Radio (*Graduate Assistant*) University of Illinois at Urbana-Champaign [Champaign, IL]

LOOKING INTO INFINITY: USING THE DEEP WEB AS A PEDAGOGICAL TOOL

ERIK RADIO

SEARCH ENGINES AND THE DEEP WEB

The evolving information landscape has created access to a plethora of resources from a diverse array of content providers. This diversity naturally necessitates considerations for how students are taught to navigate the web and evaluate information on it efficiently and thoroughly. Critical to their success is developing a deeper knowledge of how information on the web is retrieved and also of barriers to resource access that hinder retrieval systems. Information resources that are not indexed or retrieved by search engines reside in what is commonly known as the deep web. The University of Illinois at Urbana-Champaign developed a workshop on the deep web in order to equip students with a background on this growing body of buried information and how to explore it. As a framework for reuse, the structure of the workshop is detailed below as well as other ways that instruction in this essential resource can be presented to users.

The information retrieved by search engines such as Google, Bing, and Yahoo! can paint a deceptively simple picture of all the information that is available on the web for a given topic. These powerful engines remain popular tools for research given their accessibility and familiarity to users. However, the content that they provide to users belongs to only a small segment of the Internet that can be easily indexed by web crawlers. The remaining, much larger portion of the Internet, the deep web, consists of a variety of resources that for several reasons cannot be effectively crawled and indexed by search engines. Far from a collection of esoteric web resources, the deep web in large part consists of valuable information for users, even leading some to conceive of it as an 'el Dorado' of information resources (Spencer, 2007).

A general understanding of how search engines work is instructive for understanding why certain types of resources belong to the deep web at all. While the processes running search engines are quite complicated, the general functions underlying them are not difficult to grasp. One of the primary tools behind information retrieval are automated tools known as web crawlers or spiders. A crawler reads a web page, captures relevant information such as headers, metadata, and content and then, as their name suggests, crawls to the next web page through links from the initial page and then repeats the process. The information gathered in a crawl is then sent back to a centralized location for indexing. When a query is provided to a search engine, the indexed information is then used to gather potentially relevant resources that are then presented to the user.

There are several barriers that prevent web crawlers from being able to index the entirety of web resources and thereby creating the well of resources that comprise the deep web. While search engines are continuously attempting to improve their retrieval systems to incorporate these resources into their results, the work is ongoing.

Probably the greatest hindrance to crawlers is that most information resources reside in databases instead of web pages. Crawlers are very adept at indexing documents written in Hyper Text Markup Language (HTML) for use in their retrieval results but databases require a different kind of query structure to effectively retrieve documents. Given the variety and complexity of queries that can be given to databases it is not difficult to see why crawlers, as automated tools, are unable to replicate this activity themselves and index the retrieved documents or associated metadata. Even if there were an easy way for crawlers to rapidly generate queries, there remains the difficulty of ensuring that the queries are actually promising and retrieve documents in quantity for efficient crawling but also with an acceptable recall rate (Zheng, Wu, Cheng, Jiang, & Liu, 2013).

While databases by their nature create a significant barrier to the indexing of their resources, they are far from the only element that form the deep web. Metadata, while universally acknowledged as an integral element in the management of electronic resources and a prime key for discovery systems, can actually be an obstacle to a retrieval system. As Chutter Yasser notes, incorrect elements and values, information loss, and inconsistency between metadata records are just a few of the ubiquitous problems that disrupt and compromise retrieval systems (Yasser, 2011). If the splash page for a resource has bad, incomplete, or insufficient metadata, it matters little that a crawler was able to index it; an effective relevance judgement cannot be made. It is no new tale that the growing body of digital resources requires meaningful metadata for their discovery, but unless this is achieved they will languish unused in the most obscure corners of the deep web.

There are numerous other problems that hinder web crawlers from indexing a resource, but three deserve brief mention. Following links is the primary way that crawlers move from one page to another, but a page that has no links or perhaps only internal ones is essentially in isolation; a crawler could never find it unless it was registered. Second, idiosyncratic web pages that rely on uncommon markup or scripting languages to display content also pose problems for crawlers expecting HTML. Finally, pay walls represent a significant, if somewhat obvious, deterrent to indexing.

It may be justly asked just how much information actually resides in the deep web as opposed to the surface web (another name for the information that is retrieved by common search engines). Since there is no tool that can perform a federated search of all web content, it is understandably difficult to quantify what is by its nature an unknown corpus. However, the general consensus is that the deep web is several magnitudes larger than the surface web. Indeed, Marcus Zillman estimates that the deep web is in the vicinity of trillions of pages whereas common search engines can only find hundreds of billions of pages (Devine & Egger-Sider, 2014).

As web resources evolve, new barriers arise that drastically limit what kinds of information can be retrieved. As Devine and Egger-Sider note, given the diverse number of information resources available, search engine engineers must make decisions and compromises on what to crawl so as to most effectively serve their indexing (Devine & Egger-Sider, 2014). For example, a search engine that focuses primarily on social media content will effectively relegate all other kinds of information to its own deep web. Similarly, the influx of publicly available data, which itself may be indexed, will be most effectively utilized by crawlers that index information at the high level of granularity necessary to that specific resource, a necessary trade of broadness for depth.

Using Deep Web Engines

Accessing the deep web requires the use of specialty search engines that have been designed specifically for navigating it. Alternately, many deep web resources are human aggregated collections with a search interface tailored to its contents. It should be noted that oftentimes the exploration of the deep web is said to require specialty browsers such as Tor and ip2. For academic resources in the deep web these are generally not required as there are many deep web engines (DWE) that can be accessed through standard web browsers.

Given the protean nature of the deep web and its resources, access points can be somewhat ephemeral. A DWE in existence today may be retired abruptly and with no clear replacement, while some like INFOMINE have been around consistently since their creation. Regardless, standard search engines can be useful for finding and identifying new access points and particularly for finding more subject specific engines. There is an element of 'word of mouth' to finding these access points, but once found they often point to additional resources. See the appendix for a list of some widely used DWE.

Just as with the subscription databases used for academic research, interfaces may change dramatically from one DWE to the next. Yet for those that retrieve articles and other related resources the search features are generally congruent and indeed have more in common than not with the features of academic databases. Generating queries also follows the same structures commonly taught in instructional sessions (e.g., boiling down a research question into a few keywords). However, DWE that search specific types of information, such as data, usually require a different structuring, relying heavily on keywords and faceted searching to filter down to the desired datum. Generally speaking, the learning curve for any widely available DWE is not steep, especially given that the tools needed to use them are so closely aligned with those used to navigate library resources.

Deep Web Instruction

As a pedagogical tool for library instruction, the deep web is an invaluable resource. The complexity of the web grows rapidly and consequentially a deeper understanding of its landscape is required for it to be thoroughly navigated. It is no longer sufficient for students (not limited to just those in academia!) to simply use common search tools without at least a general understanding of what is happening underneath the hood. Given the importance placed on instructing students in developing robust queries for database searching, seemingly little attention is given to why the queries have to be formed in a certain way at all. A basic comprehension of how retrieval systems work, be they databases or web scale discovery engines, is crucial for students to be able to critically evaluate their own search strategies. Instruction on the deep web provides an important opportunity to educate students about the web environment and how their choice of search engine predetermines what resources they will find.

Using deep web engines provides an important exercise, particularly for students, in reinforcing search skills that they have acquired during library instructional sessions. Since so many deep web resources reside in databases, the general query structures and syntax that they have already learned are transferable to these new engines. This serves the purpose of reinforcing concepts that will aid in their immediate research strategies goals and that will also serve them later in their academic career and beyond.

There are two primary ways in which the deep web can fit into the larger context of library instruction. The first model is as a workshop or one-shot session. As a part of its Savvy Researcher series, the University of Illinois at Urbana-Champaign library developed a fifty minute workshop on using the deep web. The workshop was open to all members of the institution and generally had attendance between 7-10 people, a strong showing in the context of the series and an indication of the appeal of the subject.

The workshop was split into two sections, essentially a division of theory and practice. The first half was an explanation of the deep web, peppered intermittently with questions to the students about their own search tactics. Focus was given to explaining how search engines like Google work before moving on to discuss the barriers to these engines and how they create the deep web. Particularly instructive was the opportunity for the class to deduce why crawlers can't access certain kinds of resources like databases and unpack why other types of resources are difficult for search engines to find. This and other discussions were well participated in and showed the genuine interest in the subject by the students.

Once the mechanisms behind the deep web were explained, the instructor then did brief demonstrations on a few deep web engines, illustrating their different features but noting how all relied on similar query syntaxes. Ten to fifteen minutes were then devoted to the students using the engines on their own for their research interests while the instructor circulated to answer questions. To conclude, the class discussed what kinds of strategies worked best with the engines. Overall, students were impressed with the new tools available to them and indicated they would use them in the future.

Not all circumstances will allow for the luxury of a workshop devoted to just the deep web. However, the topic is easily included in traditional library instructional sessions. Students are often told they should not use just Google and Wikipedia, usually with the pithy explanation that there are more resources beyond the scope of what those popular services deliver. This would be a prime opportunity to introduce the deep web. Without having to necessarily go into details about how search engines work, it may be enough to explain the deep web as a vast information resource that is generally unavailable to Google and much deeper than Wikipedia. This can be reinforced by rough estimates as to the size of the deep web versus the surface web. Many students may be unfamiliar with the concept of the deep web, but once informed as to its existence it becomes a monolithic entity that is not easily forgotten.

Of course, one of the most powerful deep web engines is the library. With its rich collections of books, journals, and institutionally created content, the library is perhaps the most effective and immediate tool available to students. Since they are already being instructed on how to use library databases, these skills are inherently transferable when they move to publicly available DWE as a student and after graduation.

The deep web is a complicated but powerful pedagogical tool. As the Internet continues to develop and grow in complexity, so too will grow the amount of information resources that are available but not immediately accessible. While search engines will inevitably grow better at mining the deep web, there will always be more to find. Instructing students on this unique phenomenon and how it impacts their own ability to effectively find information is one of the more exciting and valuable lessons libraries can provide, hopefully ensuring a well-equipped and information literate population into the future.

References

Devine, J., & Egger-Sider, F. (2014). *Going beyond Google again: Strategies for using and teaching the invisible web*. Chicago, IL: ALA Neal-Schuman.

Spencer, B. (2007). Harnessing the Deep Web: A practical plan for locating free specialty databases on the web. *Reference Services Review*, *35*(1), 71–83. doi:10.1108/00907320710729364

Yasser, C. M. (2011). An analysis of problems in metadata records. *Journal of Library Metadata*, *11*(2), 51–62. doi:10.1080/19386389.2011.570654

Zheng, Q., Wu, Z., Cheng, X., Jiang, L., & Liu, J. (2013). Learning to crawl deep web. *Information Systems*, *38*(6), 801–819. doi:10.1016/j.is.2013.02.001

APPENDIX A

List of Popular Deep Web Engines

- Infomine. A virtual library of Internet resources relevant to faculty, students, and research staff at the university level. It contains useful Internet resources such as databases, electronic journals, electronic books, bulletin boards, mailing lists, online library card catalogs, articles, directories of researchers, and many other types of information.
- The WWW Virtual Library. The oldest catalog of the Web, started by Tim Berners-Lee, the creator of HTML and of the Web itself, in 1991 at CERN in Geneva. Unlike commercial catalogues, it is run by a loose confederation of volunteers, who compile pages of key links for particular areas in which they are expert; even though it isn't the biggest index of the Web, the VL pages are widely recognized as being amongst the highest-quality guides to particular sections of the Web..
- Tech Xtra. Focuses on engineering, mathematics and computing resources. Provides industry news, job announcements, technical reports, technical data, full text eprints, teaching and learning resources along with articles and relevant website information.
- Enigma. A data engine that searches open data published by government agencies, companies, and other organizations. Allows for easy filtering by keyword.
- BASE. A powerful tool for searching academic resources and providing more than 50 million documents from more than 2,900 sources. You can access the full texts of about 75% of the indexed documents. The Index is continuously enhanced by integrating further OAI and local sources into its collection.
- The Global Science Gateway. Comprised of national and international scientific databases and portals. Allows for searching in multiple languages.
- Biznar. A federated search that retrieves general business information from a variety of sources.
- Mednar. From the creators of Biznar, another federated search focusing on medical and health information.