Similarity indexes for scientometric research: A comparative analysis

Hinde Adnani, Mohammed Cherraj, Hamid Bouabid*
Faculty of Science, Mohammed V University in Rabat,
Avenue Ibn Batouta, BP: 1014 RP, Rabat, MOROCCO
e-mail: adnanihinde@gmail.com; mcherraj@yahoo.fr
* h.bouabid@fsr.ac.ma (corresponding author)

ABSTRACT
A significant number of papers in the field of scientometrics addressed the comparisons of various similarity indexes. However, there is still a debate on the appropriateness of an index compared to others, beacause of the assessment differences reported in the literature. The objective of this paper is to make a comparative analysis of the five most used similarity indexes for the three scientometric analysis types: co-word, co-citation and co-authorship. A total of 388 papers addressing similarity indexes in scientometric analysis over three decades were retrieved from the Web of Science and examined; of which 49 were retained as the most relevant according to selective criteria. The approach consisted of building cross matrices for the five indexes (Jaccard, Dice-Sorensson, Salton, Pearson, and Association Strength) for the three types of scientometric analysis. For each of these analyses, a distinction is made between papers according to their theoretical or empirical results. Furthermore, papers are classified according to the mathematical formula of the similarity index being used (vector vs non vector). In the 49 relevant papers being selected, the comparative analysis showed that there is still no consensus on the appropriateness of an index for co-word and co-authorship analyses, while for co-citation, Salton is the widely preferred one. The Association Strength is the less covered and compared to other indexes for the three analysis types. An open source computer program was developed as a tool to facilitate empirical comparative studies of indexes. It allows generating normalized matrix of any chosen index for the two mathematical variants.

KEY WORDS: Similarity index; Co-word analysis; Co-authorship analysis, Co-citation analysis; Bibliographic coupling; Proximity Index.

INTRODUCTION

Since the development of the 'similarity' concept, as a mathematical approach to quantitatively assess similarity or proximity between entities - in a group - according to specific characteristics, it continues to be widely used, particularly in the scientometrics field. A similarity index is employed to normalize the data (raw matrix) within a given group (system, network, cluster, database, etc). Numerous studies report on the use or comparison of various indexes, diffrent results, diffrent recommendations, but often with no consensus on which similarity index is the most appropriate to draw similarity (or proximity) between the entities of the group for a given domain (Hamers et al. 1989; Ahlgren, Jarneving and Rousseau 2003; White 2003; White 2004; Bensman 2004; Egghe 2009; Van Eck and Waltman 2009).

The relationship between an index and its performance is not always obvious. Schneider and Borlund (2007b) emphasized that a better understanding of proximity indexes is important so that the most accurate and efficient one can be employed. Shoaib, Daud and Khiyal (2015) noted that researchers in the field of digital libraries pay limited attention to

similarity indexes and opt for ready-made alternate indexes to estimate optimum similarity, which may not provide real picture of similarity among publications.

Similarity indexes are often applied in three types of scientometrics analysis: co-word, co-citation and co-authorship. Co-word analysis is a content analysis technique to evaluate the strength of the association (relatedness) between keywords in textual data (Rip and Courtial 1984). An advantage of co-word analysis is that relatedness can be interpreted directly according to document contents (Lu and Wolfram 2012). Co-citation analysis assesses citations made to publications. If two articles are cited by the same third article, then these two articles are co-cited. Co-citation is considered the most influential approach for assessing co-occurrences (Small 1973). Co-authorship analysis assesses multiple authoring of a publication and has been considered a method of author relatedness. Co-authorship is also referred to as co-publication, especially when the analysis takes place at aggregated levels, such as a department, an institution, or a country. Bibliographic coupling is another concept for citation structure analysis (Kessler 1963). Two papers are bibliographically coupled if they both cite one or more papers in common. As a synchronous analysis, it is the opposite of co-citation (diachronous analysis). Bibliographic coupling, is somewhat older than co-citation analysis (Small 1973). It is used mostly for limited time frames (up to five-year interval) and does not inherently identify the most important studies by citation counts as co-citation does (Zupic and Cater 2014). Bibliographic coupling is found to be mostly associated with one of the three analysis types: as co-word and compared to the new semantic analysis of Subject-Action-Object (Sternitzke and Bergmann 2009), as co-citation for journals co-citations (Thijs, Zhang and Glänzel 2013) and as author's bibliographic coupling compared to author's co-citation (Khan 2012).

While similarity indexes have been well covered in the scientific literature by empirical and theoretical studies, they are still a hot topic in scientometrics. The appropriateness of an index is in fact a complex issue. For instance, the fit of Pearson in author co-citation shows divergence when looking at studies of Ahlgren, Jarneving and Rousseau (2003; 2004) on one side, and those of White (2003; 2004) and Bensman (2004) on the other side. Similar disagreement was also reported in other studies. Association Strength was advised as a good index for the scientometric research against others such as Jaccard, Salton and the Inclusion index (Van Eck and Waltman 2009). However, the former was not found to fit fundamental properties to be satisfied by any similarity index (if a constant vector is added to both vectors for which similarity is assessed, then the similarity must increase; while if one of these vectors is added to both vectors, then the similarity must increase) compared to Jaccard and Dice-Sorenson which fulfilled these two properties (Egghe 2010a; 2010b).

The choice of an index depends on the characteristics of the data to which it is applied (Schneider and Borlund 2007b), the similarity index and its mathematical formula (vector vs. non vector) to be used, and the type of analysis to be performed. Gmür (2003) showed that the choice of a similarity index has a strong impact on the results. In addition, for the purpose of science mapping, different indexes reveal different pictures of the research landscape (Sternitzke and Bergmann 2009), which in turn lead to a range of interpretations and consequently findings, implications and recommendations.

Another difficulty encountered in scientometric assessments is the understanding and the pertinence of the mathematical formula of the similarity indexes to be used. Some of them have non-vectorial (Scalar or Euclidean) formula, while others have both variants. The non-vector formula consists of a Euclidean formula of the scalar components of the co-

occurrences matrix. Instead the vector formula consists of a geometric representation, where the similarity refers to both the vector-norm and the angle between the two vectors in the geometric space.

To make an objective guided choice of the most appropriate similarity index to use in a scientometric analysis, it is important to understand its mathematical concept, and refer to comparative studies of this index with others according to different analysis types. The other challenge defying this type of studies, is the difficulties inherent to performing comparison among indexes with conventional tool particularily in relation to data size, processing time, repeatability and accuracy. Therefore, the objectives of the present paper are to:

a) conduct a comparative analysis of five similarity indexes (Jaccard, Salton's Cosine, Dice-Sorenson, Pearson and Association Strength) used in three main scientometric analysis types (co-word, co-citation and co-authorship), and
b) develop a computer program to facilitate the analysis and automatically generate normalized matrices using one or more indexes for comparative scientometric studies.

MATERIALS AND METHODS

Selection of Papers

Papers of interest to the this study were extracted from the Web of Science Database (Clarivate Analytics) using appropriate tags[1]. The 1984 to 2015 period was used as a time span for the search. A first list of papers was obtained and was afterwards subject to a three-step filtering to select only the most relevant papers that deal effectively with the comparison of the five similarity indexes in the three scientometric analyses. These filtering steps were:

a) The paper should use at least one similarity index among those of interest to this study in at least one of the three analysis types;
b) The paper should compare at least two similarity indexes;
c) The paper should provide the mathematical formula (vetor or non-vector formula) for the similarity indexes being compared.

Similarity Indexes

The five indexes considered in this paper are Jaccard, Dice-Sorenson, Salton's Cosine, Pearson and Association Strength. Jaccard, Salton's Cosine, Dice-Sorenson, and Association Strength (known also as Proximity Index) were chosen because they are the most widely used indexes in the field of scientometrics (Van Eck and Waltman 2009). Dice-Sorenson, somewhat similar to Jaccard index, was chosen as it is one of the oldest similarity measures (1945) that is still widely used in the scientometrics field. Pearson was considered because it is an indirect similarity assessment index, a vector-space index and is the only index to describe both similarity and dissimilarity using a -1 to +1 scale (Pearson 1895). Association Strength is the only index that has no vector-variant formula and was given a probabilistic conceptualization (Van Eck and Waltman 2009).

[1] The search was done in Web os Science using 'Advanced Search' and the field tag 'TS' as : TS=(Jaccard OR Salton OR Cosine OR Pearson OR Poduvkin-Garfield OR Dice OR Sorenson OR Association-Strength OR Similarity OR Proximity) AND TS=(co-word OR co-citation OR co-publication OR co-author* OR coword OR cocitation OR copublication OR coauthor*).

The mathematical formulations of these similarity indexes are described hereafter. For all equations, ***n*** represents the number of entities within a group (words, authors, institutions, countries, fields, journals, citations, etc), ***i*** and ***j*** are two entities for which similarity is to be assessed. X_{ij} is the number of co-occurrences of both entity ***i*** and entity ***j***; the term 'co-occurrence' is used in a general sense and covers co-word, co-citations and co-authorship. Y_{it} and Y_{tj} are the total numbers of co-occurrences of entities ***i*** and ***j*** respectively, where the indice ***t*** refers to the total number of co-occurences for a given entity:

$$Y_{it} = \sum_{j=1}^{n} X_{ij} \qquad Y_{tj} = \sum_{i=1}^{n} X_{ij} \tag{1}$$

The matrix $\lfloor X_{ij} \rfloor$ may be symmetric or not. For each index, the vector and non vector formulas are presented.

Jaccard (J): this index (originally called Community Coefficient) was introduced by the Swiss Botanist Paul Jaccard (Jaccard 1901). Its original non-vector formula is written as:

$$J_{ij} = \frac{X_{ij}}{Y_{it} + Y_{tj} - X_{ij}} \tag{2}$$

The vector-variant formula of the this index, also referred to as Jaccard-Tanimoto index (Tanimoto 1957; Cha, Choi and Tappert 2009), is written as:

$$J_{ij} = \frac{\sum_{k=1}^{n} X_{ik} X_{kj}}{\sum_{k=1}^{n} X_{ik}^2 + \sum_{k=1}^{n} X_{kj}^2 - \sum_{k=1}^{n} X_{ik} X_{kj}} \qquad J_{ij} = \frac{\vec{X}_i \vec{X}_j}{\left\|\vec{X}_i\right\|^2 + \left\|\vec{X}_j\right\|^2 - \vec{X}_i \vec{X}_j} \tag{3}$$

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

$$D_{ij} = \frac{2.X_{ij}}{Y_{it} + Y_{tj}} \tag{4}$$

The vector variant of Dice-Sorenson is written as:

$$D_{ij} = \frac{2\sum_{k=1}^{n} X_{ik} X_{kj}}{\sum_{k=1}^{n} X_{ik}^2 + \sum_{k=1}^{n} X_{kj}^2} \quad \text{or} \quad D_{ij} = \frac{2\vec{X}_i \vec{X}_j}{\left\|\vec{X}_i\right\|^2 + \left\|\vec{X}_j\right\|^2} \tag{5}$$

Salton's Cosine (COS): this index was introduced by Salton (Salton and McGill 1983) to assess similarity between two 'vectors' by calculating the cosine of the angle between them. It can be applied under the vector and non-vector variants. The non-vector formula is written as:

$$Cos_{ij} = \frac{X_{ij}}{\sqrt{Y_{it}.Y_{tj}}} \tag{6}$$

and the vector formula is written as:

$$Cos_{ij} = \frac{\sum_{k=1}^{n} X_{ik} X_{kj}}{\sqrt{\sum_{k=1}^{n} X_{ik}^2} \sqrt{\sum_{k=1}^{n} X_{kj}^2}} \qquad Cos_{ij} = \frac{\vec{X}_i \vec{X}_j}{\left\|\vec{X}_i\right\| . \left\|\vec{X}_j\right\|} \tag{7}$$

Pearson (r): this index (also known as Pearson's Correlation Coefficient) was proposed by Pearson (1895) as a statistical measure. Its non-vector formula is written as:

$$r_{ij} = \frac{\sum_{k=1}^{n} (X_{ik} - \overline{X_{i.}}).(X_{kj} - \overline{X_{.j}})}{\sqrt{\sum_{k=1}^{n} (X_{ik} - \overline{X_{i.}})^2} . \sqrt{\sum_{k=1}^{n} (X_{kj} - \overline{X_{.j}})^2}} \tag{8}$$

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

The vector formula of Pearson index is written as:

$$r_{ij} = \frac{(\vec{X}_i - \overline{X_{i.}} \vec{I}).(\vec{X}_j - \overline{X_{j.}} \vec{I})}{\left\|\vec{X}_i - \overline{X_{i.}} \vec{I}\right\| . \left\|\vec{X}_j - \overline{X_{j.}} \vec{I}\right\|} \tag{9}$$

It is worth noting that Pearson's formula (equation 8) is the Cosine (equation 7) applied to the average vector which is the original vector from which the weighted average component of the original vector is subtracted.

Association Strength (AS): this index is also known as the Proximity Index (Rip and Courtial 1984). It was proposed by Van Eck et al. (2006) as a probabilistic measure. The Association Strength formula exists as a non-vector formula only, and is written as:

$$AS_{ij} = n \frac{X_{ij}}{Y_{it} Y_{tj}} \tag{10}$$

Scientometric Analysis Types

The three types of analysis considered in this paper are: co-word, co-citation and co-authorship. As stated in the introduction these types of analysis are the most used in scientometrics:

a) ***Co-word analysis:*** evaluates the strength of the association (relatedness) between keywords in textual data which directly appears from document contents.
b) ***Co-citation analysis:*** evaluates the citations co-occurrences; when two articles are cited by the same third article, they are considered co-cited. The assumption is that the appearance of two articles in the same reference list indicates likely an association between these two articles.
c) ***Co-authorship analysis:*** evaluates author relatedness. Co-authorship is also considered in this category when the analysis takes place at aggregated levels (department, institution or country). It reflects ties among the collaborating authors.

Computer Program for Calculating the Similarity Indexes

The use of the mathematical formulas of the indexes presented above is often complex, particularly when there is a need to apply more than one index at the same time (irrespective of analysis type) to generate the normalized matrices and be empirically compared. In this regard, a computer program (named **S**imilarity **I**ndex **Co**putation **P**rogram, **SICoP**), was developed to help accomplish this task. The SICoP is written with *Pro Fortran* language. It includes the set of indexes described above beside others, and offers the choice of applying both vector and non-vector formulas. The program is a free resource and can be downloaded (executable file and tutorial) from the website: http://www.fsr.ac.ma/page/ressources-de-la-fsr.

RESULTS AND DISCUSSIONS

The first run in the papers selection process yielded 388 papers. The second stage of a qualitative check, 304 papers were selected; the filtering excluded papers that

a) simply mentioned the similarity index and the analysis in their topics,
b) made a simple reference to them but not in a perspective of studying them,
c) used two indexes but respectively and incomparably in distinctive analysis types, or
d) combined 'at the same time' two indexes or more for a specific analysis type.

The third stage, which filtered papers that were both relevant to the topic by effectively using one of the aforementioned similarity indexes for an explicit scientometrics co-occurrence purpose, gave a corpus of 136 papers. The fourth stage, which sieved papers that provided the index mathematical formula (vetor or non-vector formula) and offered an explicit comparison of at least two of the aforementioned similarity indexes, gave a final list comprising 49 papers (Table 1).

Figure 1 reports a breakdown of the similarity indexes' frequency as they appear in the 49 selected papers. It shows that in the scientometric researches investigated, co-word is the dominant analysis type (46%) followed by co-citation (36%); while co-authorship was present in 18 percent of the cases. It also shows that Salton and Jaccard are the most employed indexes, with respectively 36 percent and 28 percent of the cases, totalling about two thirds of the examined studies.

One of the main findings of the overview of these 49 papers is that there are two distinct categories of papers: (a) one that comprised only very few papers that compared similarity indexes with no distinction among the three scientometric analysis types, and (b) the other comprised a large number of papers that focused only on specific scientometric analysis type.

Table 1: List of the 49 Papers in Alphabetical Order According to the Five Similarity Indexes and their used Mathematical Variant (Vector vs Non-vector)

Authors	Jaccard		Salton		Pearson		Dice - Sorenson		Association Strength	
	Vector	Non-vector	Vector	Non-vector	Vector	Non-vector	Vector	Non-vector	Vector	Non-vector
Ahlgren et al.(2003)			x		x					
Ahlgren et al.(2004)			x		x					
Al-Kharashi & Evens (1994)		x		x				x		
Bensman (2004)			x		x					
Boyack et al.(2005)		x		x	x					x
Chaudhari & Dharmadhikari (2012)				x	x					
De Meo et al. (2012)		x								
Egghe & Leydesdorff (2009)	x		x				x			
Egghe & Rousseau (2006)		x	x					x		
Egghe (2009)	x		x				x			
Egghe (2010a)	x		x						x	
Egghe (2010b)	x		x				x			
Elmacioglu et al. (2007)		x	x							
Finardi (2015)		x		x						
Froud et al. (2012)	x		x		x					
Gamallo & Bordag (2011)	x		x							
Gmür (2003)		x			x					
Hadj Taieb et al.(2013)			x		x			x		
Harmers et al. (1989)		x		x						
Jung (2015)	x						x			
Khan (2012)		x	x		x					
Klavans & Boyack (2006)		x		x	x					x
Lakshmi (2013)			x							
Leydesdorff & Zaal (1988)		x	x		x					
Leydesdorff (2008)	x	x	x		x					
Linyuan et al. (2012)				x	x			x		
Lü & Zhou (2011)		x		x				x		
Luukkonen et al. (1993)		x		x						
Naidu et al. (2013)					x					
Narayanan et al. (2013)	x		x		x					
Porter et al. (2007)			x							
Rip & Courtial (1984)		x								x
Rorving (1999)		x	x					x		
Saad & Kamarudin (2013)		x		x				x		
Schneider & Borlund (2007a)				x	x					
Schneider & Borlund (2007b)				x	x					
Singh et al. (2014)	x		x				x			
Sorkhi & Hashemi (2015)		x		x						
Sternitzke & Bergman (2009)		x		x						
Stvilia et al. (2009)			x							
Subhashini & Kumar (2010)		x		x						
Thada & Jaglan (2013)		x		x				x		
Thijs & Glanzel (2010)		x	x							
Van Eck & Waltman (2008)			x		x					
Van Eck & Waltman (2009)		x		x						x
Wagner & Leydesdorff (2003)	x		x							
Wang & Sukthankar (2013)*			x		x					
White (2003)			x		x					
White (2004)			x		x					

* does not present the formula for Salton

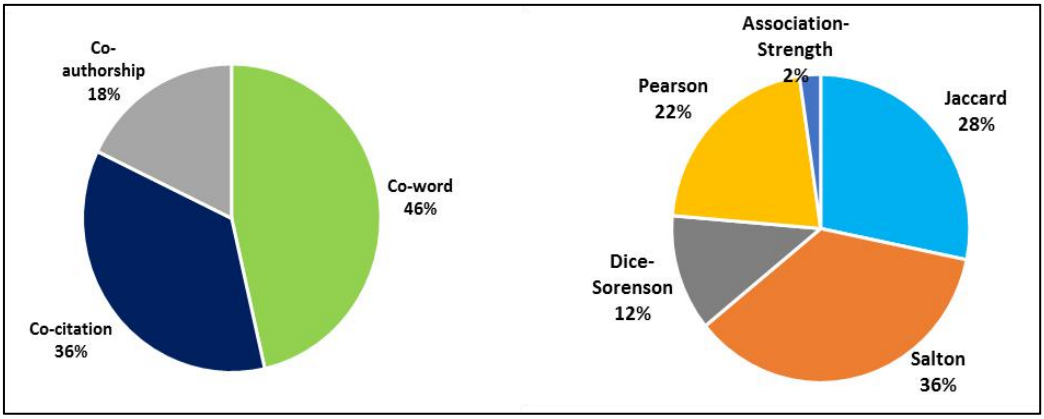

Figure 2: Frequency Ratios of Indexes and Scientometric Analysis Used in the 49 Selected Papers.

No Distinction Among Scientometric Analysis Type

This category includes only four papers (Egghe 2009; Van Eck and Waltman 2009; Egghe 2010a, 2010b). Comparing similarity indexes irrespective of scientometric analysis type allows depicting very precise understanding of the indexes used and their average results (outputs) for both the three types of analyses. The work by Egghe (2009) was theoretical and compared eight vector similarity indexes (including Jaccard, Salton and Dice-Sorenson), irrespective of the three scientometric analyses; Jaccard was found to be convexly an increasing function of Salton, but always smaller or equal to the latter. All the other studied indexes had a linear relation with the Salton one, but these indexes remain good similarity measures as well. It can be noted, that these comparisons were made with the hypothesis under which the Euclidean norm of the two vectors is related with a constant. In addition, the comparison considered only indexes which have the same numerator, hence limiting the difference to only denominators of these indexes. This may be the reason for not considering Pearson. In the study by Van Eck and Waltman (2009), that compared theoretically and empirically four similarity indexes (Association Strength, Salton, Inclusion Index and Jaccard), it was found that Association Strength was considered the best index, compared to other indexes such as Jaccard, Salton and the Inclusion index, despite their popularity.

The Association Strength was later reported to be a function of Jaccard (and also of Salton), which evolves into two stages: first convex and then concave (Egghe 2010a). Imposing two properties to be satisfied by any similarity measure in the case of Jaccard, Dice-Sorenson, Salton, Overlap and Association Strength (i: if adding a constant vector to both vectors, then the similarity must increase, and ii: if one of the two vectors is added to both vectors, then the similarity must also increase), it was shown that Dice and Jaccard satisfy the first property while Salton and Association strength did not. With regard to the second property, all of Dice, Jaccard and Salton do satisfy this property, while the Association Strength does not (Egghe 2010b) considering the same hypotheses stated earlier (Egghe 2009).

Specific Scientometric Analysis Type

This category of analysis provides explicit outputs on the relevance of one similarity index compared to others for a specific scientometric analysis type.

Co-Word Analysis

The cross matrix for theoretical and empirical studies comparing similarity indexes for co-word analysis is presented in Table 2. The results show that Jaccard, Pearson, Salton and Dice were the most addressed indexed, while Association Strength was not studied. They also show that co-word analysis has been well covered where comparisons between indexes exist in theoritical studies as well as empirical ones, except for few cases when comparing for example Pearson to Dice or Association Strength to Pearson. One of the major researches in co-word analysis is the one by Rip and Courtial (1984) that made a comparison between four indexes: Jaccard, Normalized Deviation, Inclusion and Association Strength (named as Proximity Index). This comparison consisted of exploring the relevance of an index according to the map visibility and clarity. The comparison concluded that Association Strength is more appropriate to make visible minor areas of research (less weighted areas) unlike Jaccard or Inclusion index that make major areas of the field structure more apparent. Indeed, one could see from its formula that the Association Strength enhances links of entities with less occurrences (less frequent entities) at the expense of entities with high occurrences.

The choice of an index between Jaccard, Pearson and Salton for a co-word analysis (both in title and in abstract) does not make much difference (Leydesdorff and Zaal 1988). The same statement applies to information retrieval when comparing Jaccard, Dice and Salton, even if Salton is more sensitive than Jaccard and Dice to a change in the two arrays that have been normalized (Egghe and Rousseau 2006).

Another study developed theoretical as well as geometrical analysis which compared a set of seven similarity indexes (Inner Product, Salton, Pseudo-Salton, Dice, Pearson, Covariance, Overlap and Spreading Activation) in the geometric space, based on the nature of the iso-similarity contours of each similarity index containing Query and Object vectors (Jones and Furnas 1987). This work did not recommend a specific index but opted for presenting for each index its five important attributes in the geometric space (angle monotonicity, radial monotonicity, component-wise monotonicity, unbounded single-component influence, and boundedness of similarity rating).

An interesting finding from this category is the comparaison of indexes in co-word analysis in other languages. Using Jaccard, Dice and Salton respectively for Arabic information retrieval (Al-Kharashi and Evens 1994) and for web retrieved documents (Thada and Jaglan 2013), it was found that the best results were obtained using Salton followed by Dice and then Jaccard. A similar result was reported in the work by Froud, Lachkar and Alaoui (2012) that compared Jaccard, Salton and Pearson in a specific Arabic co-words analysis and concluded that Salton and Pearson are quite similar for measuring the similarity between the Arabic words and perform better than Jaccard. It can be retained from these studies that Salton is a common index agreed upon in the case of Arabic language. Unfortunately, no other papers dealing with other languages were encountered.

Jaccard, Dice and Salton were also compared (among a set of five indexes) for calculating relative distance measures in the topic-document context (Rorving 1999). The results showed that Dice and Jaccard produced the most similar images between ordinal and interval treatments. Salton appeared (as for the Overlap measure) to perform better by recovering optimal data relationships among documents in the context of the study.

Table 2: Cross Matrix for Theoretical and Empirical Studies Comparing Similarity Indexes for Co-Word Analysis

(Theoretical research)

Jaccard	**Dice-Sorenson**	**Pearson**	**Association Strength**	
Egghe & Rousseau (2006) Egghe (2009) Egghe (2010a) Egghe (2010b) Singh et al. (2014) Sternitzke & Bergmann (2009) Van Eck & Waltman (2009)	Egghe & Rousseau (2006) Egghe (2009) Egghe (2010b) Jones & Furnas (1987) Singh et al. (2014)	Chaudhari & Dharmadhikari (2012) Jones & Furnas (1987)	Egghe (2010a) Egghe (2010b) Van Eck & Waltman (2009)	**Salton**
	Egghe & Rousseau (2006) Egghe (2009) Egghe (2010b) Singh et al. (2014)	Jones & Furnas (1987)	Egghe (2010a) Egghe (2010b) Van Eck & Waltman (2009)	**Jaccard**
		-	Egghe (2010b)	**Dice-Sorenson**
			-	**Pearson**

(Empirical research)

Jaccard	Dice-Sorenson	**Pearson**	**Association Strength**	
Al-Kharashi & Evens (1994) Froud et al. (2012) Gamallo & Bordag (2011) Lakshmi (2013) Leydesdorff & Zaal (1988) Linyuan et al. (2012) Narayanan et al. (2013) Rorving (1999) Saad & Kamarudin (2013) Singh et al. (2014) Sternitzke & Bergmann (2009) Subhashini & Kumar (2010) Thada & Jaglan (2013) Van Eck & Waltman (2009)	Al-Kharashi & Evens (1994) Gamallo & Bordag (2011) Hadj Taieb et al. (2013) Linyuan et al. (2012) Rorving (1999) Saad & Kamarudin (2013) Singh et al. (2014) Thada & Jaglan (2013)	Froud et al.(2012) Lakshmi (2013) Leydesdorff & Zaal (1988) Linyuan et al. (2012) Narayana et al. (2013) Porter et al. (2007)	Van Eck & Waltman (2009)	**Salton**
	Al-Kharashi & Evens (1994) Gamallo & Bordag (2011) Jung (2015), Linyuan et al. (2012) Rorving (1999) Saad & Kamarudin (2013) Singh et al. (2014) Thada & Jaglan (2013)	Froud et al. (2012), Lakshmi (2013) Leydesdorff & Zaal (1988) Linyuan et al. (2012) Narayanan (2013)	Rip & Courtial (1984) Van Eck & Waltman (2009)	**Jaccard**
		Linyuan et al. (2012)	-	**Dice-Sorenson**
			-	**Pearson**

In a context of document mapping, the comparaison of Jaccard, Salton and Inclusion indexes lead to the conclusion that the Inclusion index is preferred over Jaccard and Salton (Sternitzke and Bergmann 2009). However, Jaccard and Pearson were considered as better indexes instead of Salton, for more meaningful and coherent document clustering (Lakshmi 2013). They are considered the best indexes against Salton when it comes to distributed document clustering because they both generated more coherent clusters and improved the proposed algorithm, with respect to the two criteria of accuracy and clustering quality (Narayanan, Judith and JayaKumari 2013). The inappropriateness of Salton compared to Pearson is also supported in the case of 'systems recommender' in web data retrieval and among ten similarity indexes including Jaccard, Salton, Pearson and Dice-Sorenson (Linyuan et al. 2012). Salton appears also to perform less than Jaccard and Dice-Sorenson in measuring semantic similarity between sentences (Saad and Kamarudin 2013).

However, Salton was found to be the best measure for word similarity extraction in Singular Value, Baseline, Log and Filter models, while Jaccard and Dice, providing almost the same scores, were not recommended (Gamallo and Bordag 2011). Another study applying vector space model for textual information retrieval in the world wide web found that Salton besides Jaccard, Dice-Sorenson, Inner Product, and Overlap outperform other probabilistic indexes in large scale information retrievals (Singh, Singh and Chaba 2014).

Co-Citation Analysis

Table 3 presents the cross matrix for theoretical and empirical cases comparing similarity index for co-citation analysis. In the case of author co-citation (ACA), while Pearson was considered as a highly relevant similarity index (White 2003, 2004; Bensman 2004) because it gauges similarity as well as dissimilarity in partitioning authors accordingly, it was refuted in other studies (Ahlgren, Javerning and Rousseau 2003, 2004) arguing that Pearson fails (compared mainly to Salton) to fulfill two tests of stability measurement where two authors may be frequently co-cited; but the respective similarity score may decrease due to the presence of other authors who have not been co-cited with these two authors. Similar finding reported by Klavans and Boyack (2006) in analyzing co-citation for mapping science based on the four criteria of accuracy, coverage, scalability and robustness, stated that Pearson is less accurate for high coverage and Salton is the most accurate for co-citations.

Yet in co-citation analysis, and as argued by Ahlgren, Javerning and Rousseau (2003), Salton is preferred for the purpose of visualization (Leydesdorff 2008). Salton index was prefered compared to Jaccard when the occurrence matrix cannot be retrieved due to the difference that exists between the two formulas of Jaccard, vector and non-vector (Leydesdorff 2008). However, when assessing world science maps of journals co-citation and inter-citation based on two different accuracy criteria, 'scalability of the similarity algorithm' and the 'readability of layouts', both Jaccard and Salton are good similarity indexes while Pearson failed on either criteria (Boyack, Klavans and Borner 2005). Differences between Pearson and Salton indexes were reported to be marginal in practice in other studies (Naidu, Ramu and Srinivas 2013) and these two indexes were even considered related based on a theoretical comparison (Egghe and Leydesdorff 2009). However, one could note that this last relation, plotted as a cloud of points, was not a pure function but is described by a sheaf of increasing straight lines whose slopes decrease. A study by Khan (2012) argued that neither Pearson nor Salton or Jaccard are feasible for co-citation analysis when applied to detect conflict of interest. From a hypothetical example, this study concluded that the similarity score of an author and reviewer might be low if

both are frequently co-cited together, but simultaneously co-cited with a complete or partial disjoint set of other authors or authors with small co-cited values.

Table 3: Cross Matrix for Theoretical and Empirical Cases Comparing Similarity Index for Co-Citation Analysis

(Theoretical research)

Jaccard	**Dice-Sorenson**	**Pearson**	**Association Strength**	
Egghe & Leydesdorff (2009) Egghe (2009) Egghe (2010b) Hamers et al. (1989)	Egghe & Leydesdorff (2009) Egghe (2009) Egghe (2010b)	Ahlgren et al. (2004) Bensman (2004) Egghe & Leydesdorff (2009) Schneider & Borlund (2007) White (2004)	Egghe (2010b)	**Salton**
	Egghe & Leydesdorff (2009) Egghe (2009) Egghe (2010b)	Egghe & Leydesdorff (2009)	Egghe (2010b)	**Jaccard**
		-	Egghe (2010b)	**Dice-Sorenson**
			-	**Pearson**

(Empirical research)

Jaccard	**Dice-Sorenson**	**Pearson**	**Association Strength**	
Boyack et al. (2005) Khan (2012) Klavans & Boyack (2006) Leydesdorff (2008) Leydesdorff & Zaal (1988) Stvilia et al. (2009)	-	Ahlgren et al. (2003) Boyack et al. (2005) Khan (2012), Klavans & Boyack (2006) Leydesdorff & Zaal (1988) Naidu et al. (2013) Van Eck & Waltman (2008) White (2003, 2004)	-	**Salton**
	-	Boyack et al. (2005) Gmür (2003) Khan (2012) Klavans & Boyack (2006) Leydesdorff (2008) Leydesdorff & Zaal (1988)	-	**Jaccard**
		-	Egghe (2010b).	**Dice-Sorenson**
			-	**Pearson**

Co-Authorship Analysis

The cross matrix for theoretical and empirical cases comparing similarity index for co-authorship analysis is presented in Table 4. Co-authorship is probably the only scientometric analysis without theoretical research on the appropriateness of Pearson and Association Strength compared to other indexes. At an empirical level, Association

Strength does not catch yet scientometricians attention in order to be accurately compared to other indexes.

When applying the most used similarity indexes to co-authorship at countries level, Jaccard tends to underestimate the collaboration of smaller countries with larger ones, whereas Salton underestimates the collaboration of smaller countries with one another (Luukkonen et al. 1993). In another context of countries co-authorship (BRICS countries), Jaccard always produced smaller similarities than Salton, which supports the theoretical findings by Egghe (2009).

Table 4: Cross Matrix for Theoretical and Empirical Cases Comparing Similarity Index for Co-Authorship Analysis

(Theoretical research)

Jaccard	Dice-Sorenson	Pearson	Association Strength	
Egghe (2009)	Egghe (2009)	-	-	**Salton**
	Egghe (2009)	-	-	**Jaccard**
		-	-	**Dice-Sorenson**
			-	**Pearson**

(Empirical research)

Jaccard	Dice-Sorenson	Pearson	Association Strength	
Elmacioglu et al. (2007) Finardi (2015), Khan (2012) Lü & Zhou (2011) Luukkonen et al. (1993) hijs & Glänzel (2010) Sorkhi & Hashemi (2015) Wagner & Leydesdorff (2003)	Lü & Zhou (2011)	Khan (2012) Luukkonen et al. (1993) Wagner & Leydesdorff (2003) Wang & Sukthankar (2013)	-	**Salton**
	Jung (2015) Lü & Zhou (2011)	De Meo et al. (2012) Khan (2012) Luukkonen et al. (1993) Wagner & Leydesdorff (2003)	-	**Jaccard**
		-	-	**Dice-Sorenson**
			-	**Pearson**

At an institutional level, Thijs and Glänzel (2010) reported that Salton was better to reveal whether institutions that collaborate have more similar research profiles than those that do not collaborate, but advocated Jaccard instead to shape the reverse effect, that is, if collaborating institutions that have more similar profiles have a stronger collaborationship (despite the fact that there was no evidence provided in their study for both cases). The work by Lü and Zhou (2011) compared nine similarity indexes (Common Neighbours, Salton, Jaccard, Dice-Sorenson, Hub Promoted, Hub Depressed, Leicht-Holme-Newman, Preferential Attachment, Adamic-Adar) in six different real network contexts: co-

authorship, protein interaction, electrical grid, political blogs, internet, air transportation. Even if this work concluded that the Common Neighbors index performed the best accuracy in link predicting within a network for all the six studied contexts, one can note that Jaccard and Dice also experienced the same higher score for the co-authorship context (as well as for electrical grid and internet contexts) and that Salton showed a lesser but a very closer score to these two indexes in the co-authorship context. In the study by Khan (2012), Pearson, Salton and Jaccard were used in both co-citation and co-authorship analyses for detection of conflict of interest between authors and reviewers in peer review system, it was found that none of these indexes was feasible, arguing that the reason behind this is that the similarity score of an author and reviewer might be low if both are even co-cited together frequently, but simultaneously co-cited with a complete or partial disjoint set of other authors or authors with small co-cited values.

CONCLUSION

This paper provides a comparative analysis of five most used similarity indexes (Jaccad, Dice-Sorenson, Salton, Pearson and Association Strength) for the three scientometric analysis types: co-word, co-citation and co-authorship. Among 388 papers retrieved for a period of three decades, only 49 papers were retained as most relevant according to specific criteria. Cross-matrices compiling these articles according respectively to either theoretical and empirical analysis were built and examined. They provide scientometricians with practical and informative analysis comparing the five similarity indexes according to each of the three scientometric analysis types.

With the except of co-citation, no consensus exists on the best index for a specific analysis type. Our results revealed some important highlights. The use of an index is influenced by its original variant. Relative 'old' ones namely Jaccard, Dice-Sorenson and Association Strength (Proximity Index), originally built as a non-vector variants, continue to be preferred as such. However, for recent indexes such as Salton and Pearson, their vector-variant continues to mark their use. All indexes are not linked, except Jaccard which is found, in its vector variant, to be theoretically a function of Salton but tend to be smaller or equal to the latter. The relation between some indexes put forward by some works needs more in-depth studies in diverse empirical cases. In terms of analysis types, co-word was more covered both at theoretical and empirical levels, which brings to light more information about the appropriate index with respect to the purpose of the analysis, but to a large extent does not recommend Salton which appears to perform less. In co-citation analysis, Salton was the most preferred index, while Dice-Sorenson and Association Strength were less empirically compared, even if theoretical comparisons existed with almost all the other indexes. Unlike co-word and co-citation analyses, in co-authorship, Association Strength was not compared to other indexes on both theoretical and empirical basis. The computer program (SICoP) developed in this study is a valuable tool to help automatically generate the normalized matrix by employing one or more of the major similarity indexes at the same time, for both vector and non-vector variants, and to empirically compared them for a rational choice.

Beyond its profound literature analysis, this study opens up several research perspectives with regard to both theoretical and empirical comparisons of the appropriateness of specific similarity indexes (such as Pearson and Association Strength). In order to make these perspectives easier, the free computer program (SICoP) is a valuable tool.

ACKNOWLEDGEMENTS

This study was partially supported by the *Agence Universitaire de la Francophonie - Bureau Maghreb* for which the authors are very grateful. The authors would like to thank Professor L. Waltman and Professor R. Bouabid for their relevant comments and support which result in a substantial improvement of the paper. The authors would like to thank also Professor M. Gislason and Professor A. Mzerd, for their kind help. Last but not least, the authors acknowledge the anonymous referees for their valuable comments and useful suggestions on an earlier draft of this paper.

REFERENCES

Ahlgren, P., Jarneving, B. and Rousseau, R. 2003. Requirements for a cocitation similarity measure, with special reference 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. 2004. Rejoinder: In defense of formal methods, *Journal of the American Society for Information Science and Technology*, Vol. 55, no. 10: 935-936.

Al-Kharashi, I. A. and Evens, M. W. 1994. Comparing words, stems, and roots as index terms in an Arabic information retrieval system. *Journal of the American Society for Information Science*, Vol. 45, no. 8: 548-560.

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.

Boyack, K. W., Klavans R. and Borner K. 2005. Mapping the backbone of science. *Scientometrics*, Vol. 64, no. 3: 351-374

Cha, S. H., Choi, S. and Tappert, C. C. 2009. Anomaly between Jaccard and Tanimoto coefficients. *Proceedings of Student-Faculty Research Day*, CSIS, Pace university.

Chaudhari, M. P. J. and Dharmadhikari, D. D. 2012. Clustering with multi-viewpoint based similarity measure: An overview. *International Journal of Engineering Inventions*, Vol. 3, no. 1: 2278-7461.

De Meo, P., Ferrara, E., Fiumara, G. and Ricciardello, A. 2012. A novel measure of edge centrality in social networks. *Knowledge-Based Systems*, Vol. 30: 136-150.

Dice R. 1945, Measures of the amount of ecologic association between species in *Ecology*, Vol. 26, no. 3: 2976-302.

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. 2010a. On the relation between the Association Strength and other similarity measures. *Journal of the American Society for Information Science and Technology*, Vol. 61, no. 7: 1502-1504

Egghe, L. 2010b. Good properties of similarity measures 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 correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology*, Vol. 60, no. 5: 1027-1036.

Egghe, L. and Rousseau, R. 2006. Classical retrieval and overlap measures such as Jaccard's coefficient, Salton's cosine measure and the Dice coefficient satisfy the requirements

for rankings based upon a Lorenz curve, Information *Processing & Management*, Vol. 42, no. 1: 106-120.

Elmacioglu, E., Kan, M. Y., Lee, D. and Zhang, Y. 2007. Web based linkage. *In Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management* (pp. 121-128).

Finardi, U. 2015. Scientific collaboration between BRICS countries. *Scientometrics,* Vol. 102, no. 2: 1139-1166.

Froud, H., Lachkar, A. and Alaoui, O.S. 2012. A comparative study of root-based and stem-based approaches for measuring the similarity between Arabic words for Arabic text mining applications. *Advanced Computing: An International Journal,* Vol. 3, no. 6: 1212-3634.

Gamallo, P. and Bordag, S. 2011. Is singular value decomposition useful for word similarity extraction?. *Language Resources and Evaluation*, Vol. 45, no. 2: 95-119.

Gmür M. 2003, Co-citation analysis and the search for invisible colleges: A methodological evaluation, *Scientometrics*, Vol. 57, no. 1: 27-57.

Hadj Taieb, M. A., Ben Aouicha, M. and Ben Hamadou, A. 2013. Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems*, Vol. 50: 260-278.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. 1989. Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Information Processing & Management*, Vol. 25, no. 3 : 315-318.

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*, Vol. 37: 241-272.

Jones W. P. and Furnas G. W. 1987. Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science,* Vol. 38, no. 6: 420-442.

Jung, J. J. 2015. Big bibliographic data analytics by random walk model. *Mobile Networks and Applications*, Vol. 20, no. 4: 533-537.

Kessler, M. M. 1963. Bibliographic coupling between scientific papers. *American Documentation*, Vol. 14, no. 1: 10-25.

Khan S. M., 2012. Exploring citations for conflict of interest detection in peer review system. *International Journal of Computer Information Systems and Industrial Management Applications*, Vol. 3: 283-299.

Klavans, R. and Boyack, K. W. 2006. Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, Vol. 57, no. 2: 251-263.

Lakshmi, M. S. V. 2013. Correlation preserving indexing based text clustering world. *Journal of Engineering Science*, Vol. 1, no. 1: 30-37.

Leydesdorff, L. 2008. On the normalization and visualization 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 Zaal, R. 1988. Co-words and citations relations between document sets and environments. In L. Egghe & R. Rousseau (Eds.), *Informetrics*, 87/88, 105-119, Amsterdam: Elsevier.

Linyuan, L., Matus, M., Chi Ho, Y., Yi-Cheng, Z., Zi-Ke, Z. and Tao, Z. 2012. Recommender systems. *Physics Reports,* Vol. 519, no. 1: 1-49.

Lu, K. and Wolfram, D. 2012. Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, Vol. 63, no. 10: 1973-1986.

Lü, L. and Zhou, T. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, Vol. 390, no. 6: 1150-1170.

Luukkonen, T., Tijssen, R. J., Persson, O. and Sivertsen G. 1993. The measurement of international scientific collaboration. *Scientometrics*, Vol. 28, no. 1: 15-36.

Naidu, C. S., Ramu, V. and Srinivas, D. 2013. Applications of MVS on hierarchical clustering algorithms. *IJRCCT*, Vol. 2, no. 12: 1409-1415.

Narayanan, N., Judith, J. E. and JayaKumari, J. 2013. Enhanced distributed document clustering algorithm using different similarity measures. *Information & Communication Technologies (ICT), 2013 IEEE Conference*, 545-550.

Pearson, K. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London,* Vol. 58: 240-242.

Porter, A. L., Cohen, A. S., Roessner, J. D. and Perreault, M. 2007. Measuring researcher interdisciplinarity. *Scientometrics*, Vol. 72, no. 1: 117-147.

Rip A. and Courtial J.-P. 1984. Co-word maps of biotechnology: An example of cognitive scientometrics, *Scientometrics*, Vol. 6, no. 6: 381-400.

Rorving, M. 1999. Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science and Technology*, Vol. 50, no. 8: 639-651.

Saad, S. M. and Kamarudin, S. S. 2013. Comparative analysis of similarity measures for sentence level semantic measurement of text. *Proceedings 2013 IEEE International Conference on Control System, Computing and Engineering*, ICCSCE 2013, 90-94.

Salton G. and McGill, M.J. 1983. *Introduction to modern information retrieval.* Auckland, New Zealand, McGraw-Hill.

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

Schneider, J. W. and Borlund, P. 2007b. Matrix comparison, Part 2 Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics, *Journal of the American Society for Information Science and Technology*, Vol. 58, no. 11: 1596-1609.

Shoaib, M., Daud, A. and Khiyal, M. S. H. 2015. Improving similarity measures for publications with special focus on author name disambiguation. *Arabian Journal for Science and Engineering*, Vol. 40, no. 6: 1591-1605.

Singh, J., Singh, P. and Chaba, Y. 2014. A study of similarity functions used in textual information retrieval in Wide Area Networks. *International Journal of Computer Science and Information Technologies*, Vol. 5, no. 6: 7880-7884.

Small H. 1973. Co-citation in the scientific literature: A new measure of the relationship betwen documents, *Journal of the American Society for Information Science and Technology*, Vol. 24, no. 4: 265-269.

Sorenson T. 1948. a method of establishing groups of equal amplitude in plant sociology based on similarity of spicies content and its application to analyse the vegetation on Danish commons, *Biologiske krifter,* Vol. 5, no. 4: 1-34.

Sorkhi, M. and Hashemi, S. 2015. Effective team formation in collaboration networks using vertex and proficiency similarity measures. *AI Communications, Vol.* 28, no. 4: 637-654.

Sternitzke, C. and Bergmann, I. 2009. Similarity measures for document mapping: A comparative study on the level of an individual scientist, *Scientometrics*, Vol. 78, no. 1: 113-130.

Stvilia, B., Al-Faraj, A. and Yi, Y. J. 2009. Issues of cross-contextual information quality evaluation-The case of Arabic, English, and Korean Wikipedias*. Library & Information Science Research*, Vol. 3, no. 14: 232-239.

Subhashini, R. and Kumar, V. J. S. 2010. Evaluating the performance of similarity measures used in document clustering and information retrieval. *Proceedings 1st International Conference on Integrated Intelligent Computing, ICIIC*, 27-31.

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

Thada, V. and Jaglan, D. V. 2013. Comparison of Jaccard, Dice, Cosine Similarity Coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 202-205.

Thijs, B., Zhang, L. and Glänzel, W. 2013. Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes. *In 14th International Conference on Scientometrics and Informetrics*, 15-19.

Thijs, B. and Glänzel, W. 2010. A structural analysis of collaboration between European research institutes. *Research Evaluation*, Vol. 19, no. 1: 55-65.

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.

Van Eck, N. J. and Waltman, L. 2009. How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of The American Society for Information Science and Technology*, Vol. 60, no. 8: 1635-1651.

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.

Wagner, C. S. and Leydesdorff, L. 2003. Mapping global science using international co-authorships: a comparison of 1990 and *2000. Proceedings of Ninth International Conference on Scientometrics and Informetrics , ISSI 2003, 330-340.*

Wang, X. and Sukthankar, G. 2013. Multi-label relational neighbor classification using social context features. Proceedings of the *19th ACM SIGKDD international conference on Knowledge discovery and data mining*. *ACM*, 464-472.

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.

Zupic I and Cater T. 2015. Bibliometric methods in management and organization, *Organizational Research Methods*, Vol. 18, no. 3: 429-472.