# Relevancy Redacted: Web-Scale Discovery and the "Filter Bubble"

Corey Davis
*Royal Roads University*, corey.4davis@royalroads.ca



Corey Davis, "Relevancy Redacted: Web-Scale Discovery and the "Filter Bubble"" (2011). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284314965

# Relevancy Redacted: Web-Scale Discovery and the "Filter Bubble"

*Corey Davis, Technical Services Librarian, Royal Roads University Library*

Abstract:
Web-scale discovery has arrived. With products like Summon and WorldCat Local, hundreds of millions of articles and books are accessible at lightning speed from a single search box via the library. But there's a catch. As the size of the index grows, so too does the challenge of relevancy. When Google launched in 1998 with an index of only 25 million pages, its patented PageRank algorithm was powerful enough to provide outstanding results. But the web has grown to well over a trillion pages, and Google now employs over 200 different signals to determine what search results you see. According to Eli Pariser, author of "The filter bubble: what the internet is hiding from you" (Penguin, 2011), a growing number of these signals are based on what Google knows about you, especially your web history; and, according to Pariser, serving up information that's "pleasant and familiar and confirms your beliefs" is becoming increasingly synonymous with relevancy. This session will critique Pariser's concept of the 'filter bubble' in terms of collection development and the possible evolutions of discovery layers like Summon and WorldCat Local, and the challenge of providing relevant academic research results in a web-scale world where students increasingly expect the kind of personalization sometimes at odds with academia's adherence to privacy and intellectual freedom.

*The following is a critique of Eli Pariser's* The filter bubble: What the Internet is hiding from you *(2011), and attempts to capture the conversational nature of the presentation as given at the Charleston Conference in 2011.*

Not that long ago, when different people searched Google, if they used the same search terms, they got the exact same search results. Not anymore. When you search Google ("Technology overview", n.d.), over 200 signals determine relevancy, including location, and—if you're logged into your Google account, or you allow your browser to accept cookies from Google—previous search history. Two people can get two totally different results sets based on the same key words, and increasingly, these results are determined—at least in part—on what you've searched for and clicked on before. The rewards for this kind of personalization are better relevancy, but the challenges are that we will increasingly see results based on what we've looked at before, creating a kind of 'filter bubble' that isolates us from resources we might not otherwise see. The concept of a filter bubble is fairly straightforward as presented by Eli Pariser in his TED talk *Beware online 'filter bubbles'*, which has received around a million views on the TED talks website.

> As web companies strive to tailor their services (including news and search results) to our personal tastes, there's a dangerous unintended consequence: We get trapped in a "filter bubble" and don't get exposed to information that could challenge or broaden our worldview. (TED, 2011)

In 2009, Google started tailoring search results for all users—whether signed into their Google Accounts or not—based on their previous activities on the web. A wide range of websites increasingly use similar algorithms to guess what information a user wants based on what they know about that user, such as their location, previous click behavior, and search history. This kind of personalization is fairly obvious at websites like Amazon and Netflix, but it can be much more subtle on sites like Google and Facebook. The net result, however, is the same. Websites present only the information which is, in a way, similar to information previously consumed by a user. According to Pariser, people in the filter bubble are not as exposed to contradictory perspectives and can, as a result, become intellectually isolated in ways that threaten their ability to meaningfully take part in a society full of uncomfortable truths.

But what does this all have to do with Libraries and their collections? I work at Royal Roads University (RRU) in Victoria, British Columbia. We have about 2000 full-time equivalent (FTE) students but we're pursuing growth aggressively, particularly in the realm of international undergraduate students, and we expect our numbers to rise significantly in



DOI: http://dx.doi.org/10.5703/1288284314965

the coming years. Right now we focus mostly on graduate programs at the Master's level, delivered mostly online via Moodle. RRU was established by the provincial government in 1995, and took up quarters in an old military college. We have a relatively small print collection, with the focus being on our collection of ebooks and article databases. So while we have more and more students spending time with us on campus, people mostly access our collections online.

The Internet Archive's Wayback Machine http://www.archive.org/web/web.php is a wonderful resource. I can use it to see what our library website http://library.royalroads.ca looked like over a decade ago. I can see that we linked to our local catalog and an alphabetical list of article and research databases, with a rudimentary attempt to classify them by program or subject. By 2006 we started organizing things a little differently and placed a catalog search box on the homepage, but we still were experiencing the basic problem of information silos, where that multiple databases containing high-quality and highly sought-after content dispersed across dozens and dozens of different systems and platforms, with no way to effectively search across them all. For books and video, a user had to search the catalogue. For articles, they needed to choose one of many article databases. To help with this, we started creating subject guides. We also did our part during information literacy instruction to help make sense of this complex information environment.

We were suffering from three main issues, as identified by Burke (2010):

1. No clear and compelling place to start research.
2. No easy way to identify appropriate library resources.
3. A lack of awareness of library resources.

Try using the Wayback Machine to look at Google's homepage in 1998, the year the company was founded. There are no lists. The interface is easy and intuitive to use. These are the same qualities that draw us to Google today. While our library has gone through three or four major website revamps in an attempt to help our users navigate and use our resources, Google's main search site has remained remarkably stable.

In 1998 Google co-founders Larry Page and Sergey Brin published an article called *The anatomy of a large-scale hypertextual web search engine* (Brin & Page, 1998). At this time, 'human curation' was a major way that companies like Yahoo! helped people access information on the web, through the creation of directories and other lists. Page and Brin recognized that the web was growing too fast for this kind of organization to continue in a sustainable manor. Human editors simply couldn't keep up. In 1999, according to Danny Sullivan (2008) at Search Engine Land, a majority of major search engines were still presenting human-powered results. But these lists were expensive to create and they didn't scale to what the web was becoming. Lists and directories also didn't deal well with obscure topics: "Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics" (Brin & Page, 1998, p. 107). Google's creators knew humans could not organize the web effectively as it scaled, and that this organization had to be automated. One of the biggest challenges they faced was the unstructured nature of web documents, in contrast to the kind of data libraries were dealing with, such as MARC records. The web was a jumble of different shapes and sizes, not a structured catalog of information based on well-established metadata standards. "The web is a vast collection of completely uncontrolled heterogeneous documents" (Brin & Page, 1998, p. 111).

At this point in the history of the web, Google's basic key to its rapid success was the PageRank algorithm, which was powered by the nature of hyperlinks, rather that primarily by the occurrence of keywords in a particular document.

> PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at considerably more than the sheer volume of votes, or links a page receives; for example, it also analyzes the page that casts the vote. Votes cast by pages that are them-

> selves "important" weigh more heavily and help to make other pages "important." Using these and other factors, Google provides its views on pages' relative importance. (Sullivan, 2007)

Using the relationship between documents to drive relevancy was the key to Google's success. Relevancy was about relationships. Editors couldn't build directories fast enough to meaningfully provide access to the whole of the web.

> We have built a large-scale search engine which addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results. (Brin & Page, 1998, p. 108)

Although many of our individual systems at academic libraries have robust search capabilities, we couldn't until very recently bring these systems together in a meaningful and easy-to-use way. Search has come a long way since 1998. Google now indexes over a trillion pages, all accessible from a single search box. User expectations are very different now that they were in the late 1990s and early 2000s. When people search, they expect Google, not Yahoo! circa 1998, which is how many academic library websites are still organized. And because many libraries still have websites that arguably haven't in essence changed that much since the early 2000s, academic librarians rightly intuit that people are not finding our content and services as easily as they might:

> In a 2009 survey of 66 academic libraries, ProQuest found that 86 percent of libraries feel that faculty and students do not understand the breadth of their collections, and 94 percent think the collections are not explored to their fullest. (Burke, 2010).

Now, we actually have the tools to take us there. At RRU, we are using the web-scale discovery service called Summon from Serials Solutions. According to Serials Solutions:

> Through one simple search to a single unified index, the Summon service provides instant access to the breadth of authoritative content that's the hallmark of great libraries. No need to broadcast searches to other databases —it provides one search box for a researcher to enter any terms they want and quickly get credible results in one relevancy ranked-list. (ProQuest, 2011)

This is not federated or broadcast searching, where queries are sent live to disparate systems and technology such as screen scraping is used to collocate results. This is different. Summon is a pre-built index, just like Google. It searches ebooks, books, videos, theses, articles, and more, and in most cases, it searches the full-text. It is lightning-quick and really big, with a current index of over 500 million items.

And it's that "really big" that brings us back to the filter bubble. Pariser starts his book out mentioning a post on the official Google blog from the 4th of December, 2009:

> Today we're helping people get better search results by extending Personalized Search to signed-out users worldwide, and in more than forty languages. Now when you search using Google, we will be able to better provide you with the most relevant results possible. For example, since I always search for [recipes] and often click on results from epicurious.com, Google might rank epicurious.com higher on the results page the next time I look for recipes. Other times, when I'm looking for news about Cornell University's sports teams, I search for [big red]. Because I frequently click on www.cornellbigred.com, Google might show me this result first, instead of the Big Red soda company or others. (Horling & Kulick, 2009)

Google states that: "By personalizing your results, we hope to deliver you the most useful, relevant information on the Internet." (Horling & Kulick, 2009) Everybody understands that search engines are a big deal. They've changed the way we think about information. They bring the world to us. But this change was big, even in terms of Google. Danny Sullivan (2009) of Search Engine Land called it "…the biggest change that has ever happened in search engines…" According to Sullivan (2009), "until now, search engines have largely delivered the same results to everyone. Two different people could

search for Barack Obama and get back the same set of results."

> The days of "normal" search results that everyone sees are now over. Personalized results are the "new normal," and the change is going to shift the search world and society in general in unpredictable ways. (Sullivan, 2009)

This post very well could have formed the genesis for the filter bubble idea. According to Pariser (2011): "with little notice or fanfare, the digital world is fundamentally changing." (p. 6). Once an anonymous medium where anyone could be anyone, the web has become a tool for soliciting and analyzing our personal data. For example, dictionary.com, according to Pariser, places over 200 tracking cookies and beacons on your computer when you first visit the site. Search for the word "depression" on this website, and you could see ads for anti-depressants on another. "The race to know as much as possible about you has become the central battle of the era for Internet giants like Google, Facebook, Apple, and Microsoft." (Pariser, p. 6) The more personally relevant their information offerings are, the more ads they can sell, and the more likely you are to buy the products they are offering. And it works! Amazon was a pioneer of personalization. The company recorded revenues of $24.5 billion during 2009, an increase of 27.9% over 2008. (Datamonitor, "Amazon, Inc.", 2011). It makes billions by, in great part, predicting what you're going to buy. And for Google, the more relevant the results, the better they can target ads, the more money they make.

Advertising is, to understate it, a big deal for Google. Google made almost $30 billion in 2010, up 24% from 2009 (Datamonitor, "Google, Inc.", 2011). 96% of that revenue comes from ads. According to Pariser, if personalization was all about advertising, that wouldn't be so bad, but it's effecting how information flows on the web. If you get your news from Facebook (and, according to Pariser, 36% of Americans under 30 get their news from social networking sites), you may only see the things that your friends like. This is Pariser's central critique.

> …these engines create a unique universe of information for each of us—what I call the filter bubble—which fundamentally alters the way we encounter ideas and information. (Pariser, p. 9)

But why do we personalize? Too much information leads to what blogger Steve Rubel (2007) calls "attention crash". Personalization helps search providers filter through truly massive amounts of information need to get to what the user wants. If search results are not personalized, it's much more difficult for a search engine to determine what a particular user wants. And the signals users send are pretty pathetic. "A number of studies have shown that a vast majority of queries to search engines are short and under-specified and users may have completely different intentions for the same query." (Qiu & Cho, 2006, p. 1) According to Jansen, Spink, and Saracevic (2000), who analyzed over one million Web queries by users of the Excite search engine: "we found that most people use few search terms, few modified queries, view few Web pages, and rarely use advanced search features." (p. 233) The mean number of words per query was 2.21. 31% of all queries used only a single word. Most users searched for only one query and did not follow with successive searches. Silverstein, Marais, Henzinger, and Moricz (1999) analyzed an AltaVista Search Engine query log consisting of approximately 1 billion entries for search requests over a period of six weeks. This represents almost 285 million user sessions. They found that the average number of terms in a query was 2.35. For 85% of the queries only the first result screen is viewed. 77% of the sessions contain only one query. Another project (Wang, Berry, & Yang, 2003) analyzed 541,920 user queries submitted to and executed in an academic website during a four-year period, and found that 38% of all queries contained only one term, with a mean query length of two words. And analysis of Elsevier's ScienceDirect platform (Ke, Kwakkelaar, Tai, & Chen, 2002) revealed that "approximately 85.2% of queries contained one, two, or three terms, although the average query length was 2.27 terms." (p. 275)

Personalization can make search more relevant when queries are generally short and ambiguous. According to Qiu and Cho (2006), "…a user's general preference may help the search engine disambiguate the true intention of a query." (p. 727) Matthijs and Radlinski (2011) examined personalizing web

search using long term browsing history and found that "… personalization techniques significantly outperform both default Google ranking and the best previous personalization methods." (p. 34) So, when Google personalizes, they are interested in increasing the relevance of search results, and this is based on sound evidence. Personalization truly does increase relevancy.

> When you search using Google, you get more relevant, useful search results, recommendations, and other personalized features. By personalizing your results, we hope to deliver you the most useful, relevant information on the Internet. (Google, 2011)

Google looks primarily at search history. If you're signed in to a Google account, Web History is used, and if you are signed out, Google's servers link to an anonymous browser cookie that tracks your click history for up to 180 days (Google, 2011).

This is a big problem for Pariser: "what you've clicked on in the past determines what you see next." (Pariser, p. 16) This can lead to something he calls informational determinism: "in the filter bubble, there's less room for the chance encounters that bring insight and learning." (p. 11) Pariser spends a good deal of time on the importance of serendipity. He argues quite convincingly that creativity and new ideas and the solutions to our most intractable problems come from chance encounters with new or challenging people and ideas. The filter bubble threatens this. It's something that librarians and many scholars recognize the importance of too. There is a certain irony here. Before PageRank, search results could be humorously irrelevant, and the balance between locating relevant resources and discovery through serendipity was skewed to the side of chance encounters. Google has worked hard since 1998 to lessen these kinds of chance encounters, as mentioned in Brin and Page's 1998 paper: "While the results are often amusing and expand users' horizons, they are often frustrating and consume precious time." (Brin & Page, 1998, p. 116)

The need for more relevant results in order to increase advertising revenue in great part drove Google's move to personalization. Danny Sullivan from Search Engine Land, when writing about this move, used a Library metaphor to describe how it could work, and hinted at how informational personalization was done at libraries in a time before Google and other search engines:

> Imagine you're in a library—the classic metaphor for a search engine and how it interacts with a searcher, from when WebCrawler's Brian Pinkerton used to explain how they worked back in the 1990s. Someone walks in and says "travel." In a library, the librarian would ask more questions, to try and understand what they want. Early search engines didn't do this. They couldn't do this!
>
> Over time, search engines tried to do the library-style conversation by offering related searches, as a way to get searchers to refine their queries. Then Google took a huge leap last year by making use of your previous query to refine your results. That makes sense and doesn't seem to require any particular reason to ask for user opt-in. Again, imagine the librarian. It would be unreasonable to expect them to forget the last thing you said in a conversation you were having, as they tried to help you. Unreasonable and unhelpful.
>
> But would you expect the librarian to help you by remembering everything you'd asked over a half-year period? That might be helpful, sure, but it might also be eerie. But this is what Google is doing now. It remembers everything you've searched for over 180 days, and it uses that information to customize your results. To alert you about this huge change, it made a blog post on Friday afternoon. That's it. (Sullivan, 2009)

So, in some ways, Sullivan is making a connection here. Google is attempting to automate what librarians have always done, which is to use human judgment and experience to mediate access to scholarly works and other kinds of information. According to Jane Burke (2010), Vice-president of ProQuest:

> Increasingly, libraries are viewed as irrelevant to the research process, leaving them vulnerable to being cut, both financially and from the mind of the end user. However, new ways of discovering

> content in library collections holds the promise of returning the researcher to the library.

For this reason, tools like Summon hold great promise: "web scale discovery efforts aim squarely at Google as the competitor and mimic that search engine's characteristics of simple, easy, fast." (Burke, 2010)

It has taken a while to settle in with Summon, but for the most part, our users at RRU are happy with this new and important tool. Initially, relevance was a bigger problem than it is now.

Could Summon search results be personalized using cookie technology similar to that employed by Google to track users' past click history? Should we employ this kind of technology? In a 2002 thought-peice, Surprenant and Perry wrote:

> Being able to see the student allows the Cybrarian, with the aid of a diagnostic algorithm which has access to their infoprofiles, to help gauge how comfortable/secure the student is with the current skill set and to gain some insight into the level of his/her developing abilities/capabilities.

Does the future really hold a spot for a human being—a librarian—to access an 'infoprofile' of a particular individual in order to help them find the best information available? Does this kind of mediation of experience have a place in a world where complex personalization algorithms could hypothetically determine relevant without our help? What is the benefit to our users in having to interact with us, rather than an online search tool?

Surprenant and Perry (2002) also envision a high-level of personalization in the future: "Communicating through Virtual Reality helmets and V-mail, and utilizing diagnostic tools to customize resources to individual profiles, cybrarians will provide effective support for problem solving and discovery groups."

These are important concepts to ponder. They strike at the tension within academic libraries that disintermediation represents, where systems become usable enough—even in the face of increasing complexity—that reference and instruction are seen as less and less important.

Pariser ends his book with a call to 'algorithmic literacy', which means understanding the basic operating principles of the systems you rely on for information. Librarians can and should play a greater role in explaining not only the information landscape for scholarly and other resources relevant to students and faculty, but how the systems most commonly used to gather information actually work, both in terms of benefits and risks.

**References**


Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Burke, J. (2010). Discovery versus disintermediation: The new reality driven by today's end-user. Paper presented at the *VALA2010: Connections, Content, Conversations, 15th Biennial Conference and Exhibition, Melbourne,* 9-11. Retrieved from http://www.vala.org.au/vala2010/papers2010/VALA2010_57_Burke_Final.pdf.

Datamonitor. (2011). *Amazon, Inc.* Datamonitor Report. Retrieved from http://search.ebscohost.com.ezproxy.royalroads.ca/login.aspx?direct=true&db=dmhco&AN=2B52E1D8-E964-4D7F-8B1B-C48DBC97815F&site=bsi-live

Datamonitor. (2011). *Google, Inc.* Datamonitor Report. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=dmhco&AN=5B199F61-608D-4923-B4A3-F5EE15285ADE&site=bsi-live

Google. (n.d.). *Technology overview*. Retrieved from http://www.google.com/about/corporate/company/tech.html.

Google. (2011). *Personalized search basics.* Retrieved from http://support.google.com/accounts/bin/answer.py?hl=en&answer=54041.

Hoeber, O., & Massie, C. (2010). Automatic topic learning for personalized re-ordering of web search results. *Advances in Intelligent Web Mastering-2,* 105-116. Retrieved from

http://www.cs.mun.ca/~hoeber/download/2009_awic_misearch.pdf.

Horling, B., & Kulick, M. (2009). *Personalized search for everyone.* Retrieved from http://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology 52*(3), 226-234.

Ke, H., Kwakkelaar, R., Tai, Y., & Chen, L. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library & Information Science Research, 24*(3), 265-291.

Matthijs, N., & Radlinski, F. (2011). Personalizing web search using long term browsing history. Paper presented at the *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining,* 25-34. Retrieved from http://research.microsoft.com/pubs/139933/MatthijsRadlinski_WSDM2011.pdf.

Pariser, E. (2011). *Filter bubble : What the internet is hiding from you*. New York: Penguin Press.

ProQuest. (2011). *The Summon™ Service.* Retrieved from http://www.serialssolutions.com/discovery/summon/

Qiu, F., & Cho, J. (2006). Automatic identification of user interest for personalized search. Paper presented at the *Proceedings of the 15th International Conference on World Wide Web,* 727-736. Retrieved from http://oak.cs.ucla.edu/~cho/papers/qiu-ui.pdf.

Rubel, S. (2007). *The attention crash: A new kind of Dot-Com bust*. Retreived from http://adage.com/article/steve-rubel/attention-crash-a-kind-dot-bust/117325/.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. Paper presented at the *ACM SIGIR Forum, 33*(1) 6-12.

Sullivan, D. (2008). *Search 4.0: Social search engines & putting humans back in search.* Retrieved from http://searchengineland.com/search-40-putting-humans-back-in-search-14086

Sullivan, D. (2007).*What is Google PageRank? A guide for searchers & webmasters*. Retrieved from http://searchengineland.com/what-is-google-pagerank-a-guide-for-searchers-webmasters-11068.

Sullivan, D. (2009). *Google's personalized results: The "new normal" that deserves extraordinary attention.* Retrieved from http://searchengineland.com/googles-personalized-results-the-new-normal-31290.

Surprenant, T. T., & Perry, C. A. (2002). *The Academic Cybrarian in 2012: A Futuristic Essay.* Retrieved from http://www.docstoc.com/docs/15482044/Full-Text---Alphafduedu.

TED. (2011). *Eli Pariser: Beware online "filter bubbles".* Retrieved from http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.html.

Wang, P., Berry, M. W., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743-758.