ARTIGO

A bibliometria na exploração de bases de dados: a importância da Lingüística[1]

Bibliometric in databases exploration: the importance of Linguistics

Rogério MUGNAINI[2]

RESUMO

A utilização das bases de dados para levantamento do estado da arte, procedimento necessário a qualquer pesquisador, exige dos mesmos a definição de estratégias para a recuperação eficaz da informação. A Bibliometria pode ser uma ferramenta útil neste processo, permitindo a filtragem de grandes quantidades de informação. Por se tratar de uma análise estatística de dados, a qualidade destes dados é de vital importância e o procedimento de indexação para representação da informação se torna essencial. Pretende-se aclarar a importância da Estatística, Lingüística e Indexação para a Ciência da Informação, focalizando alguns aspectos de suas relações e destacando a necessidade da utilização conjunta dessas disciplinas.

Palavras-chave: bibliometria, estatística, indexação, linguagem documentária.

ABSTRACT

The use of databases to research the state of the art, the necessary procedure for any researcher, imposes on them strategies definition of efficient information

[1] Trabalho apresentado à disciplina Fundamentação Lingüística no Tratamento da Informação, sob a orientação da Professora Doutora Else Benetti Marques Válio. Curso de mestrado em Biblioteconomia e Ciência da Informação, PUC-Campinas. O trabalho foi apresentado no Seminário em Ciências da Informação (Londrina, 23-25 ago., 2001).

[2] Estatístico, Mestre em Ciência da Informação e Biblioteconomia pela PUC-Campinas. Analista de Sistemas de Informação, Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde/BIREME. Rua Botucatu, 862, Vila Clementino, 04023-901, São Paulo, SP, Brasil. *E-mail*: rogerio@bireme.br

Recebido para publicação em 10/12/2002 e aceito em 20/8/2003.

retrieving. Bibliometrics can be a useful tool in this process, allowing the filtering of great amount of information. Since it is a statistical analysis of data, its quality is vitally important and the indexing methods used by Linguistics become essential. This work intends to clarify the importance of the statistics for the Information Science focusing alll these aspects and emphasizing the joint utilization of these disciplines.

Key words: *bibliometric, statistic, indexation, documentary languages.*

INTRODUÇÃO

O acúmulo e armazenamento de informação possibilitados pelo desenvolvimento da informática suscitaram a necessidade de sua organização para posterior recuperação. A informação produzida no meio acadêmico, dita formal, passou a ter um valor agregado extra quando vista como indicador direto da produção do conhecimento humano.

O conhecimento, que só é adquirido à medida que se assimilam as informações sobre o assunto de interesse, utiliza a linguagem como veículo de comunicação, objetivando a reconstrução do pensamento através de um *conjunto de unidades de expressão*, segundo Cabré (1995), denominado terminologia.

Com o passar do tempo, a produção ininterrupta de conhecimento armazenado começa a dificultar a tarefa do pesquisador, que busca através da leitura, montar o cenário do tema relacionado à sua pesquisa. Algumas técnicas estatísticas podem ser aplicadas para filtragem de informação, com o objetivo de facilitar sua busca. A Bibliometria[3] é aqui proposta como a ferramenta capaz de medir e facilitar a análise da informação armazenada. Segundo Oliveira (1984), embora seu uso acarrete algumas dificuldades, muita informação útil concernente à transmissão de idéias, crescimento e tendências de uma disciplina científica pode ser obtida por meio da abordagem matemática.

A Bibliometria, que hoje também desfruta das regalias da era digital, utilizando como insumo a produção científica indexada nas bases de dados, é dependente direta do resultado do trabalho dos indexadores. Sendo assim, a Lingüística passa a ser cooperadora indireta na análise quantitativa da produção científica, fornecendo modelos para a indexação, que buscando representar a informação, pode ser vista como forte participante no processo de validação da mesma.

Este trabalho busca levantar argumentos sobre a importância da aplicação de técnicas lingüísticas na Indexação, para a realização de análises bibliométricas de bases de dados. Traz também algumas considerações baseadas nas técnicas estatísticas de análise, pelo fato desta área ter muito a contribuir na implementação das técnicas bibliométricas. Apresenta ainda uma breve revisão de literatura da produção científica nacional em estudos bibliométricos, sem a intenção de levantamento exaustivo.

Representação de informações de Bases de Dados através de linguagem própria

Estudos em Lingüística, voltados para a área de Documentação, vêm sendo realizados, pelo menos, desde a década de 70. Num artigo publicado em 1973, Wanderley (1973) faz a seguinte afirmação:

[3] A bibliometria consiste em analisar, globalmente, os elementos de um *corpus* documentário, com a ajuda de métodos estatísticos e matemáticos, a fim de descobrir as relações existentes entre os diversos elementos que o compõe (SANTOS, 2001).

> Então não teríamos estado, nós, bibliotecários e documentaristas, a fazer lingüística sem o saber – quase diríamos sem a saber? – ao criarmos linguagens que desejamos tão significativas quanto a natural? a traduzirmos em códigos os textos em nossa custódia? a frasear-lhes o conteúdo em resumos sinaléticos ou informativos, produzindo documentos derivados que dão acesso aos primários, numa palavra, a comunicarmos ao leitor mensagens que, no intuito de lhe responderem às questões, os autores confiaram ao nosso zelo e guarda?

O autor exprime, nestas linhas, a surpresa, como se estivesse se dando conta de estar utilizando a Lingüística em seus procedimentos de documentação. As linguagens criadas por estes procedimentos têm realmente o objetivo de representar a linguagem natural, isto é, o texto integral. E, como afirma também, cabe ao "documentarista" garantir esta representação. Para Cintra (1983), a linguagem documentária foi criada para suprir a necessidade de descrição dos documentos, por causa das dificuldades encontradas com a linguagem natural.

As bases de dados utilizam a linguagem documentária para identificação do documento, com o objetivo de possibilitar sua recuperação com eficiência. Este ato de indexar informação, tratado por Navarro (1988) em uma revisão de literatura sobre a interface da Indexação[4] com a Lingüística de 1988, era realizado na base do bom senso, passando a buscar esta interface somente no momento em que se iniciaram as tentativas de automação do procedimento.

A autora afirma, ainda, que o cientista da informação não precisa da Lingüística em geral, mas sim de algumas escolas, pela necessidade operacional do documentarista. Baseando-se no trabalho de vários autores, confere extrema importância à Semântica, por considerar a questão do significado das palavras um problema de difícil resolução; e também à Sintaxe (trata dos constituintes da frase), sobre a qual os estudos, bem como sua utilização em documentação encontram-se mais desenvolvidos que os referentes à Semântica.

Baranow (1983) destaca subáreas especializadas da Lingüística e áreas afins, interessantes à Documentação, sendo, Navarro (1988), o autor que mais indica possíveis áreas: a Terminologia, que auxilia na estruturação dos assuntos, por exemplo, a Lingüística Computacional, com o processamento automático de textos e a recuperação automática da informação; a Tradutologia, que pode servir de base para a tradução da linguagem natural para a documentária; e a Morfologia, que aponta o conteúdo do texto baseando-se na freqüência das palavras. Cintra (1983) e Navarro (1988) ainda acrescentam: a teoria dos Campos Semânticos, e por fim a Leitura do Documento, ou Análise do Discurso, expressando o momento de leitura do documento, pelo indexador.

No processo de automação, o texto passa a ser processado pela máquina (ANDREEWSKI, 1983; EVANGELISTA JÚNIOR, 1996), que exige a identificação de aspectos lingüísticos por não discernir a semântica das palavras. Já em um artigo mais recente, Mendonça (2000) embasa esta afirmação:

> A tradução do termo descritor e a problemática do empréstimo lingüístico são dois dos grandes desafios da era documental, pois prima-se por um termo que represente, de maneira clara e objetiva, a informação por meio da normalização e padronização

[4] Indexação: "é a tradução de um documento de uma para outra linguagem, visando possibilitar a recuperação da informação, sendo que estão envolvidos alguns processos, entre os quais os lingüísticos" (NAVARRO, 1988).

> dos empréstimos lingüísticos que atualmente assolam a terminologia brasileira.

Os descritores são os termos eleitos para representar o documento. Sayão (1996) denomina-os a metáfora da informação original, sendo produzidos por linguagens artificiais, e objetivando a referência a um conhecimento real. Diz ainda que, por sua artificialidade, estas linguagens são extremamente redutoras de significado, porém, alerta sobre o fato de que antes de ser uma deficiência, é o sustentáculo da identidade, do poder de ordenação e classificação, do qual a ciência não pode prescindir.

Espera-se assim que haja uma relação unívoca nessa representação, o que nem sempre é atingido na tradução das "metáforas". Algumas vezes, devido a problemas de empréstimos lingüísticos, os termos não são compreendidos da forma pretendida. Por exemplo, a utilização nas ciências sociais, de palavras com significados diversos, resultantes da atribuição realizada por determinado grupo, que nem sempre é entendida pelos de fora do mesmo, e em alguns casos, nem por seus próprios integrantes (GALVÃO, 1998).

Adotando a posição de Sayão, pode-se dizer que as linguagens das bases de dados são de extrema importância pelas características oferecidas para organização da informação. Obviamente apresenta desvantagens, como, por exemplo, a redução de significado, mas que de certa forma é compensada pelas facilidades oferecidas para sua recuperação.

A Bibliometria abrindo o caminho

O primeiro contato entre o pesquisador e a produção científica, para constituição de seu trabalho, é de fundamental importância. Sayão (1996) vê este contato como o momento em que o pesquisador reconstrói seu conhecimento. A partir das informações que ele encontra na base de dados é que poderá sanar sua necessidade de informação, até então nebulosa. O pesquisador utiliza os testemunhos dados pela comunidade científica ou acadêmica, em que ele está ou deseja estar inserido, para formular suas próprias questões.

A Bibliometria pode ser usada como um filtro de informação, podendo ser de grande utilidade para o pesquisador no levantamento do estado da arte do seu tema de pesquisa. A leitura do texto integral se torna impossível devido ao crescimento do conhecimento produzido, que acontece rapidamente, sendo de forma exponencial em alguns casos.

Rostaing (1996) compara a Bibliometria e a Demografia afirmando, que para o analista não tem sentido ler as publicações a analisar, assim como o demógrafo não conhece os indivíduos da população que estuda. Em sua comparação ele deixa claro a não dependência do conhecimento completo do objeto de estudo (num primeiro momento, no caso do analista), sendo necessário apenas conhecer algumas características do mesmo. O que se deve ao fato de lidar com quantidades, analisando o grupo em geral. Sugere-se esta "ferramenta" por sua eficácia na abordagem das bases de dados.

Wormell (1998) frisa o aumento da potencialidade da Bibliometria propiciado pelos mecanismos de busca *online* e a possibilidade da mesma ser aplicada não só a bases de dados de produção científica, mas a bases diversas, quando passa a ser denominada informetria[5] . E acrescenta que:

> O valor sofisticado do serviço de informação *online*, hoje, está no uso de bases de dados não somente para recuperar informações, mas também para anali-

[5] Informetria é o estudo dos aspectos quantitativos da informação em qualquer formato, e não apenas registros catalográficos ou bibliográficos, referente a qualquer grupo social, e não apenas aos cientistas (SANTOS, 2001).

> sar / sintetizar os resultados e combiná-los com outras informações (garimpo de dados). Assim, a técnica de busca online deve ser compreendida como um processo que agrega valor em termos dos procedimentos de seleção e refinamento realizados com base em estratégias de busca inteligentes.

Dessa forma destaca-se aqui o final do depoimento de Wormell (1998), no qual ela menciona a agregação de valor, através das buscas inteligentes. Ou seja, entende-se que as estratégias de busca booleana têm grande importância na definição do material a ser estudado, e é necessário que se utilize os devidos campos de busca, para obtenção dos documentos.

As características dos documentos indexados na base podem ser divididas em dois tipos. As descritivas, correspondentes ao "envelope" do documento, por não fazer parte de seu conteúdo. Trata-se de dados do autor, do co-autor, da instituição em que foi produzido, citação, data de publicação, editora, local de publicação. Essas características podem revelar informações importantes sobre o documento, por exemplo: sabendo que determinada instituição trabalha em determinada linha de pesquisa, fornecendo indicações sobre o tema, ou a forma de abordagem ao mesmo; a mesma coisa pode ser esperada se o autor já é conhecido, ou seu país de origem.

O outro tipo refere-se às características temáticas do documento, que consiste do título, do resumo e dos descritores, provenientes de uma linguagem documentária. Estas também evitam que o analista tenha de ler o texto, porém, fornecem informações diretamente relacionadas com seu conteúdo. A utilização destas características, numa análise bibliométrica, ajuda a limpar a grande massa retirada da base, permitindo a redução do trabalho de leitura do pesquisador, e não a eliminação.

Este trabalho atenta primordialmente para o segundo tipo de características por ter uma abordagem lingüística, e por serem estas mais indicadas quando se utiliza a Bibliometria para o fim tratado aqui. Mas as características descritivas são de extrema importância numa análise bibliométrica.

Só a Lingüística é culpada?

Vendo a análise bibliométrica como uma análise estatística, deve-se atentar primeiramente para a definição das variáveis a analisar. A qualidade dos dados será diretamente refletida no resultado da análise, o que reforça, neste caso, a questão da importância das técnicas lingüísticas para sua validação.

Os dados serão resultantes da medição da variável de interesse (nem sempre mensurável diretamente), através de uma medida indicadora do mesmo. Jannuzzi (2001) explica que a validade da medida indicadora é obtida quando se aproxima do conceito (variável de interesse), ou seja, é a sua capacidade de refletir, de fato, o conceito abstrato a que o indicador se propõe a "substituir" ou "operacionalizar".

Os descritores são os indicadores do conteúdo do documento, também responsabilizados pela qualidade da análise. A má representação do documento, por parte do descritor, resumo, ou título, significará a não inserção do mesmo no conjunto, ou a inclusão desnecessária de documentos que não deveriam constar. Um documento não inserido é chamado, na linguagem estatística, de "dado faltante".

Deming (1950), elencando uma série de possíveis fontes de erros não-amostrais, afirma que este tipo de falha no conjunto de dados pode ser destacado entre os mais prejudiciais à análise estatística, pois não só representa perda de parte da informação sobre o tema, como gera distorção no resultado final da análise. Outra fonte apontada por ele é a demora na inserção

do documento na base de dados, prejudicando o resultado final da análise.

Numa análise estatística, recomenda-se considerar o maior número possível de fontes de erro. Como são infinitas, não se pode ater aos mínimos detalhes, e saber que não existe uma melhor análise, e sim uma análise melhor que outra é importante. Dado que as fontes de erro nem sempre poderão ser extintas, as medidas serão tomadas enquanto forem consideradas no momento da interpretação dos resultados. As fontes de erro têm de ser destacadas aos olhos daquele que receber os resultados.

Baseando-se nestes fatos, o analista de informação precisa ser esclarecido de que sua amostra nem sempre é representativa da população que a proveu, pois este, muitas vezes, está lidando com uma base que não contempla exaustivamente toda a produção científica daquele tema. É, portanto, essencial que ele exponha, a exemplo desta, todas as precariedades do conjunto de dados levantado.

Um outro aspecto importante a esclarecer é a estratégia de busca das informações, como apontado por Wormell (1998). As palavras utilizadas e ordem de utilização, quais e como foram usados os operadores booleanos. Este já é um procedimento lógico que constitui a estratégia.

CONSIDERAÇÕES FINAIS

Discorrendo sobre a construção de conceitos no campo da Ciência da Informação, Galvão (1998) diz poderem ser distinguidas duas correntes teóricas que estudam a linguagem no campo da Biblioteconomia e Documentação: uma pesquisa a viabilidade de recuperação da informação via linguagem natural, e outra trabalha no desenvolvimento das linguagens documentárias. Além disso, também observa que tem buscado auxílio de outras disciplinas como a Estatística, a Ciência Cognitiva, a Lógica, a Lingüística, a Matemática, a Ciência da Terminologia, entre outras.

As duas correntes observadas por Galvão (1998) também podem ser nitidamente identificadas nos estudos bibliométricos. Rostaing (1996) diz que os tratamentos estatísticos lingüísticos foram inicialmente desenvolvidos para analisar textos literários, passando a ser utilizados pela Bibliometria para aplicação em textos científicos. E que este "parentesco" é de certa forma "confirmado", pois numerosos centros de pesquisa em Bibliometria têm como objetivo o tratamento automático de textos redigidos em linguagem natural.

A justificativa, apresentada aqui para a não utilização de recuperação por linguagem natural, é a facilidade oferecida pela linguagem documentária, que com certeza não evita a leitura, mas reduz sua quantidade.

O artigo já citado de Mendonça (2000), que busca estabelecer relações entre a Lingüística e a Ciência da Informação, recuperou os artigos que tivessem o termo lingüística no título ou resumo, em periódicos relevantes na área de Ciência da Informação. Dentre os artigos selecionados pela autora, a revista Ciência da Informação foi escolhida por apresentar o maior número de ocorrências de artigos com estas características (de 1972 a 1998). Este artigo será utilizado neste estudo por servir como uma revisão bibliográfica de interesse comum.

Dentre os grupos temáticos de classificação dos 42 artigos, está "Lingüística e Bibliometria", representando o grupo de estudos quantitativos. Apenas dois artigos foram listados: um de Maia (1973) e outro de Ribeiro (1974). Os dois utilizam a Bibliometria para análise do texto integral. Desta forma, nota-se que, pelo menos neste periódico, não foram mais publicados artigos nesta interseção de assuntos até 1998. Sabe-se, obviamente, que esta amostra está

longe de ser representativa da população completa da produção em âmbito nacional, e nem possui este objetivo, mas serve para evidenciar a carência neste campo da ciência nacional.

Complementando o estudo de Mendonça (2000) e Oliveira (1984), também já citado, afirma que nenhum trabalho sobre Bibliometria foi publicado antes de 1972, e que o pioneiro foi uma tese de Braga (1974), publicada na revista Ciência da Informação, sob orientação de Saracevic, T. em sua estada no Brasil (OLIVEIRA, 1984) aponta este fato como o precursor da Bibliometria no Brasil).

Pode-se notar que a carência nessa área ainda é clara e espera-se que com a utilização das bases de dados como vem ocorrendo, perceba-se a utilidade da Bibliometria como exposto neste trabalho. Além disso, a interdisciplinaridade que a Ciência da Informação vem apresentando, envolvendo a Lingüística e Ciências Computacionais, pode ser enriquecida com profissionais da área de Estatística, principalmente quando se tratando de Bibliometria. Este é mais um fator que expressa a necessidade da divulgação da Ciência da Informação, para conhecimento ou estudo da mesma em outras áreas.

REFERÊNCIAS

ANDREEWSKI, A. Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à Língua Portuguesa. *Ciência da Informação*, Brasília, v.12, n.1, p.61-73, 1983.

BARANOW, U.G. Perspectivas na contribuição da Lingüística e de áreas afins à Ciência da Informação. *Ciência da Informação*, Brasília, v.12, n.1, p.23-35, 1983.

BRAGA, G.M. Informação, ciência, política científica: o pensamento de Derek de Solla Price. *Ciência da Informação*, Rio de Janeiro, v.3, n.2, p.155-177, 1974.

CABRÉ, M.T. La terminología hoy: concepciones, tendencias y aplicaciones. *Ciência da Informação*, Brasília, v.24, n.3, p.289-298, 1995.

CINTRA, A.M.M. Elementos de Lingüística para estudos de indexação. *Ciência da Informação*, Brasília, v.12, n.1, p.5-22, 1983.

DEMING, W.E. *Some Theory of Sampling*. New York: Dover Publications, 1950.

EVANGELISTA JÚNIOR, C. *Metodologia para a indexação automática de textos em língua portuguesa*. 1996. 122 f. Dissertação (Mestrado em Informática) – Instituto de Informática, Pontifícia Universidade Católica de Campinas, Campinas.

GALVÃO, M.C.B. Construção de conceitos no campo da Ciência da Informação. *Ciência da Informação*, Brasília, v.27, n.1, p.46-52, 1998.

JANNUZZI, P.M. *Indicadores sociais no Brasil*. Campinas: Alínea, 2001.

MAIA, E.L.S. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. *Ciência da Informação*, Rio de Janeiro, v.2, n.2, p.99-138, 1973.

MENDONÇA, E.S. A Lingüística e a Ciência da Informação: estudos de uma interseção. *Ciência da Informação*, Brasília, v.29, n.3, p.50-70, 2000.

NAVARRO, S. Interface entre Lingüística e Indexação: revisão de literatura. *Revista Brasileira de Biblioteconomia e Documentação*, São Paulo, v.21, n.1/2, p.46-62, 1988.

OLIVEIRA, S.M. Aplicações e limitações dos processos bibliométricos. *Revista Brasileira de Biblioteconomia e Documentação*, São Paulo, v.17, n.1/2, p.43-54, 1984.

RIBEIRO, L.A. Aplicação dos métodos estatísticos e da teoria da informação e da comunicação na análise lingüística: estudo da linguagem jornalística. *Ciência da Informação*, Rio de Janeiro, v.3, n.2, p.151-154, 1974.

ROSTAING, H. *La bibliométrie et ses techniques*. Marseille: Sciences de la Société, 1996. p.7-19.

SANTOS, R.N.M. Material de aula da disciplina Recursos Informacionais na Gestão de Negócios, do curso de mestrado em Ciência da Informação da Pontifícia Universidade Católica de Campinas, Campinas, 2001.

SAYÃO, L.F. Bases de dados: a metáfora da memória científica. *Ciência da Informação*, Brasília, v.23, n.3, p.314-318, 1996.

WANDERLEY, M.A. Linguagem documentária, acesso à informação: aspectos do problema. *Ciência da Informação*, Rio de Janeiro, v.2, n.2, p.175-217, 1973.

WORMELL, I. Informetria: explorando bases de dados como instrumentos de análise. *Ciência da Informação*, Brasília, v.27, n.2, p.210-216, 1998.