WHITEPAPER

bloomreach
GET FOUND.

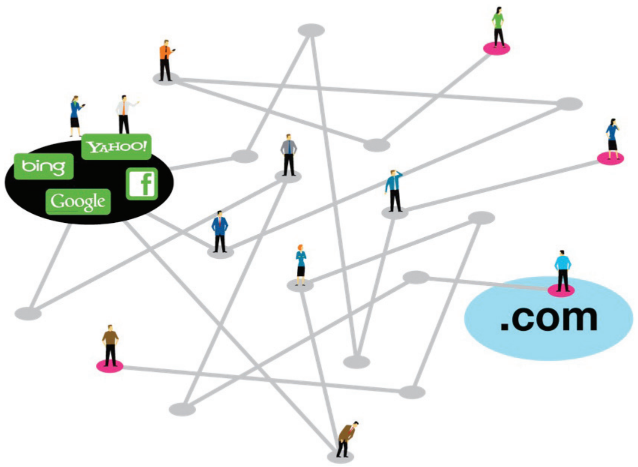
Bridging the Digital Discovery Divide

Today's world of massive information creation, distribution and consumption has created a digital discovery divide. On one side are the consumers who struggle to discover the information, products and services they seek, growing increasingly frustrated in the process. On the other side are the marketers of web businesses, who have precisely the information sought by consumers but may lack the resources to surface it effectively.

This white paper discusses the factors that have created and perpetuated the digital discovery divide and explores how web businesses can keep pace with today's discoverability challenges and take some basic steps to reveal their "invisible content" so that it is accessible to the consumers that seek it.

Executive Summary

Today's world of massive information creation, distribution and consumption has created a digital discovery divide. On one side are the consumers who struggle to discover the information, products and services they seek, growing increasingly frustrated in the process. On the other side are the marketers of web businesses, who have precisely the information sought by consumers but may lack the resources to surface it effectively.



Consumers must sift through ever-growing mounds of new content—whether it is user-generated, editorial or socially informed. It's no surprise that fully 16% percent of queries on Google.com each day are new.[1] And 18% of consumers consult social media when looking for new information online.[2] Online searchers have become increasingly specific and targeted in their queries, collaborative in their evaluation of options and forced to find novel ways to express their intent when confronted with a sea of infinite possibilities.

Web businesses have not designed customer experience with access patterns in mind and therefore must struggle with the problem of getting found. Many have a browse-centric website architecture, which virtually ensures that desired information is several clicks past the initial platforms (Google, Bing, Facebook) where shoppers begin their quests. From both a consumer and search engine perspective, the information on many websites is just too massive and too hard to crawl.

Simultaneously, the rate at which business inventory, search queries and web pages change on a daily basis is accelerating. Connecting consumer intent with relevant products has never been more challenging.

This white paper discusses the factors that have created and perpetuated the digital discovery divide. It also explores how web businesses can keep pace with today's discoverability challenges and take some basic steps to reveal their "invisible content" so that it is accessible to the consumers that seek it.

Growth in the web shows no signs of stopping

Gone are the days of the static web. The web is ineffably large and continues to grow with no sign of slowdown. Not only is content growing exponentially, the rate of multi-faceted change within this content is, nearly impossible to keep up with. Tracking the web today is equivalent to taking a still photograph of a spinning cyclone; basically, it's a completely insufficient way to measure the dynamic web.

In 1995, fewer than 15,000 domain names had been registered. Today, there are over 350 million domain names registered; another 150,000 URLs are added each day.[3] Within these sites, information continues to amass. Retailers such as Amazon.com carry as much as two billion products at any given time, while eBay can have 500 million.

[1] 25 January 2012. http://www.google.com/ads/answers/numbers.html

[2] 25 January 2012. http://socialtimes.com/the-evolution-and-future-of-content-discovery_b15766

[3] 14 December 2011. http://www.blogpulse.com

Manipulating these and other data points leads to impossibly large numbers. Consider the following examples:

There are 43,982 airports in the world.

The number of possible flight paths is $43{,}982^2$, which is just under 2 billion.

The total number of flight schedules possible is obtained by multiplying 365 days x 24 hours x 2 billion—a number that is significantly larger than all web indices put together.

There are 130 million unique books in the world. If we assume an average of 100 pages per book, there are 13-26 billion book pages out there.

Individuals are playing their own role in fueling growth in web content. The advent of user-friendly consumer web tools and the proliferation of Internet users means that the types of content being uploaded across multiple channels is increasingly varied:

- About 100,000 new blogs are added every 24 hours[4]
- Twitter users send 200 million tweets per day,[5] or 72 billion tweets a year
- 48 hours of video are uploaded to YouTube every minute[6]
- 250 million photos are uploaded to Facebook every 24 hours[7]
- eBay Marketplaces have approximately 99 million active users worldwide across 39 markets, including the U.S.[8]

On top of this user-generated content, the web is increasingly cluttered with a growing body of auto-generated content of questionable usefulness. Take the example of content farms. These sites are generally shallow in nature, yet search keyword heavy, since their goal is to maximize traffic to the sponsored ads that drive their revenues. And they often leave consumers frustrated because they appear high in search rankings but provide little relevance for a particular search.

How do search engines keep up with this explosion of content on the web? Even as search engines improve vertical searches to reach further and further (i.e. flight searches, image searches, Search Plus Your World, tweets, etc.), this exponential growth in web content renders the task of reaching—let alone surfacing—relevant deep content across multiple channels impossible.

Not all content can be crawled...

In addition to the problem of scale, the types of web content that businesses and consumers are generating and posting also makes it close to impossible for all content to be found. Web crawlers (also known as web spiders and bots among other terms) were developed and designed for yesterday's static web. Applied against today's constantly evolving and dynamic web, their design limitations result in a poor user experience. What was once a collection of static interlinked web pages comprised mostly of text has morphed into a dynamic, evolving collection of text (reviews, blog posts, news articles, books, tweets) and multimedia (images, video, etc.).

If the exact content that you seek happens to be in a non-text media format, you might be out of luck. Video, Flash, Silverlight or Ajax-based content that does not contain a text format cannot be indexed or crawled by search engines. Yet if it is not crawlable by a search engine, it is simply invisible. The feeds needed to organize this rich media content are custom to every platform **and** every web business. Classification, parsing and extraction of meaning across this vast range of content continues to be a challenge.

[5] 5 January 2012. techcrunch.com/2011/11/07/twitter-partners

[6] 14 December 2011. http://www.youtube.com/t/press_statistics

[7] 14 December 2011. http://mashable.com/2011/10/21/facebook-infographic/

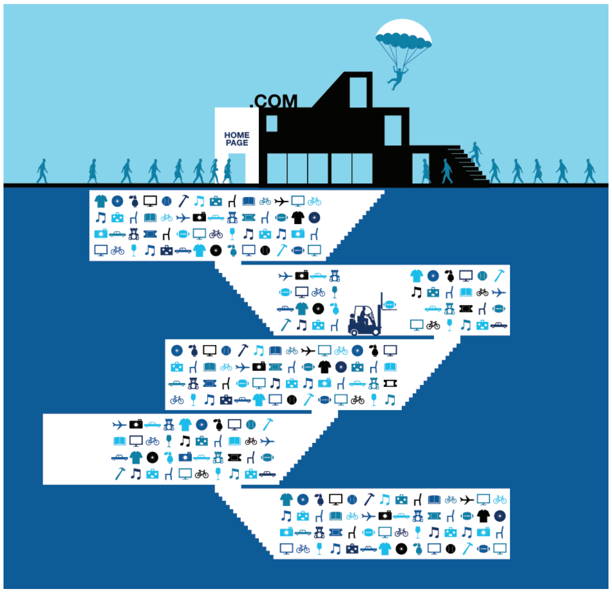
[8] 21 December 2011. http://pages.ebay.in/community/aboutebay/news/infastfacts.html

Keeping up with dynamically changing content

Outside of crawling issues, businesses must also address factors such as their own inventory churn and turnover. On the extreme side, a marketplace-driven website such as eBay.com experiences a 20% churn on a daily basis. A major apparel retailer can have as much as 30% churn on a monthly basis. As consumers, we're all familiar with labels like "Out of stock" or "Only 2 left—Order Soon" beneath our desired item. Accurately mapping dynamic inventory to shoppers is no small task, and it contributes to the discovery divide. For a web business, being able to correctly promote an overstocked item to meet seasonal margin goals or prevent an out-of-stock item from being displayed at the right time goes a long way towards maximizing revenues and keeping customers satisfied and onsite with relevant content.

The browse paradigm perpetuates the problem of invisible content

The discovery divide is further exacerbated by the way that businesses organize their content, specifically in website architectures. Before the rise of scalable search engines and social media, virtually all website visits started on a site's homepage, where businesses strived to establish and build upon a unique brand experience. This resulted in hierarchical website designs that directed visitors from a homepage to a category page, to a subcategory page, and then to a product page. Such a browse-centric paradigm operates the same way that people shop the aisles in a bricks and mortar store, where physical inventory is constrained.



But consider a young partygoer assembling an outfit for Saturday night. She might have to browse 5,000 tops, 3,000 skirts, 1,000 shoes and 500 different necklaces on a retail site. The number of potential outfits she can put together is practically infinite. And the traditional tree structure that relies on browsing simply cannot meet this shopper's needs efficiently.

You might think that this same shopper might use specific keyword searches to gather the results she wants, but it's unlikely that typing "party outfit for Saturday night" will yield the results she seeks. This is because unlike humans, search engine crawlers don't consider each website an entity but rather each document—images, media, reviews, etc.—within the website as an entity. Therefore, information that is relevant to a single user and intention is likely be distributed across multiple pages, compromising the ability of search engines to present content that matches intent. For example, if your intent is to fulfill your craving for pizza in Mountain View, California, you might search for "pizza in Mountain View" or "Italian in Mountain View," forgoing other restaurants like "Amici's" that also have the pizza you want simply because their menu and location pages are considered different entities by the search engine. To better match intent with desired content, search engines would ideally have to combine content from menu and location pages.

But due to the massive size of the web and the structure of web sites, search engine crawlers typically only scratch the surface of the tree structure (the "foliage and branches") leaving most content (in the "root structure" or on social media sites) undiscovered. "Root" content can be vastly distributed throughout the site, buried deep within the site's hierarchical structure or posted on other sites such as Facebook and YouTube. Unless a product or page is 3-4 clicks away from other relevant products or pages, you are virtually guaranteeing that it is too deep for the web spider's crawl.[9] Simply put, the internal link structures of many websites are not optimized for search engine crawlers and don't match the efficiency and speed sought by online shoppers. And these link structures must adapt as the web business's inventory turns over and consumer intentions change.

Today's online shopping reality is that every page is a potential front door for customers. Many consumers will discover merchants and products through natural search as well as social media. Consumers enter websites through countless "side doors" that relate more directly to their interests.

Consumers are highly specific

The proliferation of long tail queries is one driver that is diverting more traffic away from the home page and directly to category or product pages. Seventy percent of search queries are long tail phrases[10]—terms that are generally more than three words. As it relates to shoppers, these long tail queries usually demonstrate specific purchase intent. A shopper who types "Nikon D3100 Digital SLR AF-S VR DX 18-55mm lens" is likely to be researching a product or performing price comparisons in advance of purchase. Shoppers are also increasingly turning to their social networks to seek advice and recommendations. Seventy percent claim reviews from family members or friends exert a "great deal" or "fair amount" of influence; 53% of people on Twitter recommend companies and/or products in their tweets, with 48% of them delivering on their intention to buy the product.[11]

The long tail shopper is here to stay and is uniquely attractive to a business since he or she is much further along the purchase cycle. Yet many web businesses lack the resources to optimize for millions of uncommon, long tail phrases that are of relevance or value to these shoppers. There are just too many terms and too many combinations to manually optimize using the proven techniques that work for mainstream search queries.

Did you mean "sneakers" when you said "tennies?"

Linguistic churn is one of the reasons that 16% percent of Google queries each day have never been seen before. If you're curious what the reason behind this is, consider that the exact words someone uses to enter a search query are highly dependent on age, socio-economic status, peer group, level of education, geographic location and many more unique factors. What you might call "shoe" can be referred to as "trainers," "kicks," "sneakers," "tennies" or "rubber shoes," depending on where you grew up. Different backgrounds, exposures and needs mean that it is increasing difficult for consumers to cut through the noise. With so many ways to say the same thing, how does a web business ensure that its consumers—especially international shoppers—find exactly what they're looking for?

Separately, marketers of these businesses face an unusual conundrum: from a search perspective they need to be as similar as possible to competitors, in order to address the most popular queries that rank for traffic. Yet they also need to differentiate from these same competitors, marketing an "olive cardigan," not a "green cardigan" or a "netbook" vs. a "laptop."

[9] 9 January 2012. http://www.dotcomjungle.com/dcj-university/educate/dont-make-search-engine-spiders-crawl-too-deeply/

[10] 14 December 2011. http://www.problogger.net/archives/2011/05/13/leverage-the-long-tail-of-search-on-your-blog/

[11] 9 January 2012. http://www.bazaarvoice.com/resources/stats

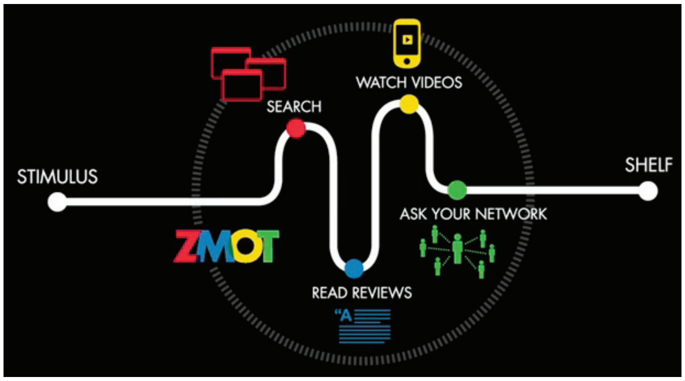
Search engines are cognizant of this problem and continue to optimize and refine their core technology to process language. In fact, Google recently made notable advances in this area, including a best-in-class spelling suggestion system, an advanced synonyms system, and a very strong concept analysis system.[12]

Synonyms are fundamental to language processing, but synonymization—coming up with semantically equivalent words, phrases, clauses, or sentences—remains highly nuanced and full of semantic loopholes.

Consider the case of a multichannel retailer who must attempt to understand user context across a massive inventory of products. A search for "baby stroller" is very straightforward, and synonyms such as "infant" or "pram" or "buggie" easily translate to this phrase in the context of child rearing. But "baby blue onesie" is not equivalent to "infant blue onesie;" in fact, the term "infant blue" is not meaningful to any product catalog.

While understanding and parsing such nuances clearly benefits consumers by making it easier to present the information they seek, the burden cannot be placed entirely on the search engines. In order to develop a robust database of synonyms and bigrams (the pairs of adjacent words that occur in a phrase), a retailer would need to "mark-up" the meta-information on each page with deduced word patterns that reflected predicted language patterns. This mark-up would make it clear to search engines that if it carries the brand Calvin Klein, the word "Calvin" is frequently followed by "Klein" and "v-neck" is likely to be followed by "sweater" or "shirt," but unlikely to be followed by "pants." Using this scientific approach to language, the retailer could exert some control in helping its consumers find the products they seek—in the way that they seek them—in the otherwise highly volatile and evolving linguistic universe of search queries. Of course, this is incredibly time-consuming work and becomes stale shortly after implementation, since language and products continuously evolve.

The browse paradigm needs to address the buy paradigm



Today, products and services have never been more available to shoppers. But wading through the morass of information—from consumer review websites to retail websites to newsfeeds to comparison shopping engines and coupon websites—on the path to purchase is more laborious and daunting than ever.[13] According to Google's Zero Moment of Truth, consumers view an average of 10.4 sources of information prior to making a purchase.[14]

As consumers continue to grow increasingly impatient in their quest for exactly what they're looking for, the onus of improving the user experience lies both on web businesses and platforms. While search engines do surface the breadth of crawlable content that is readily available high in the tree structure, web businesses must ensure visibility of the depth of the content that consumers seek, by adapting their sites in real-time to the changing interests of their consumers. They must strive to make each page a conversion-friendly front door, shifting their websites consciously to a buy paradigm that matches the modern shopper's temperament and needs and provides all of the tools and information required to convert him or her.

The gold standard of a website that is maximally optimized for conversion is Amazon.com. Each page on Amazon is loaded with shopper calls to action, ranging from information (photos, detailed product descriptions, shipping information) to promotions and special offers to user-generated content like reviews and discussions. Should the sought item be absent, Amazon offers multiple ways to navigate to the product or its alternatives, whether through

[12] 21 December 2011. http://googleblog.blogspot.com/2008/07/technologies-behind-google-ranking.html

[13] 15 February 2012. http://www.dot19.com/blog/2011/09/win-the-zero-moment-of-truth/50/

[14] 21 December 2011. http://www.zeromomentoftruth.com/

conventional navigation, recommendations, ads, sponsored links or customer tags. The multiple relevant links in and out of the site provide a “flattened” map for the site, benefiting not only search engine crawlers, but also consumers, who can hone in on their desired product or service in just a few clicks. Each page is likely to be a high quality page by the measure of platforms like Google and Bing. And so begins a self-perpetuating cycle where the sites with the richest, most actionable pages are rewarded with proportionally more indexed pages—reaping more legitimate visibility and frequency in search results.

Addressing the digital discovery divide

Growth in web content will continue to be impossibly rapid. The rate of change within this content and amidst the consumers that seek it will likewise keep pace. And while web crawlers are evolving and refining the methods by which they define, rank and index noteworthy information, web businesses must step up and play an active role as web publishers, ensuring that their hidden content gets found by an increasingly discriminating consumer—one who is looking to quickly accumulate information and make a purchase decision.

Though still in a rudimentary stage, web businesses should explore the technologies that exploit the power of big data on a web scale to adapt and stay several steps ahead of the inevitable volatility inherent to their business. The mountain of data that web businesses already amass across multiple customer touchpoints will only continue to grow and learning to effectively cull insight and deliver measurable action from this data is imperative. An unprecedented opportunity exists to conquer the digital discovery divide—one of the most exciting business challenges today and one with the potential for profound impact in positively transforming the online consumer experience.

bloomreach

BloomReach created the world’s first Web Relevance Engine dedicated to exposing the best content on the internet to the right consumers seeking it on search engines and social platforms.

BloomReach, Inc. 399 West El Camino Real, Suite 100, Mountain View, CA 94040