bloomreach
GET FOUND.

Inside the Technology: Web Relevance Engine

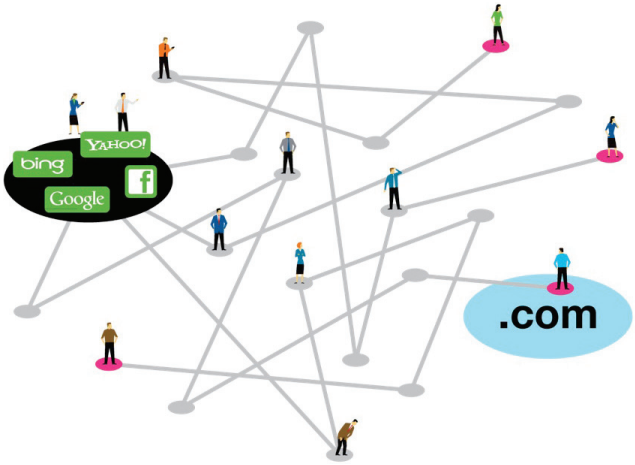
In this whitepaper, we focus on the core technology that BloomReach developed to address the Digital Discovery Divide, which is explored in a separate whitepaper, found here. This technology is evidenced in the world's first Web Relevance Engine (WRE), which harnesses the power of big data, machine learning and large-scale systems science in order to in order to target and deliver the best content given a user's intentions, while ensuring that web businesses' most relevant products and services get found.

Executive Summary

In this whitepaper, we focus on the core technology that BloomReach developed to address the Digital Discovery Divide, which is explored in a separate whitepaper found here.

The Digital Discovery Divide is a two-fold challenge: consumers with compressing attention spans can't immediately find what they seek, while web businesses are stymied by their inability to promote the right products and services for these shoppers at the right time.

BloomReach developed the world's first Web Relevance Engine (WRE), which harnesses the power of big data, machine learning and large-scale systems science in order to address the Digital Discovery Divide.



Matching consumer intent to relevant content

Consumers articulate intent in myriad ways. Explicit intentions are expressed through search engine queries, Facebook shares, follows and re-tweets on Twitter, clicks on email links or pages opened on an iPhone. Implicit intentions may be unveiled through the geo location of an iPhone, search and browse history, the time at which an email is opened, etc. For example, a relevant search result for "union square" could be highly dependent upon the geo-location of the query (whether someone is in New York vs. San Francisco).

Context and geography means that the same keyword used to express intent could mean different things. For example, "baby" and "infant" are the same in the context of food but not color. "Baby food" is equivalent to "infant food"; "baby blue" is not equivalent to "infant blue." While the keywords "baby stroller" might be extremely important for an American online retailer, it's not critical for a UK retailer whose customers would be searching for "infant pram."

Outside of correctly identifying user intent across semantic lines, there's also the challenge of dynamic data. Data changes all the time. Products go in and out of stock, fashion changes, and inventory churn can be as high as 30% a month for a major retailer. Relevant data can be hidden on pages that are hard—or impossible—for search engines to crawl if they are in Ajax or Flash without any text descriptions. What's more, ninety percent of the world's data was created in just the last two years.[1]

The evolving manifestation of user intent combined with the volume and volatility of data on the web fuels an ever-growing Digital Discovery Divide. Consumers find less of what they want. Web businesses struggle to ensure their most relevant products and services get found. And web engines continue to build a learning framework to adapt understanding in real time based on user behavior and the broader ecosystem. But the scale of the data means that whichever way you look at it—from the consumer, business or platform angle, this is a massive big data problem.

The WRE is bridging the divide

The Web Relevance Engine (WRE) was built from the ground up to target and deliver the most relevant content given a user's given explicit and implicit intentions. It does so by methodically:

1. Inspecting and gathering data available on the web
2. Building a semantic understanding
3. Scoring the content by user intention
4. Optimizing the site for relevance to the user

[1] March 2012. http://blog.nrf.com/2012/01/16/can-a-computer-named-watson-make-retail-smarter/

Once all the steps are completed, the WRE ensures constant learning from user behavior—all at web scale, using algorithmically driven machine learning and large scale systems science.

Each component of the WRE is explained in detail below.

I. Data Gathering and Inspection

In the data gathering phase, the Web Relevance Engine does a shallow crawl of the web and a deep crawl of the customer's website. Additionally, the system learns the sites that show up alongside the customer's website on web search and also fetches content. This data is combined with behavioral data collected from both the web and customer's website. This is performed on a regular basis. At last count, the BloomReach crawler had crawled upwards of 10 billion pages.

For the deep crawl of the customer site, the crawler accesses all static pages and combines this data with all the pages discovered through back-end integration—such as newly introduced product pages, and deep category pages on the site. The pages identified through backend-integration are typically "orphaned pages" (pages that are not linked from anywhere), deep pages (a page might be considered deep if it is N levels away from the homepage) or very fresh content.

Some of the data that is gathered during the deep crawl of web pages include:

Document {

- Key: Primary URL,
  Other URLs (U2, U3, U4,...)
- Title
- Body
- Links pointing into the page
- Age of the page
- Crawl frequency
- Conversion rate
- ...

In addition, during deep crawl, the WRE automatically fills out the forms on customers' website to collect all dynamic content present on the site. The content collected from the customer site is then parsed and organized. For example, on a category page, the system would extract the filters applied, list of products presented, breadcrumb, product descriptions, etc.

For competitor data, the WRE first identifies a site's competitors. In addition to the standard competitive landscape, it ascertains which other websites show up alongside web search results, on comparison shopping sites, are mentioned together on social networks, advertised alongside one another and so on. For example, Macy's may be a competitor to Bloomingdale's (even though it is owned by Macy's), but searches that include Macy's may also surface results from Nordstrom's and Dillard's.

The system maps the ontology of several competitors' websites onto the customer's website, capturing the different ways in which a product or entity can be described for the different organizations that are relevant to a website. For example, "Clothing & Shoes > Women's Clothing >Sweaters" can also be expressed as "Home > Women's Clothing > Women's Sweaters" or "Home > Outerwear > Winter Clothing > Women's > Sweaters."

Finally, the WRE marries the above collected content with anonymous user behavior data collected from both the site and through proprietary sources (query databases, etc.) and structures it in a way that can provide rich information about the data, and prepares it for the next step—semantic interpretation.

II. Semantic Interpretation

Using all the information that has been gathered, inspected and optimized, the semantic interpreter then proceeds to build a deep understanding of the content on the customer's website.

As a first step, it consumes a number of proprietary databases that BloomReach has built, compiled or collected over time. A few of these databases include a list of international celebrities, global landmarks, all the cars sold along with their make, model, etc. Examples of compiled data include cities relevant to a business based on the serving area. One key database that BloomReach make use of is its synonym database, which has been constructed by processing web content, the Google Books Ngrams database and query logs.

BloomReach has also built a user intention-to-keyword database using proprietary technology. This database delivers output similar to the below:

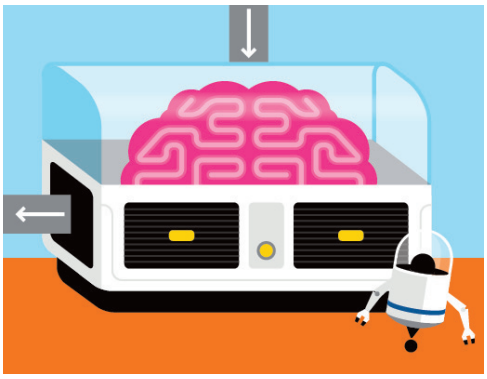
Intention	Keywords
Gift ideas for dad	Golf club, fishing poles, Craftsman tools, tools to paint room, paintbrush, roller, tray, painter's tape
Products popular in San Francisco	Cable car magnet, Golden Gate paperweight, Alcatraz striped shirt
Kid-friendly vacations	Pets, slides, roller coaster, noise OK

Using large-scale data from all its databases, the WRE semantically interprets the content of a customer's website—analyzing each page, and annotating the page with missing data. These annotations may include data like color information extracted via image processing, queries that have conversion rates, or terms that are used by competitors to define the same product. It uses the synonym map to identify missing terms from the page, the intention map to understand the intentions relevant to the product or service under consideration, and the geo map to understand the locations that will be targeted by this service.

In addition, it uses the crawl of the customers' competitors, and enriches the page with the content that competitors have used to describe the content, but are missing from the site. Behavioral data (conversion data, click data, time on page, bounce rate) is also collected from users interacting with the site and is used to identify what content people respond (or don't respond) to on the site.

Armed with a broad scope and scale of information that has been further enriched with semantic understanding, the WRE is now in a position to score the relevance of site content to user intent.

III. Relevance Scoring



The relevance scorer starts by collecting all the intentions that might be relevant for a site. These intentions might have been expressed in the form of queries on search engines, anchor texts clicked while browsing the web, the content of an email which a user clicked on, demographic data against which an ad was targeted, Facebook interest data if the user is signed-in, etc. These intentions are then mapped to keywords (see user intention-to-keyword database in the section above), against which site content is scored to identify the pages that would be most relevant to the user.

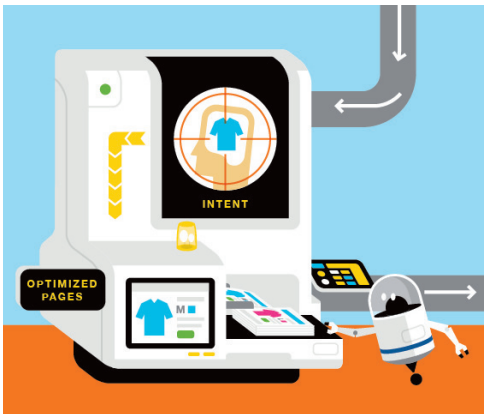
During the scoring process, the system leverages not just the content present on the page but additional data associated with the page as part of the analysis done by the semantic interpreter. Based on this score, the system predicts whether the user is going to find the page relevant or not.

A few outcomes from the work of the relevance scorer include:

- A page is relevant to the user's intention but is not optimized to be relevant based on the syntactic score
- A page is relevant but is not accessible by search engine (crawlability issue)
- A page is not relevant by itself, but combined with other data from the site, it might be able to address user's intention
- There is no content on the site relevant to users

In the first three cases, the page is fed to the site optimizer, while in the fourth case it is dropped. The site optimizer focuses on making the site's relevance available to all external sources.

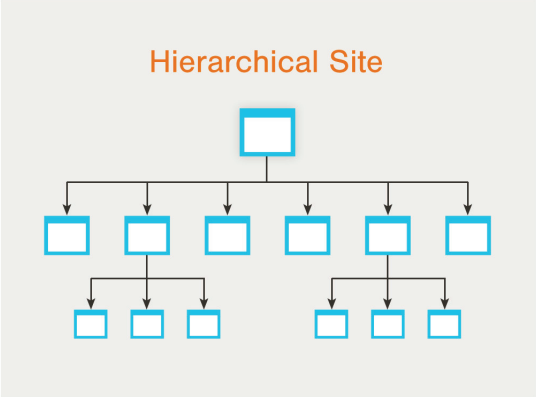
IV. Optimizing the Site

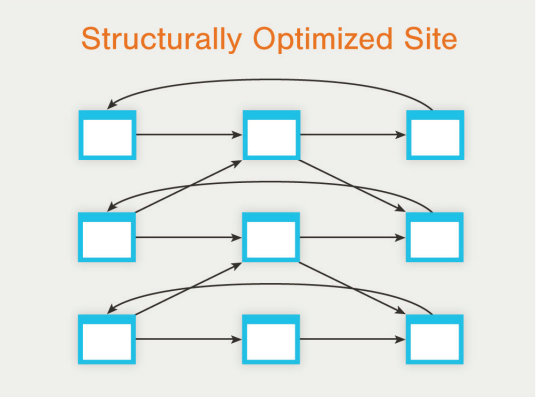


In order to optimize the relevance of each page, the WRE employs three separate techniques that work closely in conjunction to maximize discoverability:

1. **Syntactic Optimizer:** most systems use keyword match to ascertain the relevance of a document to a query. To enable matches, the WRE makes sure that a page is rich in content. From the optimization perspective, it tries to identify a small set of content from elsewhere on the site (e.g. product descriptions, address, reviews, etc.) and adds it to the page to make it syntactically more relevant to users looking for that page. The output of this is richer content on a page.
2. **Semantic and Syntactic Ontology:** as discussed earlier, site ontologies may have both syntactic and semantic holes. For example, a site may not have a good page for the query "blue cotton shirt" even though it has applicable content for it. Or it may not have any page on "gifts for Christmas" even though it sells a vast array of Christmas-related merchandise. This component automatically identifies these ontological holes and compiles the content that is most appropriate to it.

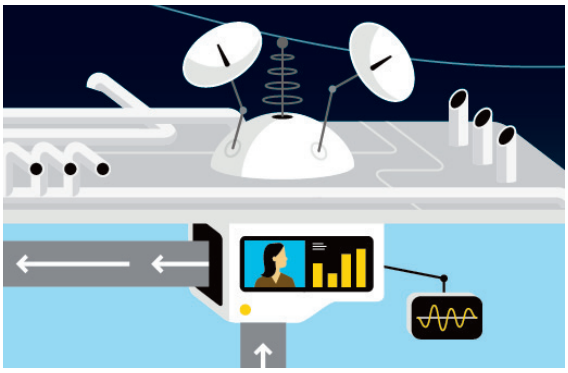
3. **Structure Optimizer:** A site's visibility is largely determined by a crawler, whose efficiency depends on site structure. In addition, the effectiveness of a landing page at converting leads is significantly impacted by relevant links on the page. This component works on making the site both crawl- and conversion-friendly. The weight of the edge is the similarity between the two nodes. Optimizing this graph for crawlability is an NP complete problem in the general case. We solve a simplified version of this problem using a maximum flow formulation along with added heuristics. The output of this is a set of edges (links that connect pages within the same site), which ensure that each "non-crawled" page is easily accessible from a set of "crawled" and "semantically relevant" set of pages.





From data gathering and semantic interpretation, to relevance scoring and site optimization the WRE is a continuous learning system that constantly measures user happiness, conversion, and site performance in order to optimize the site in question. It continuously monitors pages that are underperforming, pages or content that have gone stale and uncovers new content to feed to the optimizer to make sure that the site is relevant to dynamic consumer intentions, inventory changes and new relationships between information.

Conclusion



To bridge the digital discovery divide, web businesses need to dynamically present their information in a concise, descriptive manner relevant to every consumer while adapting constantly to new content and consumer intentions. This is a problem of massive and growing scale, which requires adoption of a real-time algorithmic approach.

Built from the ground up to address this specific challenge, the BloomReach Web Relevance Engine combines the power of big data, machine learning, and large-scale systems science to match user intent with content, enabling web businesses to get their products and services found more quickly online.