Similarity Indexes Computational Program: SICoP

WELCOME TO THIS BRIEF TUTORIAL ON SICoP. IT WILL GUIDE YOU IN ORDER TO USE IT AS SIMPLY AS POSSIBLE.

THANK YOU FOR CITING IT WHEN YOU USE IT.

Introduction

The **S**imilarity **I**ndex **Co**mputational **P**rogram (SICoP) is a useful tool for calculating several similarity indexes; it is widely used by scientometricians. The set of indexes is: Jaccard index, Salton's Cosine index, Pearson's index, Dice-Sorenson index, Pudovkin & Garfield Index, Association Strength index, and Chi2 index.

Normalization of matrix of co-occurrences is a crucial step in science mapping that is an integrative process of retrieving scientific data, pre-processing, normalizing, visualizing and analysing them. Thus, the similarity index governs the normalized output matrix in order to provide rational and objective maps and temporal evolutions, which help understanding science trends, topical analysis and networks shaping.

Set of Similarity indexes in the SICoP

Let consider the matrix $[X] = X_{ij}$ where $1 \le i, j \le n$ representing the gross matrix data of co-occurrences between the entities i and j. That may be citations (citing-cited), co-publications, co-words, bibliographic-coupling, co-authors, author co-citations, journals, etc.

Jaccard index

Jaccard: was introduced by the Swiss Botanist Paul Jaccard (Jaccard, 1901) who originally called it "*Coefficient de communauté spécifique*":

$$J_{ij} = \frac{X_{ij}}{Y_{im} + Y_{mj} - X_{ij}} \quad \text{where} \quad Y_{im} = \sum_{j=1}^{n} X_{ij} \quad \text{and} \quad Y_{mj} = \sum_{i=1}^{n} X_{ij}$$

It exists a vector-variant formula of the this index and is always referred to as Jaccard-Tanimoto index (Tanimoto, 1957, Cha et al., 2009).

$$J_{ij} = \frac{\sum_{k=1}^{n} X_{ik} X_{kj}}{\sum_{k=1}^{n} X_{ik}^{2} + \sum_{k=1}^{n} X_{kj}^{2} - \sum_{k=1}^{n} X_{ik} X_{kj}} \qquad J_{ij} = \frac{\vec{X}_i \vec{X}_j}{\left\|\vec{X}_i\right\|^2 + \left\|\vec{X}_j\right\|^2 - \vec{X}_i \vec{X}_j}$$

Salton's Cosine

It was first introduced by Salton (Salton, 1983). It is a measure of similarity between two 'vectors' by finding the cosine of the angle between them. It is applied under two variants. The non-vector one:

$$Cos_{ij} = \frac{X_{ij}}{\sqrt{Y_{im}.Y_{mj}}}$$

and a vector one:

$$Cos_{ij} = \frac{\sum_{k=1}^{n} X_{ik} X_{kj}}{\sqrt{\sum_{k=1}^{n} X_{ik}^2} \sqrt{\sum_{k=1}^{n} X_{kj}^2}} \qquad Cos_{ij} = \frac{\vec{X}_i \vec{X}_j}{\left\|\vec{X}_i\right\| . \left\|\vec{X}_j\right\|}$$

Pearson

Called also Pearson's Correlation Coefficient, it is another measure of the extent to which two vectors are related. Its formula is:

$$r_{ij} = \frac{\sum_{k=1}^{n} (X_{ik} - \overline{X_{i.}}).(X_{kj} - \overline{X_{.j}})}{\sqrt{\sum_{k=1}^{n} (X_{ik} - \overline{X_{i.}})^2} . \sqrt{\sum_{k=1}^{n} (X_{kj} - \overline{X_{.j}})^2}} \quad \text{where } \overline{X_{i.}} = \frac{1}{n} \sum_{k=1}^{n} X_{ik} \quad \text{and}$$

$$\overline{X_{.j}} = \frac{1}{n} \sum_{k=1}^{n} X_{kj}$$

It is interesting to note that the Pearson's formula is nothing than the Cosine formula applied to the averaged vectors: $X_i^{'} = (X_i - \overline{X_{i.}})$ and $X_j^{'} = (X_j - \overline{X_{j.}})$ formed by subtracting the corresponding component in the original vector by the average value of a component weight in the original vector.

Dice-Sorenson

This index is very similar to Jaccard, and was first introduced by Dice (Dice, 1945). It is also referred to as Dice-Sorenson index attributed both to Dice and Thorvald Sorenson (Sorenson, 1948):

$$D_{ij} = \frac{2.X_{ij}}{Y_{im} + Y_{mj}}$$

As to Jaccard index, the vector-variant of the Dice-Sorenson could be written as:

$$D_{ij} = \frac{2\sum_{k=1}^{n} X_{ik} X_{kj}}{\sum_{k=1}^{n} X_{ik}^2 + \sum_{k=1}^{n} X_{kj}^2} \quad \text{or} \quad D_{ij} = \frac{2\vec{X}_i \vec{X}_j}{\left\|\vec{X}_i\right\|^2 + \left\|\vec{X}_j\right\|^2}$$

Association Strength

It is a measure for normalizing co-occurrences (frequency) proposed by Van Eck et al. (2006). This measure is also known as the Proximity Index (Rip and Courtial, 1984). The Association Strength between entities *i* and *j* is given by:

$$AS_{ij} = n\frac{X_{ij}}{Y_{im}Y_{mj}}$$

where X_{ij} the number of co-occurrences of entities *i* and *j*, while X_{im} and X_{mj} are respectively the total numbers of occurrences of entities ***i*** respectively ***j*** and ***n*** is the total number of entities.

Pudovkin & Garfield

The index was created by Pudovkin (1995) and has later been called Pudovkin & Garfield index (2002). It has been initially used as an inter-citation measure for journals allowing counting for varying sizes:

$$PGR_{ij} = \frac{10^6 X_{ij}}{N_j \sum_{k=1}^{n} X_{ik}}$$

The obtained matrix is asymmetric due to the nature of the index even though the initial matrix [X] is symmetric.

Pudovkin and Garfield Factor is then the Maximum of the two indexes. That is, $PG_{ij} = Max(PGR_{ij}, PGR_{ji})$. However, they also used the arithmetic average of the two indexes PGR_{ij} and PGR_{ji} as: $PG_{ij} = \frac{PGR_{ij} + PGR_{ji}}{2}$

Chi2

Chi2 is and index for validating the independence or dependence of variables according to different observed characteristics of these variables. The whole observation is then presented as a matrix (see: http://www.univ-st-etienne.fr/lbti/biomath/Cours/chi2/Chi2.htm).

$$Ch_{ij} = \frac{(\sum_{k=1}^{n} X_{ik})(\sum_{k=1}^{n} X_{kj})}{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}}$$

If the matrix $\lfloor X_{ij} \rfloor$ is symmetric then the Chi matrix is also symmetric.

SICoP Brief tutorial

Download the directory sicop from the following website :

http://www.fsr.ac.ma/index.php/recherche/recherche-a-la-fsr.html.

To start the SICoP, one needs to have a ready gross matrix data of co-occurrences extracted from databases or built in one of the txt or xls formats. The SICoP user can then choose between these two formats of the input file.

The steps in executing the SICoP are shown below for this example: we suggest to normalize the gross matrix of scientific collaboration in Table 1.

***Table1**: Scientific collaboration matrix*

	Brazil	Canada	China	France	Germany
Brazil	0	1899	1139	3282	2963
Canada	1899	0	7758	7292	7348
China	1139	7758	0	4993	7679
France	3282	7292	4993	0	16077
Germany	2963	7348	7679	16077	0

To run SICoP, doublclick on "SICoP.exe" , then the window below (figure 1) appears prompting the user to specify the matrix size **n**. In this case 5 (5 countries or entities).

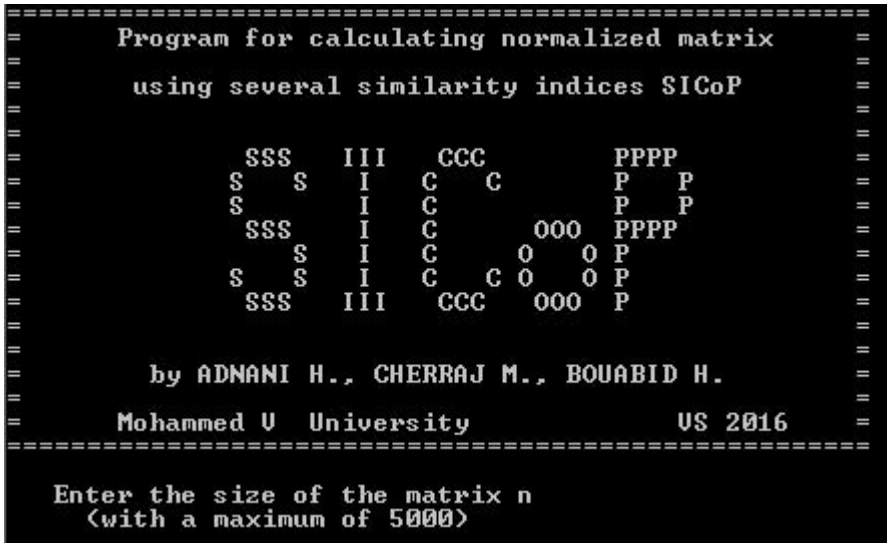

***Figure 1**: Size n of the matrix*

The size of the matrix can not exceed the maximum of 5000 columns on 5000 rows.

Afterwards, you need to choose one of the 7 available similarity indexes in the list to normalize the gross matrix as shown in Figure 2. For example, to choose Jaccard index, type

'1'. In this list 'Raw' means that the output matrix will be identical to gross matrix but produced in a txt format or .net format.

***Figure 2**: Choose a Similarity Index*

In the case of Garfield-Pudovkin index, the '*vector.txt*' should be provided. This file should contain the vector $N_j \quad 1 \leq j \leq n$, as defined above.

Once done, the user should specify the data file format, *data.txt* or *data.xls*. Both of the files *data.txt* and *data.xls* are represented below (Figure 3-a, 3-b) and are therfore accepted by SICoP .

```
Brazil
Canada
China
France
Germany
0    1899    1139    3282    2963
1899    0    7758    7292    7348
1139    7758    0    4993    7679
3282    7292    4993    0    16077
2963    7348    7679    16077    0
```

***Figure 3-a**: data.txt file*

	A	B	C	D	E	F
1	Brazil					
2	Canada					
3	China					
4	France					
5	Germany					
6	0	1899	1139	3282	2963	
7	1899	0	7758	7292	7348	
8	1139	7758	0	4993	7679	
9	3282	7292	4993	0	16077	
10	2963	7348	7679	16077	0	
11						

***Figure 3-b**: data.xls file*

Since vector and non-vector variants exists for similarity indexes (see "**Set of Similarity indexes in the SICoP**" above), the user has the choice to select one of the two variants.

When the calculation process is finished, the SICoP will ask the user for the format to generate the output file as shown in Figure 4. The user could choose between a *.txt* file format as for the input data (*data.txt*) or *.net* format that can be uploaded in most visualization softwares such as Gephi, VOSViewer, CortextManager, etc.

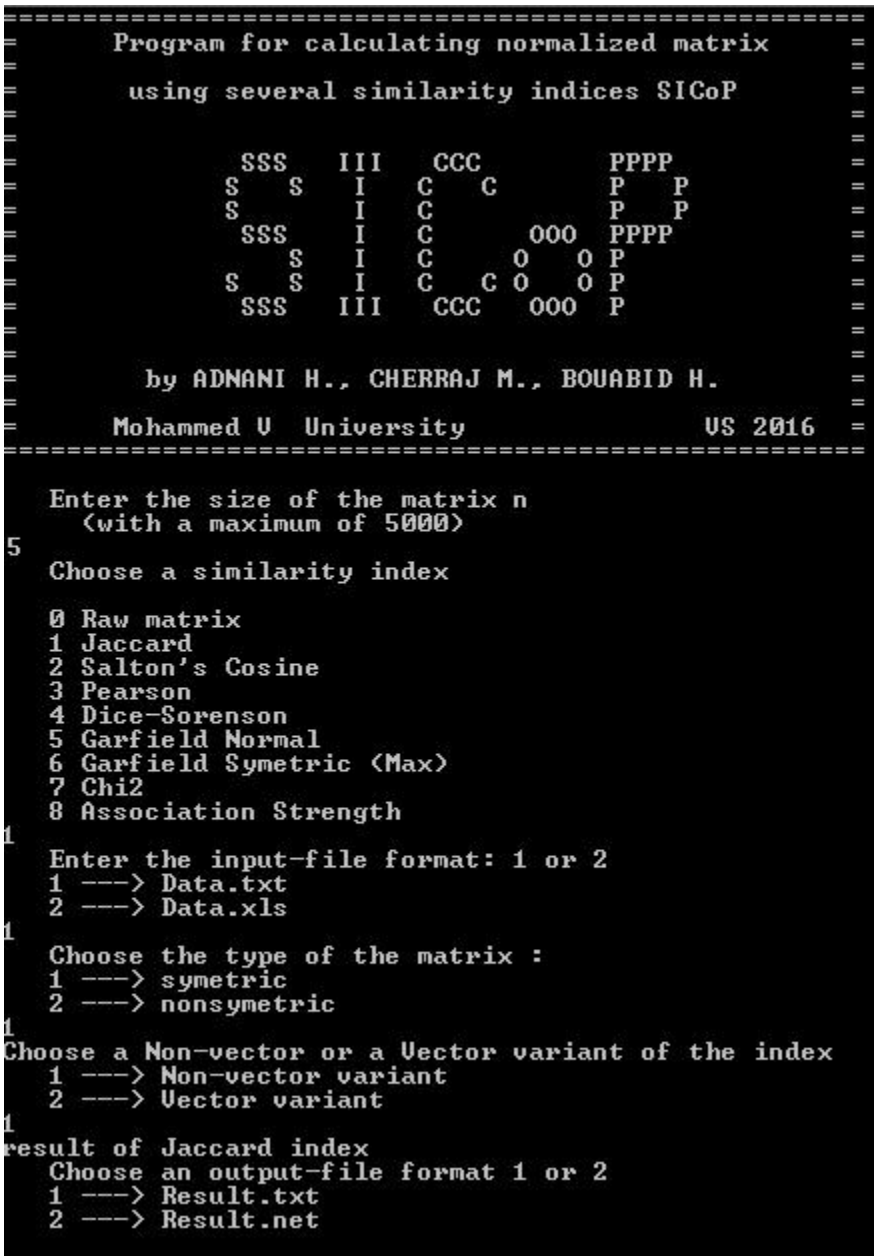

***Figure 4**: Specifying Input and Output file form*

For the example considered here, the *Result.net* from the the Jaccard index is shown in Figure 5.

```
*Vertices  5
           1    "Brazil          "
           2    "Canada          "
           3    "China           "
           4    "France          "
           5    "Germany         "

*Edges
  1     2        .0599
  1     3        .0383
  1     4        .0872
  1     5        .0734
  2     3        .2036
  2     4        .1499
  2     5        .1440
  3     4        .1035
  3     5        .1601
  4     5        .3239
```

***Figure 5**: Result.net file format*

In this brief tutorial you learn how to use the SICoP that allows you to generate a normalized matrix suitable for the science mapping from a gross matrix of co-occurrence. A set of seven similarity indexes is provided by a simple click to address your science mapping requirements.

SICoP by: Hinde ADNANI, Mohammed CHERRAJ, Hamid BOUABID

Faculty of Sciences, Mohammed V University, Rabat - Morocco.

Contact:
For any inquiry or feedback please contact **sicop@fsr.ac.ma**

Acknowledgement:
SICoP was developed with a partial support of the *Agence Universitaire de la Francophonie, Bureau Maghreb*.

References

Callon M., Courtial J.P., Laville F., (1991), Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry, Scientornetrics, 22 (1), 155-205.

Cha, S. H., Choi, S., & Tappert, C. C. (2009). Anomaly between jaccard and tanimoto coefficients. Proceedings of Student-Faculty Research Day, CSIS, Pace university.

Dice R. (1945). Measures of the amount of ecologic association between species in Ecology. 26 (3), 2976-302.

Jaccard P. (1901), Distribution de la flore alpine dans le bassin de Dranses et dans quelques régions voisines, Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 241-272.

Michelet B., L'analyse des mots associés, unpublished dissertation, 1988 (from He Q. (1999), knowledge discovery through co-word analysis, Library Trends, 48 (1), 133-159.

Pudovkin, A.I., Fuseler, E.A. (1995). Indices of journal citation relatedness and citation relationships among aquatic biology journals. Scientometrics, 32(3), 227–236

Pudovkin, A.I., Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. Journal of the American Society for Information Science and Technology, 53(13), 1113–1119.

Rip A., & Courtial J.-P., (1984), Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics, 6(6), 381-400.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. Auckland, New Zealand: McGraw-Hill.

Sorenson T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyze the vegetation on Danish commons. Biologiske krifter. 5 (4), 1-34.

Tanimoto, T.T. (1957) IBM Internal Report 17th Nov. 1957.

Van Eck N. J., Waltman L., Van den Berg J., Kaymak, U. (2006), Visualizing the WCCI 2006 Knowledge Domain. IEEE International Conference on Fuzzy Systems, 1671- 1678.