Information Services & Use 39 (2019) 205–213
DOI 10.3233/ISU-190045
IOS Press

What is Semantic Search? And why is it important?

Robert T. Kasenchak*
Director of Business Development, Access Innovations, Inc., 6301 Indian School Road, Suite 400, Albuquerque, NM, USA

Abstract. "Semantic Search" is a term used to describe a variety of approaches[1] to search using techniques beyond (or in addition to) traditional text- and keyword-matching functionality for information retrieval. This short article undertakes to explicate some of the sundry strategies encompassed under this rubric and, further, to briefly explain their potential advantages. Of particular interest is the application of these strategies in scholarly publishing, as various aspects of Semantic Search are suited to tackle the specific retrieval problems presented by a large and complex corpus. In particular, examples from scholarly publishing platforms (as well as search engines and other examples) will be used to illustrate the problems - and solutions - particular to the learned publishing industry.

Keywords: Semantic search, natural language processing, knowledge graphs, taxonomy, text analytics

1. Introduction

The term "Semantic Search" is used to describe a variety of strategies and techniques designed to enhance the functionality of the traditional search strategies based on matching an input string to strings in the text or keywords with which the text is tagged. This paper is an attempt to provide a survey of the techniques currently being described as Semantic Search and to situate them in the context of the state of search in the scholarly publishing industry.

Although Semantic Search encompasses a number of forms, they all have in common an attempt to move beyond simple text matching to take into account the context of the search to provide more accurate information retrieval. This is exemplified by the Google slogan "Things, Not Strings"[2] which, although specifically describing the (launch of the) Google Knowledge Graph, is a suitable tagline for Semantic Search in general.

To frame this discussion, it will be useful to first examine what problems Semantic Search purports to solve: *why search fails*.

*E-mail: taxobob@gmail.com.

[1]Reading material is provided at the end of this article as accessible primers for the reader; in addition, more scholarly research has been emerging since at least 2002; see https://scholar.google.com/scholar?hl=en&as_sdt=0%2C32&q=%22semantic+search%22&btnG= for examples (some may be behind paywalls), accessed July 1, 2019.

[2]Singhal, A., Google Blog "Introducing the Knowledge Graph: things, not strings" published May 16, 2012. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/ Accessed May 28, 2019.

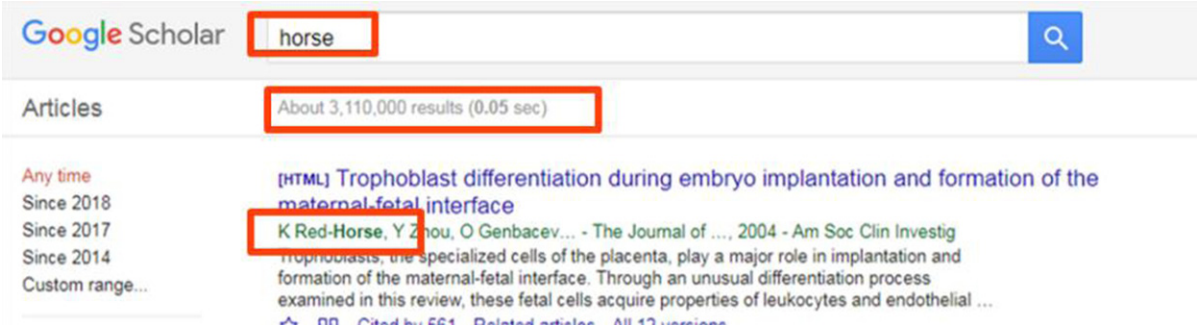

Fig. 1. Screenshot from Google Scholar, retrieved February 7, 2019 from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C32&q=horse&btnG=. Screenshot by author. Google and the Google logo are registered trademarks of Google LLC, used with permission per https://www.google.com/permissions/products/.

2. Why [traditional] search fails (in scholarly publishing)

Basic, or traditional text-string-matching, search is still widely found across scholarly publishing platforms. By "basic" or "traditional" search I mean that the user enters a keyword (or phrase) in a search box and the search application looks for that word (or phrase) in the document set. Simply put: *basic search matches text strings*. Some applications/platforms include fuzzy matching (to catch misspellings and other simple variations), but essentially the search is "dumb" – it looks for exactly the input text string without any underlying semantic logic.

For publishers of large corpora of specialized content, basic search is insufficient for researchers and other users to draw good results for information retrieval, because:

- Text-string matching does not include lexical variants, natural language processing, or other conceptual matches;
- Language is ambiguous; and
- Specialized repositories with large volumes of content covering fields with specialized vocabularies change over time.

In essence, basic search merely looks for the word(s) in the query; this is not an optimal strategy for delivering good results. Using Google Scholar (which reportedly contains over one hundred million articles), a search for the word "horse" provides two noteworthy outcomes. First, there are over three million results for this search; providing good relevancy ranking (which results appear at the top of the search) is therefore paramount. Second, and related, the very first result delivered draws the search result from the author name "Red-Horse" which, again, is not an optimal result (see Fig. 1 below).

Figure 2 shows a similar search, this time for the string "horses" in plural instead of singular. Note that there are almost half as many results (1.7 million) as the search in Fig. 1; the site search does not even recognize the equivalence of simple English plurals - only the literal text string in the search box.

Figure 3 is from the specialized repository[3] of a mid-sized scholarly publisher to illustrate a common problem in scholarly publishing: the prevalence of acronyms (and other abbreviations and commonly

[3]The author recognizes that this organization has, since the capture of these screenshots, embarked on a project to improve its website search.

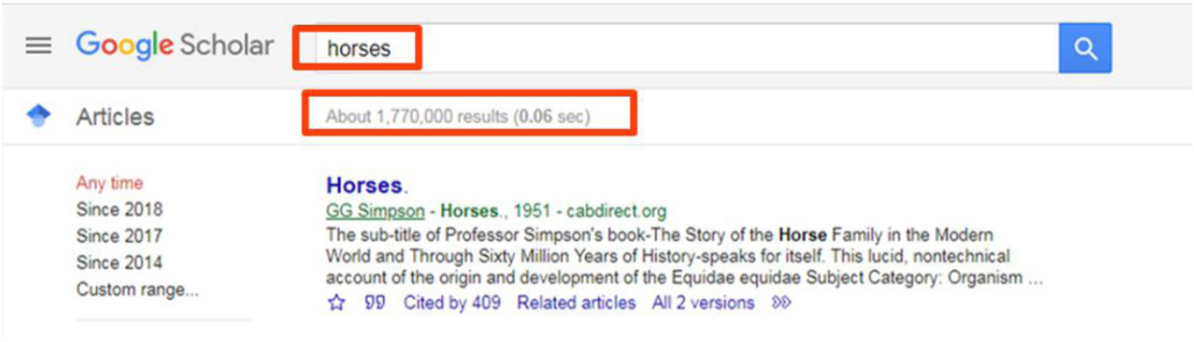

Fig. 2. Screenshot from Google Scholar, retrieved February 7, 2019 from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C32&q=horses&btnG=. Screenshot by author. Google and the Google logo are registered trademarks of Google LLC, used with permission per https://www.google.com/permissions/products/.

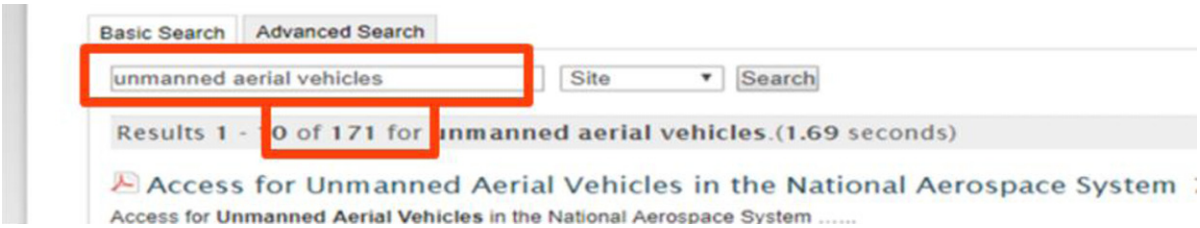

Fig. 3. Screenshot from AIAA, retrieved February 7, 2019 from https://www.aiaa.org/search-results?Keywords=unmanned%20aerial%20vehicles, Screenshot by author.

Fig. 4. Screenshot from AIAA, retrieved February 7, 2019 from https://www.aiaa.org/search-results?Keywords=UAV, Screenshot by author.

understood parlance) in every field in academia. The search shown below for "unmanned aerial vehicles" provides one hundred and seventy one search results, while a search for the ubiquitous acronym for the same concept ("UAV")… yields five hundred and ninety-two results (see Fig. 4 below). This indicates that the acronym is used some three times as often as the spelled-out concept in the content and, further, that the results for these two searches should be combined and delivered to the user as one. That is to say: the way the concept is expressed in the search has a direct bearing on the results returned.

This is precisely the problem that Semantic Search is trying to solve: providing search results around concepts instead of text strings.

3. What is Semantic Search?

Semantic Search describes a diverse set of techniques that seek to, using a variety of methodologies, go beyond simple string matching to try to examine the context of the query to drive (and improve) search relevance. These include accounting for lexical variants, incorporating taxonomy (hierarchy and synonymy), contextualizing searches through location parsing and previous searches, utilizing knowledge graphs, and similar technologies.

3.1. Lexical variants and fuzzy matching

One strategy designed to help users find relevant content without knowing the specific language or terminology used in a content set is to expose the search string to fuzzy matching and/or lexical variants via some natural language processing (NLP) operations. These techniques allow the search application to deliver near-matches rather than requiring exact string matches - so that if the user misspells "gastrointestinal stromal neoplasms" they will still divine relevant results.

The difference between lexical variants and fuzzy matching lies between structured variants - known versions of a word describing the adjectival, adverbial, and other forms (e.g., psychology, psychologically, psychological) as opposed to naïve near-matching of text strings (in which "Bob" and "Rob" are very near related words using Levenshtein distances[4]).

The drawbacks to these techniques are (1) they do not take care of the acronym problem, described above (since "UAV" and "unmanned aerial vehicles" are not very closely-related text strings), and (2) they can cause noise - false matches. However, both are easy to implement and are effective.

3.2. Query parsing

Although not new, query parsing is a kind of Semantic Search that attempts to identify the contextual meaning of a natural language search. Instead of searching for an entire text string verbatim, the parser will identify and use certain words (e.g., "where" and "when") and eliminate other words (e.g., existential verbs, articles, etc.) to deliver good results. The examples I use are from Google, for which this behavior is familiar to most users.

A Google search for the string "Harrison Ford"[5] returns, on the left-hand side of the results, links to the actor's Wikipedia and IMDB pages and similar results; the right-hand side presents results from the Google Knowledge Graph (to view click on link in footnote 5; more on the Knowledge Graph later).[6]

If instead the query "when is Harrison Ford's birthday"[7] is used, the Google search logic does not search for documents containing this text string; rather the search immediately returns the desired result (the date of his birth). Again, to be clear: this result is *not* from searching the string in the search box; that is to say:

[4]See https://en.wikipedia.org/wiki/Levenshtein_distance for a brief description; accessed July 1, 2019.

[5]As retrieved February 7, 2019 from https://www.google.com/search?q=harrison+ford&oq=harrison+ford&aqs=chrome..69i57j0l5.1607j0j9&sourceid=chrome&ie=UTF-8, accessed July 18, 2019.

[6]The author regrets that some of the illustrative screenshots intended for this paper were not suitable for rendering in the journal publishing format.

[7]As retrieved February 7, 2019 from https://www.google.com/search?ei=rUkJXaTNHImh8AP3tIrQAg&q=when+is+harrison+ford%27s+birthday&oq=when+is+harrison+ford%27s+b&gs_l=psy-ab.3.0.0i70i251j0i22i30l2j0i22i10i30j0i22i30.1761.2936..3671...1.0..0.122.361.0j3.....0....1..gws-wiz.4jFS0OpeleY, accessed July 19, 2019.

the results are not for documents containing that text string, but rather the search found the answer to the question by understanding the "when is" not as part of a text-matching query, but as a natural-language question to be answered. In this way the natural language question is parsed into a machine-readable query that answers the question.

3.3. Contextual search

The term Contextual Search comprises a set of methodologies designed to use information gathered about the user's location, recent searches, or other information to deliver relevant results.

3.3.1. Contextual search: Location

Some applications - notably Google, but also other map-based applications - use (if available) the location of the user (supplied by your IP address, geolocation of a mobile phone, or explicitly entered) to drive relevant results. To use a common example: a Google search for the string "pizza"[8] does not deliver websites and information about pizza; instead, the top results are all for nearby pizza restaurants, as if I had searched Google Maps for "pizza near me" or some similar query.[9] The contextual results are, in this case, driven by the IP-based geolocation of the device used in the search.

3.4. Google Knowledge Graph

Mentioned previously, the Google Knowledge Graph is a massive knowledge store designed to deliver results about the search *topic* over and above traditional search results in the form of *web pages matching the input query*. Launched in 2012[10], by 2016 results from the Google Knowledge graph reportedly appeared in about one-third of Google search results.[11]

For example, the search results for "empire state building"[12] again display the webpage-based results on the left and the Google Knowledge Graph results on the right; these results, typical of Google Knowledge Graph results,[13] include:

- A link to the corresponding Wikipedia page;
- Photographs, and links to more photographs;
- Physical address and contact information;

[8] As retrieved February 7, 2019 from https://www.google.com/search?q=pizza&oq=pizza&aqs=chrome..69i57j0l2j35i39j0l2.751j0j9&sourceid=chrome&ie=UTF-8, accessed July 19, 2019.

[9] It is interesting to note, however, that the right-hand sidebar result (from the Google Knowledge Graph) does provide a generic definition, a link to the Wikipedia page, and (interestingly) information about pizza company stocks in which I might like to invest.

[10] Singhal, A., Google Blog "Introducing the Knowledge Graph: things, not strings" published May 16, 2012. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/ Accessed May 28, 2019.

[11] Dewey, Caitlin. "You probably haven't even noticed Google's sketchy quest to control the world's knowledge". Published May 11, 2016, *The Washington Post*. https://www.washingtonpost.com/news/the-intersect/wp/2016/05/11/you-probably-havent-even-noticed-googles-sketchy-quest-to-control-the-worlds-knowledge/?noredirect=on&utm_term=.af4dcd824e46. Accessed May 28, 2019.

[12] As retrieved February 7, 2019 from https://www.google.com/search?q=empire+state+building&oq=empire+s&aqs=chrome.1.69i57j35i39j0l4.1920j0j7&sourceid=chrome&ie=UTF-8.

[13] For entities, at least.

- The official website;
- Reviews;
- Links to corresponding social media;
- Common Questions and Answers;
- Directions and map links;
- Hours of operation and popular/busy times for potential visitors to consider;
- Links to ticketing/ticket purchase information; and
- Links to other common searches by users ("People also searched for…").

The Graph uses standard schema.org[14] structures, which makes the data available to be consumed by other systems, and Google Knowledge Graphs also features a host of open API endpoints for use in other applications to return results from the Graph for searches for entities.[15] This, in theory, allows the entities in any corpus of content to be enriched using information from the Google Knowledge Graph.

The Graph does suffer from common-named entity disambiguation problems[16], particularly for very common names; Dave Thomas, the founder of Wendy's, did not in fact write a book on welding (which is by R. David Thomas, but listed as "Dave Thomas" as the book's author); their profiles and information are conflated on the Google Knowledge Graph page for "Dave Thomas".[17]

It is difficult to overstate the influence of the Google Knowledge Graph over the past seven years; one sign - to information professionals - that knowledge graph-type technologies are resonating beyond the information community and into the business world at-large is the recent appearance of articles in non-technical business publications such as *Forbes*.[18]

3.5. Leveraging semantic metadata: Taxonomies and tagging

If the goal of Semantic Search is to access the concepts "behind" text strings in documents, it follows in the footsteps of document tagging (or indexing) based on controlled vocabularies: taxonomy-driven subject metadata. Although this is not new technology, it merits inclusion in any discussion of Semantic Search.

Leveraging semantic tagging for search and retrieval can include tuning search to prioritize subject tags before free-text, allowing users to browse a controlled vocabulary (using various means in a graphical interface), using controlled vocabularies to power type-ahead (predictive typing) and "did you mean"-type suggestions, and using synonymy to drive relevant results for a variety of inputs.

Indeed, simple application of well-formed subject metadata could solve many of the text-string-matching search problems illustrated above: lack of synonymy, inability to equate acronyms and abbreviations to concepts within a domain, and various other ambiguity-of-language issues.

Accordingly, many scholarly publishers (and, of course, other organizations) have long employed and continue to embrace controlled subject metadata programs, often including novel applications in search interfaces.

---

[14] https://schema.org/, accessed July 1, 2019.

[15] See https://developers.google.com/knowledge-graph/ for details, accessed July 1, 2019.

[16] Although, notably, efforts seem to be underway to improve this problem.

[17] As retrieved February 7, 2019 from https://www.google.com/search?biw=1920&bih=937&ei=VPsQXaqZHYXJtQbtlI2IAQ&q=dave+thomas&oq=dave+thomas.

[18] Most recently: https://www.forbes.com/sites/cognitiveworld/2019/01/18/the-semantic-zoo-smart-data-hubs-knowledge-bases-and-data-catalogs/#75cea60c669c, accessed July 1, 2019.

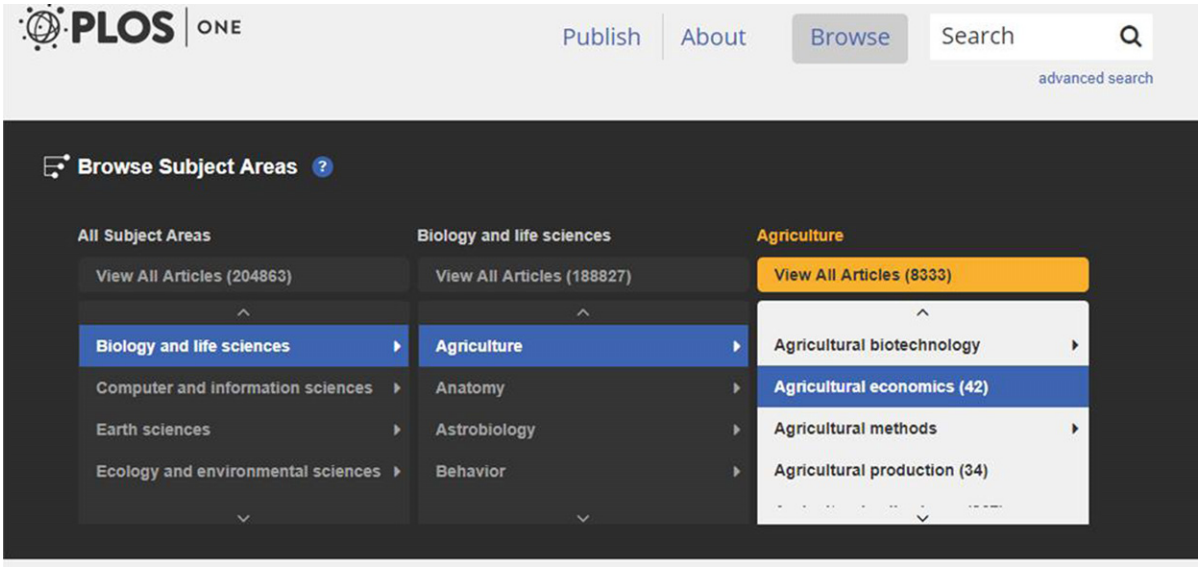

Fig. 5. Screenshot from PLOS, retrieved February 7, 2019 from https://journals.plos.org/plosone/. Screenshot by author.

The screenshot shown in Fig. 5 above is taken from the PLOS subject (taxonomy) browse interface on the main PLOS One website. In addition to a clever accordion display of the hierarchy, the interface also shows the user how many articles are to be found in the repository in each category. Clicking on a term retrieves all of the relevant articles; interestingly, the article-level display page exposes the taxonomy terms applied to that article to the user (in the bottom-right corner).[19]

In addition to displaying the tags to the user, this interface feature has two additional functions: first, clicking a tag launches a new search for articles on that topic; second, the small round buttons to the right of each topic allow the user to flag misapplied or inaccurate terms (which subsequently alerts the taxonomy team at PLOS), effectively leveraging crowdsourcing to improve the quality control of the semantics.

3.6. Novel approaches: JSTOR Text Analyzer

JSTOR has recently introduced[20] a new method for search based on real-time document indexing: the JSTOR Text Analyzer.[21] (see Fig. 6 below).

This application allows the user to provide a document (via upload, copy-and-paste, or a photograph of a document and subsequent OCR) as a search. The document is analyzed using both the JSTOR Thesaurus and naïve topic analysis, and the search engine provides results comprising documents from the JSTOR corpus with similar topics. The interface allows the user to curate the results to adjust the topics so discovered.

[19] As retrieved February 7, 2019 from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0218370. (However, any PLOS One article will display the same interface.)

[20] https://guides.jstor.org/howto-search/text-analyzer, accessed July 1, 2019.

[21] https://www.jstor.org/analyze, accessed July 1, 2019.

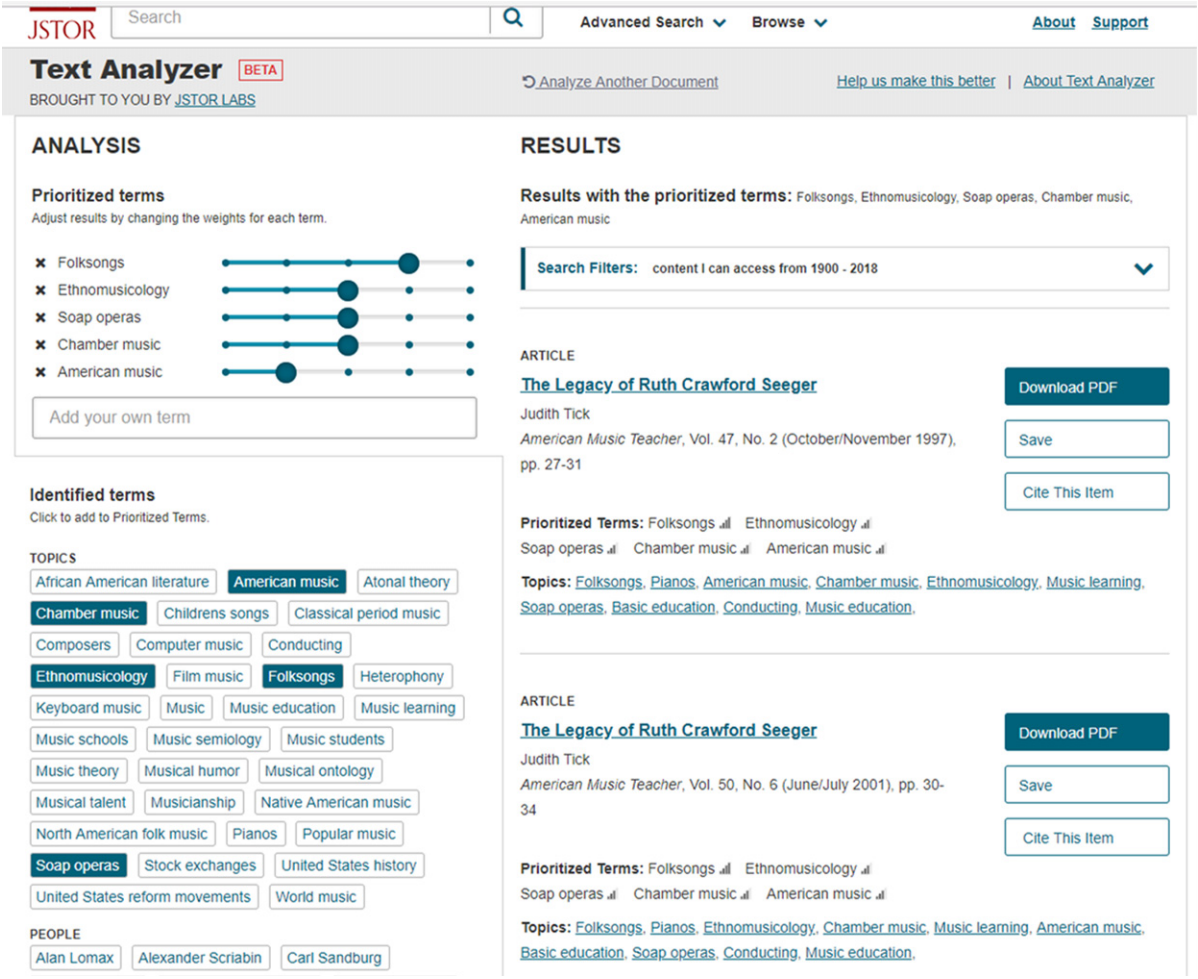

Fig. 6. Screenshot from JSTOR Text Analyzer, retrieved February 7, 2019 from https://www.jstor.org/analyze/. Screenshot by author.

4. Conclusion

As the size of document repositories continues to rapidly increase, scholarly publishers (and other organizations) are adopting a number of strategies - some from major search companies such as Google - to deliver relevant results and excellent search experiences to their users. Some of these approaches are novel, while others are based on long-standing principles of the library and information science domain. As the problem of delivering relevant search results continues to become harder - and as technologies such as voice-controlled devices require single-result answers - methods of going beyond simple text-matching for search are now emerging.

About the Author

Robert Kasenchak is the Director of Product Development at Access Innovations, Inc. and designs taxonomy and metadata projects for scholarly publishers to inform content discovery, analytics, and other initiatives with an eye toward linked data and the semantic web. E-mail: taxobob@gamail.com

Reading Material

[1] A Beginner's Guide to Semantic Search. *DeepCrawl,* https://www.deepcrawl.com/blog/best-practice/a-beginners-guide-to-semantic-search/, Accessed 1 June 2019.
[2] A. Barysevich, Semantic Search: What it is & why it matters for SEO today, *Search Engine Journal*, September 6, 2018, https://www.searchenginejournal.com/semantic-search-seo/264037/#close, Accessed 1 June 2019.
[3] A. Berasategi, Semantic search, *Towards Data Science,* https://towardsdatascience.com/semantic-search-73fa1177548f, Accessed 1 June 2019.
[4] Everything you need to know about semantic search and what it means for your website, *Daily Egg*, 14 July 2017, https://www.crazyegg.com/blog/everything-about-semantic-search/, Accessed 1 June 2019.
[5] A. Sanders, What is Semantic Search and what should you do about it?, *Moz*, 14 November 2016, https://moz.com/blog/what-is-semantic-search. Accessed 1 June 2019.
[6] Semantic Search: What is it and how it impacts your SEO, *Alexa Blog,* https://blog.alexa.com/semantic-search/. Accessed 1 June 2019.
[7] What is Semantic Search?, *BigCommerce.com,* https://www.bigcommerce.com/ecommerce-answers/what-semantic-search/, Accessed 1 June 2019.