PROCEDIMENTOS E FERRAMENTAS APLICADOS AOS ESTUDOS BIBLIOMÉTRICOS[1]

*Samile Andréa de Souza Vanz**
*Ida Regina Chittó Stumpf***

Resumo Discute os processos de avaliação da produção científica e a necessária criação de indicadores para este fim. Apresenta fontes de coleta de dados para desenvolvimento de indicadores desta produção e apresenta os procedimentos para a limpeza/padronização e organização dos dados bibliométricos. Descreve softwares livres para análise bibliométrica e a importância do uso de indicadores relativos. Discute alguns procedimentos adotados pela comunidade científica internacional para análise multivariada de dados bibliométricos.

Palavras-chave: Bibliometria. Cientometria. Análise quantitativa. Bibexcel. Medidas de similaridade.

* Professora adjunta do Departamento de Ciências da Informação da Universidade Federal do Rio Grande do Sul.
E-mail: samilevanz@terra.com.br

** Professora titular do Departamento de Ciências da Informação e do Programa de Pós-graduação em Comunicação e Informação da Universidade Federal do Rio Grande do Sul.
E-mail: irstumpf@ufrgs.br

1 INTRODUÇÃO

A avaliação da produção científica é um processo fundamental para garantir o investimento financeiro em pesquisa e a participação da Ciência na consecução dos objetivos econômicos, sociais e políticos do país (VELHO, 1986). Quanto mais ativo e produtivo o ambiente científico, mais freqüentes e rigorosas são as rotinas de avaliação vigentes. Estes processos avaliativos se fundamentam, principalmente, em duas metodologias: a avaliação qualitativa, feita pelos pares, fortemente ancorada na reputação adquirida pelo avaliado; e a que se deriva de critérios quantitativos, baseados em métodos bibliométricos e cientométricos.

As técnicas quantitativas de medição da produção científica têm algumas décadas de existência, mas não estão, ainda, completamente consolidadas (SPINAK, 1998; SANCHO, 1990). Sua utilização está em franca expansão em diversos países, e a preocupação em acompanhar a tendência mundial de avaliação de Ciência e Tecnologia (C&T) fez com que o Brasil trabalhasse na criação de diferentes tipos de bases de dados e indicadores. As bases de dados também dão suporte para a desejada visibilidade da produção científica nacional, a partir de resultados de pesquisa, pesquisadores e instituições. Entre os exemplos de amplo reconhecimento está a SCIELO, a Plataforma Lattes, o Diretório dos Grupos de Pesquisa e a Base de Patentes produzida pelo Instituto Nacional de Propriedade Industrial (INPI). Além da criação das bases de dados para coleta e organização de dados relativos à C&T, diversos pesquisadores e instituições têm trabalhado na prospecção de indicadores de *input* e *output* da Ciência nacional, como a produtividade de instituições e áreas do conhecimento, fator de impacto dos periódicos, colaboração científica e investimentos em pesquisa.

Nas nações mais produtivas da Europa e nos EUA, a avaliação da produção científica é uma prática comum adotada por agências de fomento, ministérios e organismos ligados às políticas de C&T. Tal avaliação tem-se revelado essencial para a construção de indicadores e posterior distribuição de investimentos, desenvolvimento

1 Estudo desenvolvido para realização da tese de doutorado defendida no PPGCOM/UFRGS com auxílio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

de estratégias regionais e institucionais, e é claro, a avaliação dos resultados de políticas implementadas. As práticas relacionadas aos indicadores de C&T têm despertado o interesse de outras nações, especialmente as que apresentam um rápido crescimento em relação ao *output* científico, como é o caso do Brasil.

A proposta deste relato é apresentar procedimentos da metodologia quantitativa para tratamento da produção científica e algumas ferramentas disponíveis para o desenvolvimento de pesquisas bibliométricas. O texto está organizado em três partes: a primeira aborda os procedimentos de coleta, limpeza e organização dos dados; a segunda parte descreve as ferramentas para análise bibliométrica, apresentando softwares para este fim e a importância dos indicadores relativos e, por fim, são discutidos alguns procedimentos para análise multivariada de dados bibliométricos.

2 COLETA, LIMPEZA E ORGANIZAÇÃO DOS DADOS

Os dados quantitativos referentes à produção científica estão disponibilizados em bases de dados bibliográficas gerais ou multidisciplinares e em bases de dados especializadas, dedicadas a uma grande área do conhecimento. Entre as bases de dados especializadas encontram-se o **Chemical Abstracts**, **Compendex**, **BIDS Embase**, **Pascal SciTech** e **Pubmed**, entre outros, que abrangem dados bibliográficos da área de Química, Engenharia, Ciências Biomédicas e Medicina, respectivamente (LETA; CRUZ, 2003). Além destas, os dados podem ser coletados em bases de dados multidisciplinares, como o **Web of Science**, a **Scopus** e o **Google Acadêmico**. O **Web of Science** (**WoS**), produzido pelo ISI desde a década de 60, oferece acesso à três índices de citações: Science Citation Index Expanded, Social Sciences Citation Index e o Arts & Humanities Citation Index (THOMSON CORPORATION, 2004). De forma similar, a **Scopus**, produzida pela Elsevier desde 2004, oferece ampla cobertura da literatura científica e técnica publicada a partir do século XIX em várias áreas do conhecimento (ELSEVIER, 2010). O **Google Acadêmico** disponibiliza a pesquisa em documentos não indexados em bases bibliográficas renomadas, entre eles livros, teses, dissertações, resumos, artigos e *pre-prints* de editoras acadêmicas, organizações profissionais, universidades e outras entidades (GOOGLE, 2010). Outra importante fonte de dados, o **National Science Indicators**, também é publicado pelo ISI e comercializado em CD. Seu conteúdo basei-se no Science Citation Index Expanded[2].

Os indicadores levantados em bases de dados internacionais são relevantes para monitoramento da produção científica brasileira internacional, possibilitando uma estimativa de como o Brasil contribui com a Ciência *mainstream*. O uso de bases internacionais ainda tem como vantagem a possibilidade de comparação dos resultados brasileiros com os resultados obtidos por outras nações. Porém, o levantamento de indicadores de produção científica em bases de dados nacionais é fundamental, especialmente em nações periféricas e que não possuem o inglês como língua mãe. No Brasil, a biblioteca eletrônica de periódicos Scielo desempenha um importante papel na comunicação científica nacional ao indexar e disponibilizar de forma eletrônica e gratuita o acesso a 228 periódicos. Apesar de ainda não disponibilizar publicamente o índice de citações, a Scielo oferece alguns indicadores bibliométricos consolidados, baseados na literatura científica por ela indexada. Entre eles está o indicador de citações das revistas (citações concedidas e recebidas; fator de impacto) e de co-autoria[3].

Mesmo consideradas as maiores bases de dados multidisciplinares, tanto o Google Acadêmico como a Scopus e a Web of Science apresentam inconsistências na grafia de nomes (JACSO, 2005). Entre os problemas mais comuns estão as diferentes formas de grafia de nomes dos autores. Por exemplo, identifica-se somente a inicial do primeiro nome ou então, as iniciais de todos os nomes, ou ainda o nome por completo. As homonímias – diferentes pessoas identificadas pelo mesmo nome – são comuns nestas bases, decorrentes muitas vezes do uso do sobrenome seguido apenas de uma inicial do nome. Em relação aos nomes das instituições, os problemas se referem à grafia das instituições em diferentes línguas, geralmente em português e inglês. Por exemplo, a Pontifícia Universidade Católica

2 Informação disponibilizada na lista de discussão Sigmetrics, no dia 03 de maio de 2010, por Jim Testa, da Thomsom Reuters.

3 Informação disponível em: <http://www.scielo.br>. Acesso em: 23 mar. 2010.

do Rio Grande do Sul está descrita no ISI por inúmeras formas (AUTOR, 2009), como PUCRS, PUC RS, Rio Grande Sul Pontificial Catholic Univ, Pont Univ Cat Rio Grande do Sul, Pont Univ Católica Porto Alegre, Pontifical Catholic Univ Rio Grande Sul, Pontifical University Catholic Rio Grande do Sul, entre outras. Estas diferentes grafias alteram o resultado de rankings de produtividade, estudos de co-ocorrência, como co-autoria entre pesquisadores e instituições, e como co-citação.

A constatação destas inconsistências torna necessária a padronização/limpeza de nomes de autores, instituições de filiação, títulos das obras, entre outros dados, procedimento que precisa ser realizado imediatamente após o *download* dos arquivos. Apesar de geralmente demandar o maior tempo da pesquisa bibliométrica, o procedimento vem sendo aplicado por diversos pesquisadores para garantir maior fidedignidade dos dados (MUGNAINI; JANNUZZI; QUONIAM, 2004; LETA; GLÄNZEL; THIJS, 2006; HOU; KRETSCHMER; LIU, 2008). Alguns autores, entretanto, têm dispensado o processo de limpeza/padronização de nomes de autores por considerar que o erro ocasionado pela homonímia é percentualmente muito baixo e não altera significativamente o resultado final (NEWMAN, 2001a; WAGNER; LEYDESDORFF, 2005).

Outro procedimento que pode ser realizado é a organização da produção científica em grandes áreas de publicação, de forma a evitar a sobreposição de assuntos e permitir a avaliação e comparação entre diferentes séries de dados. Uma das propostas para evitar a sobreposição é o esquema de classificação de áreas do conhecimento definido por Glänzel e Schubert (2003), que relaciona as áreas de publicação do ISI em 15 grandes áreas do conhecimento, a saber: Agricultura e meio ambiente; Biologia; Biociências; Pesquisa biomédica; Medicina clínica e experimental I; Medicina clínica e experimental II; Neurociência e comportamento; Química; Física; Geociências e ciências espaciais; Engenharias; Matemática; Ciências sociais I; Ciências sociais II; Artes e humanidades. O esquema vem sendo usado com sucesso por alguns pesquisadores em análises bibliométricas (GLÄNZEL; LETA; THIJS, 2006; MOURA, 2009; AUTOR, 2009). Outra possibilidade é adequar as categorias de assunto do ISI – presentes no campo SC – às Tabelas das Áreas do Conhecimento utilizadas pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e CAPES, a saber: Ciências exatas e da terra; Ciências Biológicas; Engenharias; Ciências da Saúde; Ciências agrárias; Ciências sociais aplicadas; Ciências humanas; Lingüística, letras e artes; Outros.

Quanto à atribuição de valores para a quantificação da produção científica, os estudos bibliométricos vem atribuindo o valor de um artigo/citação para cada autor, instituição e país envolvido no artigo. Desta maneira, os totais de ocorrência de autores, instituições, países e citações não refletem o total de artigos publicados, mas o somatório de ocorrências. Esta estratégia vem sendo utilizada por diversos autores (LUUKKONEN; PERSSON; SIVERTSEN, 1992; PACKER; MENEGHINI, 2006; LIMA; VELHO; FARIA, 2007; VILAN FILHO; SOUZA; MUELLER, 2008). A metodologia oposta, chamada de fracionamento, atribui meio artigo para cada autor no caso de um artigo publicado por dois autores, um terço de artigo no caso de três autores, e assim sucessivamente.

3 FERRAMENTAS PARA ANÁLISE BIBLIOMÉTRICA: SOFTWARES E ÍNDICES RELATIVOS

Os dados bibliográficos importados das bases de dados podem ser organizados e analisados através de diferentes softwares para análise bibliométrica, como o conjunto de aplicativos desenvolvido por Loet Leydesdorff[4] e o Bibexcel[5], desenvolvido por Olle Person. Ambos são livres e disponibilizam ferramentas para análise descritiva de produtividade e citações, entre outras, além de análise de co-autoria[6], co-citação[7], *co-words*[8], análises multivariadas e análises de redes.

O Bibexcel apresenta-se como um software flexível para o usuário, bastando,

4 Disponível em http://users.fmg.uva.nl/lleydesdorff/software.htm

5 Disponível em http://www.umu.se/inforsk/Bibexcel

6 A análise de co-autoria baseia-se nos nomes dos autores de um artigo científico. É considerada uma das formas de se medir a colaboração científica e pode se referir a pesquisadores, instituições e países, estes dois últimos através da vinculação institucional dos autores.

7 Co-citação define-se como a análise que estuda as relações e freqüências de pares de documentos que são citados por um terceiro documento. Pode ser relativa a documentos, autores e periódicos.

8 *Co-words* define-se como a análise que estuda as relações e freqüências de pares de palavras presentes em títulos e *abstracts* de documentos.

para isso, entender a estrutura básica dos arquivos e os procedimentos e comandos para as análises. Entre as suas funcionalidades está a organização de dados em arquivos de texto ou planilha, o que possibilita ao pesquisador a utilização de outros softwares para as análises e também a importação de diferentes tipos de dados, além dos bibliográficos importados da Web of Science ou Scopus. A familiaridade com registros bibliográficos é fundamental, pois o Bibexcel funciona com base nos registros e delimitadores de campos. Assim, para iniciar a análise bibliométrica é necessário informar qual é o campo a ser analisado e qual é o delimitador usado naquele campo. Para analisar autores, por exemplo, é necessário informar ao Bibexcel a sigla *AU* na janela *Old Tag* e também que o campo de autor é delimitado por ponto e vírgula, ou seja, todas as expressões presentes entre ponto e vírgula representam nomes de autores. Se o interesse do pesquisador recair na análise de citações, o campo informado é o *CD*, o qual também é delimitado por ponto e vírgula.

Outra funcionalidade do Bibexcel é a geração de rankings de produtividade e citação. As análises descritivas podem ser realizadas com o auxílio do Excel e sua ferramenta de Tabela Dinâmica. O Bibexcel oferece ainda a possibilidade de criação das matrizes de co-ocorrência que posteriormente, servem de *input* em análises multivariadas como o Escalonamento Multidimensional (EMD), Análise Fatorial, Análise de Correspondência e Análise de Agrupamentos (*Clusters*) (PERSSON, DANELL; SCHNEIDER, 2009). Algumas destas podem ser realizadas pelo próprio software. Como opção, as matrizes podem ser analisadas com auxílio de outros softwares, como o Statistics Packet for Social Science (SPSS) ou o Microsoft Excel. As mesmas matrizes também podem ser visualizadas a partir de softwares para Análise de Redes Sociais, como o Pajek[9] e Ucinet[10].

As matrizes de co-ocorrência geradas pelo Bibexcel são baseadas nos totais de ocorrência de um indicador. O uso destes indicadores absolutos (totais de ocorrência) é amplamente aceito como ferramenta útil na mensuração do desempenho científico. Entretanto, a análise bibliométrica baseada no uso de indicadores relativos e normalizados pode revelar aspectos subjacentes até então invisíveis nos dados brutos. Na análise de co-autoria, por exemplo, os números absolutos indicam o total de artigos em co-autoria sem considerar o tamanho dos autores envolvidos na colaboração − medido pelo total de artigos publicados. Para estimar propensões ou intensidade de co-autoria, faz-se necessário recorrer a indicadores relativos que levem em consideração o tamanho da produção científica de um autor, instituição ou país. Luukkonen e outros (1993), ao avaliarem a colaboração científica internacional, afirmam que o total de artigos publicados em co-autoria entre dois países deve ser analisado em relação ao total da produção científica de cada um dos países. No caso de um país ser muito produtivo e outro pouco produtivo, a colaboração entre os dois pode não ser muito significativa quando comparada à produção total do país produtivo, e ao contrário, ser bem significativa se comparada ao total publicado pelo país menor. Analogamente, a análise de colaboração entre instituições deve seguir o mesmo procedimento, visto que elas também se diferenciam quanto ao tamanho e à produtividade. Nesse sentido, Luukkonen e outros afirmam:

> [...] na análise de relações de colaboração, é essencial usar ambas as medidas absolutas e relativas. A última normaliza diferenças de tamanho dos países. Cada uma carrega tipos diferentes de informação. Medidas absolutas carregam respostas a questões como quais são os países centrais na rede internacional da ciência, se relações de colaboração revelam um centro – relações periféricas, e que países são os parceiros mais importantes de outros. Medidas relativas oferecem respostas a questões de intensidade das relações de colaboração (LUUKKONEN, 1993, p.15, tradução nossa)

Com o objetivo de possibilitar diferentes interpretações, Luukkonen, Persson e Sivertsen (1992) propuseram uma fórmula que calcula a freqüência esperada, a partir de uma distribuição aleatória dos valores da diagonal da matriz entre todas as células. A freqüência esperada é relacionada à freqüência observada, conforme a fórmula abaixo, desenvolvida no contexto de co-autoria entre países:

9 Disponível em http://www.vlado.fmf.uni-lj.si/pub/networks/pajek/
10 Disponível em http://www.analytictech.com/downloaduc6.htm

$$\frac{C_{x,y} \times T}{C_x X C_y}$$ onde,

$C_{x,y}$ = total de co-autorias entre o país X e Y
T = total de co-autorias da matriz
C_x = total de co-autorias o país X possui na matriz
C_y = total de co-autorias o país Y possui na matriz

Segundo orientação dos autores, o índice deve ser calculado com base em uma matriz completa, ou seja, com a diagonal ≠ 0. Índice igual a 1 indica uma colaboração observada de acordo com a esperada. Resultados menores que 1 indicam que a colaboração é menor do que a esperada. Os maiores que 1 indicam uma relação de colaboração mais forte do que o esperado.

Outro índice que vem sendo utilizado na literatura é o Cosseno de Salton, que pode ser calculado a partir da matriz de co-ocorrência bruta, conforme a fórmula de Luukkonen e outros (1993):

$$S_{xy} = \frac{C_{xy}}{\sqrt{C_x X C_y}}$$ onde,

C_{xy} = total de artigos publicados por x e y
C_x = total de artigos publicados por x
C_y = total de artigos publicados por y

A fórmula do Cosseno de Salton se apresenta em outros formatos. A seguir, o formato usado por Hamers e outros (1989), no contexto da co-citação de autores, e Arunachalam (2000), para análise de co-autoria entre países:

$$Ss_{(i,j)} = \frac{coc_{(i,j)}}{(cit_{(i)} \cdot cit_{(j)})^{1/2}}$$ onde,

$coc_{(i,j)}$ = total de co-ocorrências do autor i e j
$cit_{(i)}$ = total de citações recebidas pelo autor i
$cit_{(j)}$ = total de citações recebidas pelo autor j

O uso destas fórmulas pode ser feito através do Excel, com base nas matrizes de dados brutos criadas pelo Bibexcel.

4 ANÁLISES MULTIVARIADAS

Além das análises descritivas e do uso de indicadores relativos pode-se aplicar análises multivariadas aos dados bibliométricos, como o Escalonamento Multidimensional (EMD), Análise Fatorial, Análise de Correspondência e Análise de Agrupamentos (*Clusters*). Para proceder às análises multivariadas em dados bibliométricos, especialmente aqueles presentes em matrizes simétricas e assimétricas, buscou-se embasamento metodológico na literatura da área de Ciência da Informação. O periódico **Journal of the American Society for Information Science and Technology** (JASIST) publica há alguns anos uma discussão sobre a metodologia adequada à análise das matrizes de co-ocorrência, como por exemplo, co-citações, *co-words*, co-autoria, *co-membership, co-classification* e *co-participation*.

As matrizes simétricas de co-ocorrência - como as matrizes de co-autoria - são consideradas matrizes de proximidade do tipo similaridade, pois indicam o quão similar dois autores (ou instituições autoras) se apresentam (LEYDESDORFF; VAUGHAN, 2006; ECK e WALTMAN, 2007). Assim, quanto maior o número na célula de interseção entre uma linha (um autor) e uma coluna (outro autor), mais artigos publicados em co-autoria os dois autores possuem e, portanto, mais similares os dois autores se mostram.

Segundo Ahlgren, Jarneving e Rousseau (2003), a metodologia utilizada para análise de co-citações segue quatro passos. Primeiro, a matriz de dados brutos é compilada; depois, é feita uma conversão dessa matriz para uma matriz de proximidade, associação ou similaridade. O terceiro passo é a análise multivariada das relações entre os autores presentes na matriz. Nesse passo, algumas análises vêm sendo usadas: análise de agrupamentos, escalonamento multidimensional (EMD), análise fatorial e análise de correspondência. Após as análises, ocorre a última etapa do processo, a interpretação dos dados. Os autores afirmam que, apesar de existirem necessidades específicas de acordo com os objetivos da investigação, não existem diferenças teóricas e/ou matemáticas entre

análise de co-citações, *co-words*, co-autoria, *co-membership, co-classification* e *co-participation*,

A metodologia desenvolvida inclui a geração de uma matriz de similaridade a partir da matriz de dados brutos, com base em diferentes medidas. A questão de qual medida usar tem sido discutida há algum tempo e encontra respostas diversificadas na literatura. Segundo Luukkonen e outros (1993), a resposta depende do aspecto que se quer avaliar. Os autores explicam que há dois tipos de medidas de associação: as medidas de similaridade bilaterais e as multilaterais. A primeira deve ser usada se o objetivo é comparar relações entre pares de países e instituições separadamente, e, entre elas, estão a medida de Salton e Jaccard. As medidas multilaterais, como a freqüência esperada e Correlação de Person, relaciona a co-autoria entre um par de autores com todos os outros autores envolvidos na análise.

Na opinião de Ahlgren, Jarneving e Rousseau (2003), a medida de similaridade denominada Cosseno de Salton é a mais indicada quando o objetivo do pesquisador concentra-se na visualização da estrutura, seja através de Análise de Redes Sociais ou EMD, visto que é uma medida definida geometricamente. Já White (2003) defende o uso da Correlação de Pearson com o argumento de que as diferenças entre o uso de diferentes medidas de similaridade podem ser negligenciadas na prática de pesquisa. O autor testa as medidas de Correlação de Person, Cosseno de Salton e Chi-Quadrado e afirma que as três medidas podem revelar uma resposta muito parecida. Bensman (2004) também se apresenta favorável ao uso da Correlação de Person para normalização quando o objetivo são as análises estatísticas multivariadas.

Leydesdorff e Vaughan (2006) argumentam que matrizes de co-autoria são matrizes de proximidade do tipo similaridade que não requerem normalização antes de análises EMD. Para fazer a normalização, os autores afirmaram ser mais adequado usar a matriz assimétrica (matriz de ocorrência), subjacente a matriz de co-ocorrência, como base para análise multivariada. Entretanto, Leydesdorff e Vaughan (2006) divulgam a opinião de um dos avaliadores do periódico em que o artigo foi publicado, cuja sugestão é que, por razões teóricas, os pesquisadores podem continuar preferindo aplicar a medida de similaridade à matriz de co-ocorrência, com o objetivo de comparar padrões de co-autoria ao invés de comparar a contagem de artigos em co-autoria. Schneider e Borlund (2007) consideram não existir nenhum problema estatístico na prática de aplicar medida de similaridade às matrizes de co-ocorrência.

Posteriormente, no mesmo periódico, Waltman e Eck (2007) também se pronunciaram, afirmando que as análises multivariadas podem ser feitas em matrizes simétricas convertidas por diversas medidas de proximidade, sendo sugestão dos autores a Jansen-Shannon, a Bhattacharyya e o Cosseno. Os autores atribuem ao SPSS um defeito de programação, que teria levado Leydesdorff e Vaughan (2006) a concluir que o mapa distorcido era conseqüência da conversão da matriz de dados brutos para uma matriz de similaridades. Waltman e Eck (2007) avaliam que o problema pode ser contornado, e o mapa adequado é gerado a partir de uma rotina que utiliza o modelo *Spline* para análise EMD (na versão 14.0 ou inferior do SPSS).

Dando continuidade à discussão no **JASIST**, Leydesdorff (2008, p. 79, tradução nossa) afirma que: "Em princípio, pode-se normalizar tanto matrizes simétricas quanto assimétricas através de várias medidas". Formalmente, Person e Cosseno são equivalentes, com exceção de que Pearson normaliza através da média aritmética, enquanto o Cosseno utiliza como parâmetro a média geométrica. Ou seja, o Cosseno mede a similaridade entre dois vetores usando o ângulo entre eles. Eck e Waltman (2008) concluem que a Correlação de Pearson não apresenta resultados satisfatórios quando usada para medir a similaridade entre padrões de co-citação de autores porque é uma medida apropriada para medir a correlação linear entre duas variáveis. O Cosseno e, também, a medida de divergência Jensen-Shannon e a de Distância de Bhattacharyya, são as medidas mais adequadas na opinião dos autores. Além disso, Eck e Waltman (2008), contrariando opiniões anteriores, defendem que a escolha de uma medida de similaridade apropriada tem relevância prática e não só teórica, visto que os resultados encontrados divergem, especialmente quando mapeados através de técnicas EMD. Egghe e Leydesdorff (2009) dão

a última palavra na discussão, dizendo que, apesar das diferenças entre Salton e Pearson serem mínimas, ninguém pode estimar a sua significância, e indicam a preferência pelo Cosseno de Salton para análise e visualização de similaridades.

Como relatado, a literatura não apresenta conclusões sobre o assunto e os procedimentos, apesar de estarem em uso desde o início dos anos 1990, ainda não estão plenamente consolidados. Muitos estudos ainda estão sendo feitos para definir a necessidade de conversão da matriz de dados brutos para uma matriz de similaridades, e, a partir daí, definir qual medida é mais adequada; e ainda, para definir qual a matriz mais adequada para análise, se a de ocorrência ou a matriz quadrada de co-ocorrência.

Alguns tipos de análises bibliométricas não resultam em matrizes de co-ocorrência, como por exemplo, a análise de citações. Nesta análise o pesquisador tem duas variáveis: um documento citante e o autor ou documento citado. Neste caso, por conter uma variável nominal (autor), a Análise de Correspondência é indicada e vem sendo usada para medir possíveis relacionamentos e proximidades entre citantes e citados (AUTOR, 2008).

5 CONSIDERAÇÕES FINAIS

A pesquisa bibliométrica e o uso de indicadores da produção científica vem sendo alvo do trabalho e das pesquisas de diversos autores. Inúmeras discussões vem sendo propostas entre a comunidade científica mundial e, sem dúvida, elas fundamentam e contribuem com as pesquisas realizadas aqui no Brasil. Entretanto, temos consciência da necessidade de desenvolvermos indicadores e metodologias adequadas a realidade nacional.

A criação de bancos de dados contendo a produção científica nacional, além de informações sobre pesquisadores, instituições e grupos de pesquisa brasileiros já iniciou há alguns anos e a comunidade científica e agências de fomento contam hoje com ferramentas consolidadas. Agora, é preciso desenvolver indicadores a partir destes bancos de dados. Assim, de maneira simultânea ao desenvolvimento das ferramentas de pesquisa, é fundamental que os pesquisadores brasileiros aprofundem o conhecimento sobre os procedimentos aplicados na pesquisa bibliométrica para que a área avance e tenhamos condições de propor indicadores mais adequados para medir a produtividade científica nacional.

PROCEDURES AND TOOLS APPLIED TO BIBLIOMETRIC STUDIES

Abstract *Discusses the process of scientific prodution evaluation and the necessary development of indicators for this purpose. Presents sources for data collection for development of scientific production indicators and presents the procedures for cleaning/standardization and organization of bibliometric data. Describes free softwares for bibliometric analysis and the importance of using relative indicators. Discusses some procedures adopted by the international scientific community for multivariate analysis of bibliometric data.*

Key-words*:* *Bibliometrics. Scientometrics. Quantitative analysis. Bibexcel. Similarities measures.*

Artigo recebido em 09/06/2010 e aceito para publicação em 07/09/2010

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 of Information Science & Technology**, New York, v. 54, n. 6, p. 1616-1628, 2006.

BENSMAN, S. J. Person's r and author cocitation analysis: a commentary on the controversy. **Journal of the American Society of Information Science & Technology**, New York, v. 55, n. 10, p. 935-936, 2004.

ECK, N. J.; WALTMAN, L. Appropriate similarity measures for author cocitation analysis. 2007. **Journal of the American Society of Information Science & Technology**, New York, v. 59, n. 10, p. 1653-1661, 2008.

EGGHE, L.; LEYDESDORFF, L. The relation between Pearson's correlation coefficient *r* and Salton's cosine measure. **Journal of the American Society of Information Science & Technology**, New York, v. 60, n. 5, p. 1027-1036, 2009.

ELSEVIER. **Scopus**. 2010. Disponível em: < http://www.scopus.com/home.url>. Acesso em: 13 abr. 2010.

GLÄNZEL, W.; SCHUBERT, A. A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. **Scientometrics**, Amsterdam, v. 56, n. 3, p. 357-367, 2003.

GLÄNZEL, W.; LETA, J.; THIJS, B. Science in Brazil. Part 1: a macro-level comparative study. **Scientometrics,** Amsterdam, v. 67, n. 1, p. 67-86, 2006.

GOOGLE. Google Acadêmico. 2010. Disponível em: <http://scholar.google.com.br/intl/pt-BR/scholar/about.html>. Acesso em: 13 abr. 2010.

HAMERS, L. et al. Similarity measures in Scientometric Research: the Jaccard Index versus Salton's Cosine formula. **Information Processing & Management**, New York, v. 25, n. 3, p. 315-318, 1989.

HOU, H.; KRETSCHMER, H.; LIU, Z. The structure of scientific collaboration networks in Scientometrics. **Scientometrics**, Amsterdam, v. 75, n. 2, p. 189-202, 2008.

LETA, J.; GLÄNZEL, W.; THIJS, B. Science in Brazil. **Scientometrics**, Amsterdam v. 67, n. 1, p. 87-105, 2006.

LEYDESDORFF, L. The mutual information of university-industry-government relations: an indicator of the Triple Helix dynamics. **Scientometrics**, Amsterdam, v. 58, n. 2, p. 445-467. 2003.

LEYDESDORFF, L. Similarity measures, author cocitation analysis, and Information Theory. **Journal of the American Society of Information Science & Technology**, New York, v. 56, n. 7, p. 769-772, 2005.

LEYDESDORFF, L.; VAUGHAN, L. Co-occurrence matrices and their applications in Information Science: extending ACA to the Web Environment . **Journal of the American Society of Information Science & Technology**, New York, v. 57, n. 12 p. 1616-1628, 2006.

LEYDESDORFF, L. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard Index. **Journal of the American Society of Information Science & Technology**, New York, v. 59, n. 1, p. 77-85, 2008.

LIMA, R. A.; VELHO, L. M. L. S.; FARIA, L. I. L.. Indicadores bibliométricos de cooperação científica internacional em bioprospecção. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 50-64, jan./abr. 2007.

LUUKKONEN, T.; PERSSON, O.; SIVERTSEN, G. Understanding patterns of international scientific collaboration. **Science, Technology & Human Values**, Thousand Oaks, v. 17, n.1, Winter, 1992, p. 101-126.

LUUKKONEN, T. et al. The measurement of international scientific collaboration. **Scientometrics**, Amsterdam, v. 28, n.1, p. 15-36, 1993.

MOURA, A.M.M. **A interação entre artigos e patentes**. 2009. 269 f. Tese (Doutorado) - Programa de Pós-Graduação em Comunicação e Informação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

MUGNAINI, R.; JANNUZZI, P.; QUONIAM, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. **Ciência da Informação**, Brasília, v. 33, n.2, p. 123-131, maio/ago. 2004.

PERSSON, O.; DANELL, R.; SCHNEIDER, J.W. How to use Bibexcel for various types of bibliometric analysis. In: ASTROM, F. et al (ed.). **Celebrating scholarly communication studies:** a festschrift for Olle Persoon at his 60th birthday. ISSI, 2009. p. 9-24.

SANCHO, R. Indicadores Bibliometricos Utilizados en la Evaluación de la Ciência y la Tecnologia: revision bibliográfica. **Revista Española de Documentación Científica**, Madrid, v. 13, n. 3-4, p. 842-65, 1990.

SPINAK, E. Indicadores cienciométricos. **Ciência da Informação**, Brasília, v. 27, n.2, p.141-148, maio/ago. 1998.

THOMSON CORPORATION. **Web of Science 7.0**: education program. 2004. 96 p.

VELHO, L. A avaliação do desempenho cientifico. **Cadernos USP**, São Paulo, n. 1, out. p. 22-40. 1986.

VILAN FILHO, J. L.; SOUZA, H. B.; MUELLER, S. Artigos de periódicos científicos das áreas de informação no Brasil. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 13, n.2, p. 2-17, maio/ago. 2008.

WALTMAN, L.; ECK, N. J. Some comments on the question whether co-occurrence data should be normalized. **Journal of the American Society of Information Science & Technology**, New York, v. 58, n. 11, p. 1701-1703, 2007.

WHITE, H. D. Author cocitation analysis and Pearson's r. **Journal of the American Society of Information Science & Technology**, New York, v. 54, n. 13, p. 1250-1259, 2003.

ZIMBA, H.F.; MUELLER, S.P.M. Parcerias na ciência. **Datagramazero**, Rio de Janeiro, v. 5, n. 1, art.4, 2004.