U.S. patent citation data documentation

Bhaven N. Sampat
Assistant Professor
Mailman School of Public Health
Columbia University
bns3@columbia.edu

August 5, 2011

1 Summary

The file `USCITES` includes U.S. patent citations in utility patents issued from 1/1/1975 to 12/31/2010. Each observation is a citing-cited pair. The dataset is based on information from a custom extract DVD generated by the Electronic Information Products Division of the USPTO on 4/29/2011.

Non-utility patents were eliminated from the *cited* patent list. That is, citations to design, plant, and reissue patents, defensive publications, and statutory invention registrations were eliminated. Please contact me if you require this information.[1] In addition, cited U.S. patent *applications* were removed from the file, though are also available on request. These patents were removed since citations to sources other than U.S. utility patents are reported in haphazard formats, and difficult to merge with other sources. As a practical matter, most analyses use utility patent - utility patent citations, so these dropped observations are benign. The field `cited` thus includes all U.S. utility patents cited by the patent indicated in the `patent` field.

However, citing patents include all types of patents, so one could calculate "forward" citation counts even for non-utility patents, e.g. for re-issues in

[1] See http://www.uspto.gov/web/offices/ac/ido/oeip/taf/patdesc.htm for a description of patent types.

the BASICBIB file.

2 Codebook

```
. codebook
------------------------------------------------------------------------------
patent                                                          (unlabeled)
------------------------------------------------------------------------------

                  type:  string (str7)
         unique values:  4158354                  missing "":  0/45856453
              examples:  "5265725"
                         "6171296"
                         "6889631"
                         "7460738"
------------------------------------------------------------------------------
cited                                                           (unlabeled)
------------------------------------------------------------------------------

                  type:  numeric (float)
                 range:  [0,9989019]                  units:  1
         unique values:  4985123                  missing .:  0/45856453
                  mean:   4.9e+06
              std. dev:   1.3e+06
           percentiles:        10%       25%       50%       75%       90%
                            3.2e+06   4.1e+06   5.0e+06   5.8e+06   6.4e+06
```

3 Notes

Both patent and cited can be merged to the basic bibliographic files, inventor files, and assignee files. Note that these files do not have information on *cited* patents issued before 1975. The file genyear.do can be used to create issue years for cited patent numbers for these earlier patents. Class information for older patents will also be posted soon.

Counts from this file were merged into the basic bibliographic file to create variables n_fcites and n_bcites, counts of "forward" and "backward" citations in that dataset. Note that these constructed variables do not exclude "self" citations, which is useful to do in some analyses. Users should use this citation file, and the assignee information in the bibliographic file, to create citation counts net of self-citations. Please contact me for code to do this.

Information on the source of citations (examiner versus other) will be posted soon, for patents issued between 2001 and 2010.