ESTUDOS BIBLIOMÉTRICOS SUPORTADOS POR TECNOLOGIA DE BANCO DE DADOS COM INTERFACE GRÁFICA, DESENVOLVIDOS A PARTIR DE SOFTWARES LIVRES

Jorge Fernando Maxnuck Soares[1], Silvio Popadiuk[2], Takato Kurihara[3],Tiago Novaes Mathias[4]

Abstract – *This paper presents some concepts about bibliometric studies based on data base technoclogy with graphic interface, both developed from free softwares. Such Project discusses the main bibliometric indexes, either under the aspect of data input, as well as of data output, brought by article about Balanced Scorecard (BSC) based on JSTOR, creating an experimental data base. The studies priorize the evaluation of the use of bibliometric indexes in the area of citation and co-citation. The application developed in this project to store and recover bibliometric data, with some adaptation may be used for several other purposes.*

Index Terms — bibliometria, Lei de Bradford, interface gráfica, software livre

1. INTRODUÇÃO

A Bibliometria pode ser definida como um conjunto de medições e princípios aplicados mediante análise estatística, os quais procuram mapear a contribuição científica em uma determinada área, autor, periódicos ou publicações. Para cada necessidade são utilizados, por exemplo: análise de citações, co-citação, vida média de citações, dentre outros. O tratamento destes indicadores são suportados pelas leis de Bradford, Lotka e Zipf, onde cada uma dela procura abordar um aspecto dos indicadores bibliométricos.

A Lei de Bradford, define o grau de relevância de periódicos em dada área do conhecimento, partindo do princípio que os periódicos que produzem o maior número de artigos sobre dado assunto formam um núcleo de periódicos, supostamente de maior qualidade ou relevância para aquela área.

A Lei de Lotka considera que alguns pesquisadores, supostamente de maior prestígio em uma determinada área do conhecimento, produzem muito e muitos pesquisadores, supostamente de menor prestígio, produzem pouco.

As Leis de Zipf, permitem estimar as freqüências de ocorrência das palavras de um determinado texto científico e tecnológico e a região de concentração de termos de indexação, ou palavras-chave**,** que um pequeno grupo de palavras ocorre muitas vezes e um grande número de palavras é de pequena freqüência de ocorrência [1].

No presente projeto busca-se estudar o uso de *softwares* livres no estudo e apresentação gráfica de dados com o uso da Lei de Bradford, para um maior entendimento quanto à disseminação, evolução e tendências, da produção acadêmica sobre o tema *Balanced Scorecard* (BSC) nos últimos anos, que em última análise é o objeto de pesquisa.

Este tipo de trabalho serve para estabelecer relações, não tão evidentes, de um dado artigo(s) e/ou autor(es) com outros autores do referencial bibliográfico do mesmo. Este incremento do conhecimento científico compõe a chamada análise transversal, dada pelo método da citação e co-citações de autor(es), sendo uma ferramenta robusta na análise bibliométrica utilizando-se a Lei de Bradford [2].

Este projeto foi desenvolvido com o apoio financeiro do Mackpesquisa da Universidade Presbiteriana Mackenzie, (UPM).

2. BREVE REVISÃO SOBRE BIBLIOMETRIA

A Bibliometria é um campo da ciência da informação, que infere sobre a produção bibliográfica de um determinado autor, tentando assim medir a produtividade de cada autor e assim criar métodos de comparação entre vários autores. Outra medida que é utilizada também pode ser medida pelo numero de citações que foram feito do artigo original que demonstra mais sobre a qualidade do artigo. Ou ainda a Bibliometria é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada.

A Bibliometria desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisões [3].

[4] afirmam que os indicadores científicos são cada vez mais necessários para os que formulam as políticas científicas no plano nacional, os quais devem determinar as prioridades de pesquisa entre e dentro dos mais diversos campos científicos.

[1] Jorge Fernando Maxnuck Soares, Professor da Faculdade de Computação e Informática da Universidade Presbiteriana Mackenzie, Rua da Consolação, 930, Prédio 13, 01302-907, São Paulo, SP, Brasil, maxnuck@mackenzie.br

[2] Silvio Popadiuk, Professor de Mestrado/Doutorado em Administração de Empresas da Universidade Presbiteriana Mackenzie, Rua da Consolação, 930, Prédio João Calvino, 01302-907, São Paulo, SP, Brasil, spopadiuk@mackenzie.br

[3] Takato Kurihara, Professor e Coordenador de Extensão da Faculdade de Computação e Informática da Universidade Presbiteriana Mackenzie, Rua da Consolação, 930, Prédio 13, 01302-907, São Paulo, SP, Brasil, takato@mackenzie.br

[4] Tiago Novaes Mathias, Aluno de Mestrado em Engenharia Elétrica da Universidade Presbiteriana Mackenzie, Rua da Consolação, 930, Prédio 13, 01302-907, São Paulo, SP, Brasil, novaes06@gmail.br

Já para [5] as vantagens da Bibliometria consistiriam em amenizar os elementos de julgamento e produzir resultados quantitativos que tendessem a ser a soma de muitos pequenos julgamentos e apreciações realizados por várias pessoas. Como uma grande parte da produção científica torna-se conhecida através da sua publicação, fica mais fácil a avaliação das atividades de pesquisa por meio desta.

Além da tradicional avaliação por pares, [6] destaca que uma forma de avaliar a qualidade de uma publicação consiste em verificar o nível de interesse dos outros pela pesquisa. O método mais simples para obter esta medida se dá por meio da quantidade de citações dessa pesquisa na bibliografia ulterior.

Conforme [7] entende-se como Bibliometria o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. Desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisão.

Os dados de citações categorizados por periódicos e publicados em forma de indicadores no *Journal Citation Reports* (JCR) do *Institute for Scientific Information* (ISI) passaram a ser usados como parâmetro de avaliação de pesquisadores e instituições. São publicados anualmente no JCR três indicadores, por título de periódico: o Índice de Citação Imediata (*Immediacy Index*), a Meia-Vida das Citações (*Cited Half-Life*) e o Índice Bibliométrico mais conhecido e utilizado, o Fator de Impacto (*Impact Factor*).

À medida que o fator de impacto (FI) foi se estabelecendo como critério de avaliação de publicações nas mais diversas instâncias, foi crescendo igualmente o número de pesquisas que passaram a investigá-lo sob variadas formas e, invariavelmente, a criticá-lo como indicador de influência das publicações. [8] afirma que o FI representa os resultados finais de um conjunto complexo de variáveis e as pessoas que consultam o FI não podem desconsiderar esta complexidade.

De acordo [9] pode-se definir a Bibliometria como: o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. A Bibliometria desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisões. [7],

Autores como [10], que buscam reinstalar a supremacia da Bibliometria, não ajudariam muito a resolver a confusão teórica e o caos terminológico no que diz respeito ao âmbito, ações e inter-relações entre as três especialidades. Seria fundamental ter-se bem esclarecidas as suas funções, formas de utilização de acordo com os conceitos, amplitude, objeto de estudo, métodos, qualificação e características próprias. Como apregoa [11], devemos consolidar a idéia de que todas estas especialidades encontram sua ligação a partir da sua origem comum: todas elas indubitavelmente constituem-se em subconjuntos da Ciência da Informação. Considerando-as desta maneira, permitem compreender mais satisfatoriamente a aplicabilidade de cada uma e a utilidade que oferecem para as mais diversas áreas do conhecimento.

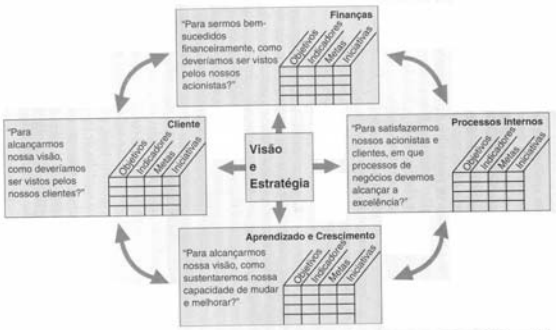
3. BREVE REVISÃO SOBRE O BALANCED SCORECARD (BSC)

As empresas atualmente estão inseridas em um ambiente competitivo muito complexo, onde é fundamental que exista uma perfeita compreensão de seus objetivos e metas e dos métodos para alcançá-los.

O Balanced Scorecard (BSC) é uma metodologia que visa orientar a estratégia das empresas. Trata-se de um modelo de gestão que auxilia as organizações a traduzirem a sua missão e estratégia em objetivos, indicadores, metas e iniciativas que direcionem as decisões corporativas e a condução dos negócios da empresa para um fim comum. É um conjunto abrangente de medidas de desempenho que serve de base para um sistema de medição e de gestão estratégica.

O desempenho organizacional das empresas é então medido sob as quatro perspectivas, conforme pode-se observar na FIGURA 1, e tais perspectivas formam a estrutura do *Balanced Scorecard* [12].

FIGURA 1: Estrutura do *Balanced Scorecard*



O *Balanced Scorecard* deve traduzir a missão e a estratégia de uma empresa em objetivos e medidas tangíveis.

4. BREVE REVISÃO SOBRE BANCO DE DADOS

Segundo [13] software não é apenas um programa de computador, mas sim toda a configuração de sistema e documentação de software. A engenharia de sistemas é o meio de especificar, projetar, implementar, validar, implantar e manter o sistema em operação. Neste projeto, os componentes de comunicação e interface são implementados pelo próprio *framework* CakePHP. O componente de computação é a classe ReferencialComponent e os componentes de sensores podem ser consideradas as classes que incluem algo no sistema que são chamadas pelo *framework* de *Controllers*.

Para [14], a meta principal de um *data warehouse* é a criação de uma visualização lógica única de dados que residem em diversos bancos de dados físicos. Isso fornece aos desenvolvedores e usuários um modelo de trabalho

único dos dados da empresa, algo que está ausente em virtualmente toda organização.

Segundo [15], as ferramentas necessárias para o povoamento do *data warehouse* podem ser divididas em:
a) Extração, filtragem, transformação e migração.
b) Transferência de dados e replicação.
c) Gerenciamento e administração.

5. DESENVOLVIMENTO DO SOFTWARE

Neste trabalho foi utilizado o *framework* CakePHP, baseado na linguagem de programação PHP. Segundo o *site* oficial, CakePHP é um *framework* para desenvolvimento rápido que utiliza padrões *Model View Controller* (MVC) e *Object-Relational Mapping* (ORM).

Segundo [16], o CakePHP foi projetado para se utilizar mais convenções do que configurações. Por isso, foi seguido o padrão, que será explicado melhor mais adiante, na construção do banco de dados garantindo, assim, um desenvolvimento mais ágil, seguro e fácil de dar manutenção conforme necessário.

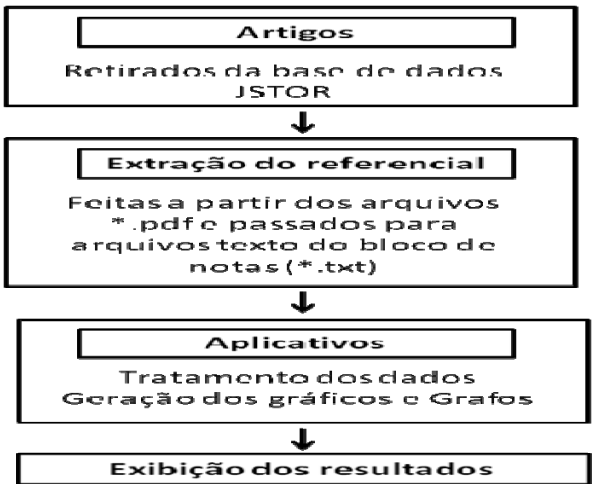
Para gerar os gráficos, foi utilizado a biblioteca PHPLot, distribuída gratuitamente sob a licença *GNU Public License* (GPL) e desenvolvida por L J Bayuk, Miguel de Benito e Afan Ottenheimer.

Nos grafos, foi utilizada uma classe chamada Image_Graphviz, também gratuita disponível no *PHP Extension and Application Repository* (PEAR). Segundo [17], o PEAR contém centenas de classes e extensões que ajudam os desenvolvedores a adicionar novas funcionalidades em suas aplicações. Esse reuso é importante pois [13] diz que com o reuso, menos componentes têm de ser especificados, projetados, implementados e validados. Após ser passado esses parametros, a classe Image_Graphviz gera um código intermediário e chamada ao programa Graphviz, instalado na máquina onde o código está rodando, que irá gerar a imagem com o grafo.

Segundo [18], a modelagem de um banco de dados é muito importante, pois toda a aplicação será dependente dessa estrutura. Qualquer alteração no banco de dados deverá ser seguida por alterações na aplicação. Portanto, o banco de dados foi modelado no início do projeto para que a aplicação fosse feita sem muitas alterações no decorrer do seu desenvolvimento.

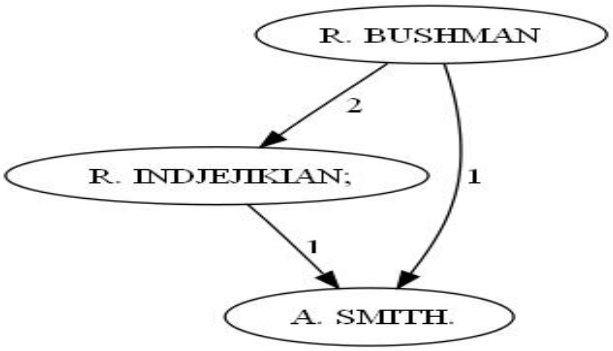
Para o desenvolvimento dos aplicativos seguiu o diagrama da FIGURA 2. Nele pode-se identificar quatro grupos, sendo estes:
a) Entrada, composto pelos artigos pertinentes ao assunto BSC;
b) Extração de informação, efetuada a partir dos arquivos de tipo *.pdf e transportados para arquivos texto (*.txt);
c) Aplicativos, onde foram desenvolvidos alguns programas e utilizados softwares livres para o tratamento dos dados retirados dos artigos;
d) Exibição dos resultados, onde pode-se verificar os indicadores bibliométricos por meio de imagens do tipo grafo e gráficos.

FIGURA 2: Fluxo de informações



A FIGURA 3, mostra um exemplo de como fica a imgem final:

FIGURA 3: Grafo de Indicadores Bibliométricos



A modelagem do banco de dados para este projeto foi feita visando uma separação entre artigos e referenciais. Também foi adotado o seguinte padrão na construção do mesmo, baseado nos padrões para utilização do CakePHP:

Todos os itens do referencial devem conter seções ,nessa ordem: Autor(es); Título do artigo; Ano do artigo.

Segundo [19], dados são fatos brutos para a construção de blocos de informações. Todos os dados obtidos através do sistema JSTOR são guardados no banco de dados na sua forma mais básica, para que posteriormente possa ser recuperado e tratado.

Ainda segundo [19], o objetivo principal de um SGBD é guardar e recuperar informações de maneira eficiente. O banco de dados deste projeto foi feito pensando nessa eficiência, fazendo uma separação entre trabalhos cadastrados e trabalhos referenciais. Com essa separação, nenhuma tabela fica muito sobrecarregada de informações o

que resultaria numa demora muito grande para a recuperação dos dados na hora de gerar os grafos.

[20] diz que é preciso especificar como será determinado as linhas que serão recuperadas, e isso é feito especificando uma condição que será aplicada a cada linha.

No sistema, a geração de grafos é de responsabilidade da classe *GraficoController*, nela é executada as *querys* necessárias para pegar e tratar os dados, somente após o tratamento deles, é colocado num *array*, este *array* vai para uma *view*, que com estes dados já tratados apenas gera os grafos.

Após a execução do programa, ao selecionar no sistema a visualização de, por exemplo, autor referencial, o sistema executa a *query* para pegar todos os autores, e em cima do ID de cada autor, assim executa uma segunda *query* para pegar os autores relacionados a ele; assim tem-se uma posição principal no *array* com o autor dentro dessa mesma posição, um outro *array* com os autores relacionados, resumindo, uma matriz, onde será cruzada as informações e mostrado na tela do usuário.

Os exemplos demonstrados são para recuperação dos autores das referências. Para recuperação dos autores dos trabalhos cadastrados no banco de dados, é feita da mesma maneira apenas trocando o nome dos campos e tabelas.

No gráfico de Quantidade de artigos x Tempo e Quantidade de referencias x Tempo, é exibido um gráfico em linha demonstrando a evolução da quantidade de artigos e referencias, através dos anos, FIGURA 6.

6. METODOLOGIA

A maioria dos estudos bibliométricos analisa a informação das bases de dados elaboradas pela ISI. Estas bases de dados caracterizam-se por registrar todas as citações diretas dos autores assim como as citações que estes recebem posteriormente de forma indireta. Estes conteúdos são os que distinguem e tornam únicos os dados do ISI em comparação com outras bases de dados mais utilizadas. Os indicadores bibliométricos refletem a atividade científica e a publicação científica é uma etapa chave de todo o projeto de investigação, uma etapa em que os resultados se submetem a validação pública. O estudo foi realizado utilizando artigos presentes na bases de dados JSTOR disponíveis na Universidade Presbiteriana Mackenzie (UPM).

Neste trabalho foram considerados os artigos publicados eletronicamente na base JSTOR, tendo como no campo específico de busca o tema *Balanced Scorecard*, no título, no texto ou em palavra-chave, cobrindo o período compreendido entre 1992 a 2007. A exatidão de todos os artigos selecionados foi verificada mediante um processo de leitura dos mesmos.

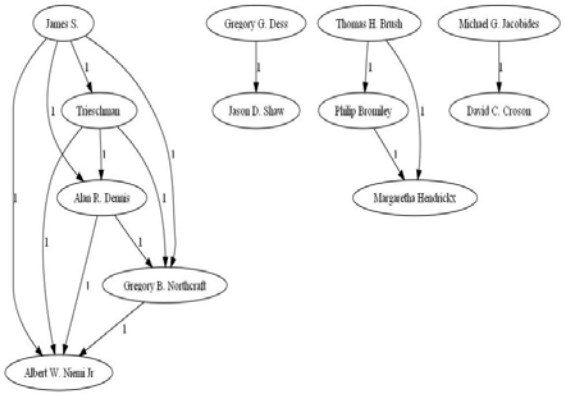
Os indicadores bibliométricos diretos observados são os seguintes:

a) Número de autores e colaboradores por artigos;
Número de referências;
b) Distribuição das referências no tempo;
c) Produção e repercussão da atividade científica do artigo e do (s) autor (es), mediante o número de artigos publicados pelo autor (es).

7. TRATAMENTO DOS DADOS E RESULTADOS

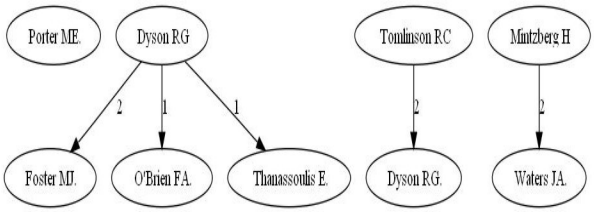
Tendo como resultado um relacionamento entre os autores, a partir dos arquivos de referencial, foram gerados grafos, onde se pode notar com maior clareza a relação entre os diversos autores e referências, a figura 4 ilustra uma parte do relacionamento existente entre os dados inseridos na base criada, mostrando as ligações entre os autores, assim como a quantidade que cada um cita o outro.

FIGURA 4: Relacionamento autor x referencial bibliográfico



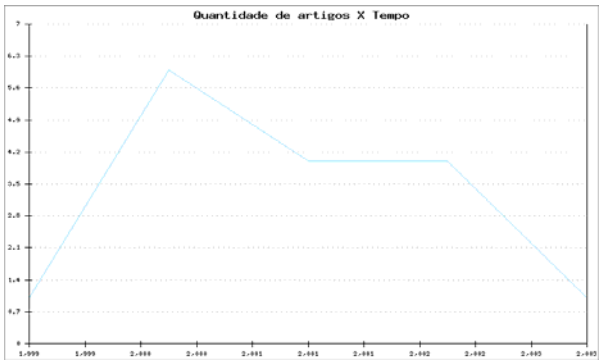
No caso da FIGURA 4, os vértices superiores do grafo são os autores do artigo, sendo ligados pelas arestas aos outros autores do referencial, onde são marcados numericamente quantas vezes foram utilizados. Também para o indicador de co-citação foi gerado um grafo, indicando as ligações entre autores, mostrado na FIGURA 5.

FIGURA 5: Grafo de co-citação



Neste grafo, FIGURA 5, é mostrado quantas vezes determinados autores (vértices) desenvolvem artigos com outros autores (vértices), sendo identificada na numeração posta em suas arestas. Na presente pesquisa foi desenvolvido também, um gráfico de meia vida, onde se nota a produção literária em um determinado período de tempo FIGURA 6.

FIGURA 6 – Gráfico de meia vida



8. CONCLUSÃO

O *framework* CakePHP mostrou-se de difícil manipulação por sua documentação online ser muito dispersa; no entanto, por outro lado é um dos mais utilizados no mercado, o que justifica o seu uso em trabalhos futuros. Dificuldades na criação dos grafos e mapas bi-dimensionais foram contornadas com a utilização do programa GraphViz, porém o programa mostrou-se pouco robusto para tal. Em trabalhos futuros seria necessário uma investigação maior ou outras classes dentro de GraphViz ou até mesmo o GIMP ou XARA 3D (gera gráficos animados).

A grande desvantagem do GraphViz foi o seu método de chamada, tendo que ser efetuado no código fonte (o que não ocorre no PAJEK); por outro lado, esse aplicativo proporciona facilmente a geração dos grafos, sendo que sem este as bibliotecas de GD *Graphics Library* (biblioteca gratuita acoplada ao PHP) teriam que ser desenhadas manualmente antes da inserção dos dados, inviabilizando a criação dos grafos. O PAJEK apresenta uma solução mais fácil e robusta, devendo ser preferido em desenvolvimentos futuros em vez do GraphViz. .

O desenvolvimento do banco de dados do sistema foi realizado em MySQL Workbench, o qual mostrou-se de fácil e suficientemente robusto para a atividade , mas outros softwares livres como o WEKA e PAJEK (já citado anteriormente), também podem ser considerados em trabalhos futuros.

O PHPLot mostrou-se como bastante viável, principalmente por ser bastante utilizado e documentado, devendo ser utilizado em trabalhos futuros com a mesma linha de pesquisa.

Quanto as base de artigos propostas, o projeto foi fixado na base JSTOR; no entanto, todas as bases foram analisadas de forma a se poder concluir que nenhuma base de artigos isoladamente possibilita a extração de todos os dados necessários que compõem os diversos indicadores da Bibliometria, sempre falta ao menos um tipo de dado, por exemplo origem ou instituição do(s) autor(es). Isto leva a uma conclusão importante, de que é necessário, em trabalhos futuros, um sistema de mineração de dados (*Datamining)* bastante robusto para tal processamento.

REFERÊNCIAS

[1] GUEDES, Vania e BORSCHIVER, S. Bibliometria: Uma ferramenta estatística para a gestão da informação e do conhecimento, em Sistemas de Informação, de Comunicação e de Avaliação Científica e Tecnológica. In: CINFORM - Encontro Nacional de Ciência da Informação, VI. Salvador. Anais ... Salvador: UFBA, 2005.

[2] ZITT, M., e BASSECOULARD, E. Reassessment of co-citation methods for science indicators: effect of methods improving recall rates. Scientometrics, 37 (2); pp. 223 – 244, 1996.

[3] MCGRATH, W. What bibliometricians, scientometricians and informetricians study; a typology for definition and classification; topics for discussion. In: *International Conference on Bibliometrics, Scientometrics and Informetrics*, 1989, Ontario. Second Conference.. Ontario: The University of Western Ontario, 1989.

[4] IRVINE, J.; MARTIN, B. R. International comparisons of scientific performance revisited. *Scientometrics*. Amsterdam, v. 15, n. 5-6, p. 369-392, 1989.

[5] CASTRO, Cláudio de Moura. *Ciência e universidade*. RJ, Zahar, 1985.

[6] MEADOWS, A. J. *A comunicação científifi ca*. Brasília: Briquet de Lemos, 1999.

[7] TAGUE-SUTCKIFFE, Jean. An introduction to informetrics. *Information Processing & Management*, Oxford, v. 28, n. 1, p. 1-3, 1992.

[8] HUTH, Edward J. Authors, editors, policy makers and the impact factor. Croatian *Medical Journal*, Lengerich, v. 42, n. 1, p. 14-17, 2001.

[9] MACIAS-CHAPULA, César A. O papel da infometria e da Cienciometria e sua perspectiva nacional e internacional. *Ciência da Informação*, Brasília, v. 27, n. 2, p. 134-140, maio/ago. 1998.

[10] GLÄNZEL, W.; SCHOEPFLIN, U. Little scientometrics, big scientometrics and beyond? *Scientometrics*, Amsterdam, v.30, p. 375-384, 1994.

[11] RUSSELL, Jane M. Back to the future for informetrics? *Scientometrics*, Amsterdam, v.30, n.1, p.407-410, 1994.

[12] KAPLAN, R. S., NORTON, D. P.. A *Estratégia em Ação: Balanced Scorecard*. Rio de Janeiro: Campus, 1997

[13] SOMMERVILLE, Ian. Engenharia de Software. 6 ed. São Paulo: Pearson Wealey, 2005.

[14] SINGH, Harry. "Data Warehouse: conceitos, tecnologias, implementação e gerenciamento". São Paulo: Makron Books,2001.

[15] OLIVEIRA, Wilson J. "*Data Warehouse*". São Paulo: Visual Books, 2002.

[16] BARI, Ahsanul e SYAM, Anupom. CakePHP Application Development. Packt Publishing, Mumbai, India. 2008.

[17] VASWANI , Vikram. PHP Programming Solutions. McGraw-Hill, 2007.

[18] POWELL, Gavin. Beginning Database Design. Wiley Publishing, Inc. 2006

[19] SUMATHI, S. e ESAKKIRAJAN, S. Fundamentals of Relational Database Management Systems. 2007.

[20] CHURCHER, Clare. Beginning Database Design. Springer. 2007.