IR Systems

Adele Chase

School of Information Science, University of South Carolina

ISCI 706: Information Organization and Access

Dr. Ehsan Mohammadi

July 27, 2023

IR Systems

Learning effective methods of information retrieval (IR) is of utmost importance for librarians. IR refers to the process, methods, and procedures of searching, locating, and retrieving data and information from a file or database. In libraries and archives, modern IR is performed by locating items from search engines, full-text databases, bibliographic databases, and other information systems (Mohammadi, 2023b).

Google Books is a service that allows users to search, preview, buy, borrow, and read books, magazines, and newspapers. Google Books has scanned, converted to text, and stored millions of publications from libraries and publishers worldwide. The advanced search function allows users to refine their search by using various filters and criteria, including full-text keyword searching, phrase searching, and Boolean searching; searching all available documents; searching a subset of publications that offer some combination of limited previews, full views, or books that are available to purchase through Google eBooks; limiting the search to either books, magazines, or newspapers; and searching by the publication's language, title, author, publisher, subject, date, ISBN, or ISSN. In addition to the fields available via the advanced search form, Google Books also collects metadata about the publication's format (for example, eBook, paperback, hardback), original publication date, original publication language, page count, date of digitization, table of contents, series name and volume, edition, awards and nominations, genres, and characters (See Appendix 1). Not all metadata fields are available for every publication, and the metadata is sometimes machine-generated and may be incorrect (James & Weiss, 2012).

When performing IR, it is important for librarians to know search terminology to perform effective searches and identify which type of search would be most appropriate for the information system and search task at hand. Keyword searching, in which users enter words or

phrases they want to find information about, is a basic but important skill. It is a straightforward process of searching for information anywhere in a document or database without knowing in which specific field the information might be found. This type of searching can be employed in the early stages of IR, but it is imprecise and often returns either too many or not enough results (Hildreth, 1997). Field-based keyword searching is similar, but it allows users to specify in which field they want a keyword to be found. A field is a segment of metadata that IR systems keep for each item, such as the title, author, subject, date, etc. (APUS Librarians, 2023). Searching by field can increase the precision of search results (Hildreth, 1997).

Controlled vocabulary field searching is a type of search that uses a predefined list of terms or phrases to describe the content of an item. It is a type of field search that can help users find relevant and consistent information that matches their query, regardless of variations or synonyms in natural language (Mohammadi, 2023a).

An example of a controlled vocabulary is subject headings, such as the Library of Congress Subject Headings (LCSH) (Harpring, 2010). Subject searching is a controlled vocabulary search that searches only the subject field, as in the LCSH. As useful as controlled vocabularies are, they do have drawbacks. Controlled vocabularies are created by humans and, therefore, rife with biases (Mohammadi, n.d.).

Boolean searching uses logical operators such as AND, OR, and NOT to combine search terms. For example, one might want to search for information about pets, but not dogs or cats. Different search engines have different syntactical requirements for Boolean searches; using Google Books or the Google search engine, the syntax for this search would be "*pets" -dog -cat*. In contrast, the Open Library does not correctly parse multiple Boolean operators on the same field in full-text searches; the only way to perform a similar search would be to do a metadata search for *pets AND NOT cat AND NOT dog*; however, because it only searches the metadata,

results that contain the words *dog* or *cat* in the full text can be found in the results. Boolean search methods are simple but may not capture all the possible synonyms or variations of the search terms (Hildreth, 1997).

When search precision is important, utilizing authority records can help. An authority record is a type of controlled vocabulary that establishes the authoritative, preferred form of a term and notes variations (Harpring, 2010). For example, the Library of Congress (LOC) lists six distinct authority headings for the name "James Baldwin"; they are distinguished by the date(s) of birth (and death). The famous African American author by that name is listed under the authorized personal name heading, "Baldwin, James, 1924-1987"; the authority record for this author is available at https://lccn.loc.gov/n79076619.

I decided to search for *Treasure Island* by Robert Louis Stevenson, a classic, well-known work of fiction, in both the Library of Congress catalog (LCC) and Google Books (GB) to highlight the differences between the two IR systems. The first and most obvious difference is that the book is much easier to find in GB, where the book was the first result in the search for *treasure island*. In contrast, I had to do an advanced search by title and author, and still, an edition of the book was only the fourth result in LCC. Google's use of social indexing ensures that the most popular result, the one which has historically garnered the most clicks for the words *treasure island*, will appear high in the search results.

While both GB and LCC utilize free-text indexing – searching the full text of their databases – LCC's database is limited to metadata. In contrast, GB often indexes the full texts of publications, even if users cannot view any of the text or can only view limited previews or snippets. Both services also use controlled vocabulary indexing; for example, GB users can perform the search *inauthor:"Robert Louis Stevenson"* while LCC users can visit the authority record for *Stevenson, Robert Louis, 1850-1894.*

Despite controversies and disputes between libraries, publishers, and Google over GB, primarily due to disputes over copyright and fair use (Janes, 2006), I find it one of the most useful and easy-to-use information resources in existence. GB is not only a convenient and comprehensive source of information but also a challenging and rewarding one, as it requires me to apply my knowledge and skills of IR to get the best results. It is very easy to perform both basic and advanced searches on Google Books, but I like to go beyond the advanced search form and formulate my own advanced search queries by using search operators to create searches that are not available on the advanced search form. For example, putting two dots between numbers retrieves documents that contain numbers in a certain numerical range (Hardwick & Oh, 2023).

In all the IR systems I used, I would like to see further research and improvement, such as improvements in the quality and accuracy of the metadata, the availability of publications, and the accessibility and usability of the service.

References

APUS Librarians. (2023, June 5). *What is "field searching"?* Richard G. Trefry Library LibAnswers. https://apus.libanswers.com/faq/2420

Hardwick, J., & Oh, S. (2023, April 25). *Google search operators: The complete list (44 advanced operators)*. SEO Blog by Ahrefs. https://ahrefs.com/blog/google-advanced-search-operators/

Harpring, P. (2010). *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. Los Angeles, CA: Getty Publications*.

Hildreth, C. R. (1997). The use and understanding of keyword searching in a university online catalog. *Information technology and libraries*, *16*(2), 52.

James, R., & Weiss, A. (2012). An assessment of Google Books' metadata. *Journal of Library Metadata*, *12*(1), 15-22. https://doi.org/10.1080/19386389.2012.652566

Janes, J. (2006). Internet librarian; Google Book Search: evil or misunderstood? *American Libraries*, *37*(1), 74-75.

Mohammadi, E. (n.d.). Bias in information systems [Power Point slides]. University of South Carolina, Information Organization and Access. Blackboard: https://blackboard.sc.edu

Mohammadi, E. (2023a, June 18). ISCI 706-Unit Six: Unit Five: Document Representation-L1 [Video]. YouTube. https://youtu.be/ioPK5Q1YAtI

Mohammadi, E. (2023b, July 14). ISCI 706-Unit Seven: Information Retrieval L1 [Video]. YouTube. https://youtu.be/2bvuvnWxzr8