Controversies arising from which similarity measures can be used in co-citation analysis

Eustache Mêgnigbêto
Bureau d'Etudes et de Recherches en Science de l'information (BERSI)
09 BP 477 Saint Michel, Cotonou, REPUBLIC OF BENIN
e-mail : eustachem@gmail.com

ABSTRACT

Pearson's r has been used as similarity measure in author co-citation analysis since the introduction of this technique in the 1980s. However, some scientists supported that Pearson's r coefficient does not fulfill mathematical conditions of a good similarity measure, and therefore should not be used in co-citation analysis. They proposed alternatives measures that yield much precisions. Other scientists defend the use of Pearson's coefficient and supported that it does well the job author co-citation analysis is for. This article makes the point of the controversy that rose several years ago and fed in the Journal of the American Society of Information Science and Technology (JASIST) *about what similarity measure to use in author co-citation analysis and similar techniques.*

Keywords : Author co-citation analysis; Informetrics; Pearson's r coefficient; Similarity measures; Citation analysis

INTRODUCTION

Author Co-citation analysis (ACA) was introduced by White and Griffith (1981) and used to analyse the intellectual structure of a given scientific field. McCain (1990) standardized the procedure that has since been adopted as a standard worldwide (Ahlgren, Jarneving and Rousseau, 2003) and laid on the use of the Pearson's correlation coefficient (Pearson's r) as a similarity measure. However, Ahlgren, Jarneving and Rousseau (2003) criticized this choice and they formulated two natural requirements a similarity measure should satisfy, stated and demonstrated that Pearson's r fails in fulfilling that test, and then proposed some more relevant measures that lead to "objectively better results". These proposals have resulted in controversies and debates about the pros and cons of using Pearson's r in ACA. The debate lasted several years and, certainly, is not yet closed. The objective of this paper is to go back over the pros and cons arguments by giving the key points with defence for each point.

A GLANCE AT THE STATEMENTS FROM THE ORIGINAL PAPER

Ahlgren, Jarneving and Rousseau (2003) first recalled the four main steps identified by McCain while doing ACA:

a) compilation of raw data matrix,
b) matrix conversion to a proximity, association, or similarity matrix,
c) multivariate analysis of the relations between the authors represented in the matrix,
d) interpretation and validation of the results.

They recalled that a co-citation matrix data is symmetric, that the element on the intersection of row *k* and column *l* denotes the number of times authors A_k and A_l are co-cited; hence a problem occurs about what to put in the matrix diagonal cells. They noticed that scientists did not treat diagonal as missing values, but used to put citation frequencies or a combination of other co-citations frequencies. They concluded that "this method for creating diagonal values is at best inelegant and at worst completely arbitrary". Later, they proposed a better solution in the use of the number of times an author has been co-cited with him/herself, excluding self citation as diagonal values which is "the number of articles in the pool under study that cite at least two different works (co)authored by the author".

Secondly, Ahlgren, Jarneving and Rousseau, (2003) recalled that Pearson's r measures the strength and direction (decreasing or increasing, depending on the sign) of a linear relationship between two variables. Citing Dominich (2001), they stated that a similarity measure should be non negative, that is, not in the case of Pearson's r which may vary from -1 to 1; they noticed however that Pearson's r may be easily transformed into a positive value. They stated that similarity between two objects can be measured in two approaches: local approach e.g. direct similarity between the two objects, or global approach based on relatives values, e.g. the way two objects related to other objects in the population or data. They concluded that Pearson's r may be considered as an approach based on relatives values.

Ahlgren, Jarneving and Rousseau (2003) then formulated mathematically, before explaining in expressive words, two requirements called "test of stability of measurement" by White (2003), which a similarity measure should fulfil.

Requirement 1: "Assume that A and B belong to a group of authors for which an association measure has been calculated. Assume next that this group is expanded by a new set of Authors. If A as well as B are now never co-cited with this new group of authors, then we require that the association measure between A and B does not decrease."

Requirement 2: "Assume that A, B, C and D belong to a group of authors for which an association measure has been calculated. Assume next that this group is expanded by a new set of authors. It so happens that none of these four authors is co-cited with this new group of authors. Then we require that if, before the expansion, the association between A and B was smaller than that between C and D, it stays smaller after the expansion".

Ahlgren, Jarneving and Rousseau (2003) demonstrated that Pearson's r correlation coefficient does not fulfil any of these two requirements because it leads to "the absurd situation that two perfectly similar authors showing the same behaviour with respect to a group of new authors are not perfectly similar anymore" and that "vectors that are unrelated to two vectors under consideration can have an influence on the mutual association on them, and hence, on the resulting mapping". They identified and demonstrated that the Chi-square distance and Salton's Cosine measure fulfil the two requirements. Mathematically, Ahlgren, Jarneving and Rousseau (2003) found no difference between ACA, co-word analysis and other forms of social interaction research; therefore, these requirements also apply to those techniques. Based on statistical literature, they argued that cosine and Pearson's r should not be used in case of ordinal variables; they considered co-citation data as measured on an ordinal scale and concluded that "using Pearson's r is out of question as it is only meaningful for data measured on (at least) on interval scale". Citing Spiegel and Casterman (1988), they stated that Pearson's r requires bivariate normal distribution.

Up to this point, the arguments were theoretical (mathematical). Ahlgren, Jarneving and Rousseau (2003) supported their statements with real-life data, by building the author co-citation matrices for 12 information retrieval specialists and 12 scientometricians. They noticed that when the two matrices were merged, r values changed with more than 0.5.

THE CONTROVERSY: *WHY R?*

White (2003) was the first author who replied to Ahlgren, Jarneving and Rousseau's critique, on behalf of the Drexel ACA team. He recalled that the primary aim of ACA is to visualize the broad patterns of a field and not to interpret the underlying pair-wise coefficients. He emphasized that, the team, in a sense, were indifferent to the r's as long as one can make quick sense of the maps and defended that the recurrent intelligibility of result across a variety of studies is the main reason they have continued to use r over the years. White (2003) informed that McCain (1984) acknowledged doubts about r as a similarity measure and supported that others similarity measures yields the same analysis. He then recalled the context in which "Drexel style" authors have chosen Pearson's r against raw co-citation counts and justified the use of r: r was mentioned as one of the two basic proximity measures in multidimensional scaling on the first page of Davison's (1981) textbook on the subject, and, it is also used in an illustration of multidimensional scaling in older SPSS manuals. He then listed the advantages in using r: (a) it is widely taught and understood; (b) common statistic software can easily read and convert author co-citation matrices to it; (c) such kind of matrices could be used as input to principal components analysis, multidimensional scaling and hierarchical clustering routines; (d) in all 3 routines, r produces a highly intelligible results. White (2003) indicated that these advantages have not lessened to date, but technological advances have made other measures and routines easier. He also recognized that depending of the context, r may be replaced by another measure, but added: "the motivation should yields a clear gain in interpretation, especially in visualization that the two requirements couldn't justify".

Is r a similarity measure?

Ahlgren, Jarneving and Rousseau (2003) supported that a similarity measure should not be negative. Egghe (2009) defined mathematically "good similarity measures" and confirmed. He gave some examples of similarity measures such as Jaccard, Cosine, and Dice, but did not mention Pearson's. Van Eck and Waltman (2009) dealt with "well known similarity measures" but did not refer to Pearson's r; van Eck and Waltman (2008) asserted that Pearson's r is not a very satisfactory measure for co-citation profiles. Egghe (2010) formulated two mathematical conditions a similarity measure should satisfy; on the basis of Ahlgren, Jarneving and Rousseau's (2003) critique, he declared not to consider Pearson's r in his study.

Even though r may be transformed into positive value, Bensman (2004) found no logic in the requirement that a similarity measure should be positive. He claimed that a similarity measure is expected to not only measure similarity but also dissimilarity, the job Pearson's r does well, because it partitions sets starting from 0 (similarities rise from 0 to 1 and dissimilarities from 0 to -1). White (2003) claimed that the behaviour of r resulting from adding 0 to a co-citation matrix should not be presented as a drawback but as a virtue. Bensman (2004) affirmed that any measure filling the requirements of Ahlgren, Jarneving and Rousseau (2003) lacks the partitioning clarity of Pearson's r and indicated that by doing so, Ahlgren, Jarneving and Rousseau demonstrated unintentionally the robustness of the Pearson's r in partitioning sets, and therefore, as a similarity measure. Then, he

pointed out a problem occurring if, as required by Ahlgren, Jarneving and Rousseau, a similarity measure must be positive and asked: Where does partitioning begin? Ahlgren, Jarneving and Rousseau (2004b) discarded: if Pearson's r may be transformed into a positive value, with the formula (r+1)/2, the partition that starts at 0 before the transformation then starts at 0.5 after. On the contrary, Egghe and Leydesdorff (2009) seemed to rescue Bensman (2004) as they declared that the cut-off level is no longer given naturally in the case of cosine and that the choice of the thresholds remains arbitrary. They demonstrated that there is a relation between cosine and r and that "a cosine can never correspond with an $r < 0$".

Even though the difference between using Pearson's r correlation coefficient and alternatives measures (like Cosine) is neglected in practice (White 2003; White 2004; Leydesdorrf 2005; Egghe and Leydesdorff 2009), some scientists have preferred not to use Pearson's r for the analysis and visualisation of similarities due to the critique of Ahlgren, Jarneving and Rousseau (e. g. Egghe and Leydesdorff 2009).

Diagonal Treatment: a Point of Agreement

White (2003) agreed with Ahlgren, Jarneving and Rousseau (2003) that diagonal could be the most meaningful if it contains a count of the publications in which two or more of the author's works are jointly cited. He further noticed that such data are difficult to collect from the Institute of Scientific Information (ISI) databases and gave the example of Hopkins (1984), who treated diagonal as missing data but resulted in the same conclusion. He concluded that the strongest argument against treating the diagonal as missing data would be technical, because most statistical programmes do not run unless matrix diagonals are filled.

"A theoretical statement with no relation with practice"

Bensman (2004) reproached to Ahlgren, Jarneving and Rousseau's (2003) paper that their arguments were overly mathematical and could not be applicable into information science; White (2003) affirmed that the problem Ahlgren, Jarneving and Rousseau (2003) posed is remote from both theory and practice in traditional ACA. White (2004) accused Ahlgren, Jarneving and Rousseau (2003) of correcting the taste of the "Drexel style" rather than their science. Even though Ahlgren, Jarneving and Rousseau (2004a) recognized their observation is purely theoretical and proven mathematically; they confirmed that something is wrong in the standard procedure of ACA, namely the use of the Pearson's r, and that is possible to correct it. Bensman (2004) regretted that Ahlgren, Jarneving and Rousseau (2003) did not give anywhere a "logical explanation for the reasons of their two axioms".

White (2003) also expressed disapproval to Ahlgren, Jarneving and Rousseau (2003). In the way they built data on co-citation matrix between 12 information retrieval and 12 scientometricians; he argued that one should not normally study literatures known to be disjointed and advised not to map authors that produce large "zero blocks" in the raw count matrix. He also criticized that Ahlgren, Jarneving and Rousseau (2003) did not cluster nor map their data to demonstrate how fluctuations in r's will mislead the analysis. He then visualized the data and found that the maps presented two distinct points at opposite poles; and he advised to do two different mappings to interpret two negatively correlated groups. Ahlgren, Jarneving and Rousseau (2004b) replied that they decided against publication of the maps because they did not bring any new arguments to their claim; they asserted that White's (2003) map essentially have proven this point.

White (2003) also used Ahlgren, Jarneving and Rousseau's (2003) data in multidimensional scaling and clustering routines and showed that "despite r's fluctuation, clusters based on it are much the same for the combined groups as for the separate groups". With different treatments for diagonal values and with various proximity measures in multidimensional scaling and clustering, he also concluded that there are small variations in cluster membership within groups.

SHOULD CO-CITATION MATRICES BE NORMALIZED?

Normalization refers to changing the scale of numbers in the data matrix, either for the matrix as a whole, or for each row (or column) separately; it makes the data comparable (Borgatti, Everett and Freeman 2002). While discussing this controversy, the authors opened an adjacent debate to Pearson's r use in ACA. It is related to normalization of co-citations matrices. The question has been formulated as followed by Leydesdorff (2007): "Should co-occurrence data be normalized?".

Leydesdorff and Vaughan (2006) argued that co-citation matrices are symmetrical matrices, and hence contains proximities data (either similarity or dissimilarity), and as such can be input into multi-dimensional software directly, to generate a map. They asserted that proximities data should first be derived from an asymmetrical matrix before analysis. They warned that when a symmetrical matrix is normalized, it should lead to the observations Ahlgren, Jarneving and Rousseau (2003) denounced. They gave an illustration taken from the SPSS software (Leydesdorff and Vaughan 2006). Waltman and van Eck (2007) did not approve this point of view; they pointed out an error in the SPSS software application version Leydesdorff and Vaughan (2006) used that had lead to incorrect multidimensional scaling map and mislead to such a conclusion. But in a rejoinder, Leydesdorff (2007) claimed that Waltman and van Eck (2007) did not distinguish sufficiently between the two types of matrices. Citing Burt (1982) and Schneider and Borlund (2007), he required normalization of authors attributes if one wishes to show the similarities between authors.

CONCLUSION

What similarity measure should be used in ACA and similar techniques? The debate is not perhaps closed, and the answer surely not easy. The primary aim of ACA is to know "who is close to whom in the eyes of citers" (White, 2004) and, as such, does not require much precision in visualisation. In this sense, even though r is not a "good similarity measure", it does the job well; indeed, the results it outputs is very similar to those of other similarities (White 2003; Bensman 2004; Leydesdorff 2008). The "Drexel team" agreed that this measure could be replaced, but rejected the theoretical arguments advanced by Ahlgren, Jarneving and Rousseau (2003).

Ahlgren, Jarneving and Rousseau (2003) presented the Cosine and the Chi-Square distance as alternatives to Pearson's r. Leydesdorff and Vaughan (2006) found that Pearson's r should not be used while normalizing symmetrical matrix, but rather to derive proximity data from an asymmetrical matrix. Egghe (2010) defines two mathematical conditions a similarity measure should fulfil and found that Cosine fails in satisfying the first but not the second. In the web environment, Leydesdorff and Vaughan (2006) advised to use the Jaccard index instead of Cosine. The other alternatives proposed to Pearson's r are Dice

(Egghe 2010), the Jensen-Shannon divergence, the Bhattacharya distance (van Eck and Waltman 2008), and the Euclidian distance (Leydesdorff 2008). van Eck and Waltman (2008) suggested more research on what similarity measure should be used in ACA and similar techniques.

REFERENCES

Ahlgren, P., Javerning, B. and Rousseau, R. 2003. Requirements for a co-citation similarity measure, with special references to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, Vol. 54, no. 6: 550-560.

Ahlgren, P., Javerning, B. and Rousseau, R. 2004a. Author co-citation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology*, Vol. 55, no. 9: 550-560.

Ahlgren, P., Javerning, B. and Rousseau, R. 2004b. Rejoinder: in defense of formal methods. *Journal of the American Society for Information Science and Technology*, Vol. 55, no. 10: 935-936.

Bensman, S.J. 2004. Pearson's r and author co-citation analysis: a commentary on the controversy. *Journal of the American Society for Information Science and Technology*, Vol. 55, no. 10: 935.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. *Ucinet 6 for Windows*. Harvard: Analytic Technologies.

Burt, R.S. 1982. *Toward a structural theory of action*. New York: Academic Press.

Davison, M.L. 1983. *Multidimensional scaling*. New York: John Wiley & Sons.

Dominic, S. 2001. *Mathematical foundations of information retrieval*. Dordrecht: Kluwer.

Egghe, L. 2009. New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology*, Vol. 60, no. 2: 232-239.

Egghe, L. 2010. Good properties of similarities measure and their complementarity. *Journal of the American Society for Information Science and Technology*, Vol. 61, no. 10: 2151-2160.

Egghe, L. and Leydesdorff, L. 2009. The relation between Pearson's r and Salton's cosine measures. *Journal of the American Society for Information Science and Technology*, Vol. 60, no. 5: 1027-1036.

Hopkins, F.L. 1984. New causal theory an enthnomethodology: co-citation patterns across a decade. *Scientometrics*, Vol.6, no.1: 33-53.

Leydesdorff, L. 2005. Similarity measures, author co-citation analysis, and information theory. *Journal of the American Society for Information Science and Technology*, Vol. 56, no. 7: 768-772.

Leydesdorff, L. 2007. Should co-occurrence data be normalized?: a rejoinder. *Journal of the American Society for Information Science and Technology*, Vol. 58, no 14: 2411-2413.

Leydesdorff, L. 2008. On the normalization and visualisation of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, Vol. 59, no. 1: 77-85.

Leydesdorff, L. and Vaughan, L. 2006. Co-occurrence matrices and their applications in information science: extending ACA to the web environment. *Journal of the American Society for Information Science and Technology*, Vol. 57, no. 12: 1616-1628.

McCain, K.W. 1984. Longitudinal author co-citation mapping: The changing structure of macroeconomics. *Journal of the American Society for Information Science*, Vol.35: 351-359.

McCain, K.W. 1990. Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science*, Vol. 41, no.6: 433-443.

Schneider, J.W. and Borlund, P. 2007. Matrix comparison, part 1: motivation and important issues for measuring the resemblance between proximity measures and coordination results. *Journal of the American Society for Information Science and Technology*, Vol.58, no.11: 1586-1595.

Spiegel, S. and Castellan, J.J. 1988. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

van Eck, N.J. and Waltman, L. 2008. Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology*, Vol. 59, no. 10: 1653-1660.

van Eck, N.J. and Waltman, L. 2009. How to normalize co-occurrence data? An analysis of some well known similarity measures. *Journal of the American Society for Information Science and Technology*, Vol. 60, no. 8: 1635-1650.

Waltman, L. and van Eck, N.J. 2007. Some comments on the question whether co-citation data should be normalized. *Journal of the American Society for Information Science and Technology*, Vol. 58, no. 11: 1701-1703.

White, H.D. 2003. Author cocitation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology*, Vol. 54, no. 13: 1250-1259.

White, H.D. 2004. Replies and a correction. *Journal of the American Society for Information Science and Technology*, Vol. 55, no. 9: 843 – 844.

White, H.D. and Griffith, B.C. 1981. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, Vol. 32, no 3: 163-171.