Educating for Search: Understanding the Past and Present Search Technology to Teach for Future Resilience

Dave Crusoe
Educating for Search: Understanding the Past and Present Search Technology to Teach for Future Resilience

Dave Crusoe

Who taught you web search? What do you remember learning? My formal introduction to web search was made in the mid 1990s, when I was in high school. I was taught to think about search as a way to excavate resources that included or excluded specific characteristics. My instruction was linked to set theory and Boolean algebra, which are the visual, mathematical, and conceptual languages that enable us to illuminate objects’ relationships. We might use a Venn diagram, for example, to visually separate sweet from sour fruit, or distinguish those fruit that display both sweet and sour characteristics from others. In the United States, students are typically introduced to the Venn diagram in middle school. More sophisticated Boolean algebra, such as written notation, may be part of a high school math class.¹

Central tenants of Boolean algebra and set theory were adopted and applied by computer scientists in their quest to solve problems generated by the volume and messiness of digital information. Early search pioneers – and developers since – have applied Boolean algebraic logic to help users and software identify relevant information. Early search engines, for example, required users to apply literal Boolean as a means to define their queries. Search engine Excite.com, for example, allowed searchers to apply AND, OR, AND NOT logic, and even to bracket terms parenthetically to apply term precedence.²

This symbiosis – a relationship between search logic, search engine function, and education – has led to a situation in which set theory and some Boolean concepts are taught alongside search engine instruction. The result is a situation in which education about search logic, including keyword structure and syntax, presents a false concep-

---

¹ The Common Core references algebraic set theory in its high school appendices. See http://www.corestandards.org/assets/CCSSI_Mathematics_Appendix_A.pdf for more information.

Fig. 1. Venn diagram.
tual depiction about how search engines work at a time when common search engine function is far less about explicit algebraic operation and increasingly about how a user taps prior knowledge to construct, manipulate, and rephrase a series of queries.

While observing a number of U.S. middle and high school class sessions focused on teaching web search to students, I learned that educators’ understanding of how search engines work were often reflective of an earlier or different era. As a result, I concluded that educators would benefit from a brief, but holistic depiction of how search engines have historically functioned and evolved since, due to a number of pressing information challenges, and a discussion of why, as a result of shifting technologies, instruction must shift from teaching logic to teaching search function and practices. The result is, I hope, an experience that engages educators in cognitive dissonance over the nature of the topical education we provide, with the hope that a stronger educational experience will emerge.

A Starting Point: Boolean in the Digital Space
To have a productive discussion about the use of Boolean logic in the digital space, it’s important that we share an understanding of what Boolean algebra actually is, and how its logic functions. Boolean algebra includes many of the familiar conceptual, linguistic, and symbolic tools we use daily: ANDs, ORs, and NOTs, among others. These tools can be expressed linguistically (‘AND’), notationally (∧ for AND), and programmatically (‘&’ in many programming languages).

There are also a number of logic operations that may be less familiar, save to programmers. Some search engines may still utilize precedence grouping, which we know as brackets (and). Bracketing allows certain set operations to happen before other set operations. Another operation, a logical disjunction, (‘XOR’, expressed symbolically as ⊕), indicates that an operation validates as true if A or B are true, but not both A and B simultaneously. Other logic structures in Boolean Algebra include equivalence, tautology, and contradiction. However, the user-facing syntax utilized by most search engines is less technically powerful than the true syntax of full Boolean algebra. For example in the searches we conduct, we may utilize:

– **AND**: A conjunction implies a necessary relationship between keywords such that subsequent words are required to appear in results listing the initial term. For example, show me all fruit that are [sweet AND sour].

– **Implied AND**: When an AND is implied, keywords are implicitly linked with a conjunction. For example, parsing [sweet sour] with an implied AND would be recognized as [sweet AND sour].

– **Implied phrase**: Another interpretation of keywords is by implied phrase. This is a transformation of the multiple keywords in a query into an explicit phrase for the purposes of search matching. For instance, were [sweet salty sour] an implied phrase, we’d receive results that literally contained the specific words ‘sweet salty sour’ together, in that order.

– **Exact phrase**: Explicit indication of a phrase for the purposes of search matching.
Yet depending on the search engine, the effect our syntax has upon results may differ based upon how developers have stated that the engine should act. For example, for many years, Google utilized an implied ‘AND’ between keywords that were entered without a phrase. According to online discussions, that is no longer the case.3

It’s helpful to think about the syntax we’re provided as one layer in a series of interpretations. Programmers have provided the user with Boolean functionality that they interpret in managing the communication between input (what we provide) and output (what they have to instruct the engine to do, to provide us with results). For example, let’s pretend that we’re searching a simple literary database for books by Jules Verne. As that query translates directly into MySQL, a popular database engine, it might look something like this:

    Select title,year,dewey-id from books where author = (Jules Verne AND language = English)

In this case, we’re asking the database to tell us the title, year, and the Dewey Decimal information from a table of books, when the author of the book is Jules Verne and the language is English. Yet search engine creators have found that related or similar results might also be helpful – but not included by the literal query. As a result, the search we type is further interpreted. Imagine we’re searching for a book by Jules Verne, but we misspell his name as ‘Julez Vern’. Search engine programmers have crafted their algorithms not only to show us all strict matches, but also those matches that sound the same. Their algorithms automatically translate our query into something like:

    Select title,year,dewey-id from big-search-database where SOUNDEX(author) = SOUNDEX(Julez Vern)

This is, of course, a gross oversimplification of how a search engine might locate information, but it serves to illustrate how queries we have expressed in simple terms evolve into queries that search engine operators might use to develop and display results. Understanding this shift puts us on a good footing to explore the pressures that have forced the relationship between information seeking practices and search to evolve past the ‘Boolean’ phase. In this next section, we’ll investigate those pressures.

**Evolution: Pressures to ‘Do Better’**

Perhaps the most fundamental goal of any engine is to provide users quickly with high quality content that responds accurately to the information need. To accomplish this feat, major search technologies break various challenges into separate, interconnected systems. Query pages help support and guide users through the process of creating accurate queries for the engine to parse. The query engine probes an index, built by an indexing engine typically spread across tens, hundreds, or thousands of machines, to identify individual results. In turn, the indexing engine builds its results with input from a crawler, which identifies and retrieves content to feed into the index.

---

Along each step of the process, various algorithms are applied to rank, weigh, judge, and vet content on various quality dimensions. At each step of the way, internal and external pressures provide impetus for technological evolution. Perhaps the most important factor is economic – a business imperative. If the user isn’t satisfied, then he or she will move search activity elsewhere. Yet the pressures are primarily technological.

One primary pressure is the need to eliminate spam. Search engineers face the dual task of immediately surfacing query-relevant documents, while reducing or eliminating documents that have been made to seem relevant to indexing algorithms through search engine optimization techniques. Search Engine Optimization (SEO) is the practice of artificially enhancing the ranking or rating of a website to increase traffic. The central challenge for search engine creation is to determine not only what is legitimate, but also what is relevant and high quality. You can expect that, for any given search, a significant body of content has been excluded for its potential to be spam.

Alternatively, there are many websites or resources that present high-quality information but have not been optimized in any way. As a result, these sites are tough or even impossible to discover. Publications, particularly content published prior to digitization or accessible beyond a paywall, are an especially common example. Historical research or dialogue, including materials published by the web’s precursor, Gopher, are also hidden from search engines’ view. Similarly, not all discussion forums are entirely public. Microcommunities, including Howard Rheingold’s Brainstorms bulletin board system, which has supported fifteen years of discussion about the future of technology, are accessible only through a membership process. This important material can be classified as part of the ‘deep’ web.

Either way, what’s important to understand is that search engine providers work to balance the results they provide so that the results reflect a minimum of ‘spammy’ returns, and a maximum of high-quality results for any query. The result of their work is that certain content may or may not actually appear for any given search. Another pressure is to ensure freshness. Freshness relates to the newness of content and is based upon how quickly crawled items can be indexed and made available to searchers through search results. As technologies improve, freshness increases. For instance, users expect to find the latest news articles or the latest version of a website when they visit a site. Just as with spam filtering, content is filtered for its freshness. A particular news article, for instance, may not appear on the first or second results page if much newer news articles appear. Alternatively, a particularly well-liked older article may appear above much fresher news or content.

Beyond serving spam-free, fresh results, search engine providers want to understand the intention beyond the query. Is a user conducting a navigational query (e.g., to find


a specific website), informational query (to ‘acquire some information assumed to be present on one or more web pages’), or a transactional query (‘to perform some web-mediated activity’)?

Delivering results that match the user’s desires will ensure satisfied and returning customers.

In his testimony before Congress in 2011, Google CEO Eric Schmidt related that in 2010, his company undertook thousands of algorithmic tweaks and, ultimately, implemented ‘516 changes’ to improve search quality. This volume of changes virtually ensures that search results do change over time. Is this important? I believe that it is, and that it’s vital we view search engines not as stable technologies, but as vibrant and frequently evolving tools. The content we retrieve has been filtered and ranked, although we may not retrieve all there is to know about a topic, and what we do retrieve will not be the same for long. These are quiet, non-obvious changes.

Yet one of the most profound advances doesn’t have to do with spam filtering, speed, freshness, or even content ranking algorithms. It has to do with providers’ desire to read far beyond the simple textual query that a user has entered – the search provider wants to understand and interpret what the user really wants to know. What the provider utilizes to produce this portrait of intent is called signals.

**The Result: From Interpreting a Query to Interpreting Intent**

Writing about the future of search, a 2002 research team proposed: ‘We need a new generation of web searching tools based upon a more thorough understanding of human information behaviors. Such tools would assist users with query construction and modification, spelling, and analytical problems that limit their ability or willingness to persist in finding the information they need.’ Ten years later, the query (in the formal sense) remains relevant; it is a primary means through which a search engine divines what a searcher seeks. So what is the query? The text of the query is only a small part of what we might conceive of as the entire request. The query is a combination of who and where the searcher is, what the searcher tells the engine to retrieve, and what the searcher may have told the engine in the past. These are ‘signals’, and increasingly search engines collect signals to interpret user intent.

Geospatial reference information – location – is a powerful signal, particularly for many commercial queries. For instance, a searcher interested in movie times may actually intend to search for movie times local to their current location. As I type ‘movie time’ into my search bar, I receive a suggestion: “movie time” atlanta’, which is my current location. Were I to wish for movie times in New York, I’d have to craft my query to be more specific.

---

Some search engines may also tap the history of our searches to predict the kind of information we’d like to gain through our current or future searches. Were I conducting a research project about a particular historical figure, this tailoring could be helpful – particularly if it were to help me refine my keywords (the syntax I’d enter). For example, Google

[...] now considers over 200 factors in assessing site quality and relevance. When a user types a query into Google Search, Google’s proprietary technology analyzes these signals to provide a determination as to what the user is looking for. Google uses this ever-improving technology to organize information, rank sites, and present results to users. Google’s search results are ultimately a scientific opinion as to what information users will find most useful.10 [emphasis added]

We can read this to mean that the language we use to express our query, whether typed, spoken, or otherwise communicated, may not be interpreted literally. Search engines take into account a variety of linguistic devices to better match our intent with quality results. Stemming, for instance, is used to match queries like ‘child’s’ to ‘children’s’, or ‘runner’ to ‘run’ and ‘running’. So, a search for one word may yield other, very similar stems. Additionally, the engine might match for synonyms. A search for ‘stationary’ may also yield information about ‘office supplies’ and ‘staples’ since these are a popularly used, similar terms.11

The way that Boolean syntax is applied is also shifting, although the language of the syntax has not. By now, we’re familiar with AND, NOT, and OR, which are the basic Boolean search operators. Search engine providers frequently modify how these operators interact with their search technology to display matching results. For example, until fall of 2011 Google search software applied AND to link keywords, and utilized a ‘+’ to indicate essential and exact keywords.12 In fall of 2011, the syntax evolved to reduce the utilization of implied AND and transform the ‘+’ operator syntax to quotes, “”.

Furthermore, it is only possible to receive exact responses to queries via its Verbatim tool, which provides search results ‘using the exact keywords you typed’.13 Queries conducted through its standard tool may include items not specifically designated by AND or + syntax.14

While writing this essay, I was also surprised to find that Google’s interpretation of the long-standing quotation mark syntax, “”, which request exact-phrase matching,

14.  For example, at one point the help text of the Google Basic Search page was amended to read: ‘A particular word might not appear on a page in your results if there is sufficient other evidence that the page is relevant.’ This change was cited by respondents in help discussions in Google forums, including http://productforums.google.com/forum/#!topic/websearch/x3Pt5XB29Pc.
may also be shifting in application. These searches now return results with characters between the elements (‘standardization test’ yields matches for ‘standardization: test’) and searches with a singular, plural, or passive voice return some ‘corrected’ results (‘world serie’ returns results for ‘world series’). Autocorrection is also enforced. Exact phrase searching for ‘thsi American life’ using Google’s standard tool returns results for ‘This American Life’, yet searching for the same incorrect phrase using Google’s Verbatim tool yields a significantly different results set. As a point of comparison, Bing’s search results reflect a similar approach to exact phrase auto-correction. A Bing search for ‘literar device’ returns only results for ‘literary device’, despite that Bing’s syntax guide indicates that quotes should return ‘results that contain the specified phrase, exactly’.15

The increasing emphasis on geolocation, an increase in the use of fuzzy matching and synonyms, the rapidity with which the index is updated, and the evolving nature of how our precise syntax is interpreted points to the shifting nature of our query syntax. Our search is no longer about the keywords we enter. It is nothing less than an interpretation of the sum of all signals we send to the engine. In sum, we must think of the query as the sum of all signals we send to the engine and not simply the words that we type into the search box.

Search Education in a Signals World
To respond to these shifts, search education must focus on what’s appropriate, of high value, and leads learners to develop an understanding responsive and resilient to technological change. Where is instructional time best spent?

First: Teach About How Search Engines Work
First and most fundamentally, it’s important for searchers to possess a basic understanding of what a search engine is and how it works. It’s essential to teach that a search engine may crawl large portions of what we can conceive of as the internet but that not all portions will be visible to us through the searches we conduct. Engines’ parsing or ranking may be beneficial, for instance, in the filtering of spam, but it may also be detrimental. In which cases would it be detrimental, and what are some ways to overcome artificial barriers? What tools might be built into search engines that can allow a searcher to take more advantage of what the technology has to offer? By helping learners understand what a search engine is, how it works, and how and where to learn more about it, we provide them with tools to adapt more easily to the unforeseen.

Second: Teach Learners to Use Keywords, Phrases, or Sentences that are Likely to Occur in a Page
On major search engines, keywords are the primary signal a user shares with the search engine to drive the search engine’s results. But what should those keywords be for any particular search? How do we know? In some engines, particularly in academic engines like EBSCO, keywords may have to match pre-defined keywords that relate to an article or specific entry fields.16 But these keywords are decreasingly correlated

with success when using major search technologies. In the case of Google and other providers, it’s more important that keywords or phrases target the language that a web page may use, or even attempt to replicate a sentence that might be found on a particular page.

What does this mean for the teaching of Boolean keywords like AND, OR, and NOT? It is decreasingly important that we educators teach young people specific keyword syntax, and much more important to teach about the overall technological function that search engines fulfill.

Third: Teach Patience and Refinement
Educators must also teach patience. Fellow educators and I have described this as refinement thinking. It’s quite common for searchers to select one of the first four options on a page, but far less frequent for one of the first four options to be ‘the best’ or ‘the only’. Social psychologists call this ‘cognitive miserism’, in which ‘the basic idea is that people do not like to take a lot of trouble thinking if they do not have to’. Yet it’s an important habit to break, particularly because the quick path is so very quick. How often have you observed the same practices? What were your responses, and how might they be improved?

One good response is to engage learners actively in a discussion about the merits and drawbacks of information from any particular source, and to challenge them to synthesize information from multiple sources. But don’t treat this as a passive assignment by simply grading the number of sources a learner might cite. Instead, ask for elaboration and thinking – how do sources corroborate one another, and where do they differ? Which sources seem to be of higher quality than others, and why? What can one take from this knowledge to future research? These are all valuable questions to ask of learners.

Finally, for older learners, we must at least mention the code-switching necessary to make the transition from successfully searching sophisticated web engines to the far simpler search technologies used by libraries and research databases. Quite often, research databases function with explicit Boolean operators. Negotiating these databases is an essential skill for students of secondary and higher education institutions. Yet we must recognize and differentiate between the instruction required for success in one ‘code’, that of web engines, and the other, or older Boolean-based databases.

Fourth: Teach About Content Sources
No discussion would be complete without relaying the importance of recognizing high-quality sources. High-quality sources for information provide a level of vetting, and potential detail, beyond what many readily available sources might yield. A search for ‘how a cat purrs’, for example, yields a large number of results from sources as divergent as content aggregators (http://ask.com and http://answers.com), government websites (http://www.loc.gov, the Library of Congress), and publications (such

---

17. This term emerged through discussions with D. Abilock and others on the ALA’s Information Literacy listserv.
as *Scientific American*) – all within the first dozen Google search results. The content and probable quality of the results is as diverse as the result sources themselves, but knowing and understanding which locations are most likely to be detailed and of high quality (such as the Library of Congress or *Scientific American*) is important.

Yet, we must ask: how do learners come to identify which sources are of the highest quality and which are of dubious quality? At school? Home? Whose responsibility is it to ensure that a young person can identify a small handful of high-quality resources? As educators, I believe it is our responsibility. A librarian-educator recently remarked to me that she wished subject educators would not only grade classroom assignments, but would also grade the overall quality of resources that a learner had utilized to derive the information in their report. Additionally, what if it were an assignment to improve upon resources that had already been identified? Perhaps, as educators, we could work harder to engage learners in critiquing their own sources – and at an earlier age.

**Conclusions**

In the past decade, search engine technology has become immeasurably more sophisticated. Paradigms that were once helpful, including the user paradigms of Boolean logic as a tool for searching, have faded in importance as more sophisticated paradigms emerge. In crafting this essay, I wanted to provide a grounding for how search engines function and the challenges that search providers face. By sharing a common understanding and respect for these challenges, I believe we can provide ourselves with a firm footing through which to engage our learners in understanding the same.

Yet there’s more. I hope this has prompted thinking about how search education must be delivered at a time when the search technologies themselves have moved beyond the straightforward and logical tools that we may have once learned. Many of us educators must evolve our conceptual understanding of what it is to teach search. To teach search now, we are best teaching how search engines actually work. What do they find – ‘index’ – and why? What’s missed in the process? With that understanding in place, it’s trivial to move forward with teaching about why patience, ‘refinement thinking’, is an important practice. Yet that practice won’t be embedded unless we educators take more care with our students – not only by teaching them how to synthesize information and about sources they can trust and reference, but by helping them to critique the sources they’ve already found. As a whole, these practices, well learned, will enable future researchers to thrive now and to be resilient through inevitable technological evolution.

Special thanks to Mary Roth, IBM Research, and Sara Armstrong, OnCUE, for their feedback.
References


Google. ‘Search for Exact Words or Phrases’, http://support.google.com/websearch/bin/answer.py?hl=en&p=g_verb&answer=1734130.


