

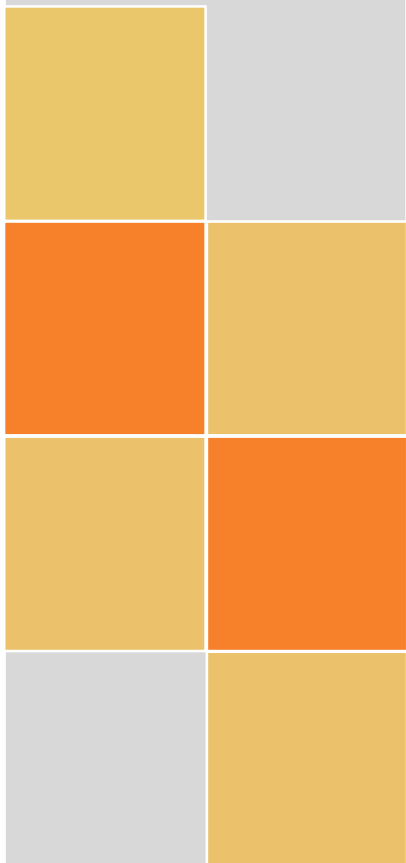


Discover what's popular on the web

The Attention Frontier

Moving beyond crawlers with the wisdom of crowds

The "Attention Frontier" is the set of pages that people care about and are looking at on the web at any given point in time. At Wowd, we believe that the Attention Frontier can be used to help people have a much better experience finding things on the web. The Attention Frontier can be monitored in real-time to discover new content that is popular with real people, and it can provide a powerful source of critical ranking data to deliver high-quality search results.



Background

The search industry has been relying on web crawlers as the principal means of acquiring content since its inception. It is actually quite amazing to consider that crawlers have not changed very much over the last 15 years.

A web crawler is a simple program, endlessly repeating the same sequence of tasks: fetch a page, find all links on the page, eliminate duplicates, put the new links in a queue to be crawled; repeat. Of course, real-world crawling is more complicated. The crawler has to respect the restrictions specified in each site's "robots.txt" file. But, for the most part, that is all there is to it. If you do all of these things well, you'll have a world-class web crawler.

"So what's the problem?", one might ask. "Crawling has been working just fine, why rock the boat?"

There are actually many problems and limitations associated with web crawling. Here are a few:

- Crawlers lack judgment. Crawlers are, by definition, trying to find all the good web pages they can, but crawlers aren't very skilled at distinguishing a good page from a bad one. Thus, crawlers will happily take in huge amounts of useless content, including millions of pages created by search spammers in their "black-hat SEO" attempts to manipulate and influence search results. The search industry likes to promote the perception that this problem is under control, but the reality is quite different. One might say that the mere existence of a large SEO industry shows that influence of organic results is alive and well.
- Crawler's can't read or type. There are huge amounts of high-quality web content that are inaccessible to crawlers. Much of this is in the so-called "deep web," where content is hidden behind online forms that only users know how to activate. A similar challenge is found at the growing number of AJAX-driven sites, where JavaScript-powered user interactions in the browser determine what content is displayed.
- The speed at which content changes presents a problem for crawlers. When content changes quickly and pervasively, crawlers cannot keep up. There are two critical factors here. First, when something new appears, it must be accessed and indexed as quickly as possible. Second, given the overwhelming volume of new material on the web, there is an important question of figuring out what new content matters most, and thus what new content should be dealt with first.
- Crawlers present a problem for many publishers. A publisher must maintain enough server capacity to handle both their real (human) users and also the un-real (robot) crawlers that visit their site. Some publishers have content that changes with such volume, and at such a high rate, that they simply

can't afford and don't want to provide the server capacity required to allow crawlers to find new content at an acceptable rate.

The nature of crawling also relies on a crucial assumption. This assumption is that all links from page-to-page and all paths through those links are equally likely and equally valuable. In this respect, crawlers and the systems that use them are incredibly unsophisticated. Affordances in a page that are obvious, even to an inexperienced web user, are completely beyond reach of even the most sophisticated page analysis algorithms. The history of Artificial Intelligence (AI) has consistently demonstrated that human judgment beats computer algorithms for tasks such as identifying the topic and quality of text every time.

In light of these serious crawler limitations, it's amazing that the search industry has been relying on them exclusively for all these years. But, if crawlers aren't the best way to find good content, what is? We believe that tremendous advances can be made by tapping into the wisdom of the crowds.

Scaling the wisdom of crowds

It seems obvious that leveraging human judgment is a better approach. So why hasn't it been done? We believe one of the principal reasons for the current and lasting infatuation with crawlers has been the perception of the problem of scale. One might imagine scores of computers in data centers continuously fetching millions of pages every second. It seems inconceivable that any group of actual users, even ranging into the millions, could possibly fetch and process content at such rates. But, there is much more going on here than meets the eye.

Consider a world-class search index, say one that includes tens of billions of indexed pages. As an example, the Google index is reported to be in excess of 40 billion pages, with the Yahoo and Bing indexes considerably less, on the order of 10 billion pages. In reality, those size numbers do not matter as much as numbers that characterize index *quality* and *freshness*.

Consider next a simple question – how quickly does the entire useful content of a major search engine index actually change? To get a better idea you may want to ask yourself the following question: does a page of results from Google change completely from year to year? We believe it is pretty clear that many, if not most results, across wide ranges of queries, will not change much in a year. As a consequence, the rate of change would appear to not be that high. In fact, we would argue that the entire corpus of results from typical major search engines is really quite stationary.

However, for the sake of argument, let's make the conservative assumption that the entire content of Google changes completely every year. Assuming 40 billion pages in Google's index, there will be approximately 110 million pages changed every day.

This is a large number, but one can envision a large group of users, say five million, that will visit a comparable number of pages during their normal browsing activities. A group of five million users would need to make 22 clicks, on average, every day, to cover all of this material. We believe that such a number is quite reasonable and agrees quite well with actual user behaviors.

Hence, we have just shown that a large group of users could, in fact, keep up with the *actual* rate of change of the web. The key point is that the theoretical rate of, say, millions of pages crawled per second is, in fact, irrelevant since the useful content *does not change* that quickly.

It might still feel surprising that a large group of people could keep up with crawlers. Another way of examining this situation is to consider that, since useful web page content does not change that much, the crawlers are repeatedly fetching the same content, just to see if anything has changed. This is actually quite wasteful and energy inefficient!

Attention Frontier

We have argued in the preceding section that the Attention Frontier of a large group of web users is more than sufficient to unveil content at the same scale as crawlers.

But, the quality of the pages that those web users are looking at, at any given moment, is at least as important as the sheer number of pages. Almost by definition, the relevance of what those people are looking at is very high, as people are much better at discerning high quality and relevant content than are machines. In addition, there is the temporal aspect of what those people are looking at, *right now*. This point is incredibly important in terms of being able to always provide the freshest and most relevant content. What these people are looking at helps us know what is actually popular at any moment in time.

As we said at the outset, we call this notion of what people are looking at, and what they care about, the *Attention Frontier*. It is the anonymized aggregation of such attention data that has huge value in terms of content discovery, index freshness, and search results relevance.

The emergence of Twitter has been very important in demonstrating the power of the Attention Frontier. Note that Twitter's own Attention Frontier is rather limited as Twitter is not really aimed at web page discovery. Only a fraction of tweets contain links. The number of tweets with embedded links ranges in the low ones of millions per day. Still, it is amazing how powerful this attention data can be.

The best indicator to-date of the value of this data is the recent licensing deals between Twitter and (each of) Google and Bing. The enthusiasm with which these large search players have embraced this third-party source of attention data clearly demonstrates how valuable these data are.

The insertion of links in tweets is an example of an explicit action where users generate signals that can be used for discovery and ranking. Such signals require specific actions to be taken by users, in which they create a tweet and insert a link. In terms of such signals, Twitter is clearly not aimed at, nor optimized for such actions. The most valuable information to be mined from Twitter is the human attention data embodied in the Tweets that contain URL references. A user tweeting a web page is a strong signal of the quality of that page.

On the other hand, a system can be created in which users need not take any explicit actions to indicate the focus of their attention, but instead the signals can be automatically inferred from a user's browsing behavior and from the pages that they choose to visit and the links they chose to click.

We believe that such a system is more efficient than Twitter and other "explicit action" systems in terms of discovery of web content and use of attention data for ranking. Wowd has embraced this idea since inception. This idea is also shared by others; for example, Wolf Garbe of Faroo has written on this topic <http://blog.faroo.com/?p=292>.

Using the Attention Frontier while keeping private information private!

Wowd uses the Attention Frontier of real human beings to discover content. Doing this overcomes all of the crawler-oriented problems outlined above.

Wowd employs Attention Frontier data without compromising our users' privacy by processing that data locally on each user's personal computer. In order for a centralized system to use attention data, the data must first be gathered from all the users and stored and processed in that central system. This raises serious privacy concerns, as many users are not comfortable sharing their browsing history with a company.

Wowd is designed around a decentralized architecture that allows each user's computer to participate as an equal, anonymous member of the cloud. This distributed approach avoids the potential problem of sharing personal data with centralized servers. Wowd has no central index or search logs. Anonymous attention statistics are gathered and stored in the distributed index and used in ranking algorithms that run on each user's computer. Anything of a personally identifiable nature stays with the person that it might identify. Wowd makes it safe to embrace the Attention Frontier as a solution to the problems inherent in the crawler-based approach, without sacrificing user privacy.

It's a wonderful version of the old parable of Mohammed and the Mountain: rather than bringing the user's data to the central computer for processing, we bring the computation to the user's local data and computer.

This approach is fantastically scalable as it ensures that the resources required for processing attention frontier data organically grow with the size of the frontier itself.

Conclusion

In summary, there are inherent limitations in the dominant crawler-based approach used to index the web, and a new approach based on the Attention Frontier offers many attractive benefits.

Wowd uses the Attention Frontier approach. In this way, Wowd taps into the scale and power of real web users to find and present timely and relevant search results while also making it easy to discover popular pages based on attention data.

More on Wowd

For more information, visit <http://www.wowd.com> or <http://blog.wowd.com>