Bibexcel – Quick Start Guide to Bibliometrics and Citation Analysis

Alan Pilkington
a.pilkington@rhul.ac.uk

Introduction and Installation

I have been using bibexcel for a few years and keep recommending it to people. However, I keep getting questions from people on how to do get started, and also have to reread my own hand written notes every time I start some more analysis. So I thought it was time to write something more structured on how to get going with bibliometrics using bibexcel. As such, I hope you find these notes useful. If you have any suggestions or notice any mistakes please let me know.

Bibexcel is a great tool for helping with bibliometric analysis, and citation studies in particular. You can get the latest version from the web at: www.umu.se/inforsk. Installation is easy, just copy the files to a directory on your hard drive and be sure to put the help files in the same directory. Read the web page for more help if this doesn't work as there are some exceptions.

Using Bibexcel for Citation Analysis

The first step is to get some source data to analyse. In cittion analysis, this invariably means finding a selection of source articles from the Social Science Citation Index/Science Citation Index. These are commercial databases and are part of the Web of Science or ISI data services to which your university probably subscribes.

Using Social Science/Science Citations Index

Identify your source articles using the WoS/ISI search functions as you would in the normal way. It is important to understand what your study is about before you rush into downloading data. I've undertaken a few studies based on the contents of one journal and so my source is easy to identify. More elaborate projects may at the citations of one author or university department. Whatever your plans, if you get the data from the Science/Social Science Citations Index, then you need to follow the same steps in downloading and preparing the data.

In WoS, you have to make a marked list before downloading. Then you can proceed to download the selected papers, being sure to select that you want the citations as well. You can either do this as a "download for future analysis" or by emailing them to yourself. Both produce a plain text file.

The download process may result in the data service timing out if you ask for too many citations. You can check this by opening the file you got (either in bibexcel using the box at the top left to find and view the file which will then appear in the

bottom left or using a text editor) and looking at the last few lines. If they contain an error formatted in HTML then it has timed out, if it just looks like the end of one of the records then you have everything. If it times out the solution is to redo the download but by reducing the amount to download, possibly by changing the number of years in the original search. If you have to download in pieces, you have to remember to put the separate files back together again before continuing - just open them in a text editor and cut and paste, but be sure that there is only one header:

```
FN ISI Export Format
VR 1.0
```

at the top of your file, and not at the start of each section you downloaded.

The data comes as plain text and so it is easy to look at the files using any text editor, but be wary of using a wordprocessor such as MS Word as they tend to add characters, reformat lines and other things which can cause problems later.

One aspect to watch is that unix and windows end of line/line feeds are different and bibexcel works with the windows style. If you open the source file in bibexcel (using the top left area) and view it (in the bottom right) and see that it contains only one line of text instead of neat columns, then you need to convert the line feeds to windows. I use "editpad lite" for this which is a free download from the internet from JGSoft - look for it under google. It has a menu option to change the line endings – but I dare say there are many other ways of doing it.

To go further with the preparation and analysis, your raw data file should look something like this when viewed in bibexcel or a text editor:

```
FN ISI Export Format
VR 1.0
PT Journal
AU Brown, S
   Blackmon, K
TI Aligning manufacturing strategy and business-level competitive
   strategy in new competitive environments: The case for
   strategic resonance
SO JOURNAL OF MANAGEMENT STUDIES
NR 190
CR 1998, IND WEEK       1207, V247, P22
   YOUNDT MA, 1996, ACAD MANAGE J, V39, P836
   ZAJAC EJ, 2000, STRATEGIC MANAGE J, V21, P429
   ZAJAC EJ, 1989, STRATEGIC MANAGE J, V10, P413
BP 793
EP 815
PG 23
JI J. Manage. Stud.
PY 2005
PD JUN
VL 42
IS 4
GA 929TJ
J9 J MANAGE STUD-OXFORD
UT ISI:000229369000004
ER

PT Journal
AU Brown, S
   Cousins, PD
TI Supply and operations: Parallel paths and integrated strategies
```

```
SO BRITISH JOURNAL OF MANAGEMENT
NR 105
CR ANDERSON JC, 1991, INT J OPER PROD MAN, V11, P86
   BADRI MA, 2000, OMEGA, V2, P155
   BEACH R, 2000, INT J OPER PROD MAN, V20, P7
   WOMACK J, 1996, LEAN THINKING
   WOMACK J, 1990, MACHINE CHANGED WORL
   ZAIRI M, 1992, INT J OPER PROD MAN, V12, P34
BP 303
EP 320
PG 18
JI BRIT. J. MANAGE.
PY 2004
PD DEC
VL 15
IS 4
GA 874LZ
J9 BRIT J MANAGE
UT ISI:000225353200002
ER

PT Journal
AU Laycock, M
TI Transforming Rover, renewal against the odds 1981-1994 -
   Pilkington,A
SO LONG RANGE PLANNING
NR 1
CR PILKINGTON A, 1994, T ROVER RENEWAL ODDS
BP 738
EP 739
PG 2
JI Long Range Plan.
PY 1996
PD OCT
VL 29
IS 5
GA VW288
J9 LONG RANGE PLANN
UT ISI:A1996VW28800021
ER
```

Now you are ready to start in bibexcel...

Starting the Analysis

Bibexcel is very powerful because of its flexibility and so as a result it can be a little confusing at the start as there are ways of ding multiple things in one step or combining several different data sets together to process one file. There is help available, press F1 when bibexcel is active for the help system, but it is probably more for the advanced user who knows what they want to do and need some pointers as to how to do it in bibexcel. Hopefully these notes fill the gap of a tutorial/quick start guide.

The first thing is to work out what data you wish to analyse. Your downloaded text file from the steps above will have a field identifier of CR (or CD) to denote the citations (you needed to specify downloading of citations when you captured the data – do you have some entries starting CR (or CD) when you view the data file?). As this is the area which bibliometrics is most interested in, much analysis uses this data, but you can also use the software to study the other interesting data fields...

Converting to Dialog Format

To get the data in a format you can apply to bibexcel, you need to follow the a few steps to prepare your data. There is more on this in the help files for bibexcel - press F1 when bibexcel is active for the help system. Look at the index and entries for: downloading, convert to dialog, and preparing the data, first for the best introduction before looking at the analysis steps.

I'll summarise the steps to get the data ready here:

First check that your file has windows style end of lines (see above).

To convert, view the data file you got from SSCI by using the top left boxes to navigate and the view will appear in the window labelled "The List" on the right. In bibexcel, you normally select the file to be worked on using the top left boxes and select a menu item to carry out the task, or click one of the start/prep buttons.

Doing the "Misc/ Converttodialog/ convertfromWebofScience" menu item whilst your data file is selected is the first step to get it into the right format for bibexcel.

If you haven't already done so, this import is achieved by selecting your raw data in the top left (use the view file button to check). Then run the menu command: Misc/ Converttodialog/ convertfromWebofScience. This should give you a .doc file (the same file name as your original, just with a .doc file ending) which you can select and then view before pulling out the fields you want to use for further analysis.

View the .doc file and notice how you get nice tags (PT-, AU-, SO-, CD-, PY- etc.) at the start of each line showing what the information in the record is about, and neat end of line "|" and end of record flags "ER ||" as shown. Notice also how bibexcel has put semi-colons between the entries in the fields which can have multiple entries such as authors and citations. This helps when it comes to splitting them out later.

PT- Journal|
AU- Brown S; Blackmon K|
TI- Aligning manufacturing strategy and business-level competitive strategy in new competitive environments: The case for strategic resonance|
SO- JOURNAL OF MANAGEMENT STUDIES|
NR- 190|
CD- 1998, IND WEEK 1207, P22, V247; 1998, IND WEEK 1207, P24, V247; ADLER PS, 1990, P55, CALIFORNIA MANAG SPR; ANDERSON J, 1991, V1, P86, INT J PRODUCTION OPE; ZAJAC EJ, 2000, V21, P429, STRATEGIC MANAGE J; ZAJAC EJ, 1989, V10, P413, STRATEGIC MANAGE J|
BP- 793|
EP- 815|
PG- 23|
JI- J. Manage. Stud|
PY- 2005|
PD- JUN|
VL- 42|
IS- 4|
GA- 929TJ|
J9- J MANAGE STUD-OXFORD|
JN- JOURNAL OF MANAGEMENT STUDIES, 2005, V42, N4, P793-815|
UT- ISI:000229369000004 ER ||

Extracting Simple Fields

When you view the .doc file, notice that it has a field called TI- (for title, and you can see the names of the others, such as AU- for authors, PY- pub year, CR- or CD- for citations, etc.). Each of these and all the others can be pulled out for the file and analysed further.

As an example of an easy analysis, let's say you want to analyse the title words from your papers. These could be thought of as looking for the keywords which relate the different articles together. Which are the most popular words?

In this case you need to pull out the contents of the TI- field. Begin by selecting the .doc file (as before when viewing), and put the tag (TI) in the old tag box (bottom left) and select the right style for the data (blank separated field to treat each word alone) from the drop down box next to PREP (top middle). Then just press the PREP button to perform the operation and this should give you a new file called .out.

This .out file is the one to use as you go further with the analysis...look at the help pages by pressing F1 to get more information on the things you can do to manipulate this data.

View the .out file using the boxes in the top left, and note how it has kept the words you wanted, but with a link to which source article they came from (the numbers in the first column). This is what makes the programme so powerful, as you can look at the links between the different source articles easily. Here is an example of a title .out file:

1	Aligning
1	manufacturing
1	strategy
1	business-level
1	competitive
1	strategy
1	case
1	strategic
1	resonance
2	Supply
2	operations
2	Parallel
2	paths
2	integrated
2	strategies
3	conceptual
3	synergy
3	model
3	strategy
3	formlation
3	manufacturing
4	Technology
4	portfolio
4	alignment
4	commercialisation
4	investigation
4	fuel
4	cell
4	patenting

There is nothing to stop you making your own .out type file from some other source such as a database or excel and using bibexcel to perform the next steps of analysis. Just be sure it has the same format and it plain text.

Basic analysis

Frequencies of the items in the .out file (or if it has been updated it will be called .oux) are generated by selecting and viewing the file (top left of screen), then in the middle left window use: "whole string, sort descending, start" to give a .cit (citation) file of frequencies.

Now you can view the .cit file and see which was the most popular word in the titles of the source articles. The file I am using now shows manufacturing appeared 9 times followed by strategy and strategic:

9	manufacturing
8	strategy
6	Strategic
4	management
3	operations
3	competitive
3	investigation
2	learning
2	Literature
2	relationships
2	links

Citation Analysis

One of the most popular methods in bibliometrics is citation analysis, and bibexcel makes the steps to getting the data ready and performing the analysis relatively easy. The biggest problem is often to extract from the raw data just the parts of the citation information you want.

The first step is to pull out all the citation information from the .doc file, so we repeat the steps above but using the CD tag in the old tag box and select "any ; separated field." This will give a .out file listing each citation with its source article number:

1	ADLER PS, 1990, P55, CALIFORNIA MANAG SPR
1	ANDERSON J, 1991, V1, P86, INT J PRODUCTION OPE
1	ANDREWS KR, 1971, CONCEPT CORPORATE ST
1	ANSOFF HI, 1965, CORPORATE STRATEGY A
1	PILKINGTON A, 1998, V41, P31, CALIF MANAGE REV
1	ZAJAC EJ, 2000, V21, P429, STRATEGIC MANAGE J
2	BEACH R, 2000, V20, P7, INT J OPER PROD MAN
2	BESSANT J, 2003, V23, P167, INT J OPER PROD MAN
2	BRAGLIA M, 2000, V28, P195, OMEGA-INT J MANAGE S

We could work with the full citations as shown here, but it is often better to pull them apart for separate analysis of the authors and titles, or do some cleaning of the data (such as standardising on one initial) before using just the author, title and year for analysis.

To extract the authors from the .out file, view the .out file, and in the middle left panel select cited author, remove duplicates and make new out file. Pushing start gives a .oux file listing just the authors (or at least the entries which should be authors if the file is in the correct format).

1	ADLER PS
1	ANDERSON J
1	ANDREWS KR
1	ANSOFF HI
1	BAHRAMI H
1	BAIN JS
1	BARNEY J
1	BARNEY JB
1	BATES KA
1	BEACH R
1	BERRY WL
1	BESSANT J
1	BOEKER W
2	COUSINS PD
2	CROSBY P
2	DANGAYACH GS
2	DSOUZA DE
2	DURAY R
2	DYER JH
2	ELLRAM L
2	ELLRAM LM
2	FARMER D
2	FEITZINGER E
2	FLYNN BB
3	GRANT RM
3	HAKSEVER C
3	HAMMER M
3	HART SL
3	HAX AC
3	HAYES RH
3	HENDERSON JC
3	HEWLETT CA

You may want to use excel or something to strip away the second initial and so standardise this data before going further. I would do this using the excel “text to column” menu to separate the initials from the surnames, and then the function LEFT to pull off the first initial which can then be CONCATENTEd with the surname. The data can then be put back together in a text editor or excel into the same plain text format file as the .out/.oux file for bibexcel to work with.

You can do the frequencies procedure on this to give a .cit file as above, in this case to see the most cited authors:

27	PILKINGTON A
11	HAYES RH
11	SKINNER W
9	HILL T
7	PRAHALAD CK
6	LEONG GK
6	MINTZBERG H
6	PORTER ME
6	STALK G
6	SWINK M
6	VOSS CA
6	BARNEY J
6	WOMACK J
6	HAYES R

The same steps can be used to extract different elements from the .out file such as the publication titles, or even some combined elements as you require. Bibexcel uses the way that entries are formatted in the SSCI to identify which parts to extract. So if you ask for journal entries, you get only those with valid volume and page information. These tools need to be treated with care as the data in the SSCI is often not in the correct format.

Co-occurrences and Networks

After looking at frequencies for the different fields in the source articles or the citations, an interesting approach is to look at the relationships and networks/maps between citations or phrases. This is termed coocurrance in bibexcel and is covered in the help file page on make a matrix.

You can use any data you want to build the coocurrance data. Some meaningful ones are title words, authors, journal titles, or combined entries such as "author, journal, year" to identify individual publications. I often use bibexcel to analyse coocurrance for patent data which came from a separate database by making a .out like file by hand for to feed the analysis.

Essentially the steps in coocurrance involve making a .cit file of frequencies to help select the terms to analyse, then using this index to analyse the .oux/.out file and produce the coocurrance data in a .coc file. This can then be turned into a matrix much like an excel pivot table, with the cells containing the frequencies of the column and row headers.

It is normally best to take the extra step of removing multiple entries when doing this type of analysis as we are often only concerned with if the link exists rather than whether there are many citations to the same work in one file. This can be done on the .out or .oux file using the middle left boxes and remove duplicates flag to make a new file.

To make a coocurrance or .coc file, view your .cit file and select (make blue in the main window) the words/authors/titles/citation strings you want to analyse. Once you have the entries you want highlighted in the .cit file, then do: "Analyse: Coocurrance: slectunits via list box", to get just those terms in the "the list" window. Next select your .out file in the top left (do not view this file as you want to keep the selected words highlighted). Then do "Analyse: Coocurrance: make pairs via listbox", to give you a .coc or co-occurance file. View the file to see the results.

The .coc file will contain the frequency of occurrence and then the two terms matched. For example an author coocurrance file:

17	PILKINGTON_A	HAYES_R
16	VOSS_C	HAYES_R
15	HAYES_R	HILL_T
14	MEREDITH_J	HAYES_R
14	VOSS_C	MEREDITH_J
14	HAYES_R	SKINNER_W
13	VOSS_C	HILL_T
12	PILKINGTON_A	HILL_T

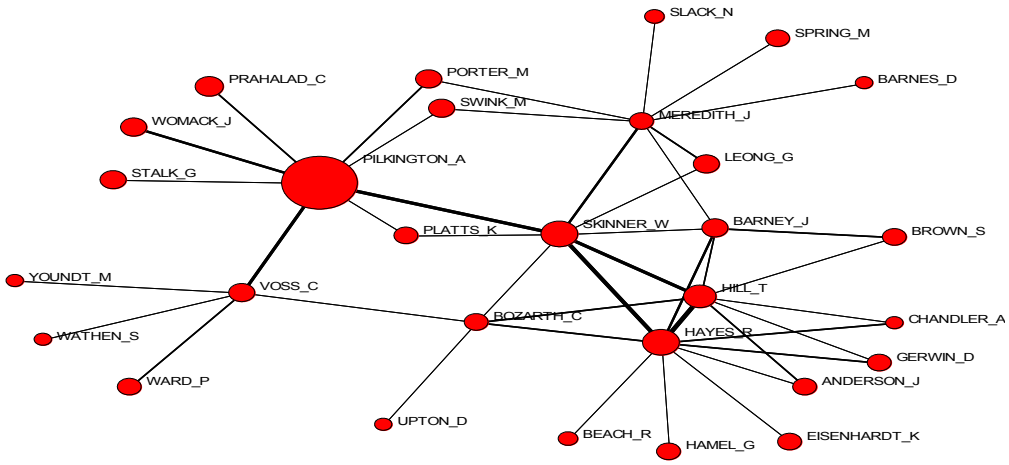
Or a title word coocurrance file:

```
6       manufacturing       strategy
4       Strategic   Management
3       strategy    competitive
2       strategy    new
2       mass        customisation
2       manufacturing       study
2       strategy    case
2       manufacturing       competitive
2       strategy    strategic
2       competitive         case
2       competitive         strategic
```

Personally, I often use this data to go further with some form of network analysis in a programme like UCINET. As the .coc file resembles a .DL format data file with labels, but with the frequencies in the left most instead of right most column it is relatively easy to move the data into UCINET. If this is what you want to do, then read the UCINET help files for more on how to get the data into the analysis software. The steps I use involve importing the .coc file to excel, cut and paste the left frequency column to move it to the right then cut and paste all three columns to the text editor and adding a header to the file which turns it into a DL format, such as:

```
dl n = 5000 format = edgelist1
labels embedded
data:
Hayes_R;RESTORING_OUR_COMPET;1984       Hill_T;MANUFACTURING_STRATE;1985        25
Hayes_R;RESTORING_OUR_COMPET;1984       Skinner_W;HARVARD_BUS_REV;1969  24
Hill_T;MANUFACTURING_STRATE;1985        Skinner_W;HARVARD_BUS_REV;1969  19
Eisenhardt_K;ACAD_MANAGE_REV;1989       Yin_R;CASE_STUDY_RES;1984       19
Barney_J;J_MANAGE;1991      Wernerfelt_B;STRATEGIC_MANAGE_J;1984    18
Barney_J;J_MANAGE;1991      Prahalad_C;HARVARD_BUS_REV;1990 15
Dixon_J;NEW_PERFORMANCE_CHAL;1990       Kaplan_R;HARVARD_BUS_REV;1992   14
Hayes_R;DYNAMIC_MANUFACTURNG;1988       Hayes_R;RESTORING_OUR_COMPET;1984       14
```

The result you can get from UCINET often provides a very clear view of what is happening in the data matrix, as shown, and also allows many more analysis tools to be used.



A Map of the Co-cited Authors in the Ego Net of A.Pilkington

Often you want a square matrix of the .coc file terms. To turn the list data of the .coc file into a matrix, select the same words from your .cit file as before by highlighting them using “analyse: coocurrance: select units via listbox”, and then select your .coc file and do: “analyse: make a matrix”. This gives a .ma2 matrix file of the results which can be used elsewhere as it is still a plain text file.

One catch to getting these matrix files into other programmes is that they only contain the labels at the top of the columns and not down the side. To solve this, you can import the file into excel, insert a new empty first column, then copy the top row and do an “edit: paste special: transpose” to add the labels to the start of the rows. This gives a fully labelled square matrix of coocurrances much like a pivot table which can then be import into SPSS for factor analysis to group the terms statistically.

Bibliometric Coupling

There is some debate as to the value of citation coocurrance or co-citation analysis in mapping the links between literature and some people recommend bibliometric coupling. Here instead of looking at the linkages between different cited works, the links between the source articles are exposed and analysed. Needles to say this can also be achieved using bibexcel using the shared units routine.

Further Possibilities

This is about as far as I need to go for my own work in bibexcel, but as you look at the menus and help files it becomes clear that it can do far more…

Alan Pilkington
9.1.06