2
Bibliometria e representação do conhecimento como instrumentos para avaliação da produção científica

Apresenta-se neste capítulo o referencial teórico sobre o tema central da dissertação – bibliometria e a representação do conhecimento como instrumentos para avaliação da produção científica de um determinado campo ou área científica. Na sequência, abordam-se os métodos e ferramentas bibliométricas mais adotadas em nível internacional, com destaque para as que se mostram mais adequadas para a análise pretendida na fase de pesquisa aplicada. Ressalta-se, ao final, a oportunidade de se utilizar indicadores bibliométricos complementares aos tradicionalmente adotados em avaliações de programas de pós-graduação, com ênfase para aqueles gerados pela análise de co-ocorrência de palavras-chave e termos significativos de uma ou mais áreas de conhecimento correlatas.

A base conceitual sobre bibliometria, juntamente com o entendimento sobre as Leis de Bradford, Lotka e Zipf, fundamentam a discussão central deste capítulo em torno das seguintes questões:

- Qual a contribuição da bibliometria e dos métodos de representação do conhecimento para a avaliação da dinâmica da produção científica de um programa de pós-graduação?
- Quais são as ferramentas de escolha para a avaliação da dinâmica da produção científica do Programa PósMQI, desde sua criação?

2.1.
Bibliometria: conceitos e as leis de Lotka, Bradford e Zipf

Padrões estatísticos encontrados em bases de dados, periódicos, livros e demais formas de comunicação científica podem ser medidos através da bibliometria, cientometria e informetria por meio de variáveis distintas, tais como citações, palavras-chave, autores, temas, localidades, dentre outras (Lotka, 1926; Price, 1965; Potter, 1981; King, 1987; Narin, 1987; Potter, 1988; Sancho, 1990;

Leydesdorff, 1991; Quoniam, 1992; Shapiro, 1992; Cambrosio et al., 1993; Spinak, 1996; Rostaing, 1996; Hood e Wilson, 2001; Maltrás Braba, 2003; Moed et al., 2005).

Em particular, a bibliometria refere-se a estudos de natureza teórico-conceitual, quando associados a estudos sobre o avanço do conhecimento da própria temática, propondo novos conceitos e indicadores, bem como reflexões e análises relativas à área em foco. Reúne um conjunto de conhecimentos relacionados à avaliação da informação produzida e alicerçados na Sociologia da Ciência, na Ciência da Informação, Matemática, Estatística e Ciência da Computação (Potter, 1981; King, 1987; Quoniam, 1992; Cambrosio et al., 1993; Wasserman, 1994; Rostaing, 1996; Hood e Wilson, 2001; Maltrás Barba, 2003; Hanneman e Riddle, 2005).

Um histórico da evolução da bibliometria elaborado por Oliveira (2013) revela que a análise estatística da literatura científica começou quase 50 anos antes que o termo "bibliometria" tivesse sido cunhado (Glänzel e Schubert, 2003). Um estudo pioneiro sobre a distribuição de frequência de produtividade científica de autores de diversos campos da Ciência, compreendendo o período entre 1909 e 1916, foi publicado por Lotka em 1926 (Lotka, 1926). Nesse estudo, o autor ressaltava que uma significativa proporção da literatura científica era produzida por um pequeno número de autores e que um grande número de autores com poucas publicações se igualava, em produção, ao reduzido número dos autores mais profícuos.

De acordo com Oliveira (2013), oito anos após a divulgação da Lei de Lotka, Bradford em 1934 publicou um estudo sobre a distribuição de frequência de trabalhos em periódicos, também chamada lei da dispersão bibliográfica de periódicos (Bradford, 1961; Brookes, 1977; Garfield, 1980; Pinheiro, 1983).

Na sequencia, a terceira lei bibliométrica foi formulada em 1949 e é chamada Lei de Zipf. Essa lei relaciona a frequência das palavras em um texto e a ordem de série dessas palavras (Quoniam, 1992).

No Brasil, as abordagens teóricas e aplicações bibliométricas e informétricas entre pesquisadores brasileiros especialistas em metrias da comunicação científica apontam para diversidade, convergências, fortalecimento e expansão desse campo

de estudos (Pinheiro e Silva, 2008). A seguir, apresentam-se as três leis bibliométricas e os principais indicadores adotados no campo da bibliometria.

2.1.1.
Lei de Lotka: produtividades de autores

A Lei de Lotka analisa a produção científica dos autores, ou seja, determina a contribuição de cada um deles para o avanço do campo científico em análise. A Lei de Lotka data de 1926 e é também conhecida como Lei do Quadrado Inverso, devido à seguinte premissa: o número de autores que tenham publicado exatamente (n) trabalhos é inversamente proporcional a (n^2). Maltrás Barba (2003), em sua revisão sobre indicadores bibliométricos, ressalta que a cada 100 autores com um trabalho somente, haverá 25 autores com dois trabalhos, 11 autores com três trabalhos e assim sucessivamente.

A Lei de Lotka também pode ser vista com uma função de probabilidade da produtividade. Quanto mais se publica, mais parece que se facilita publicar um novo trabalho e os pesquisadores que publicam resultados mais interessantes ganham mais reconhecimento e acesso a recursos para melhorar sua pesquisa. (Maltrás Barba, 2003). O enunciado da Lei de Lotka encontra-se na caixa de texto a seguir.

Enunciado da Lei de Lotka

O número de autores que produzem n trabalhos corresponde a $1/n^2$ daqueles que produzem apenas um trabalho. E a proporção de todos os autores que fazem apenas um trabalho fica em torno de 60% (Lotka, 1926).

2.1.2
Lei de Bradford: produtividade de periódicos

A Lei de Bradford concentra sua descrição no comportamento repetitivo das ocorrências em um determinado campo ou área científica (Bradford, 1961; Brookes, 1977; Garfield, 1980). Segundo Quoniam (1992) e Quoniam et al. (2001), Bradford escolheu o periódico como unidade de análise, devido às suas características de incidência de assuntos e tendências, tendo observado que poucos periódicos produziam muitos artigos e muitos periódicos produziam poucos artigos. Como ressalta Pinheiro (1983), as primeiras observações de Bradford

sobre a dispersão de artigos ocorreram em 1934 em um trabalho pioneiro, mas somente em 1948 recebe o status de lei, depois de sintetizadas. Apresenta-se na caixa de texto abaixo o enunciado da Lei de Bradford.

Enunciado da Lei de Bradford

Se os periódicos forem ordenados em ordem de produtividade decrescente de artigos sobre um determinado assunto, poderão ser distribuídos em um núcleo de periódicos mais particularmente devotados a esse assunto e em diversos grupos ou zonas contendo o mesmo número de artigos que o núcleo (Bradford, 1961).

A Lei de Bradford pressupõe que após a publicação de alguns artigos sobre um determinado novo tema de um campo científico em periódicos qualificados, os mesmos periódicos polarizarão artigos sobre este novo tema durante um tempo. Em paralelo, diferentes periódicos apenas iniciam a publicação dos primeiros artigos sobre o referido tema. No caso de uma consolidação do tema, surgirá um núcleo gravitacional dos periódicos que mais publicaram sobre o tema em foco (Brookes, 1968; 1977).

2.1.3
Lei de Zipf: frequência de ocorrência de palavras

A Lei de Zipf é também chamada de lei quantitativa fundamental da atividade humana (Hood e Wilson, 2001; Spinak, 1996; Quoniam, 1992). Segundo Quoniam (1992), a Lei de Zipf subdivide-se em:

- primeira Lei de Zipf: diz que o produto da ordem de série de uma palavra multiplicado pela frequência de ocorrência é aproximadamente constante. É regida É regida pela expressão matemática: **K**= **R** x **F**, onde **K**= constante; **R**=ordem das palavras; **F**=frequência das palavras;
- segunda Lei de Zipf que enuncia que, em um determinado texto, várias palavras de baixa frequência de ocorrência (alta ordem de série) têm a mesma frequência.

Assim, a Lei de Zipf calcula uma constante em relação às frequências das palavras em um texto. Por exemplo: se a palavra com maior frequência for citada mil vezes, a décima palavra da lista decrescente de frequências terá cem citações e a centésima palavra da lista terá dez citações. Para fins desta dissertação, adota-se a curva de Zipf, conforme representada na Figura 2.1, a seguir.

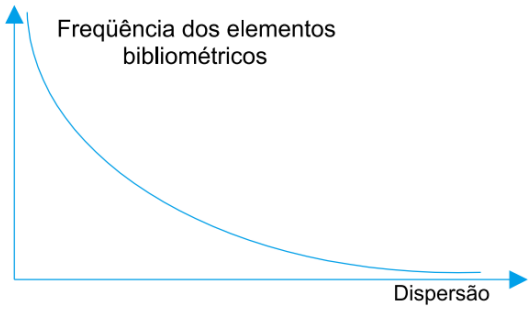

Figura 2.1 – Curva de Zipf
Fonte: Quoniam (1992).

Conforme classificação apresentada por Quoniam (1992), podem ser delimitadas na curva Zipf três zonas distintas, a saber:

- Zona I – informação trivial ou básica: define os temas centrais da análise bibliométrica;
- Zona II – informação interessante: localiza-se entre as Zonas I e III e mostra ora os temas periféricos, ora a informação potencialmente inovadora;
- Zona III – sinais fracos ou informação de ruído: essa zona tem como característica conter conceitos que podem ser emergentes (sinais fracos) ou apenas ruído estatístico.

A Figura 2.2 ilustra as três zonas de distribuição na curva de Zipf.

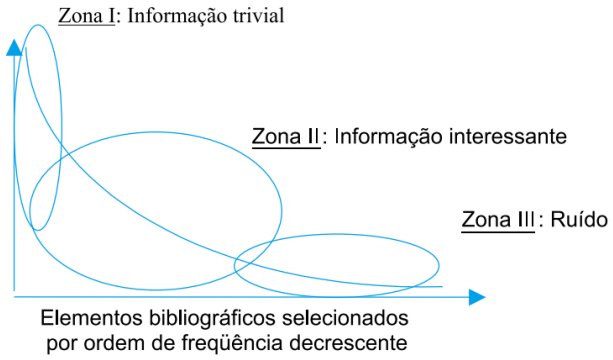

Figura 2.2 – Zonas de distribuição na curva de Zipf
Fonte: Quoniam (1992).

Para fins da presente pesquisa, a zona III da curva de Zipf representa sinais fracos referentes a tópicos emergentes e não ruído estatístico, conforme pode ser constatado no estudo de caso apresentado no capítulo 4, seção 4.6.

2.2.
Indicadores bibliométricos

A discussão sobre indicadores bibliométricos baseou-se fundamentalmente nos trabalhos de Potter, 1981; King, 1987; Sancho, 1990; Leydesdorff, 1991; Shapiro, 1992; Moed et al., 1995; Spinak, 1996; Robredo e Cunha, 1998; Dahal 1998; Maltrás Barba, 2003; Glänzel e Schubert, 2003; Moed et al. 2005; Hood e Wilson, 2001; Saes, 2005; Jung, 2008; Yi e Choi, 2012; e Jang et al., 2012.

Segundo Saes (2005), existem três grupos principais de indicadores:

- indicadores de tamanho e de características da produção científica;
- indicadores de impacto das publicações;
- indicadores relacionais de primeira e segunda geração.

Desses três grupos de indicadores, os indicadores relacionais – um dos focos desta pesquisa – permitem gerar os mapas representativos do conhecimento ou mapas tecnológicos (Yi e Choi, 2012). Já os dois grupos anteriores, denominados em conjunto como indicadores de atividade, expressam a contagem e a distribuição do número de publicações, instituições ou países e representam os indicadores bibliográficos mais triviais. Assim como a contagem dos autores, os organismos, revistas, artigos, empresas, patentes, temas ou datas de publicação também expressam a produtividade do elemento bibliográfico.

Independentemente da trivialidade, as evoluções temporais dessas medidas são interessantes, pois geram uma visão macro da aceleração desses indicadores, considerada de fundamental importância para processos decisórios e, em especial, como apoio a decisões estratégicas.

Vale observar que os indicadores de tamanho e de características da produção científica também são designados por indicadores de publicação e são responsáveis pela mensuração da qualidade e do impacto das publicações (Spinak, 1996; Dahal 1998; Sancho 1990).

Integram o terceiro grupo, os indicadores relacionais de primeira e segunda geração, incluindo citações, cocitações e co-ocorrência de palavras (King, 1987).

Buscando obter uma compreensão abrangente dos indicadores bibliométricos e de sua aplicabilidade, pesquisadores em vários campos do conhecimento têm utilizado análises bibliométricas, particularmente *"co-citation analysis*" (Price, 1965; Garfield, 1972; Small, 1973; Chen, 2004) e "*co-word analysis"* (Callon et al., 1983; Callon et al., 1991; He, 1999; Ding et al., 2001; e Chung e Han, 2009, dentre outros).

A "*co-citation analysis*" tem sido aplicada em vários campos para identificar a estrutura intelectual na qual o conhecimento científico interdisciplinar é compartilhado entre os diferentes campos. No entanto, a partir dos resultados de uma "*co-citation analysis*", é difícil visualizar a imagem global da pesquisa e compreender a estrutura do conhecimento em maior nível de detalhe (Jung, 2008). Já a "*co-word analysis"* permite a retenção das informações contidas nos dados e a geração de gráficos que constituem a base para a elaboração dos mapas representativos do conhecimento (Jang et al., 2012).

2.2.1. *"Co-word analysis"*

De acordo com Saes (2005), o "*co-word analysis*" é a análise construída a partir da co-ocorrência do valor de uma variável, ou seja, a ocorrência simultânea de dois ou mais valores de uma mesma variável tais como: palavras-chave, autor, ano ou até mesmo uma área de texto.

As palavras-chave são palavras utilizadas em indexação para indicar um acervo ordenado de informações e conhecimentos. Em co-ocorrência, a ocorrência simultânea indica uma relação de proximidade entre elas. Ou seja, os conceitos que representam estão conexos. É importante notar que uma fraca intensidade de co-ocorrência pode também sugerir que uma nova tecnologia ou um novo "*cluster*" esteja surgindo (Escorsa e Maspons, 2001).

Como comentado anteriormente, nos últimos anos diversos autores aplicaram a "*co-word analysis*" a estudos da situação ou da evolução de diversas áreas da ciência. Dentre eles, merecem destaque os trabalhos de Rip e Courtial, 1984; Law et al., 1988; Whittaker, 1989; Leydesdorff, 1991; Law e Whittaker, 1992; Peters e Van Raan, 1993; Cambrosio et al., 1993; He, 1999; Ding et al., 2001; Saes, 2005; Su e Lee, 2010; Kwon et al., 2010; Kim et al., 2011; Liu et al., 2011; Yi e Choi, 2012; Zong et al., 2013; e Cho, 2014.

A co-ocorrência compreende a detecção das palavras que caracterizam o conteúdo de um tema e a contagem das ocorrências dessas palavras nos trabalhos sobre o tema em foco. Em seguida, verifica-se a co-ocorrência dessas palavras, ou seja, que palavras aparecem simultaneamente nesses trabalhos e com que frequência. Dessa mensuração, extrai-se informação de alto valor agregado sobre a proximidade ou a distância entre essas palavras e, mediante o uso de *software* específico para esse tipo de análise, geram-se gráficos que constituem a base para a elaboração dos mapas representativos do conhecimento.

2.3. Ferramentas e *software* de análise bibliométrica

É importante contextualizar e observar que, genericamente, as matrizes de co-ocorrências são compostas por uma massa muito ampla de dados que, na maioria das vezes, são impossíveis de serem identificadas ou mesmo mapeadas por um ser humano. Em termos cognitivos, reconhecem-se as limitações humanas para filtrar informações dentro de uma expansiva massa de dados numérica ou textual. Assim, os gráficos surgem como uma resposta para essa carência humana facilitando a interpretação de dados e fornecendo informações valiosas.

Um exemplo da riqueza dos gráficos fica clara em uma hipotética análise das relações entre nós de milhões de dados dispostos em uma simples tabela e a mesma análise sendo obtida pela representação da informação em forma de gráfico que revela os vínculos e os nós.

Nesse contexto, um gráfico ou ainda um infográfico em conjunto com um pequeno texto que sublinha as relações estruturais, entrega um amplificado entendimento dos dados anteriormente complexos e sem sentido.

Saber se representações visuais de co-ocorrências são ferramentas úteis é determinar se, dentro do universo informacional, há relações essenciais que se deseja visualizar entre os nós. O objetivo é figurar dados associados às co-ocorrências e não conceber a estrutura em si. A estrutura e a conectividade não são importantes, já os dados associados com os links e os nós são a meta principal.

Vale ressaltar que esta pesquisa não busca um desenho da estrutura das co-ocorrências somente. Ela busca concluir sobre o que a representação gráfica das co-ocorrências revela sobre a estrutura e seus componentes.

Os *software* de análise bibliométrica proporcionam uma observação mais clara desses relacionamentos e uma melhor visualização dos resultados, produtos das relações geradas por dados em co-ocorrências.

No âmbito desta pesquisa, o *software* mais indicado foi o Pajek desenvolvido por Batagelj e Mrvar (1996). Embora não gere grafos automaticamente a partir da *Web*, o Pajek amplia os processos de representação e análise de co-ocorrências das palavras-chave ou "*co-word analysis*".

O Pajek 3.11 tem a capacidade de gerar mapas tecnológicos ou mapas representativos do conhecimento e fornece uma visualização da centralidade e da densidade dos *clusters* do PósMQI (Pajek, 2013).

2.4. Representação do conhecimento científico: léxico básico e mapas conceituais

Um léxico básico não se configura como um simples vocabulário controlado (Robredo e Cunha, 1998). No léxico de um campo ou área científica, as palavras não descrevem, mas significam. Cada palavra é um conceito. Desta forma, as palavras-conceitos não são mais palavras, mas sim "termos significativos". Essa é linguagem que será adotada na apresentação do caso do PósMQI no capítulo 4.

Os mapas representativos do conhecimento fornecem uma visão da estrutura de relações que existe em um domínio. Estes mapas evidenciam um cenário detalhado das linhas de pesquisa em uma área desejada, por meio da análise do que se está produzindo. Comparando mapas de diferentes períodos, permitem seguir a evolução, no tempo, das tecnologias e das linhas de pesquisa de um determinado Programa de Pós-graduação – foco desta pesquisa.

Os mapas baseados nas ocorrências permitem detectar as áreas, "clusters" ou projetos de pesquisa em que se subdivide um domínio, mas não nos permitem entrar no conteúdo dos documentos. Assim, é indispensável aproximar-se dos termos significativos, já que esses explicam o conteúdo dos próprios documentos. Por meio da "*co-word analysis*", do léxico básico e dos mapas representativos do conhecimento, é possível visualizar a estrutura do conhecimento que está por trás do documento e o que o tornou possível.

2.5. Reflexão sobre os instrumentos para avaliação da produção científica de programas de pósgraduação

Buscou-se neste capítulo descrever a base conceitual sobre representação do conhecimento e bibliometria, com o objetivo de fundamentar a discussão em torno das seguintes questões:

- Qual a contribuição da bibliometria e dos métodos de representação do conhecimento para a avaliação da dinâmica da produção científica de um programa de pós-graduação?
- Quais são as ferramentas de escolha para a avaliação da dinâmica da produção científica do Programa PósMQI, desde sua criação?

Essa base conceitual, principalmente o entendimento sobre as três leis da bibliometria e os fundamentos da "*co-word analysis'*, propiciou uma reflexão sobre a aplicação das leis e do referido método no contexto de programas de pós-graduação. Foi possível concluir que padrões estatísticos encontrados em bases de dados, periódicos, livros e demais formas de comunicação científica podem ser medidos por ferramentas bibliométricas. Podem ser medidos por variáveis distintas, tais como: citações, palavras-chave, autor, tema, localidade, etc. Logo, a bibliometria, como um método de pesquisa quantitativa, de análise e de estatística, permite gerar indicadores e modelos representativos de uma determinada área científica, além daqueles baseados em estatística descritiva, amplamente adotados para avaliar a produção científica de programas de pós-graduação. Destacam-se para fins de desta pesquisa as três leis da bibliometria, consagradas na literatura especializada. São elas: Lei de Lotka (produtividades de autores); Lei de Bradford (produtividade de periódicos); e Lei de Zipf (frequência de ocorrência de palavras). Especialmente, a Lei de Zipf mostra-se bastante interessante para identificação de termos significativos das publicações produzidas por um determinado programa de pós-graduação. Pretende-se, na fase de pesquisa aplicada, ilustrar e demonstrar sua importância na construção de uma léxico básico a ser proposto para o Programa PósMQI.

Não menos importante para fins da avaliação da dinâmica da produção científica de programas de pós-graduação é a "*co-word analysis*", que diversos

autores da área da Ciência da Informação têm aplicado para estudos da situação ou da evolução de diversas áreas da Ciência, como já citado.

A análise da co-ocorrência de palavras-chave das publicações geradas por um programa de pós-graduação consiste na detecção das palavras que caracterizam o conteúdo de um tema ou linha de pesquisa e a contagem das ocorrências dessas palavras nos trabalhos sobre esse tema ou dessa linha. Em seguida, verifica-se a co-ocorrência dessas palavras, ou seja, que palavras aparecem simultaneamente e com que frequência. Dessa mensuração, é possível extrair informação sobre a proximidade ou a distância entre essas palavras e, mediante o uso de *software* específico para esse tipo de análise, geram-se gráficos que constituem a base para a elaboração dos mapas representativos do conhecimento do referido programa.

Em relação à segunda questão, destaca-se para fins do estudo de caso do PósMQI a seguinte ferramenta: Pajek 3.11.

O Pajek 3.11, será o *software* de escolha para gerar resultados significativos na fase da pesquisa aplicada, pois permitirá construir mapas representativos do conhecimento, possibilitando a visualização da centralidade e da densidade dos *clusters* dos termos significativos das dissertações do PósMQI defendidas no período de 1999 – 2013.