XI Encontro Nacional de Pesquisa em Ciência da Informação
Inovação e inclusão social: questões contemporâneas da informação
Rio de Janeiro, 25 a 28 de outubro de 2010

GT 7: Produção e Comunicação da Informação em CT&I
Modalidade de apresentação: Comunicação Oral

ANÁLISE DE COCITAÇÃO DE AUTORES: QUESTÕES METODOLÓGICAS

Ana Maria Mattos
Universidade Federal do Rio Grande do Sul
Eduardo Wense Dias
Universidade Federal de Minas Gerais

RESUMO: Descreve-se a origem, a evolução e os debates atuais em torno da análise de cocitação de autores, visando estimular as discussões entre os bibliometristas brasileiros sobre os procedimentos metodológicos envolvidos na seleção, tabulação, interpretação e validação dos dados deste método. Arrolam-se as discussões na literatura acerca da obtenção dos dados, da definição das unidades de análises, da geração e transformação de dados brutos e matrizes de proximidade bem como a escolha da medida de proximidade. Conclui-se que o assunto apresenta grande potencial de pesquisa como demonstra o intenso debate na literatura internacional.

Palavras-chave: Ciência da Informação. Bibliometria. Análise de cocitação de autores.

1 INTRODUÇÃO

Cada indivíduo contribui para o conhecimento construindo sobre o que os outros já edificaram. Neste processo, citações são objetos que nos permitem conectar as contribuições que foram publicadas, desde que se utilize os métodos bibliométricos, que podemos separar em duas categorias: (i) a citação, que agrega os dados pela contagem de frequência de um documento, ou conjunto de documentos, sem se considerar suas ligações intelectuais e que utiliza basicamente a estatística descritiva; e (ii) a cocitação, que agrega os dados para análise em uma ordenação multidimensional, utilizando a análise multivariada de dados, procurando identificar as ligações intelectuais entre os documentos (OKUBO, 1997).

Desde sua introdução, ao final da década de 1960, a análise de cocitação de autores (ACA) tornou-se uma técnica popular. No entanto, recentemente emergiu um debate sobre vários dos procedimentos metodológicos em ACA. Entende-se que este debate deve integrar a agenda de pesquisas dos bibliometristas brasileiros, justificando-se assim, esta revisão de literatura. Objetiva-se descrever a origem, a evolução e as questões atuais em ACA, pretendendo motivar as discussões acerca dos procedimentos metodológicos envolvidos na seleção, tabulação, interpretação e validação dos dados em ACA.

2 ORIGEM E EVOLUÇÃO

A ACA encontra-se entre os métodos bibliométricos, e segundo Eom (2009), a idéia original da análise de cocitação de múltiplos autores e documentos foi desenvolvida por Rosengren em 1968 com o nome de análise de començões (*co mentions analysis*).

Na década de 1970 houve um aumento no número de estudos bibliométricos, consagração do *Science Citation Index (SCI)*, fundado em 1963 por Garfield, que abriu caminho para todos aqueles que procuravam medir a ciência materializada em publicações. Na mesma época, Small desenvolveu técnicas de mapeamento de cocitação no *Institute of Scientific Information*, entidade responsável pela publicação do *SCI*. Destacam-se ainda os trabalhos realizados no *College of Information Studies* da *Drexel*

University que ajudaram a criar o interesse na ACA. Nascia assim uma nova disciplina, a Ciência da Ciência (MCCAIN, 1990; NOYONS, 1999; OKUBO, 1997; PRICE, 1976).

Fortes críticas sobre os pressupostos em relação ao comportamento da citação surgiram na década de 1990. Porém, paralelamente emergiram dois novos campos de investigação: a webometria, que estuda o fenômeno *world wide web*; e a cibermetria, que pesquisa todas as aplicações da Internet. Nestes ambientes realiza-se a análise de *colinks* na *web* (*web colink analysis*) (EOM, 2003). Björneborn e Ingwersen (2004), além de introduzirem a estrutura básica para a análise webométrica, proporcionaram uma visão ampla das relações entre a Ciência da Informação e a Infor/Biblio/Ciencio/Ciber e Webometria (Figura 1). Por causa das críticas, ou apesar delas, a qualidade das pesquisas Bibliométricas aumentou, enquanto sua quantidade, diminuiu (EOM, 2003).

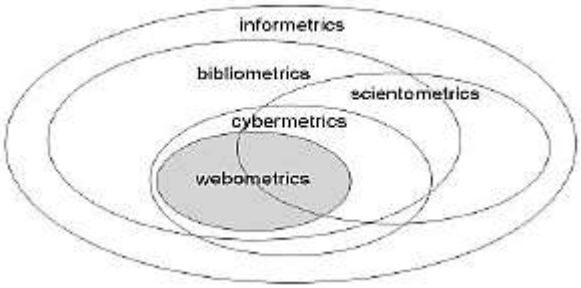

Figura 1: Relações entre a Ciência da Informação e a infor/biblio/ciencio/ciber/webo/metria
Fonte: Björneborn e Ingwersen (2004, p. 1217).

Incluem-se entre as limitações dos estudos de citação a qualidade do documento citado, os erros originários da prática de referência desleixada, os autores que trabalham em colaboração, o uso da auto citação e as diferentes motivações para citar; que podem criar um viés no conjunto de dados. Além disso, encontram-se erros de inclusão, omissão ou ambos, pois um documento pode citar obras que não foram utilizadas ou não citar obras que estão sendo invocadas (SMITH, 1981).

Evidentemente todos estes problemas também podem influenciar a análise de cocitação. Dentro de uma especialidade, um grupo de cientistas pode responder por uma quantidade desproporcional de contatos inter pares, citações e produtividade e os autores podem evitar citar documentos consagrados, uma vez que os consideram universalmente aceitos ou conhecidos. Por outro lado, documentos podem vir a ser citados por motivos diferentes: um documento pioneiro pode ser citado no futuro para lastrear a metodologia ou as conclusões (SMITH, 1981). Inclui-se também entre as limitações dos estudos de cocitação sua dependência da comunicação formal. Felizmente existem incentivos para o

registro das percepções compartilhadas *inter alia,* pois apenas documentos publicados e acessíveis se tornarão parte da base de uma disciplina (OKUBO, 1997).

Mas, apesar das limitações, as citações configuram-se como importantes documentos para investigar o estado atual de um campo do conhecimento. São abundantes e não requerem interação com os sujeitos estudados, que já tomaram suas decisões *a priori*, quando escreveram o documento publicado. As citações encontram-se universalmente presentes nas publicações e proporcionam uma visão parcial sobre o que é considerado como necessário para entender ou criar o conhecimento (SMITH, 1981).

Pode-se afirmar que, em termos de conhecimento, um documento científico é um conglomerado de unidades menores, as citações, relacionadas entre uma parte ou a totalidade dos documentos anteriores e uma parte ou a totalidade do documento atual. Para Smith (1981), aceitar as limitações inerentes aos estudos de citação nos permite teorizar que a atividade de citação é legível. Na prática, apesar das limitações, as tendências e padrões de comunicação podem ser observados analisando-se os dados agregados das cocitações e considera-se que estes estudos são representações válidas da estrutura intelectual (EOM, 2009; OKUBO, 1997).

De fato, citações funcionam como símbolos de conceitos e, frequentemente, a cocitação revela interesses comuns sobre assuntos que serão esclarecidos por esses símbolos de conceitos. Com a análise de cocitação mede-se a semelhança percebida da articulação conceitual ou da relação cognitiva entre dois documentos cocitados (Figura 2).

A medida da relação entre documentos que geram e recebem citação em uma ordenação multidimensional possibilita a análise de acoplamento bibliográfico (Figura 3) desenvolvida por Kessler em 1963. Destaca-se que o acoplamento é extrínseco ao trabalho original, eles podem não ter uma conexão intrínseca, mas seus respectivos conteúdos podem ser necessários para a investigação em curso (EOM, 2009; GARFIELD, 2001; GLÄNZEL, 2003; SMITH, 1981).

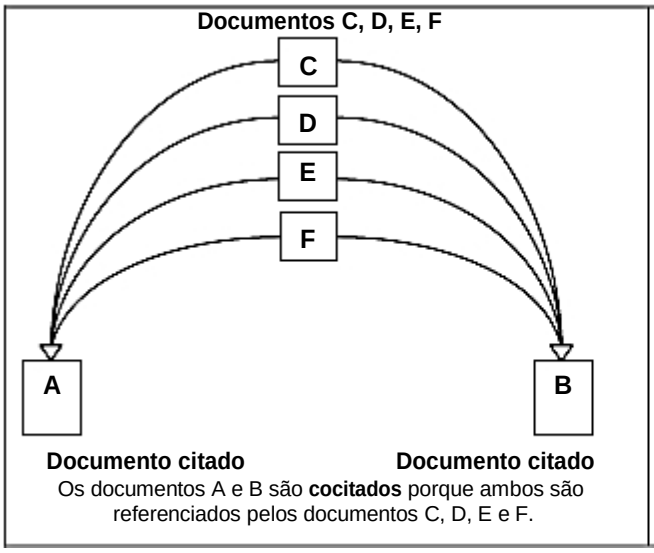

Figura 2: Cocitação

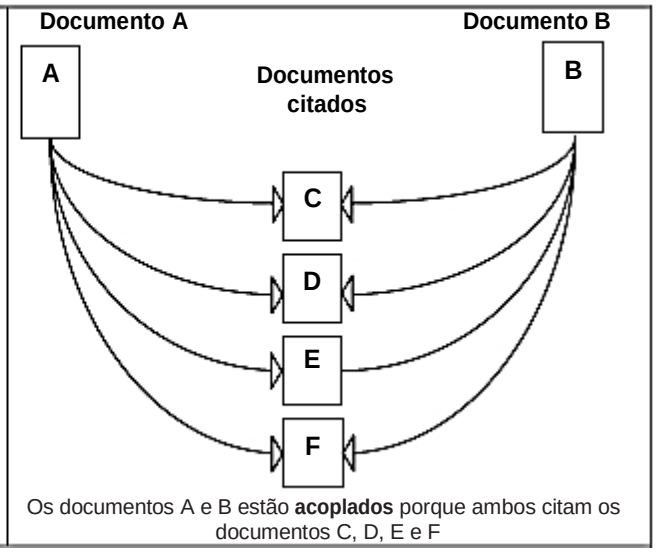

Figura 3: Acoplamento bibliográfico

Fonte: Adaptados de Garfield (2001, p. 3)

Por outro lado, na análise de coocorrência de palavras identifica-se quando certos termos e palavras chave aparecem juntos em um grupo de documentos. Busca-se descrever as relações entre as agendas de pesquisa nas quais os temas estão sendo abordados em conjunto. As limitações incluem a natureza relativamente subjetiva do conjunto de dados a ser analisado, uma vez que palavras chave são, com frequência, conservadoramente atribuídas e podem não representar fielmente o assunto do documento (EOM, 2009).

Utiliza-se a análise de correlação entre artigos científicos e patentes para medir o nível de macro cooperação através da medida de correlação de produtividade entre artigos de periódicos e patentes. Aqui a análise de coautoria pode fornecer informações sobre as tendências na cooperação formal entre pesquisadores. Entre as limitações incluem-se as dificuldades de identificar a instituição ou país de um determinado autor. A coautoria também apresenta problemas quanto ao crédito aos autores de um documento. Pode-se dividir a autoria entre todos ou atribuir mais peso ao autor principal, entretanto, encontra-se dificuldade em fazer essa definição (EOM, 2009; GARFIELD, 2001; GLÄNZEL, 2003).

2.1 A ANÁLISE DE COCITAÇÃO DE AUTORES (ACA)

A ACA é um "conjunto de dados compilados, analisados e representados graficamente, que podem ser usados para produzir mapas empíricos dos autores importantes em várias áreas de pesquisa" (MCCAIN, 1990, p. 433). Em geral, pesquisadores com problemas de pesquisa semelhantes citam fontes informacionais similares, e a ACA tende a mostrar a estrutura social deste domínio de investigação. O desenho desta estrutura social representado nas citações simultâneas de dois autores em um mesmo documento constitui-se nos dados iniciais através dos quais as inferências em ACA serão feitas (NOYONS, 1999).

Nos anos 1960, pesquisadores do *College of Information Studies* da *Drexel University* fizeram da ACA um método de pesquisa popular. Nos anos 1980, White e Griffith (1981), introduziram uma abordagem mais geral, conhecida como Abordagem Drexel, para identificar, analisar e traçar a estrutura intelectual de uma disciplina acadêmica. A evolução e as diferenças, nos passos necessários para a realização de ACA podem ser visualizadas na Figura 4: à esquerda os passos segundo McCain (1990) e à direita os passos de acordo com Eom (2009).

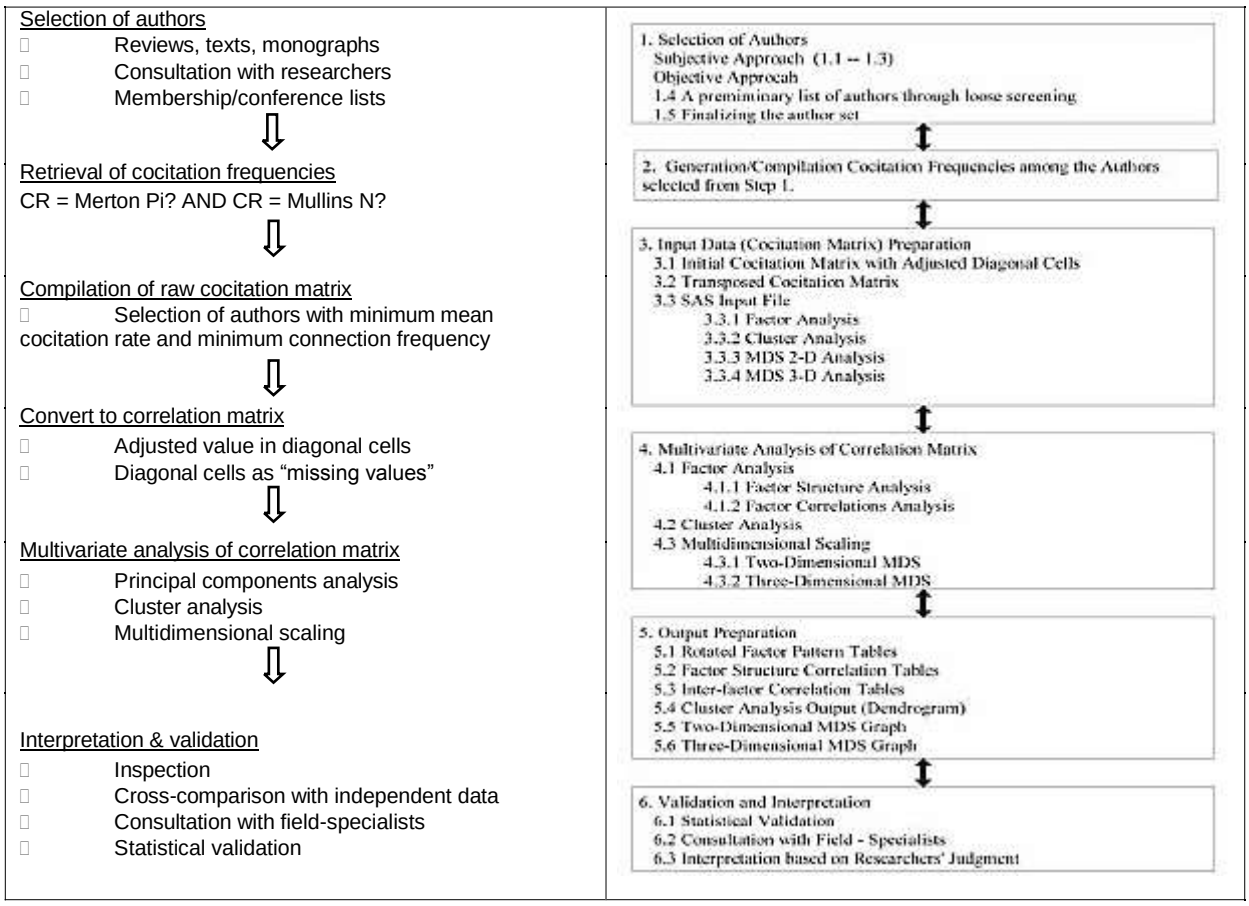

Figura 4: Seis passos para a realização da ACA: 1990 e 2009
Fonte: McCain (1990, p. 434) e Eom (2009, p. 146)

Segundo Smith (1981), examina-se o documento de citação no nível macro, de um título para outro título, e isso pode repercutir em um erro grave quando dois autores são associados por cocitação. Argumenta ainda, que a ACA não oferece temas de pesquisa porque, com freqüência, dois autores podem ser citados juntos sem necessariamente apresentar conteúdo semelhante.

Por outro lado, Eom (2009) assevera que a ACA baseia-se na suposição que os autores que foram citados juntos tem estreitas relações, que as citações bibliográficas são um substituto aceitável para a real influência e que a maior frequência relativa de cocitação indica que os autores estão relacionados com o conteúdo da investigação. Em suma, a ACA pode sim, revelar a estrutura intelectual de um domínio de investigação.

Contudo, McIntire (2006) conclui que a afirmação emitida por Eom (2003) não define totalmente o conceito de "relacionadas com o conteúdo da investigação". Segundo McIntire (2006), a cocitação indicaria que as pesquisas que estão sendo conduzidas

naquele documento avançam pela análise de uma parte dos documentos citados. Portanto, os autores cocitados não são arbitrariamente relacionados com o conteúdo daquela investigação, mas são considerados parte da base de conhecimentos necessários naquele determinado momento para avançar em uma frente de investigação.

3 AS QUESTÕES ATUAIS EM ACA

O foco do debate atual encontra-se basicamente em quatro questões metodológicas relacionadas à Abordagem Drexel: (i) a obtenção dos dados; (ii) a definição das unidades de análises, ou seja, quais autores cocitados se devem incluir nos estudos; (iii) a geração e transformação de dados e matrizes de proximidade e como isso influencia o agrupamento e mapeamento dos autores co-citados; e (iv) a escolha da medida de proximidade, ou seja, sua influência sobre o agrupamento e mapeamento dos autores cocitados.

3.1 A OBTENÇÃO DOS DADOS

A principal fonte de dados para ACA são as bases de dados comerciais ou institucionais. Podemos citar o *Chemical Abstracts, Compendex, Inspec, Pascal, IEEE, Scopus, Medline, Latindex, Scielo* ou a *Web of Science* (WoS), esta a mais utilizada. As práticas de indexação e os comandos de busca de informação disponíveis nas respectivas *interfaces* de consulta da maioria destas bases limitam a recuperação das frequências de cocitação ao primeiro autor cocitado.

Como se observa na Figura 4, o primeiro passo metodológico em ACA é a seleção dos autores, seja em uma abordagem subjetiva (*top-down*) ou objetiva (*bottom-up*). Na abordagem subjetiva utiliza-se bases de dados comerciais ou institucionais tendo-se como principal desvantagem a dificuldade em identificar os estudiosos emergentes. Na abordagem objetiva selecionam-se os autores em um banco de dados customizado especialmente desenvolvido para uma análise mais aprofundada. Esta abordagem torna os estudiosos emergentes mais suscetíveis de serem selecionados e facilita a

identificação das disciplinas (EOM, 2009).

A maioria das bases de dados comerciais ou institucionais recupera apenas o primeiro autor do documento, independente do número de autores, na contagem de cocitação e esta tem sido uma crítica na sua utilização. Nestas bases também se pode encontrar dificuldades relacionadas à cobertura (documentos, assunto, temporalidade, idioma, etc.), a alterações em nomes próprios ou homógrafos (mesmo sobrenome e iniciais), aos sinônimos, aos erros de escrita e à padronização dos nomes de instituições e títulos de periódicos (EOM, 2009).

Com todas estas limitações, os bibliometristas se obrigam a longas conferências dos dados importados diretamente das bases de dados comerciais ou institucionais, mesmo quando utilizam programas que importam os dados como o BibExcel ou o EndNote, visando padronizá-los e dar-lhes consistência. Naturalmente estas dificuldades conduziram os investigadores em ACA para a construção de bases de dados customizadas como sugerido por Eom (2009), Glänzel (2003), McIntire (2006) ou Zhao (2006), ou quando é possível, como no estudo de Schneider, Larsen e Ingwersen (2009), utilizam-se documentos estruturados em XML que permitem a construção de índices de citação com todos os autores.

Em resumo, quanto à captura dos dados para ACA, a discussão metodológica gira em torno do acesso limitado por restrições que podem variar desde aquelas inerentes à constituição das bases de dados comerciais ou institucionais, que não foram elaboradas para este fim, até restrições econômicas de acesso por parte do pesquisador ou instituição de pesquisa. A opção pelo uso das bases de dados comerciais ou institucionais pode limitar o estudo como ocorreu com Persson (1981), que ignorou mais de 90% das referências feitas a trabalhos não indexados pelo *ISI* como documentos fonte em suas análises e conclusões.

Nesta perspectiva, ao se construir uma base de dados customizada pode-se utilizar as informações dos bancos de dados comerciais ou institucionais para saber quais documentos precisamos reunir. Os documentos indisponíveis ao pesquisador por limitações diversas, como por exemplo, o período de assinatura de um título de periódico; podem ser obtidos em outros locais (Internet, estante da biblioteca ou comutação bibliográfica). O acesso livre, os repositórios institucionais eletrônicos e o documento eletrônico tendem a tornar a construção de uma base de dados customizada mais

acessível. Esta opção pode ser válida principalmente quando se examina a coleção de um título de periódico específico como fez McIntire (2006).

3.2 A DEFINIÇÃO DAS UNIDADES DE ANÁLISES

O problema da definição de coautoria iniciou-se com os estudos de cocitação (LINDSEY, 1980). Pesquisadores em ACA definem autor como um corpo de escritos ou conjunto de contribuições de uma pessoa (*author's oeuvres*). Pessoa pode se referir a qualquer um dos autores do documento (EOM, 2009; SCHNEIDER; LARSEN; INGWERSEN, 2009). Ainda hoje, o "autor cocitado" não foi definido e as imprecisões podem induzir a erros nas contagens dos valores da matriz. Visualiza-se no Quadro 1 os métodos de contagem de cocitação: conta-se só o primeiro autor; todos os autores excluindo-se as coautorias ou todos os autores incluindo-se as coautorias (ROUSSEAU; ZUCCALA, 2004; ZHAO, 2006).

<table>
<tr><th rowspan="2">Os autores Lee a Hair sendo cocitados em uma lista de referências</th><th colspan="3">MÉTODOS DE CONTAGEM DE COCITAÇÃO</th></tr>
<tr><th>Só o primeiro autor (pure first-author co-citations)</th><th>Todos os autores (general co-citations)</th><th>Todos os autores inclusive (pure co citation)</th></tr>
<tr><td>Artigo 1</td><td></td><td></td><td></td></tr>
<tr><td>…</td><td rowspan="4">Não</td><td rowspan="4">Sim</td><td rowspan="4">Sim</td></tr>
<tr><td>Lee, K. (2000). XSLT and XML. XML Journal</td></tr>
<tr><td>Lee, K., & Hair, S. (1998). RDF and OWL. The Semantic Web</td></tr>
<tr><td>...</td></tr>
<tr><td>Artigo 2</td><td></td><td></td><td></td></tr>
<tr><td>...</td><td rowspan="4">Sim</td><td rowspan="4">Sim</td><td rowspan="4">Sim</td></tr>
<tr><td>Lee, K. (2000). XSLT and XML. XML Journal</td></tr>
<tr><td>Hair, S., & Lee, K. (2003). RDF-Schema. Ontologia</td></tr>
<tr><td>…</td></tr>
<tr><td>Artigo 3</td><td></td><td></td><td></td></tr>
<tr><td>…</td><td rowspan="3">Não</td><td rowspan="3">Sim</td><td rowspan="3">Não</td></tr>
<tr><td>Lee, K., & Hair, S. (1998). RDF and OWL. The Semantic Web</td></tr>
<tr><td>…</td></tr>
<tr><td>Número total de cocitação entre Lee e Hair</td><td>1</td><td>3</td><td>2</td></tr>
</table>

Quadro 1: Métodos de contagem de cocitação
Fonte: Adaptado de Zhao (2006) e Rousseau e Zuccala (2004)

Na Figura 5 visualizam-se quando os métodos de contagem de cocitação (círculos externos) incorporaram outros métodos (círculos internos), aumentando a contagem das

cocitações de autores, sendo que a contagem de cocitação/coautoria engloba todas as definições (ROUSSEAU; ZUCCALA, 2004).

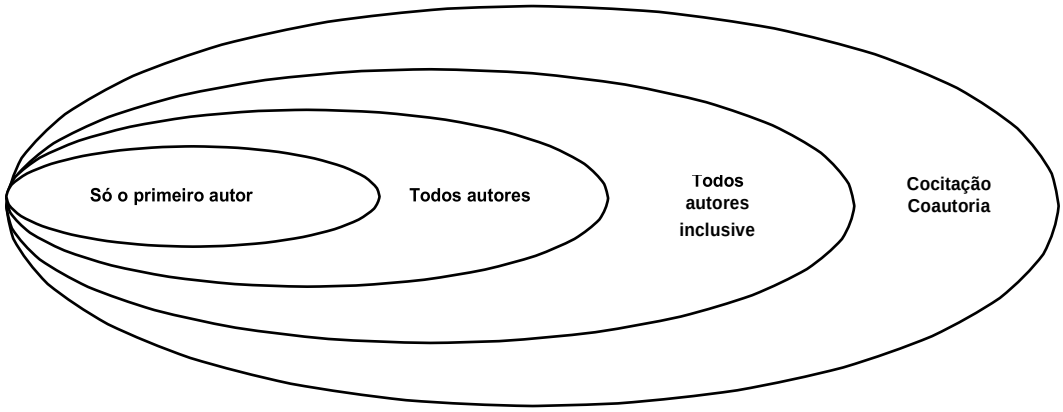

Figura 5: Classificação hierárquica das definições de cocitação
Fonte: Adaptado de Rousseau e Zuccala (2004, p. 516)

Persson (2001) produziu o primeiro estudo empírico em ACA que compara as potenciais diferenças com base no primeiro autor e em todos os autores inclusive. Apesar da limitação na obtenção dos dados, o estudo demonstra que se deve preferir a contagem de todos os autores para visualizar a estrutura dos campos de pesquisa porque capta a maioria dos pesquisadores influentes. Porém, a captura da estrutura das especialidades tende a ser a mesma em ambos os métodos.

Rousseau e Zuccala (2004) sugerem que, independente da classificação do autor na autoria geral, sua contribuição pode ser substancial para o desenvolvimento de uma área de investigação. Assim, ACA com todos os autores inclusive apresenta um retrato mais preciso da contribuição do autor individual para a área.

Zhao (2006) utilizou a Abordagem Drexel e os resultados indicaram que a inclusão de todos os autores inclusive, que ele limitou aos cinco primeiros, cria grupos mais coerentes e, portanto, consideravelmente mais claros para se identificar e interpretar. Os resultados indicaram também que quando o mesmo número de autores *top* é selecionado e analisado, todos os autores inclusive podem levar à identificação de menos especialidades em um campo de investigação em relação a ACA que utiliza só o primeiro autor.

Jarneving (2008) apresenta um método experimental de cálculo da frequência de cocitação só com o primeiro autor comparando-o com a Abordagem Drexel. Ele verificou que seu método experimental descreveu mais detalhadamente a estrutura da

especialidade; que um mapeamento completo e em profundidade exige que todos os autores inclusive estejam presentes na análise e que uma parte considerável dos autores mapeadas no estudo ocorrem com frequência como coautores, de acordo com observações anteriores relatadas por Persson (2001) e Zhao (2006).

Destaca-se que Jarneving (2008) assim como Rousseau e Zuccala (2004) tem o foco da investigação no impacto dos autores sobre o artigo no qual são citados, ao contrário de outras pesquisas recentes sobre métodos de contagem como Persson (2001) ou Zhao (2006), onde a questão principal tem sido a definição de como se deve contar os autores.

Schneider, Larsen e Ingwersen (2009), contribuíram para debate comparando análises só com o primeiro autor e com todos os autores inclusive, e confirmaram que estudos de ACA com todos os autores inclusive produzem agrupamentos mais coerentes da estrutura dos campos de pesquisa. Contudo, não confirmaram que os estudos de ACA somente com o primeiro autor identificam mais especialidades.

Os resultados encontrados por Eom (2009) apóiam os achados de Persson (2001) no que diz respeito a adoção da definição de todos os autores inclusive para a melhor identificação dos pesquisadores influentes em um campo. Porém, quanto à identificação das sub especialidades ser independente da definição de coautoria adotada na investigação, os resultados de Eom (2009) não confirmam Persson (2001) e Zhao (2006).

O cenário encontra-se indefinido. Torna-se imperativo que as pesquisas futuras utilizem os mesmos pressupostos e critérios no que diz respeito à seleção dos autores em ACA, pois se não padronizarmos os métodos de contagem de cocitação os resultados serão incomparáveis e perderão o significado.

3.3 A GERAÇÃO E TRANSFORMAÇÃO DE DADOS E MATRIZES DE PROXIMIDADE

Uma matriz consiste em um quadrado com o conjunto de membros organizados na mesma ordem nas linhas e colunas, e cada célula, em seguida, registra a transação de um membro com o outro. A dificuldade essencial na análise de todas as matrizes quadradas deste tipo deve-se ao fato de que as auto transações, dadas na diagonal, são muitas vezes indefinidas.

O importante papel desempenhado pelas matrizes na análise de cocitação tem recebido atenção de Leydesdorff e Vaughan (2006). Eles demonstram a diferença fundamental entre dados assimétricos (ocorrências) que geram matrizes ***n×m***; e a proximidade simétrica (coocorrências) que geram matrizes ***n×n***. Argumentam que matrizes simétricas de contagens de coocorrência são matrizes de proximidade e devem ser tratadas como tal.

Na abordagem Drexel para ACA as contagens de cocitações são inseridas em uma matriz simétrica quadrada pré construída, considerada uma matriz de proximidade. Por outro lado, na tradicional análise multivariada, as matrizes de proximidade são derivadas dos dados de uma matriz assimétrica resultante das variáveis *versus* observações No entanto, ao se aplicar a análise fatorial como ferramenta exploratória para identificar as estruturas latentes e agrupamentos intelectuais necessita-se de uma matriz de proximidade simétrica de covariância ou coeficiente de correlação (SCHNEIDER; LARSEN; INGWERSEN, 2009).

Quanto à geração da matriz, Schneider, Larsen e Ingwersen (2009) testaram duas diferentes abordagens, a Drexel e a multivariada convencional, utilizando em ambas, duas definições de autor: só o primeiro e todos os autores inclusive. As duas definições de autoria deram início a uma matriz simétrica de proximidade ***n×n***, que corresponde à Abordagem Drexel e outra matriz de dados assimétrica ***m×n***, que corresponde à análise convencional de dados multivariados. Os valores das diagonais foram tratados como dados perdidos.

Fundem-se aqui a questão dos métodos de contagem de cocitação com a geração da matriz. Ocorre que quando se interpreta os resultados em análise multivariada de dados, o importante é o significado de cada fator. Assim, contando-se apenas o primeiro autor diminuem-se as frequências de cocitação e o conjunto final de autores definidos para ACA também deve ser diminuído mediante algum critério, podendo-se utilizar a taxa média de cocitação, a cocitação com pelo menos um terço do conjunto de todos os autores ou restringindo-se o conjunto final de autores para os 20% que recebem o maior número de citações (EOM, 2009).

Tal matriz não se encontra disponível na Abordagem Drexel, onde uma matriz de proximidade simétrica ***n×n*** é gerada pela contagem dos pares de autores cocitados. O resultado é concebido em um processo heterodoxo, onde se obtém a matriz de

proximidade na contagem direta das cocitações transformadas em uma matriz de proximidade adicional derivada dos coeficientes de correlação dos primeiros autores cocitados entre os autores incluídos. Note-se que uma transformação linear de uma matriz de proximidade simétrica não é simples, além disso, há problemas em relação a algumas matrizes de transação (PRICE, 1981; SCHNEIDER; LARSEN; INGWERSEN, 2009).

As transformações na matriz causam um problema fundamental em relação à interpretação e ao tratamento dos valores de contagens das cocitações não processadas na diagonal da matriz de proximidade original; apesar de White (2003) afirmar que o tratamento dos valores diagonais é um problema menor. Isso pode ser verdade, dependendo do conjunto de dados que se dispõe, mas o problema genérico surge onde a matriz de proximidade obtida diretamente é tratada como uma matriz de dados para fins de transformação. Os problemas finais poderiam ser evitados aplicando-se a abordagem convencional multivariada para a geração da matriz e de sua transformação (SCHNEIDER; LARSEN; INGWERSEN, 2009).

Para viabilizar a análise estatística multivariada, encontram-se na literatura oito abordagens em ACA para gerar o valor das células na diagonal da matriz, a saber:

a) realizar-se a contagem de cocitação pura;
b) considerá-las dados perdidos;
c) utilizar-se a média da contagem de cocitação para cada autor;
d) atribuir-lhes valor igual a zero;
e) designar-lhes as maiores contagens de cocitação de fora da diagonal para cada autor;
f) ajustar-lhes os valores com os valores de fora da diagonal;
g) designar-lhes a contagem de cocitação bruta; e
h) reconstruir os elementos da diagonal.

Em ACA os valores que correspondem a contagem de cocitação entre o autor contra si próprio, excluindo-se as auto citações, são chamados de cocitação pura (AHLGREN, JARNEVING, ROUSSEAU, 2003). O problema é que a contagem de co-citação pura é um processo extremamente moroso e complexo de ser programado em computador. É difícil, por exemplo, desenvolver um sistema inteligente o suficiente para distinguir uma citação de uma auto citação, tanto em bases de dados comerciais quanto

em bases de dados customizadas como as desenvolvidas por Eom (2009) ou McIntyre (2006).

Na abordagem que considera os valores na diagonal da matriz como dados perdidos, deve-se substituí-los por um único ponto (**.**). Nesta abordagem, a análise fatorial não pode ser aplicada ao conjunto de dados, antes eles devem ser processados para criar uma matriz de correlação (MCCAIN, 1990). As abordagens que utilizam a média da contagem de cocitação para cada autor, zero ou as maiores contagens de cocitação de fora da diagonal para cada autor em substituição aos valores da diagonal da matriz, fazem exatamente o que expressam, ou seja, preenchem a diagonal com os respectivos valores enunciados (EOM, 2009).

Após algumas reflexões sobre a maneira pela qual tais valores podem ser distribuídos em uma matriz, White e Griffith, (1981) decidiram que, tomando-se os três valores maiores de fora da diagonal dividido por dois, é possível gerar valores para as diagonais que harmonizariam as pontuações mais próximas da distribuição, indicando de forma geral, a importância relativa de um determinado autor dentro do campo.

Na contagem de cocitação bruta, experimentada por Eom (2009), considera-se um o número total de cocitações que contém uma ou mais contribuições do autor (autoria ou coautoria) por documento, e não o total exato de vezes que ele é citado. A Figura 6 ajuda a entender a utilização desta abordagem. No exemplo, a contagem de cocitação bruta é dois e não três, pois nas referências do documento 1, Eom é contado como um.

Referências do documento 1:

EOM, S. B. A survey of operational expert systems in business (1980-1993). **Interfaces**, v. 26, n. 5, p. 50-70, 1996.

EOM, S. B.; LEE, S. M.; KIM, J. K. The intellectual structure of decision support systems (1971-1989). **Decision Support Systems**, v. 10, n.1, p. 19-35, 1993.

FARHOOMAND, A. F. Scientific progress of management information systems. **Data Base**, v. 18, n.4, p. 48-56, 1987.

Referências do documento 2:

SMITH, K.; THOMAS, T. The k-procedure. **The Gamma Journal**, 1992

THOMAS, T.; SMITH, K. More details about the k-procedure. **The Gamma Journal**, 1992.

MIN, H.; EOM, S. B. An integrated decision support system for global logistics. **International Journal of Physical Distribution and Logistics Management**, v. 24, n. 1, p. 29-39, 1994.

Figura 6: Exemplo de contagem de cocitação bruta
Fonte: Adaptado de Eom (2009, p. 97)

Pode-se também reconstruir dos elementos na diagonal da matriz, objetivando preenchê-la a partir de cálculos que envolvem multiplicações, divisões e extração de

raízes dos coeficientes de linhas e colunas para finalmente se proceder à análise da matriz resultante, como um produto de coeficientes de linhas e colunas na forma usual (PRICE, 1981).

Eom (2009) realizou o primeiro estudo que analisa o impacto de seis das diferentes abordagens para designar valores na diagonal da matriz sobre os resultados em ACA. Sua principal conclusão é que o valor verdadeiro produzido pela abordagem de contagem de cocitação pura pode ser menor do que o sugerido por alguns pesquisadores. Por outro lado, as abordagens alternativas, que sugerem valores mais elevados na diagonal, resultaram soluções fatoriais que melhor explicam a variância total.

Os resultados também sugerem que, caso a contagem de cocitação pura não se encontre disponível, as melhores alternativas para o preenchimento da diagonal, na ordem da maior variância total explicada são: considerá-las dados perdidos, utilizar-se a média da contagem de cocitação e designar-lhes as maiores contagens de cocitação de fora da diagonal. Dois métodos produziram os piores resultados: atribuir-lhes zero e designar-lhes a contagem de cocitação bruta (EOM, 2009).

Percebe-se que serão necessários muitos estudos comparativos entre as diferentes abordagens de geração de matriz, como o de Schneider, Larsen e Ingwersen, (2009) e entre as diferentes abordagens de preenchimento da diagonal, como o elaborado por Eom (2009). Também neste caso se deve perseguir a padronização do método visando atribuir significado às pesquisas com o objetivo de torná-las comparáveis.

3.4 A ESCOLHA DA MEDIDA DE PROXIMIDADE

Analisam-se os dados de ACA em *softwares* estatísticos como o *Statistical Analysis System* (*SAS*) e o *Statistical Package for the Social Sciences* (*SPSS*), que são os mais utilizados, ou o *IDAMS,* disponibilizado gratuitamente pela UNESCO. Em ACA, a matriz de cocitação bruta é compilada normalizando-se a medida de similaridade para então aplicarmos três técnicas de análise multivariada: análise fatorial, análise de cluster e escalonamento multidimensional, com o objetivo de agrupar/classificar todas as variáveis em diversos subgrupos, com base nas estruturas subjacentes e nas características e/ou atributos comuns (EOM, 2009).

É importante observar que embora estas técnicas procurem sintetizar e simplificar um grande número de variáveis existe importantes diferenças em seus resultados. Os fatores na matriz de cargas fatoriais correspondem às especialidades de investigação ou escolas de pensamento. A análise de cluster reduz os dados em agrupamento de entidades em vários grupos semelhantes no que diz respeito a alguns critérios de seleção pré definidos, gerando o histórico dos agrupamentos e o dendograma. Como resultado do escalonamento multidimensional pode-se produzir imagens tridimensionais da configuração de cada autor em espaços multidimensionais revelando o panorama geral das relações central e periférica nas especialidades.

O primeiro passo no mapeamento ou agrupamento de autores cocitados é a conversão dos dados brutos da matriz em uma matriz de valores de proximidade, que indicam a relativa semelhança ou dessemelhança de pares de autores. Em vários estudos ACA, a correlação de *Pearson (r)* é utilizada como medida de similaridade, a preferida na tradicional Abordagem Drexel. Os *softwares* estatísticos podem gerá-la e a escolha pode depender das análises a serem executadas e as limitações dos programas utilizados (MCCAIN, 1990).

O debate sobre o uso da correlação de *Pearson (r)* teve início com Ahlgren, Jarneving, Rousseau (2003), que a questionaram e criticaram como medida de similaridade padrão adotada desde a publicação do artigo de McCain (1990). Em ACA, a matriz de cocitação bruta deve ser compilada para se normalizar a medida de similaridade para então se realizar a análise fatorial, a análise de cluster e o escalonamento multidimensional. O ponto focal da crítica de Ahlgren, Jarneving, Rousseau (2003) em relação ao uso da correlação *Pearson (r)* é a sua sensibilidade aos elementos de valor zero do vetor, expandindo o conjunto de autores a ser analisado, adicionando mais autores que nunca foram cocitados ao conjunto original de autores e, portanto, acrescentando zeros à matriz de cocitação.

Em resposta, White (2003) afirma que o problema é irrealista na tradicional abordagem de ACA e que o uso da correlação *Pearson (r)* é suficiente, pois produz mapas e clusters semelhantes aos produzidos pelas duas medidas alternativas possíveis:

o *Cosseno*[1] para semelhanças e o *Chi-Quadrado*[2] para dessemelhança. A Abordagem Drexel não se concentra na mudança não temporal e seleciona o conjunto de autores que desde o início tem um alto grau de interconexões. A matriz de cocitação é fixada desde o início e a matriz de correlação produzida por *Pearson (r)* é fixa. Portanto, White (2003) considera altamente irrealista, em situações hipotéticas, expandir o conjunto definido de autores originais para incluir um novo conjunto de autores com zero cocitações, considera ainda que o tratamento dos valores na diagonal da matriz é um problema, mas sem muita relevância.

Bensman (2004) contribuiu para o debate ao criticar Ahlgren, Jarneving, Rousseau (2003) por apresentarem um sistema construído em espaços matemáticos que não demonstrou na prática a força das medidas de similaridade propostas. Leydesdorff e Vaughan (2006) realizaram experimentos para investigar os resultados do uso do coeficiente de correlação de *Pearson (r)* e as medidas de *Cosseno* e concluíram que as diferenças entre os dois são mínimas e marginais.

Segundo Eom (2009), há dois importantes desenvolvimentos recentes em ACA: o uso do coeficiente de correlação de *Pearson (r)*, como uma medida de similaridade e as ferramentas de visualização de redes, como o *Pathfinder* (WHITE, 2003), o *Authorlink* (LIN; WHITE; BUZYDLOWSKI, 2003), e o *VxInsight* (BOYACK; WYLIE; DAVIDSON, 2002), além do *Pajek* (BATAGELJ; MRVAR, 2010) e do *CiteSpace* (CHEN, 2006; 2004).

As ferramentas de visualização de redes fogem ao escopo desta discussão, mas favorecem a compreensão de estruturas complexas, tais como as redes de comunicação científica, facilitando a análise de um grande volume de dados. Junto com os mapas de visualização, os *softwares* geram cálculos de centralidade, frequência, proximidade, intermediação e densidade. Entretanto, é importante conhecer e observar as potencialidades e limitações das ferramentas disponíveis, para escolher a mais adequada à investigação que se pretende empreender, e se possível, pensando em adotá-la em futuros estudos comparativos para validar a aplicação do método.

Ao que parece, o uso do coeficiente de correlação de *Pearson (r)* tem preferência entre os pesquisadores em ACA, contudo acredita-se que a ampliação das investigações

[1] Medida de distância entre a variável e o fator. Valores próximos de 1 indicam um nível maior de representação da variável por meio dos fatores.
[2] Valor da dispersão para duas variáveis de escala nominal que nos diz em que medida os valores observados se desviam do valor esperado, caso as duas variáveis não estivessem correlacionadas.

comparando todas as possibilidades de medição pode contribuir fortemente para a discussão. De fato, a discussão é recente e há muito a debater.

4 CONCLUSÃO

Tomando-se a ACA como método, torna-se imperativo ampliar a produção de evidências empíricas visando aperfeiçoá-la. Sugere-se que a discussão, extremamente atual, tenha a colaboração dos bibliometristas brasileiros, investigando a ciência produzida no Brasil e suas redes de colaboração. Percebe-se um amplo campo para investigação em todas as etapas do método e que as decisões sobre elas repercutirão fortemente nas análises e resultados

A recente inclusão de uma grande massa de periódicos brasileiros em bancos de dados internacionais nos permite avaliar se a importação dos dados sobre autores selecionados, ou a construção de bases de dados customizadas, podem ser mais adequadas para as análises em ACA. A questão fundamental em ACA, inclusão ou não de todos os autores, também se tornou um campo de pesquisa viável para os bibliometristas brasileiros, seja por meio da construção de bases de dados customizada ou utilizando-se documentos estruturados em XML disponíveis, por exemplo, no Scielo.

Notam-se reclamações recorrentes entre os bibliometristas quanto à trabalhosa necessidade de padronização e consistência dos dados brutos para análises bibliométricas, reclamação esta que se considera exagerada. Outras metodologias de pesquisa também apresentam etapas trabalhosas, como por exemplo, a elaboração e aplicação de questionários ou entrevistas. O objeto de investigação das métricas da informação encontra-se manifesto em documentos publicados e com acesso relativamente facilitado, eliminando-se assim, grande parte da tarefa inicial de outros métodos. No que tange ao manejo, a consistência e a padronização dos dados, não há como escapar seja qual for o método adotado.

Atribui-se a computação, que se tornou inerente aos bancos de dados e estudos bibliométricos, o desejo legítimo dos bibliometristas de obter dados mais padronizados e consistentes. Entretanto, não se pode esquecer que na cadeia produtiva documental, apesar do processo automatizado/informatizado, encontra-se o ser humano como

gerador, organizador ou usuário da informação publicada, gerando equívocos inerentes às nossas limitações. Conclui-se que a ACA apresenta grande potencial de pesquisa como demonstra o intenso debate na literatura internacional.

ABSTRACT: The paper describes the origin, evolution and current debates surrounding the author cocitação analysis, to stimulate discussions between the Brazilian bibliometricians on the methodological procedures involved in the selection, tabulation, interpretation and validation of this method. We list the discussions in the literature about the acquisition of data, definition the units of analysis, generation and processing of raw data and arrays of proximity, and the choice of the measure of closeness. We conclude that the subject has great potential for research as evidenced by the intense debate in the international literature.

Key-words: Information Science. Bibliometrics. Author cocitation analysis.

REFERÊNCIAS:

AHLGREN, P.; JARNEVING, B.; ROUSSEAU, R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. **Journal of the American Society for Information Science and Technology**, v. 54, n. 6, p. 550-560, 2003.

BATAGELJ, V.; MRVAR, A. **Pajek**: program for analysis and visualization of large networks reference manual, list of commands with short explanation version 1.26. 2010. Disponível em: http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf. Acesso em 02 maio 2010.

BENSMAN, S. J. Pearson's R and author cocitation analysis: a commentary on the controversy. **Journal of the American Society for Information Science and Technology**, v. 55, p. 10, p. 935, 2004.

BJÖRNEBORN, L INGWERSEN, P. Toward a basic framework for webometrics. **Journal of the American Society for Information Science and Technology**, v. 55, n. 14, p. 1216-1227, 2004.

BOYACK, K. W.; WYLIE, B. N.; DAVIDSON, G. S. Domain visualization using VxInsight for science and technology management. **Journal of the American Society for Information Science and Technology**, v. 53, n. 9, p. 764-774, 2002.

CHEN, C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. **Journal of the American Society for Information Science and Technology**, v. 57, n. 3, p. 359-377, 2006.

CHEN, C. Searching for intellectual turning points: progressive knowledge domain visualization. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, supl. 1, p. 5303- 5310, April 6, 2004. Disponível em: http://www.pnas.org/content/101/suppl.1/5303.full.pdf. Acesso em 02 maio 2010.

EOM, S. **Author cocitation analysis**: quantitative methods for mapping the intellectual structure of an academic discipline. Hershey: IGI Global, 2009.

EOM, S. **Author co-citation analysis using custom bibliographic databases**: an introduction to the SAS approach. Lewiston: Edwin Mellen Press, 2003.

GLÄNZEL, W. **Bibliometrics as a research field**: a course on theory and application of bibliometric indicators. 2003. Disponível em: http://www.norslis.net/2004/Bib_Module_KUL.pdf. Acesso em 05 maio 2010.

GARFIELD, E. **From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography:** a citationist's tribute to Belver C. Griffith. 2001. http://www.garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf. Acesso em 02 maio 2010.
JARNEVING, B. A variation of the calculation of the first author cocitation strength in author cocitation analysis. **Scientometrics**, v. 77, n. 3, p. 485–504, 2008.

LEYDESDORFF, L.; VAUGHAN, L. Co-occurrence matrices and their application in information science: extending ACA to the web environment. **Journal of the American Society for Information Science and Technology**, v. 57, n. 12, p. 1616–1628, 2006.

LIN, X.; WHITE, H. D.; BUZYDLOWSKI, J. Real-time author co-citation mapping for online searching. **Information Processing & Management**, v. 39, n. 5, p. 689-706, 2003.

LINDSEY, D. Production and citation measures in the sociology of science: the problem of multiple authorship. **Social Studies of Science**, v. 10, p. 145-162, 1980.

MCCAIN, K. Mapping authors in intellectual space: a technical overview. **Journal of the American Society for Information Science**, v. 41, n. 6, p. 433-443, 1990.

MCINTIRE, J. S. **The clothing and textile research base**: an author cocitation study. 2006, 172 f. Dissertação (Mestrado) – Faculty of the Graduate School, University of Missouri, 2006.

NOYONS, E. C. M. **Bibliometric mapping as a science policy and research management tool**. Leiden: CWTS, Universiteit Leiden, 1999.

O'CONNOR, D. O.; VOOS, H. Empirical Laws, Theory Construction, and Bibliometrics. **Library Trends**, v. 30, n. 1, p. 9-20, 1981.

OKUBO, Y. **Bibliometric indicators and analysis of research systems**: methods and exemples. Paris: OECD Science, 1997.

PERSSON, O. All author citations versus first author citations. **Scientometrics**, v. 50, n. 2, p. 339-344, 2001.

PRICE, D. S. The analysis of square matrices of scientometrics transactions. **Scientometrics**, v. 3, n. 1, p. 55-63, 1981.

PRICE, D. S. **O desenvolvimento da ciência**: análise histórica, filosófica, sociológica e econômica. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

ROUSSEAU, R.; ZUCCALA, A. A classification of author co-citations: definitions and search strategies. **Journal of the American Society for Information Science and Technology**, v. 55, n. 6, p. 513-529, 2004.

SCHNEIDER, J. W.; LARSEN, B.; INGWERSEN, P. A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses. **Scientometrics**, v. 80, n. 1, p. 105–132, 2009.

SMITH, L. C. Citation analysis. **Library Trends**, v. 30, n. 1, p. 83-106, Summer 1981.
WHITE, H. D. Author cocitation analysis and Pearson's r. **Journal of the American Society for Information Science and Technology**, v. 54, n. 31, p. 250–259, 2003.

WHITE, H. D.; GRIFFITH, B. Author cocitation: a literature measure of intelectual structure. **Journal of the American Society for Information Science**, v. 32, n. 2, p. 163-171, 1981.