Gest. Prod., São Carlos, v. 22, n. 1, p. 107-118, 2015
http://dx.doi.org/10.1590/0104-530X792-13

Mapeamento de um núcleo de partida de referências em *Data Mining* a partir de periódicos publicados no Brasil

Mapping a core starting of references in Data Mining from journals published in Brazil

Glauco Barbosa da Silva[1]
Helder Gomes Costa[1]

Resumo: O presente trabalho tem por objetivo principal estabelecer um mapeamento da produção científica publicada no Brasil e definir um núcleo inicial de referências bibliográficas para desenvolvimento de pesquisas em *Data Mining*. O mapeamento contemplou 48 artigos, distribuídos em 28 diferentes periódicos, que foram selecionados a partir do conteúdo disponível na base SCIELO. O objetivo secundário é apresentar algumas reflexões no viés da Engenharia de Produção. A amostra avaliada teve cerca de 8% de artigos publicados em periódicos indicados pela ABEPRO.

Palavras-chave: *Data Mining*. Bibliometria. *Webibliomining*.

Abstract: *The main objective of this paper is to establish a mapping of the scientific production published in Brazil and a starting core to a set of references in the development of research in Data Mining. Mapping included 48 articles, distributed in 28 different journals, which were selected from the content available on the SCIELO database. At the end, some comments are presented on the bias of Production Engineering.*

Keywords: *Data Mining. Bibliometric. Webibliomining.*

1 Introdução

Nas últimas décadas, em face dos avanços tecnológicos aliados a um declínio contínuo nos custos de armazenamento, tem ocorrido um rápido crescimento na geração, coleta e armazenamento eletrônico de dados. Os grandes e numerosos repositórios de dados têm ultrapassado a capacidade humana de interpretá-los, muitas vezes condenando os dados a um simples arquivamento de pouca utilidade. Como reflexo, decisões importantes são tomadas sem levar em consideração as informações imersas nos dados. Para reduzir o distanciamento entre a grande quantidade de dados e a pequena quantidade de informação, técnicas estatísticas e computacionais são aliadas numa área denominada Descoberta do Conhecimento em Banco de Dados (KDD - *Knowledge Discovery from Data*). Equivocadamente, é comum encontrar o termo *Data Mining* sendo usado com o mesmo significado de KDD.

Dada a necessidade de transformar dados em informação e a ampla gama de aplicações, *Data Mining* tem atraído muita atenção dos pesquisadores. Entretanto, conforme afirma Steiner et al. (2006), as etapas iniciais do KDD, referentes à análise exploratória dos dados que fazem uso de ferramentas estatísticas, não têm recebido a mesma atenção. Além disso, é uma percepção dos autores, a existência de poucos trabalhos na área de Engenharia de Produção, constituindo uma pergunta básica a pesquisar.

O presente trabalho tem por objetivo principal estabelecer um mapeamento da produção científica publicada no Brasil e definir um núcleo inicial de referências bibliográficas para desenvolvimento de pesquisas em *Data Mining* por meio da aplicação do método de *Webibliomining*. Para tal, está organizado da seguinte maneira: conceitos e definições são apresentados na seção 2; na seção 3, é descrito o método *Webibliomining;* na seção 4, - metodologia, o método é a aplicado para geração do núcleo de referências em *Data Mining* e são apresentados os resultados; a seção 5 apresenta as conclusões e considerações acerca do trabalho.

2 Conceitos e definições

Nesta seção, são descritos os fundamentos teóricos e conceituais, além de citar trabalhos prévios que utilizaram o modelo *Webibliomining*, que serviram de base para a pesquisa.

[1] Departamento de Engenharia de Produção, Universidade Federal Fluminense – UFF, Rua Passo da Pátria, 156, Bloco D, São Domingos, CEP 24210-240, Niterói, RJ, Brasil, e-mail: glauco.barbosa@me.com; Helder.uff@gmail.com

Recebido em Ago. 13, 2013 — Aceito em Out. 8, 2014

Suporte financeiro: CAPES e CNPQ.

2.1 KDD e Data Mining

Segundo Steiner et al. (2006), o processo de KDD é um conjunto de atividades contínuas que compartilham conhecimento descoberto a partir de bases de dados. É um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados, composto das seguintes etapas: Seleção, Pré-processamento e Limpeza, Transformação, *Data Mining*, Interpretação e Análise Exploratória dos dados (FAYYAD et al., 1996). Por tratar-se de um termo amplamente utilizado na literatura, optou-se por não traduzir o termo *Data Mining.* A Figura 1 apresenta as etapas do KDD.

Segundo Han et al. (2011), *Data Mining* é uma das etapas do processo de descoberta do conhecimento em banco de dados (KDD - *Knowledge Discovery from Data*). Consiste em um processo essencial, no qual técnicas estatísticas e computacionais são aplicadas para identificar padrões em grandes volumes de dados. As tarefas de *Data Mining* podem ser classificadas em duas categorias com diferentes aplicações: descritivas (sumarização, análise de associações, segmentação) e preditivas (previsões e classificação).

2.2 Bibliometria

Guedes & Borschiver (2005) definem Bibliometria como uma ferramenta estatística que permite mapear e gerar diferentes indicadores de tratamento e gestão da informação e do conhecimento, necessários ao planejamento, avaliação e gestão da ciência e tecnologia de uma determinada comunidade científica ou pais. A Bibliometria é norteada por três leis: Lei de Brandford (foco na dispersão dos periódicos), Lei de Lotka (foco na produtividade dos autores) e Lei de Zipf (foco na frequência das palavras), que são detalhadas em (ARAÚJO, 2006). Em Hood & Wilson (2001) e Guedes (2012), são discutidos outros termos correlatos como: Informetria – termo mais recente que faz uso dos princípios bibliométricos a contextos e bases não acadêmicos; Cienciometria, que faz uso dos métodos bibliométricos ao mapeamento de todos os aspectos da ciência e tecnologia; e Webmetria, como a área emergente na Ciência da Informação que consiste na aplicação de métodos infométricos a *World Wide Web*. Um método que integra conceitos de Pesquisa Operacional, Bibliometria e Webmetria denominado *Webibliomining* é proposto por Costa (2010).

Concluem Guedes & Borschiver (2005) que a Bibliometria constitui um instrumento quantitativo que permite minimizar a subjetividade inerente à indexação e recuperação das informações, produzindo conhecimento em determinada área, dessa forma auxilia na tomada de decisão da gestão da informação.

Não faz parte dos objetivos deste trabalho aprofundar-se nas questões referentes à Ciência da Informação. Portanto, de uma maneira geral, assumimos que a Bibliometria compreende o estudo de técnicas e métodos para o desenvolvimento de métricas para documentos e informações visando associar estatísticas à pesquisa bibliográfica.

Costa (2010) ressalta que o uso da Bibliometria no âmbito da Engenharia de Produção no Brasil não é usual, o que justifica a pesquisa do tema. Como alternativa, o modelo *Webibliomining* proposto é descrito na seção seguinte. Por ser um modelo ainda pouco difundindo, poucas aplicações encontram-se disponíveis. Sendo assim, como trabalhos correlatos, pode-se citar: Rodriguez et al. (2013), que estabelecem um mapeamento da produção científica sobre a aplicação de Métodos de Auxílio Multicritério à Decisão (AMD) em Problemas de Planejamento e Controle da Produção; Neves et al. (2013), que apresentam um estudo no âmbito da aplicação de métodos multicritério ao

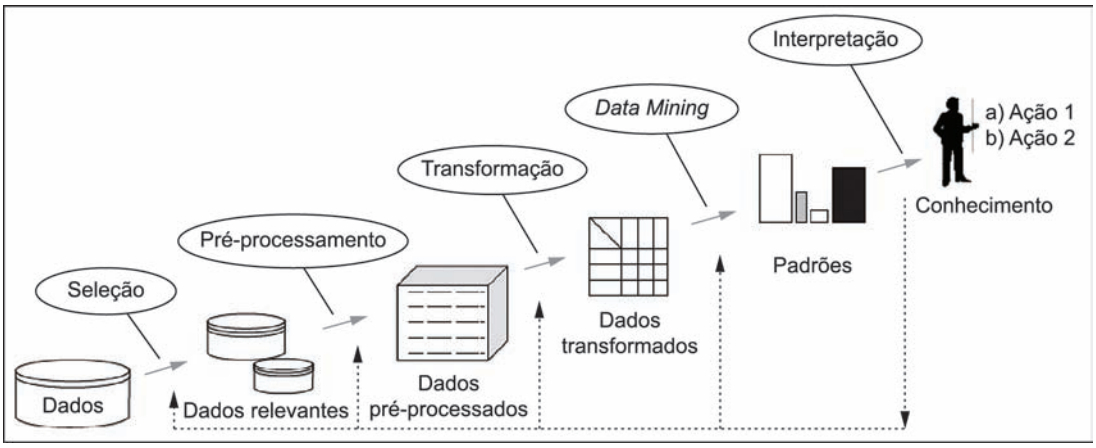

Figura 1. Etapas do *KDD* (Fonte: (Steiner et al., 2006)).

planejamento e gestão da indústria de petróleo e gás; e Dias & Costa (2012), que apresentam um mapeamento da produção científica no domínio da Ontologia.

2.3 *Webibliomining*

O modelo proposto por Costa (2010) tem por objetivo fornecer a seleção de um núcleo inicial de artigos para pesquisa bibliográfica; contempla conceitos de Pesquisa Operacional, Bibliometria e Webmetria. O Webibliomining está estruturado em seis etapas:

- Definição da amostra da pesquisa
- Pesquisa na amostra, com palavras-chave
- Identificação dos periódicos com maior número de artigos publicados sobre o tema
- Identificação dos autores com maior número de publicações
- Levantamento cronológico da produção, identificando os ciclos de maior produção
- Seleção dos artigos para a composição do *núcleo de partida* para a pesquisa bibliográfica, que deve considerar:

Os artigos mais relevantes

Identificação dos primeiros e últimos autores a escreverem sobre o tema

Identificação dos textos mais relevantes em cada ciclo de maior produção.

As etapas descritas nesta seção são aplicadas na definição de um referencial de partida para a pesquisa no tema *Data Mining*, conforme detalhado na seção seguinte.

3 *Webibliomining em Data Mining*

3.1 Definição da amostra e pesquisa na amostra com palavras-chave

A amostra definida corresponde aos artigos indexados na Base de Dados SciELO Brasil (Scientific Electronic Library Online), uma biblioteca eletrônica que abrange uma coleção de periódicos científicos brasileiros. O objetivo geral da SciELO é contribuir para o desenvolvimento da pesquisa científica nacional, por meio do aperfeiçoamento e da ampliação dos meios de disseminação, publicação e avaliação dos seus resultados por uso intensivo da publicação eletrônica. A escolha desta base deu-se pela acessibilidade e porque os critérios de admissão dos trabalhos na coleção são representativos. Contempla os periódicos com melhores avaliações segundo o critério WebQualis (Engenharias III) listados pela Associação Brasileira de Engenharia de Produção (ABEPRO) - Gestão & Produção (B3), Revista Pesquisa Operacional (B2); e Revista Produção (B2). O Quadro 1 lista os periódicos divulgados pela ABEPRO e a respectiva classificação conforme sistemática CAPES.

Quadro 1. Lista de periódicos de Engenharia de Produção.

Periódicos de Engenharia de Produção	***WebQualis***	**Periódicos Online de Engenharia de Produção**	***WebQualis***
Ciência & Tecnologia	**B5**	Revista Científica Eletrônica Produção Online	**B4**
Educação & Tecnologia	**-**	Revista Carioca de Produção	**B5**
Revista De Design, Inovação e Gestão Estratégica(Redige)	**B5**	Pesquisa & Desenvolvimento Engenharia de Produção	**B5**
Exacta	**B5**	Sistemas & Gestão	**B5**
Gestão & Produção	**B3**	Rio´*s International Journal on Sciences of Industrial and Systems Engineering and Management*	**B5**
Gepros	**B5**	Revista Gestão Industrial	**B5**
Product: Management & Development	**B4**	Revista de Gestão da Tecnologia e Sistemas de Informação	**B5**
Revista Abenge	**B4**	Revista Eletrônica Produção & Engenharia	**B4**
Revista O Mundo Da Usinagem	**B5**	Boletim Técnico Organização & Estratégia (O&E)	**B5**
Revista Pesquisa Operacional	**B2**	INGEPRO	**B5**
Revista Produto E Produção	**B4**	Revista TN Petróleo OnLine	**-**
Revista Produção	**B2**	GEINTEC - Gestão, Inovação e Tecnologias	**-**
Revista Tecnologia	**B5**		
Revista Da Universidade Do Amazonas	**B4**		

Fonte: ABEPRO e Portal CAPES.

Quanto ao recorte temporal, a pesquisa foi realizada em junho de 2013, contemplando todos os anos que estavam disponíveis na base.

Inicialmente foi definida como palavra-chave para a pesquisa "*Data Mining*", sendo obtido como resultado 45 artigos. Complementando a pesquisa inicial, buscando reduzir possíveis *gaps* ocorridos devido a erros de inserção, uma nova pesquisa com a palavra-chave "*Datamining*" foi realizada. Três resultados foram obtidos para a pesquisa complementar. Cabe ressaltar que durante a pesquisa optou-se pelo uso do conectivo "e" entre as palavras *Data* e *Mining* para evitar artigos de mineração (extração mineral).

Considerando os termos em Português e Inglês, uma lista de palavras-chave /*keywords* encontradas (228 palavras) foi processada gerando uma nuvem de palavras conforme apresentado na Figura 2. As maiores ocorrências foram: *Data Mining* (19), Mineração de Dados (11), Inteligência Artificial (4), *Bioinformatics* (3), KDD (3), *Artificial Intelligence* (2), Árvore de Decisão (2), Avicultura (2), Base de Dados (2), Bases de Conhecimento (2), *Clustering* (2), Descoberta de Conhecimento (2), *Knowledge Discovery* (2) e *Mortality* (2).

Conforme apresentado na Figura 3, Os resumos dos artigos selecionados também foram processados e geraram uma nuvem de palavras.

Figura 2. Nuvem das palavras-chave dos artigos pesquisados. (http://www.wordle.net/).

Figura 3. Nuvem dos resumos dos artigos pesquisados. (http://www.wordle.net/).

3.2 Identificação de periódicos

Os 48 artigos resultantes da pesquisa estão distribuídos em 28 diferentes periódicos. A partir do Quadro 1, dos periódicos listados pela ABEPRO e constantes da base SciELO (Gestão & Produção; Revista Pesquisa Operacional; e Revista Produção), pode-se observar que dois deles fizeram-se presentes entre os resultados detalhados no Quadro 2, que apresenta a quantidade de trabalhos publicados por periódico e o *WebQualis* (Engenharias III).

A partir da quantidade de artigos listados no Quadro 2, utilizando uma classificação do tipo PARETO é possível identificar os periódicos que merecem maior atenção e que podem ser classificados e agrupados como no Quadro 3, a seguir:

3.3 Identificação de autores

Este trabalho não faz distinção entre autoria e coautoria. Dessa forma, 186 autores foram identificados na pesquisa. O Quadro 4 apresenta os autores e periódicos em que os trabalhos foram publicados, limitados àqueles que têm pelo menos dois trabalhos indexados, num total de 20 autores listados.

A consulta aos currículos Lattes dos autores permitiu refinar a lista dos 20 autores, relacionando-os às áreas de atuação informadas, como apresentado no Quadro 5.

3.4 Levantamento cronológico e ciclos de produção

A fim de possibilitar a observação da evolução da produção científica em *Data Mining* no Brasil, os dados quantitativos referentes à distribuição de registros dos artigos publicados, agrupados por ano de publicação, encontram-se consolidados na Figura 4. Há indícios de que houve uma retomada na produção científica no tema a partir de 2008.

A Figura 5 apresenta um diagrama de dispersão do número de citações dos trabalhos publicados, procurando apontar, em termos de citações, as publicações de maior relevância. A partir do diagrama da Figura 5 é possível identificar que os artigos com maiores números de citações foram publicados nos anos de 1999, 2001 e 2008.

3.5 Composição do núcleo de partida

Com base nas regras estabelecidas pelo método *Webibliomining* proposto por Costa (2010) para a composição do núcleo de partida, algumas adaptações nos percentuais e na medição da relevância foram realizadas para adequação à amostra, tem-se:

- Seleção dos três trabalhos mais antigos, de autores diferentes, representando cerca de 6% da amostra. Esta regra busca abranger diferentes linhas de pensamentos nas discussões iniciais; como resultados da aplicação desta regra foram obtidos os seguintes trabalhos:

Zhu et al. (1999);
Agnez-Lima et al.(2001); e
Andrietta et al.(2001).

- Seleção de seis trabalhos mais recentes, de autores diferentes, representando aproximadamente 12% da amostra. Esta

Quadro 2. Lista de periódicos resultante dos pesquisados e quantidade de artigos publicados.

Periódico	QTD	*WebQualis*
Ciência e Agrotecnologia	1	**A2**
Pesquisa Agropecuária Brasileira	2	**A2**
Scientia Agricola	1	**A2**
Acta Paulista de Enfermagem	1	**B1**
Brazilian Journal of Oceanography	1	**B1**
Cadernos de Saúde Pública	1	**B1**
Ciência Rural	3	**B1**
Journal of the Brazilian Chemical Society	1	**B1**
Revista Brasileira de Ciência Avícola	2	**B1**
Revista de Saúde Pública	1	**B1**
Transinformação	1	**B1**
Ciência da Informação	2	**B2**
Engenharia Agrícola	4	**B2**
Pesquisa Operacional	2	**B2**
Gestão & Produção	2	**B3**
Revista de Administração Pública	1	**B3**
Journal of the Brazilian Computer Society	1	**B4**
Revista Brasileira de Meteorologia	2	**B4**
Revista de Administração Contemporânea	1	**B4**
Fisioterapia em Movimento	1	**B5**
JISTEM - *Journal of Information Systems and Technology Management*	1	**B5**
Genetics and Molecular Biology	9	-
ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)	2	-
Boletim de Ciências Geodésicas	1	-
Dental Press Journal of Orthodontics	1	-
Estudos de Psicologia (Natal)	1	-
Revista Brasileira de Hematologia e Hemoterapia	1	-
Revista de Administração Mackenzie	1	-

Fonte: O autor.

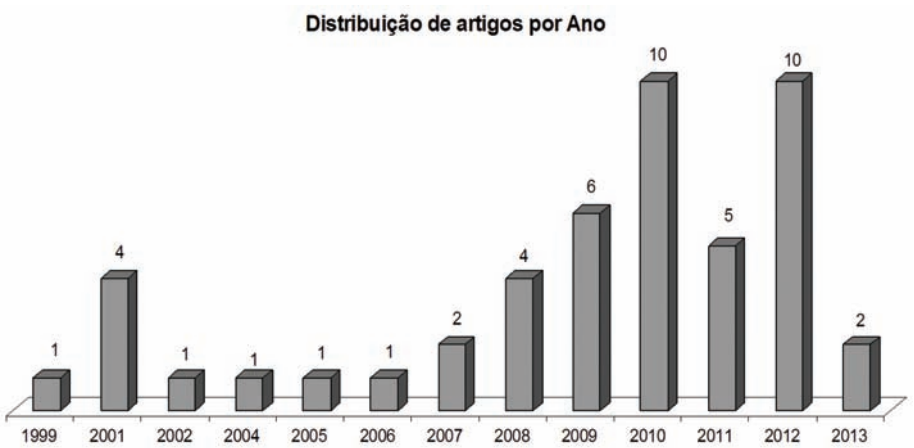

Figura 4. Distribuição de trabalhos publicados por ano.

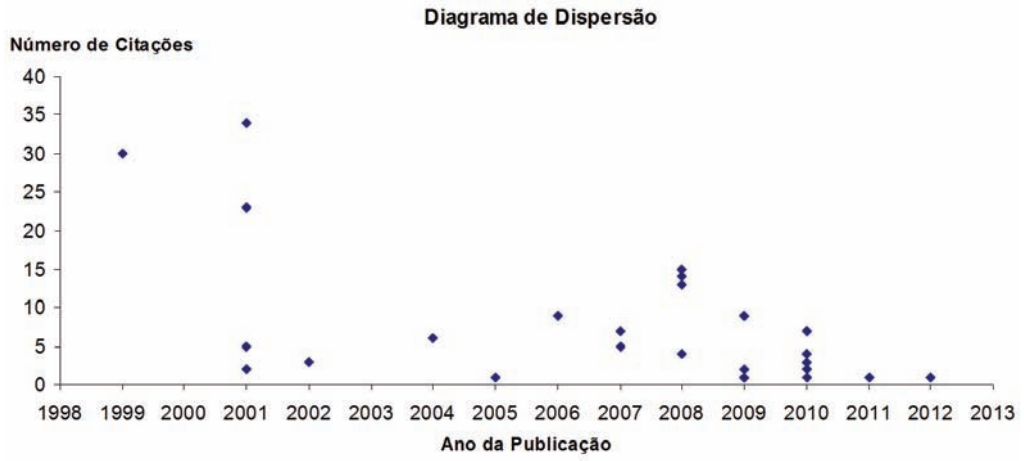

Figura 5. Diagrama de dispersão do número de citações de trabalhos publicados por ano.

Quadro 3. Distribuição de periódicos conforme Classificação ABC.

Classe A
Genetics and Molecular Biology
Engenharia Agrícola
Ciência Rural
Classe B
ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)
Ciência da Informação
Gestão & Produção
Pesquisa Agropecuária Brasileira
Pesquisa Operacional
Revista Brasileira de Ciência Avícola
Revista Brasileira de Meteorologia
Classe C
Acta Paulista de Enfermagem
Boletim de Ciências Geodésicas
Brazilian Journal of Oceanography
Cadernos de Saúde Pública
Ciência e Agrotecnologia
Dental Press Journal of Orthodontics
Estudos de Psicologia (Natal)
Fisioterapia em Movimento
JISTEM - *Journal of Information Systems and Technology Management*
Journal of the Brazilian Chemical Society
Journal of the Brazilian Computer Society
Revista Brasileira de Hematologia e Hemoterapia
Revista de Administração Contemporânea
Revista de Administração Mackenzie
Revista de Administração Pública
Revista de Saúde Pública
Scientia Agricola

Figura 6. Nuvem das palavras-chave dos artigos do núcleo de partida. (http://www.wordle.net/).

regra busca abranger diferentes tendências nas discussões mais recentes, o intuito de um percentual maior para os trabalhos recentes é dar maior ênfase aos trabalhos mais recentes. Como resultado da aplicação desta regra foram obtidos os seguintes trabalhos:

Montes-Grajales et al. (2013);
Ferreira et al.(2013);
Pessoa et al. (2012);
Osorio et al.(2012);
Leiva-Mederos et al. (2012); e
Lamparelli et al. (2012).

- Seleção de seis trabalhos com maior grau de relevância, aqui considerada como o número de citações fornecidas pelo Google Acadêmico, observando a interseção de trabalhos já selecionados nas regras anteriores, esta regra representa aproximadamente 12% da amostra. Resultado obtido:

Telles et al.(2001);
Quoniam et al.(2001);
Vale et al. (2008);
Pereira et al.(2008);

Cardoso & Machado (2008); e
Steiner et al. (2006).

Consolidando os resultados obtidos da aplicação das regras aos artigos componentes da amostra inicial, o Quadro 6 apresenta a lista dos artigos selecionados para compor o núcleo de partida.

A análise do núcleo de partida revelou um conjunto de setenta e cinco palavras-chave/*keywords*, que, após processadas, geraram uma nuvem, como apresentado na Figura 6, a seguir:

Comparando-se as Figuras 2 e 6, pode-se perceber que foi preservado o destaque das palavras *Data*, *Dados*, *Mining e Mineração*. O mesmo ocorre na comparação das Figuras 3 e 7, o que nos fornece indícios de que o núcleo é uma amostra representativa dos artigos pesquisados.

Quadro 4. Principais autores identificados na pesquisa e quantidade de artigos publicados.

Autor	QTD	Periódicos
Nääs, Irenilza de Alencar	5	Ciência Rural Engenharia Agrícola Revista Brasileira de Ciência Avícola Scientia Agricola
Oliveira, Stanley Robson de Medeiros	4	Engenharia Agrícola Pesquisa Agropecuária Brasileira Scientia Agricola
Vale, Marcos Martinez do	4	Ciência Rural Revista Brasileira de Ciência Avícola Scientia Agricola
Carvalho, Deborah Ribeiro	3	Cadernos de Saúde Pública Fisioterapia em Movimento Revista de Saúde Pública
Moura, Daniella Jorge de	3	Ciência Rural Revista Brasileira de Ciência Avícola Scientia Agricola
Rodrigues, Luiz Henrique Antunes	3	Ciência e Agrotecnologia Revista Brasileira de Ciência Avícola Scientia Agricola
Alvarenga, Samuel Mazzinghy	2	Pesquisa Agropecuária Brasileira *Genetics and Molecular Biology*
Caixeta, Eveline Teixeira		
Maciel-Zambolim, Eunize		
Sakiyama, Ney Sussumu		
Hufnagel, Bárbara		
Cordeiro, Alexandra Ferreira da Silva	2	Engenharia Agrícola
Kuretzki, Carlos Matias, Jorge Eduardo Fouto Moraes, Roberto da Silveira Oliveira, Mateus Martinelli de Pinto, José Simão de Paula	2	ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)
Pereira, DF	2	Revista Brasileira de Ciência Avícola
Nievola, Julio Cesar	2	Gestão & Produção Cadernos de Saúde Pública
Steiner, Maria Teresinha Arns	2	Gestão & Produção

Figura 7. Nuvem dos resumos dos artigos do núcleo de partida (http://www.wordle.net/).

Figura 8. Áreas de atuação dos autores dos trabalhos componentes do núcleo de partida.

Quadro 5. Principais autores identificados por área de atuação.

Autor	Área de Atuação
Nääs, Irenilza de Alencar	Engenharia de Produção; Zootecnia; Engenharia Agrícola; e Agronomia.
Oliveira, Stanley Robson de Medeiros	*Data Mining*; Banco de Dados; Ciência da Computação; Engenharia de *Software*; e Simulação
Vale, Marcos Martinez do Cordeiro, Alexandra Ferreira da Silva Moura, Daniella Jorge de	Zootecnia; e Ciências Agrárias.
Carvalho, Deborah Ribeiro	Ciência da Computação; e Ciência da Informação.
Rodrigues, Luiz Henrique Antunes	KDD; *Data Mining*; Ciência da Computação; Aprendizagem de Máquina; Engenharia Agrícola; Pesquisa Operacional; e Engenharia de Produção.
Alvarenga, Samuel Mazzinghy	Ciências Biológicas; Genética; Bioinformática; e Biologia Molecular.
Caixeta, Eveline Teixeira	Ciências Agrárias; Biotecnologia; Genômica; e Genética.
Maciel-Zambolim, Eunize	Agronomia; Fitotecnia; e Biotecnologia.
Sakiyama, Ney Sussumu	Ciências Agrárias; e Agronomia.
Hufnagel, Bárbara	Ciências Biológicas; Bioquímica; Biologia Molecular; Bioinformática; e Genética.
Pinto, José Simão de Paula	Ciência da Computação; e Banco de Dados.
Nievola, Julio Cesar	Ciência da Informação; Ciência da Computação; e Probabilidade e Estatística.
Steiner, Maria Teresinha Arns	Engenharia de Produção; Pesquisa Operacional; Reconhecimento de Padrões; e Engenharia de Transportes.

Quadro 6. Núcleo de partida bibliográfica para pesquisa em *Data Mining*.

Titulo	Autor	Palavras-chave
A process for mining science & technology documents databases, illustrated for the case of “knowledge discovery and data mining”	Zhu et al. (1999)	-
Base excision repair in sugarcane	Agnez-Lima et al.(2001)	-
Identification of sugarcane cDNAs encoding components of the cell cycle machinery	Andrietta et al.(2001)	-
DDT and derivatives may target insulin pathway proteins	Montes-Grajales et al. (2013)	DDT; AutoDock Vina; AKT-1; insulin; docking;
Classificação de características produtivas fenotípicas de diferentes raças de poedeiras por meio da mineração de dados	Ferreira et al.(2013)	Bioclimatologia; ovos marrons; melhoramento animal; ambiência;
Mineração de dados meteorológicos para previsão de eventos severos	Pessoa et al. (2012)	Mineração de dados; Previsão meteorológica; Eventos convectivos
Identification and in silico characterization of soybean trihelix-GT and bHLH transcription factors involved in stress responses	Osorio et al. (2012)	Drought; gene expression; Glycine max; phylogeny; plant-microbe interactions
PuertoTex: un software de minería textual para la creación de resúmenes automáticos en el dominio de ingeniería de puertos y costas basado en ontologías	Leiva-Mederos et al. (2012)	Automata; Disambiguation of scientific texts; Data mining; Ontologies; Automatic summarization;
Use of data mining and spectral profiles to differentiate condition after harvest of coffee plants	Lamparelli et al.(2012)	Monitoramento de cultura; comportamento espectral; manejo; sensoriamento remoto
Bioinformatics of the sugarcane EST project	Telles et al. (2001)	-

Quadro 6. Continuação...

Titulo	Autor	Palavras-chave
Inteligência obtida pela aplicação de *data mining* em base de teses francesas sobre o Brasil	Quoniam et al.(2001)	*Data Mining*; Bibliometria; Análise bibliométrica; Teses francesas; Brasil; Descoberta de conhecimento; Base de dados; Lei de Zipf;
Data mining to estimate broiler mortality when exposed to heat wave	Vale et al. (2008)	ITU; frango de corte; dados ambientais; THI; broiler production; environmental data;
Data mining for environmental analysis and diagnostic: a case study of upwelling ecosystem of Arraial do Cabo	Pereira et al. (2008)	Mineração de dados; Sistemas inteligentes; Diagnóstico ambiental; Gerenciamento de ecossistemas; Integridade biológica;
Gestão do conhecimento usando *data mining*: estudo de caso na Universidade Federal de Lavras	Cardoso & Machado (2008);	Gestão do conhecimento; descoberta de conhecimento em bancos de dados; *data mining*; plataforma Lattes;
Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados	Steiner et al.(2006)	Mineração de dados, processo KDD, análise exploratória dos dados.

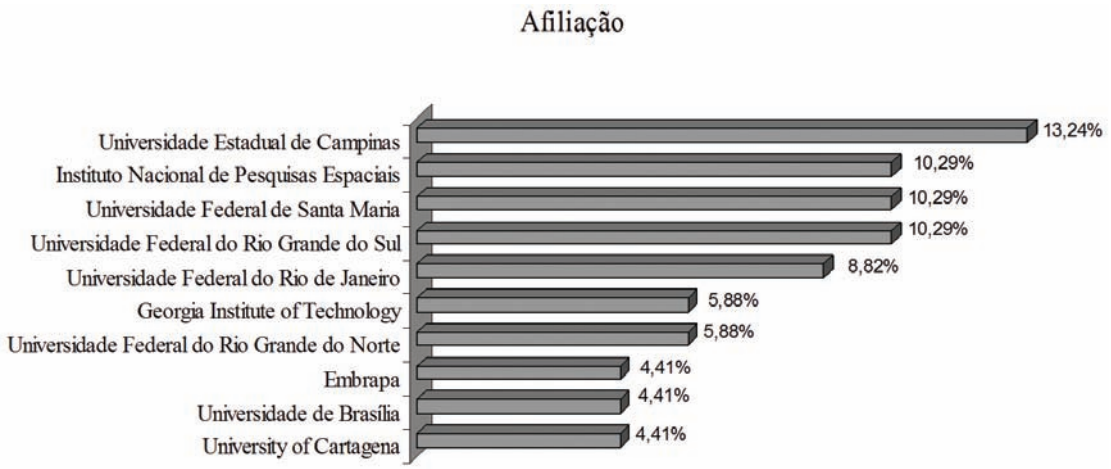

Figura 9. Distribuição dos autores por afiliação.

Analisando a distribuição das áreas de atuação dos autores dos trabalhos que compõem o núcleo de partida, foi possível a divisão dos trabalhos por área de aplicação, conforme apresentando na Figura 8, a seguir:

A legenda *outras* apresentada na Figura 8 representa as áreas de atuação: Administração e Economia; Geociências; Oceanografia e Química, que assim foram agrupadas por apresentarem, isoladamente, menos de 3% dos trabalhos dos componentes do núcleo.

A Figura 9 consolida o mapeamento das afiliações dos autores dos trabalhos componentes do núcleo de partida, representando cerca de 80% das instituições, as demais instituições não aparecem representadas na figura. Tal exclusão deve-se ao fato de representarem, isoladamente, menos de 3% do número total das instituições identificadas.

Numa classificação por regiões geográficas, cerca de 50% dos autores concentram-se em instituições da região sudeste, 30% na região sul, 11% na região centro-oeste e 7% na região nordeste. Não foram identificados trabalhos de autores afiliados às instituições da região norte.

O Quadro 7 apresenta uma matriz que relaciona as instituições às quais os autores são afiliados e suas áreas de atuação, que serve para consolidar, em outra visão, a distribuição dos trabalhos selecionados.

Quadro 7. Instituições x Áreas de atuação dos autores para pesquisa em *Data Mining*.

	Ciência da Computação	Ciências Biológicas	Ciência da Informação	Zootecnia	Engenharia Agrícola	Engenharia de Produção	Administração e Economia	Química	Geociências	Oceanografia
INPE	x								x	
ITA						x				
Unicamp	x				x					
UF Santa Maria				x						
Beijing Institute of Technology						x				
Embrapa	x	x			x					
Georgia Institute of Technology			x							
Inmetro	x									
Instituto de Estudos Paulo Moreira										x
PUC-PR	x					x				
UNIOESTE					x					
Universidad Central de Las Villas			x							
UnB			x							
USP						x				
UF Lavras							x			
UFPR							x			
UFRJ	x	x								
UFRN		x								
UFRS		x								
UniP				x						
Université du Sud -Toulon - Var, Ingémédia			x							
University of Cartagena		x						x		
University of Granada			x							

4 Conclusões

A partir da aplicação do método *Webibliomining*, foi possível gerar um núcleo de partida para a pesquisa em *Data Mining* em periódicos no Brasil, que abrange trabalhos distribuídos por diferentes momentos da produção científica do tema, propondo uma cobertura ampla e consistente.

A etapa dois proposta no método garante ao núcleo a presença de trabalhos recentes, que não apresentam, em geral, um alto número de citações, consequência do pouco tempo de publicação destes trabalhos.

Aproximadamente 228 palavras-chave foram encontradas na amostra. Entretanto, observa-se que, dentre estas, uma quantidade razoável de palavras pouco contribuem para a seleção e identificação dos trabalhos. Isso pode ser observado no Quadro 6, que apresenta palavras como *ovos marrons* e *frango de corte*. Como solução alternativa para este problema, alguns periódicos já adotam como norma a seleção de palavras-chave a partir de um conjunto pré-definido de palavras constantes da base do periódico, ou seja, garante e facilita o direcionamento dos trabalhos.

Uma comparação entre os Quadros 5 e 6 permite afirmar que o método tem foco na seleção do trabalho, pois a quantidade de trabalhos publicados pelo autor não influencia o resultado do núcleo. Nesse viés, a partir do núcleo, uma nova lista de autores pode ser construída.

Como o número de artigos analisados foi relativamente pequeno, os percentuais da amostra utilizados dentro das regras do método *Webibliomining* foram ajustados e arbitrados pelo pesquisador. Em termos gerais, o núcleo gerado contempla cerca de 30% da amostra inicial.

A partir da observação e análise da área de atuação dos autores, é possível identificar que cerca de 20%

de todos os autores são pesquisadores com atuação em Engenharia de Produção. A amostra avaliada teve cerca de 8% de artigos publicados em periódicos indicados pela ABEPRO. Os resultados obtidos possibilitam inferir que há indícios de um número pequeno de publicações no tema *Data Mining* em Engenharia de Produção.

Como sugestões para trabalhos futuros, a inclusão de trabalhos de teses disponíveis no Banco de Teses da Capes deve incrementar a abrangência dos resultados. Um estudo semelhante incluindo bases internacionais como a *Scopus*, *Science Direct*, *Esmerald*, *entre outras*, pode fornecer indicadores capazes de ampliar o detalhamento da produção de pesquisadores brasileiros em *Data Mining* e Engenharia de Produção.

Referências

AGNEZ-LIMA, L. F. Base excision repair in sugarcane. **Genetics and Molecular Biology**, v. 24, n. 1-4, p. 123-129, 2001. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572001000100017&nrm=iso>.

ANDRIETTA, M. H. Identification of sugarcane cDNAs encoding components of the cell cycle machinery. **Genetics and Molecular Biology**, v. 24, n. 1-4, p. 61-88, 2001. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572001000100010&nrm=iso>.

ARAÚJO, C. A. Á. Bibliometria: evolução histórica e questões atuais. **Em Questão**, v. 12, n. 1, p. 21, 2006. Disponível em: <http://revistas.univerciencia.org/index.php/revistaemquestao/article/viewFile/3707/3495>.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de Administração Pública**, v. 42, n. 3, p. 495-528, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-76122008000300004&nrm=iso>.

COSTA, H. G. Modelo para webibliomining: proposta e caso de aplicação Model for webibliomining: proposal and application. **Revista da FAE**, v. 13, n. 1, p. 115-126, 2010. Disponível em: <http://www.unifae.br/publicacoes/v.13_01-2010.pdf#page=119>.

DIAS, E. A. V.; COSTA, H. G. Mapeamento da produção científica no escopo da Ontologia. **Sistemas & Gestão**, v. 6, n. 4, p. 481-507, 2012. http://dx.doi.org/10.7177/sg.2011.v6.n4.a6

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. American Association for Artificial Intelligence, 1996.

FERREIRA, P. B. et al. Classificação de características produtivas fenotípicas de diferentes raças de poedeiras através da mineração de dados. **Ciência Rural**, v. 43, n. 1, p. 164-171, 2013.

GUEDES, V. L. D. S. A Bibliometria e a Gestão da Informação e do Conhecimento Científico e Tecnológico: uma revisão da literatura. **Revista Ponto de Acesso**, v. 6, n. 2, p. 74-109, 2012.

GUEDES, V. L. S.; 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. Guaratinguetá: UNESP, 2005.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2011.

HOOD, W.; WILSON, C. The Literature of bibliometrics, scientometrics, and informetrics. **Scientometrics**, v. 52, n. 2, p. 291-314, 2001. http://dx.doi.org/10.1023/A:1017919924342

LAMPARELLI, R. A. C. Use of data mining and spectral profiles to differentiate condition after harvest of coffee plants. **Engenharia Agrícola**, v. 32, n. 1, p. 184-196, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-69162012000100019&nrm=iso>.

LEIVA-MEDEROS, A.; DOMÍNGUEZ-VELASCO, S.; SENSO, J. A. PuertoTex: un software de minería textual para la creación de resúmenes automáticos en el dominio de ingeniería de puertos y costas basado en ontologías. **Transinformação**, v. 24, n. 2, p. 103-115, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862012000200003&nrm=iso>.

MONTES-GRAJALES, D.; OLIVERO-VERBEL, J.; CABARCAS-MONTALVO, M. DDT and derivatives may target insulin pathway proteins. **Journal of the Brazilian Chemical Society**, v. 24, n. 4, p. 558-572, 2013. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-50532013000400006&nrm=iso>.

NEVES, R. B.; PEREIRA, V.; COSTA, H. G. Auxílio multicritério à decisão aplicado ao planejamento e gestão na indústria de petróleo e gás. **Production**, 2013. No prelo. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65132013005000060&nrm=iso>.

OSORIO, M. B. et al. Identification and in silico characterization of soybean trihelix-GT and bHLH transcription factors involved in stress responses. **Genetics and Molecular Biology**, v. 35, n. 1, p. 233-246, 2012.

PEREIRA, G. C.; COUTINHO, R.; EBECKEN, N. F. F. Data mining for environmental analysis and diagnostic: a case study of upwelling ecosystem of Arraial do Cabo. **Brazilian Journal of Oceanography**, v. 56, n. 1, p. 1-12, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1679-87592008000100001&nrm=iso>.

PESSOA, A. S. A. et al. Mineração de dados meteorológicos para previsão de eventos severos. **Revista Brasileira de Meteorologia**, v. 27, n. 1, p. 61-74, 2012. http://dx.doi.org/10.1590/S0102-77862012000100007

QUONIAM, L. et al. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. **Ciência da Informação**, v. 30, n. 2, p. 20-28, 2001. http://dx.doi.org/10.1590/S0100-19652001000200004

RODRIGUEZ, D. S. S.; COSTA, H. G.; CARMO, L. F. R. R. S. D. Métodos de auxílio multicritério à decisão aplicados a problemas de PCP: mapeamento da produção em periódicos publicados no Brasil. **Gestão & Produção**, v. 20, n. 1, p. 134-146, 2013. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-530X2013000100010&nrm=iso>.

STEINER, M. T. A. et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gestão & Produção**, v. 13, n. 2, p. 325-337, 2006. http://dx.doi.org/10.1590/S0104-530X2006000200013

TELLES, G. P. et al. Bioinformatics of the sugarcane EST project. **Genetics and Molecular Biology**, v. 24, n. 1-4, p. 9-15, 2001. http://dx.doi.org/10.1590/S1415-47572001000100003

VALE, M. M. D. et al. Data mining to estimate broiler mortality when exposed to heat wave. **Scientia Agricola**, v. 65, n. 3, p. 223-229, 2008. http://dx.doi.org/10.1590/S0103-90162008000300001

ZHU, D. et al. A process for mining science & technology documents databases, illustrated for the case of "knowledge discovery and data mining". **Ciência da Informação**, v. 28, n. 1, p. 7-14, 1999. http://dx.doi.org/10.1590/S0100-19651999000100002