ADVANCED SEARCH IN THOMSON REUTERS' WEB OF SCIENCE

RONALD ROUSSEAU
University of Antwerp IOIW-IBW
KU Leuven, Dept. of Mathematics
ronald.rousseau@uantwerpen.be
ronald.rousseau@kuleuven.be

Abstract: A short overview of search tools in Advanced Search is provided, with special attention to lemmatization.

INTRODUCTION

Documentary searches in Thomson Reuters' Web of Science are best performed using the Advanced Search functionality. Within this frame one can adapt the language of publication and the type of document (filtering). A set of tags are available as prefixes (see Table 1). As these are not the topic of this contribution I refer the interested reader to Thomson Reuters' documentation.

Moreover, the following Boolean operators can be used: AND, OR, and NOT. The AND-operator is used when it is required that both (or more) terms be present in retrieved records; the OR-operator is used when at least one term from the OR-search string should be present; the NOT-operator is used when it is required that all records containing the term after NOT are excluded from the search.

SAME and NEAR/x are proximity operators. SAME is used within the address field, requiring that search terms refer to the same address. NEAR/x is used to find records where the terms joined by the operator are within a specified number of words (indicated by the x in /x) of each other. This is true even when the words are across different fields. The symbol x specifies the maximum number of words that separate the terms. Without this specification the system finds records where the terms joined by NEAR are within 15 words of each other.

TS = Topic	TI = Title	AU = Author
AI = Author Identifiers	GP = Group Author	ED = Editor
SO = Publication Name	DO = DOI	PY = Year Published
CF = Conference	AD = Address	OG = Organization—Enhanced
OO = Organization	SG = Suborganization	SA = Street Address
CI = City	PS = Province/State	CU = Country
ZP = Zip/Postal Code	FO = Funding Agency	FG = Grant Number
FT = Funding Text	SU = Research Area	WC = Web of Science Category
IS = ISSN/ISBN	UT = Accession Number	

Table 1. Tags available when using the Advanced Search functionality

MORE FUNCTIONALITY

Besides the search options clearly shown on the Advanced Search webpage, more options are available by truncation and the use of wildcards. This gives more control in the retrieval of plurals and variant spellings.

- `term*` results in the retrieval of records in which this term occurs followed by zero to many characters;
- `term?` results in the retrieval of records in which this term is followed by exactly one character;
- `term??` results in the retrieval of records in which this term is followed by exactly two characters;
- `term$` results in the retrieval of records in which this term is followed by zero or one character(s);
- The symbols `?` and `$` can also be used as wildcards inside a search term: using `t??th` in a search results in records including the terms teeth or tooth, but also truth (for instance).

SPELLING VARIANTS AND LEMMATIZATION

Now we come to the main purpose of this note. In the Quick Reference Guide for the Web of Science (webofscience_qrc_en.pfd) we find the following:

> *"British/English spellings are searched automatically."*

Under the heading lemmatization it is explained that

> *"Lemmatization automatically helps find variations by stemming for plurals (even complex plurals like tooth/teeth) and searching different verb tenses (run/running) and degrees of comparison (big finds bigger and biggest). Lemmatization can be turned off by enclosing terms in quotation marks."*

Let us try this by performing a search; we note that the numbers of retrieved records depend on the version of the WoS which is locally available and on the date of the search (here May 22, 2014). They are just given to illustrate the retrieval process and refer to the WoS as available in Flanders at that moment.

First, we do a check to see if indeed British and American spelling are searched for automatically. As an example we consider:

`TS=harbor` yields 80,894 results (Set #1).

The search `TS=harbour` also yields 80,894 results (Set #2)

Clearly the WoS automatically unifies English and American spelling. Of course, one example does not prove anything, but at least there is no reason to doubt that retrieval works as announced by Thomson Reuters.

Next we check the behaviour of lemmatization.

`TS="harbor"` yields 23,960 results (Set #3) and

`TS=("harbor" OR "harbour")` yields 33,665 results (Set #4).

Trying to find the content of Set #1 we continue:

`TS=(“harbor” OR “harbour” OR “harbors” OR “harbours”)` yields 43,297 results (Set #5) and `TS=(“harbor” OR “harbour” OR “harbors” OR “harbours” OR “harboring” OR “harbouring”)` yields 69,541 results (Set #6)

Finally: `TS=(“harbor” OR “harbour” OR “harbors” OR “harbours” OR “harboring” OR “harbouring” OR “harbored” OR “harboured”)` yields the required 80,894 items (Set #7).

As a check we see that #1 NOT #7 yields an empty set. So, lemmatization works.

Using `$` does not turn lemmatization for these terms off. Hence more records are found.

Concretely, `TS=(harbor$ OR harbour$)` yields 80,923 records (Set #8), namely all those in Set #1 plus other ones containing e.g. the term Harbord.

Yet, using `?` does turn lemmatization off. `TS=(harbor? OR harbour?)` yields 10,692 records (Set #9), while `TS=(“harbors” OR “harbours”)` yields 10,663 records (Set #10). Again records containing Harbord are in Set #9 and not in Set #10.

Moreover #9 NOT #10 yields the same result as #8 NOT #1.

Finally using * does turn off lemmatization. `TS=harbor*` (set #11) yields 59,605 but not including those with the term harbour (unless also harbor is included, which happens when, to an article written in British English KeyWords Plus added keywords in American English).

Finally, we note that:

`TS=(harbor* OR harbour*)` yields 81,284 records (Set #12) including all those in Set #1 (#1 NOT #12 is empty) but also including records with the terms Harbourage or Harborth.

Similarly, a search `TS=organize` (Set #13) yields the same result as:

`TS=(“organize” OR “organise” OR “organizes” OR “organises” OR “organized” OR “organised” OR “organizing” OR “organising”)`,

while `TS=organise$` yields a larger set, including all records of the previous set, but also including records that contain, e.g. the term organizer (and none of the terms of Set #13).

We further note that `TS=formula` yields the same result as

`TS=(“formula” OR “formulas” OR “formulae”)`.

The same kind of lemmatization occurs for the prefix `TI=`. For an author search it is somewhat more complicated: the retrieval result for a search `AU=Zhang` is the same as that for `AU=”Zhang”`, but `AU= Zhang J` yields the same result as `AU=Zhang J*`. If one wants to search for Zhang J and not for Zhang JW (for example), then one needs quotation marks: `AU=”Zhang J”`.

However, `TS=shakespeare` yields the same set as `TS=”shakespeare”`, while `TS=shakespeare$` yields more than 400 extra records, some containing the plural shakespeares.

CONCLUSION

Lemmatization is probably useful in a large percentage of searches. Yet, occasionally one might want to know the exact content of one's search. Hence I suggest informetricians always to use quotation marks “.” which leads exactly to the set one wants to retrieve.

In cases where completeness is essential one may combine (use OR) a lemmatized search with one consisting of search terms between quotation marks. By doing so one does not forget terms that should have been included, e.g. the American or the English spelling variant.

More information about lemmatization can be found e.g. at: http://images.webofknowledge.com/WOKRS53B4/help/WOS/hs_current_limits.html