts from affected areas.

Twitter integrated real-time search last year, enabling users to receive the most recent unfiltered tweets right away. Other real-time search engines have emerged, and even established players such as Google, Microsoft, and Yahoo are working on the technology.

Although real-time search has clear potential, it also must address numerous issues before it can be widely adopted.

REAL-TIME SEARCH

Tobias Peggs, general manager of real-time search engine OneRiot, said 20 percent of searches are performed to access a specific website and 40 percent are performed to find static data such as how-to articles or corporate contact information. Neither requires real-time technology.

However, he added, "Some 40 percent of searches would best be served by real-time results." A traditional search for Barack Obama would yield, for example, a Wikipedia page, but a real-time search would yield up-to-date news and commentary.

Twitter Search was the first real-time search engine.

The technology

Traditional search engines aggregate pages from across the Web over periods of days and index them based on multiple factors, none related to real-time access. The engines then use keyword matches, relevance algorithms, and other factors to return query results.

Real-time search engines cull new data streams from across the Web. The search services subscribe to social-networking websites for notifications of new content. They retrieve this content via the social site's API using HTTP, FTP, or whatever protocol it uses.

The engines index material by subject matter and then filter and organize the data based on the time it was posted. The search architectures can index millions of pages per hour, according to Peggs.

Real-time engines retrieve and index data so fast because they don't crawl for, aggregate, and index infor-

Published by the IEEE Computer Society

mation like traditional search engines do. Instead, they get most of their data via direct feeds from social websites such as Twitter. They can thus index and filter the material immediately.

Real-time engines use algorithms that rank and index the content according to, for example, immediacy of interest, relevance to queries, credibility of authors based on factors such as the number of regular followers they have, and the reputation of a link based on factors such as the number of times it is forwarded by readers.

Implementations

There are a number of real-time search engines. Twitter integrated real-time search for tweets into its social-networking service last year. Twitter is currently the biggest source of real-time data, according to Mike Grehan, vice president and global content director, for Incisive Media, owner of Search Engine Watch, a news and information website.

OneRiot. This social engine updates its results in real time with content from Delicious; Digg; FriendFeed; Twitter; and OneRiot's browser toolbar, which searches Facebook and MySpace.

The system—which can yield raw results or answers filtered for spam—prioritizes its findings based on an algorithm that considers 26 factors.

One filter measures *hotness*, the rate at which a link is shared on the social Web within the preceding minute. This indicates whether a piece of content is increasing or decreasing in popularity from one minute to the next.

Users' online reputations—their credibility based on the number of followers they have and how often their posts are forwarded by others—also help rank their content's social relevance.

OneRiot determines a link's relative popularity based on factors such as the number of followers the sender has and how fast and how many times links are shared. This lack of reliance on one factor keeps content on major websites with high link reputations from automatically ranking higher than more relevant results on minor sites.

OneRiot's search volume grew after it launched an API that lets other sites and applications—120 so far—tap into its real-time search capabilities, said the company's Peggs.

OneRiot generates revenue via real-time advertising related to queries and search results.

An estimated 40 percent of searches would benefit from real-time results.

Google. A feature the company recently launched—accessed by clicking on "Show options" at the top of a page of search results—permits filtering of results by time, including categories marked "Latest", "Past 24 hours," "Past week," "Past year," and "Specific date range." The "Latest" findings—including Flickr, FriendFeed, Twitter, and blog posts—are the real-time results.

"We have announced new agreements with Facebook and MySpace to more efficiently index their public content," said Google Fellow Amit Singhal. Facebook's and MySpace's privacy controls would let users prevent Google from indexing their material if they so desire.

Google's real-time search automatically scrolls new relevant information within a few seconds after it appears in the Web index. Traditionally, users have had to make additional search requests to see new information.

"The new technologies include algorithms that allow us to assign accurate relevance to real-time content," said Singhal.

The algorithms enable real-time search with updates and quality assessments. They use filters for characteristics like the probability of a result's relevance to a query, the number of times readers forward posts to other people, and author quality based on factors such as the number of followers a content producer has, according to Singhal.

Google's crawlers now index and display virtually any web page as it appears. The company has also deployed new technologies that monitor slightly more than a billion documents a day for fresh updates, according to Singhal.

One algorithm uses a language model that compares sequences of words in updates to sequences it has seen in other documents to determine whether the updates contain new information. Other algorithms address semantics to clarify the meaning of content, cache past queries and responses to improve performance, measure the relevance of a search result to the query topic, and look at patterns of aggregate search results to determine the importance of recent information.

Collecta. This search engine monitors the update streams of real-time blogs and sites like Flickr, Twitter, and WordPress, and can show material that addresses queries as soon as it is posted.

The engine works with the XML-based Extensible Messaging and Presence Protocol, which enables near-real-time instant messaging of results to searchers

The service, shown in Figure 1, also uses long polling. Thus, if there is no new data on the server belonging to one of the websites that Collecta checks, the request remains until there is fresh information. When there is new data, Collecta returns it immediately and sends another

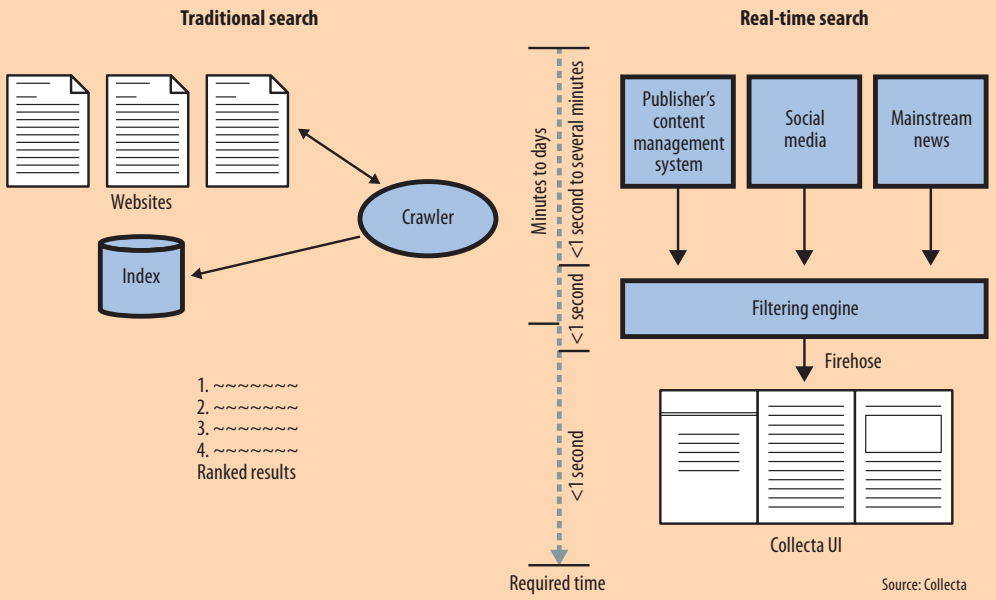


Figure 1. Traditional search engines aggregate content via web crawlers and then index the material before ranking query results based on relevance, a time-consuming process. Collecta's real-time search engine receives data from various types of external sources, including content publishers' content-management systems, social-media sites, and blogs. These sources push content directly to Collecta's čltering engine. The engine quickly člters content based on whether it constitutes spam and is relevant to queries. The content eventually runs via a master route, called a črehose, to the UI, which displays results.

request, thereby providing information to users in an ongoing stream, noted company founder Jack Moffitt.

Collecta doesn't rank results but instead shows a chronological stream, filtered for spam and relevance.

Other sites. According to Microsoft, the company has given its Bing search engine access to Twitter's real-time data feed and has signed an agreement to license Facebook's API.

Yahoo's real-time search results appear on its regular search page. They include material from Twitter and are slated to add content from Facebook this year. Yahoo is using its own algorithms to identify new material and ensure that content is relevant, according to Larry Cornett, the company's vice president of consumer products.

He added, "We are working with Microsoft on a Microsoft-Yahoo search alliance."

Scoopler provides live, automatically updating, real-time search results across multiple social-networking services, including Delicious, Digg, Flickr, and Twitter.

In real time in one column, Scoopler provides lists of the most popular links, videos, and images related to a query, ranked by how recently the content was produced and by how much it has been shared on social-networking sites. Other popular related content from around the Web appears in another column. "The delay on data in the pipeline is 30 seconds maximum," said company CEO AJ Asver.

FriendFeed lets users submit search queries and create their own real-time response feed based on links their friends are posting on its website. The service also pulls in updates from other sites such as Flickr, Twitter, and YouTube.

Topsy shows real-time results based on searches for Web links posted on Twitter and ranked based on the number of times the link appears in tweets.

Also, said company vice president of research Rishab Ghosh, "The more people who retweet a [user's posts], the more weight Topsy gives to that person's tweets."

CrowdEye uses its CrowdRank real-time algorithm for ranking search results from Twitter. CrowdEye lets users sort results by relevance, arranging them primarily by the number of followers the people who post information have, as well as how often their posts have been forwarded by others.

REAL CHALLENGES

Real-time search is still in its infancy. "The technology didn't exist three years ago," said OneRiot's Peggs. Not many people thus know about and use real-time search yet.

Users don't understand how real-time search works, stated Erika Brown, executive vice president of

corporate strategy for search and social media for Frost & Sullivan, a market research firm.

Moreover, real-time search filtering is new and less than perfect. "Results can be loaded with irrelevant or repetitive data, particularly for broad searches," explained Danny Sullivan, editor-in-chief of Search Engine Land, a news and information website.

And because real-time search involves fast, complex calculations based on many factors, he added, it requires considerable computational overhead.

Incisive Media's Grehan said that because Twitter makes up the bulk of real-time information, it tends to dominate real-time search results.

According to Peggs, making money from the real-time search market won't work with the advertisements that appear on traditional search pages. That's because real-time searchers will want ads that let them take action right away based on query results.

Real-time search will need to become popular enough to generate large volumes of queries—and to have the ability to match queries with advertisements—if it hopes to generate significant ad revenue, said Scoopler's Asver.

In the future, explained Collecta's Moffitt, real-time search could immediately provide advertising for products that would appeal to users conducting searches on related topics. For example, an ad could promote a DVD starring an actor about whom a user is searching.

But, said Toby Bell, an analyst with market research firm Gartner Inc., "Real-time search has significant prospects for pollution with spam and advertising. People could be paid to pollute Twitter or add comments on blogs to the benefit of one company."

ON THE HORIZON

Real-time search will have many uses, according to Asver. People will be able to create custom pages on topics about which they are passionate or get information about events they are attending, he explained.

In the future, real-time search will focus more on location-related issues, said social-media author John Havens, a founding member of the Association for Downloadable Media.

For example, mobile-device users might announce traffic jams via the social Web. People could then use real-time search to find this information.

Companies could use real-time search to monitor online commentary about their products, said Gartner's Bell. "The technology gives you a quick sense of your online reputation and the effect it is having on your brand," he explained.

"It's about companies ensuring that their brand is part of the conversation going on in social media in a positive way," added Frost & Sullivan's Brown. They could become part of the conversation by responding to customer needs and complaints, and steering the conversation in a positive direction, she said.

Users might even be able to use real-time search to make on-the-spot financial and other decisions, noted Paul Sondereger, chief strategist for search-technology vendor Endeca Technologies.

Said OneRiot's Peggs, "Real-time search will become more pervasive. Consumers are moving to an always-on Web."

However, cautioned Yahoo's Cornett, there is a difference in use cases for real-time and traditional search. Poorly filtered but near-instantaneous information is helpful for some searches, such as efforts to gain a qualitative assessment of user sentiment on various timely subjects, he explained. The filtered, organized, slower results of traditional engines are superior for other searches, such as attempts to find information on a past event or historical topic, he said.

Nonetheless, added Topsy's Ghosh, "Real-time search technology will grow as a share of the total search base as people become more interested."

David Geer is a freelance technology writer based in Ashtabula, Ohio. Contact him at david@geercom.com

Editor: Lee Garber, *Computer*, l.garber@computer.org