Savvy searching
Citedness scores for filtering information and ranking search results

Péter Jacsó

Abstract

There are many options available to savvy searchers to help focus topical searches. These include limiting the search to one or more specific fields, such as title, descriptor or abstract in full-text databases to eliminate items where the search term(s) occur only in the full-text, often just mentioned in passing. In a few databases (such as ERIC, MEDLINE, PsycINFO), it has been possible to limit the search to the major descriptor field, which offers a good chance to retrieve only those records where the topic identified by the descriptor(s) in the query is the focus of the primary document. In newspaper databases restricting the search to the lead paragraph has been a powerful option for filtering the search. As discussed in a previous Savvy Searching column (“Citation-enhanced indexing/abstracting databases”, Online Information Review, Vol. 28 No. 3), in several databases there are additional alternatives available for restricting the search by document type (such as cover story), journal type (such as scholarly journal), the treatment of the topic by the source document (practical) or the length of document. These data elements added by indexers started out as descriptive record enhancers and later became search criteria. We see a similar trend emerging through the inclusion of cited references in records. The difference is that the smartest or most user-centric online services add extra value to this by calculating and including in the records the citedness score of the documents. These in turn can help to filter and/or rank large result sets.

Citedness score

The most innovative feature for filtering search results is the use of citedness scores of primary documents. These numbers show how many times the items which appear in the result list were cited by documents within the database or database group. The scores are dynamically calculated each time the database is updated. The citedness score is not yet a direct search option in any online services (although it could and should be made one). It has become, however, a sort criterion (as in Elsevier’s upcoming Scopus, and in the ISI Web of Science systems), or at least a prominently displayed value-added information element in the short result lists (as in many databases hosted by the CSA system). The citedness score can be a remarkably useful piece of data for showing users which are the most cited (and presumably the most respected) documents on the search topic.

Of course, calculating the citedness score assumes that the indexing/abstracting databases (such as the Science Citation database or
PsycINFO), or the full-text databases (such as the full text archives of Elsevier or the American Psychological Association) have the cited references in unambiguously identified, clearly tagged and structured format.

The tacit assumption for this filter is that the more often a document is cited, the more validated and useful it is. The absolute citedness scores, of course, may be misleading, as an article published four years ago has had a better chance of being cited than one published a year ago. Calculating relative citedness scores (adjusted to the age of the documents) can provide a level playing field, although this requires some extra steps from users, because currently none of the online services provide these scores.

A search scenario

It is easy to envisage that psychologists, psychiatrists and other health care practitioners and theoreticians will have to brush up their knowledge of the diagnosis and treatment of military personnel (and their family members) suffering from war-related post-traumatic stress disorder in the coming years. A quick and dirty search in the discipline-specific databases such as MEDLINE, Embase, PsycINFO, CINAHL brings up several hundred records on this topic. Science Citation Index, Social Science Citation Index, the mega databases of Ebsco, ProQuest, Gale, H.W. Wilson and Elsevier, would find thousands of matching records for the subject. Choosing a few of the best is a daunting task, especially when the results are presented in reverse chronological order, the most common default sort value. Some systems offer relevance ranking in presenting the results, but many of the ranking algorithms are of questionable logic and are not transparent for the end-users.

A few online services now create on the fly the citedness score of the items in the result list and offer a filtering/ranking tool which is transparent and easy to grasp by the end-users. Using the citedness score for filtering and/or ranking the results would certainly please them as well as most information specialists who deal with scholarly publications and do not settle for a Google search. Many may even realize that Google’s ranking algorithm has borrowed a lot from the results of citation analysis research of scholarly literature. I discuss three approaches to illustrate how the citedness scores are presented. Ebsco has an approach which is similar to that of CSA in the PsycINFO database, and it started to extend this feature also to some of its own mega-databases, such as Academic Search Elite.

The Scopus approach

Scopus is the new online service of Elsevier to be launched in autumn 2004 with some innovative features which can make your work much more efficient. (I used a preliminary beta version in early July 2004 for this article.)

From the perspective of this column, the most valuable feature of Scopus is the use of the citedness score as shown earlier in Figure 1. It appears in the rightmost column of the novel and almost perfect matrix display of the short result list. It displays the publication year, the title of the document, its author(s), the journal name and the citation score (labelled “cited by”) of each items retrieved by the query. I have been begging online information services for such a layout for a long time. Dialog has the REPORT command to create such a layout, but it works in only in a very few directory databases.

The matrix format facilitates the quick scanning of the result list, and the columnar layout makes the bibliographic elements stand out clearly. It could be better only if an additional column labelled “Cited by/year” were added, and if the volume and issue numbers were omitted. I would prefer to allow the users to decide whether to include in the matrix the subtitles, all the authors, etc.

The citedness score is not only prominent in Scopus, but very importantly, by clicking on the header label the result set can be sorted by it. Surprisingly, Scopus seems to have no limit on the size of the result set, and it sorts blazingly fast. I could make it sort a result list of almost 50,000 records in less than 30 seconds. Clicking on the citedness score itself will bring up the records of the citing documents (which is true in all the three online servies). These are also displayed in a matrix format and can be sorted by their citedness score (Figure 2).

The CSA approach

CSA deserves credit for pioneering the display of citedness score in the short result list more than a year ago. It does not stand out as prominently as in Scopus, but it is good enough to give a hint for the user about the potential importance of the items by virtue of their citedness. Regrettfully, sorting is not possible by citedness score (Figure 3).

CSA calculates and includes citedness scores also for cited references in records enhanced by the content provider with their bibliographic data. Scopus has an almost identical feature, but the cited references are the last data elements, and
Figure 1: Search results presented in a matrix form and sorted by their citedness.

| Figure 2: Citing references sorted by their citedness score in Scopus |

<table>
<thead>
<tr>
<th>Results: 49</th>
<th>Search within results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Document (Sort by relevance)</td>
</tr>
</tbody>
</table>
there is no customization of the display of record content, while in CSA it is possible, and the citedness score of the cited articles are more prominent than in Scopus.

If a cited reference has no citedness score, it does not mean that it was not cited by journals covered in PsycINFO. The American Psychological Association claims to have comprehensive coverage of cited references from 2001, but in my experience there are many records even from that year onward which have not been enhanced by references. PsycINFO has inexplicable gaps in indexing and abstracting for some of its own journals, so there are no records to be enhanced with cited references. This may distort some citedness scores (Figure 4).

The Institute for Scientific Information (ISI) has far the most comprehensive coverage of cited references both in terms of retrospective coverage and depth of indexing. WoS does not display the citedness score in the short result list. It shows it only when the user clicks on a record to display its full content. Then it appears prominently (Figure 5).

To its credit, WoS offers a sort option to sort the short result list by the citedness score, although it still would display the score only when clicking on a record for details. The sort feature works accurately but is available only if there are fewer than 300 items in the result set (which is reasonable for the typical search but not for bibliometric searches) (Figure 6).

To be fair, in its very nicely redesigned and richly enhanced new interface launched in July, 2004, ISI offers a highly customizable set of output options, including the saving of any user-selected data elements of the result set in a format which
can be uploaded directly in Excel. You are limited to 500 records here, but the result set can be segmented using data ranges, for example. With a simple macro a spreadsheet can be generated instantly. It is important because the citedness score is one of the elements which can be chosen (other systems do not offer this when saving records), so with a few extra steps the result list can be displayed in decreasing order of the absolute or the calculated relative citedness score.

You can see from the excerpt of the spreadsheet below that the items with the highest citedness (TC) and citedness/year (TC/Yr) score would show up as the 476th, 393rd, 260th, 474th and 258th records on the result list by the default reverse chronological sort order (SQ) (Figure 7).

**Caveats**

You must take citedness scores with a grain of salt for several reasons. One is that not all the authors cite purely for scholarly reasons. Not all the citations are accurate, and thus not all of them can be awarded to the paper which would deserve it.

Books receive far more citations than journal articles. Books about methodological issues are far the most cited ones. Many cited references deal with only a particular aspect of your search topic, or have an overly general approach to the topic.

Only a fraction of the records may be enhanced with references and only for a limited time frame, say for the past 4-5 years. Knowing that most papers get the majority of their citations 2-3 years later...

Full Record

**Title:** POSTTRAUMATIC-STRESS-DISORDER AND WAR-ZONE EXPOSURE AS CORRELATES OF PERCEIVED HEALTH IN FEMALE VIETNAM-WAR VETERANS

**Author(s):** WOLFE J, SCHMURR PP, BROWN PJ, FUREY J

**Source:** JOURNAL OF CONSULTING AND CLINICAL PSYCHOLOGY 62 (6): 1235-1240 DEC 1994

**Document Type:** Note

**Language:** English

**Abstract:** Previous studies have identified traumatic exposure and posttraumatic stress disorder (PTSD) as predictors of physical health complaints without considering the relationship between exposure and PTSD. This study examined the unique associations of war-zone exposure and PTSD with perceived physical health outcomes in a non-treatment-seeking sample of 109 female veterans of the Vietnam War who responded to a series of psychological, exposure, and health...
after publication, this limited retrospectivity obviously penalizes the older articles. While citedness scores may not be effective weapons in fighting infoglut, they certainly are a step in the right direction. It is very likely that other publishers and aggregators will also hop on the bandwagon, facilitating the selection of the few most promising items from a result list of hundreds or thousands of items.