Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil

Luc Quoniam
Doutor em ciência da informação e comunicação pela Université de Aix-Marseille III - França. Diretor do Centro Franco-Brasileiro de Documentação Técnico-Científica (Cendotec).
quoniam@univ-tln.fr

Kira Tarapanoff
Doutora em ciência da informação pela Universidade de Sheffield - Inglaterra. Professora de pós-graduação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília.
kira@ibict.br

Rogério Henrique de Araújo Júnior
Mestre e doutorando em ciência da informação pela Universidade de Brasília.
patricia.cormier@caixa.gov.br

Lillian Alvares
Mestre em ciência da informação pela Universidade de Brasília e especialista em inteligência competitiva. Secretaria de Tecnologia Industrial do Ministério do Desenvolvimento, Indústria e Comércio Exterior.
lilliana@mdic.gov.br

Resumo

*O assunto Brasil foi analisado na base de teses francesas **DocThèses**, compreendendo os anos de 1969 a 1999. Utilizou-se a técnica de **Data Mining** como ferramenta para obter inteligência e conhecimento. O **software** utilizado para a limpeza da base **DocThèses** foi o **Infotrans,** e, para a preparação dos dados, empregou-se o **Dataview**. Os resultados da análise foram ilustrados com a aplicação dos pressupostos da Lei de Zipf, classificando-se as informações em trivial, interessante e ruído, conforme a distribuição de freqüência. Conclui-se que a técnica do **Data Mining** associada a **softwares** especialistas é uma poderosa aliada no emprego de inteligência no processo decisório em todos os níveis, inclusive o nível macro, pois oferece subsídios para a consolidação, investimento e desenvolvimento de ações e políticas.*

Palavras-chave

Data Mining; Bibliometria, Análise bibliométrica, Teses francesas, Brasil, Descoberta de conhecimento, Base de dados; Lei de Zipf.

Intelligence obtained with the application of data mining analysing the French DocThéses on subjects about Brazil

Abstract

The subject Brazil was analysed within the context of the French data base DocThéses, comprising the years 1969 up to 1999. The data minig technique was used to obtain intelligence and infer knowledge. The software used to do the cleaning of the base DocThèses was Infotrans; and for the preparation of the data was Dataview. The results of the analysis were illustrated by making use of the assumptions of the Zipf Law, on bibliometrics, classifying the information in trivial information; of interest; and "noise", according to the distribution of frequency. The conclusion is that the Data Mining technique associated with specialist software is a powerful ally for the competitive intelligence applied on all levels of the decision making process, including the macro level. It can enhance the consolidation, investment and development of actions and policies.

Keywords

Data mining; Bibliometrics, Bibliometrical analysis, French thesis, Brazil, Knowledge discovery; Data bases; Zipf Law.

INTRODUÇÃO

A capacidade de armazenamento em banco de dados, assim como sua utilização, vem crescendo na mesma proporção dos avanços em novas tecnologias de informação e comunicação. A atividade de extrair informações relevantes, por conseqüinte, está se tornando bastante complexa. Este processo de "garimpagem" é chamado de *Knowledge Discovery in Databases* – KDD (Descoberta de Conhecimento em Bases de Dados).

O KDD pode ser visto como o processo da descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados estocados. Este processo se vale de tecnologias de reconhecimento utilizando padrões e técnicas estatísticas e matemáticas. O *Data Mining*[1] é uma das técnicas utilizada para a realização de KDD. Aspectos específicos incluem investigação e criação de conhecimento, processos, algoritmos e mecanismos de recuperação de conhecimento potencial de estoques de dados (Norton, 1999).

A descoberta de conhecimento em bases de dados, KDD, é vista como uma disciplina mais ampla, e o termo *Data Mining* (mineração ou garimpagem de dados) como o componente que trata dos ***métodos*** do descobrimento do conhecimento (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996).

A aplicação do *Data Mining* torna possível comprovar o pressuposto da transformação de dados em informação e posteriormente em conhecimento. Esta possibilidade torna a técnica imprescindível para o processo de tomada de decisão. Para chegar-se a este resultado, é preciso investigar o uso efetivo do conhecimento obtido pelo *Data Mining* no processo de tomada de decisão, bem como os impactos que teve na solução efetiva de problemas e ações propostas e concretizadas.

[1] *Data Mining*, *DM* – ou mineração de dados: tarefa de estabelecer novos padrões de "conhecimento", geralmente imprevistos, partindo-se de uma massa de dados previamente coletada e preparada para este fim (Tarapanoff, (Org.), 2001).

Propõe-se, por meio deste trabalho, demonstrar a aplicação da técnica de *Data Mining*, usando como estudo de caso a base *DocThéses*, um catálogo de teses francesas. O foco foram as teses que tiveram o Brasil como assunto, estando incluídos também os trabalhos de brasileiros defendidos na França. O período em estudo compreendeu os anos de 1969 a 1999. Os indicadores levantados foram:

1) ocorrência de teses com o assunto Brasil por áreas do conhecimento;

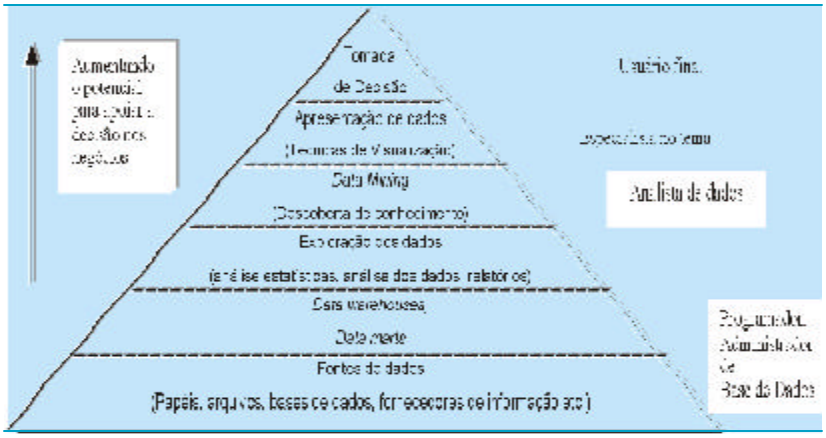
2) orientadores de teses que se destacaram por áreas do conhecimento;

3) distribuição geográfica das cidades que hospedaram as instituições onde as teses foram defendidas;

4) incidência, por áreas do conhecimento, dos anos de defesa das teses.

FIGURA1
Esquema do Data Mining

FIGURA 2
Evolução do valor estratégico de base de dados



Fonte: baseado em Cabena *et alii*, 1997 & Tyson, 1998

O PROCESSO DE DATA MINING

São chamadas de *Data Mining* (DM) todas as técnicas que permitem extrair conhecimento de uma massa de dados que, de outra maneira, permaneceria escondido nas grandes bases. Na fase anterior ao processo do DM, temos o pré-processamento, no qual são executadas as fases de coleta, armazenagem e "limpeza"[2] dos dados. Para realizá-lo com sucesso, é necessário conhecimento da base, incluindo o entendimento dos dados, a limpeza e sua preparação para não haver duplicação de conteúdo através de erros de digitação, abreviações diferentes, valores omissos, entre outros.

As ferramentas *Data Mining* identificam todas as possibilidades de correlações existentes nas fontes de dados. Através das técnicas para exploração de dados, pode-se desenvolver aplicações que venham a extrair, dos bancos de dados, informações críticas, com o objetivo de subsidiar plenamente o processo decisório de uma organização.

O conceito de *Data Mining* (ou mineração de dados), segundo Cabena *et alii* (1997), é a técnica de extrair informação, previamente desconhecida e de máxima abrangência a partir de bases de dados, para usá-la na tomada de decisão (figura 1).

A figura 2 mostra o posicionamento lógico de diferentes fases da tomada de decisão com seu valor potencial para as dimensões tática e estratégica. Em geral, o valor da informação para apoiar a tomada de decisão aumenta a partir da base da pirâmide. Uma decisão baseada em dados nas camadas mais baixas, onde há tipicamente milhões de registros de dados, não possui muito valor agregado[3]; já aquela apoiada em dados altamente resumidos nas camadas superiores da pirâmide tem probabilidade de alto valor estratégico.

Da mesma forma, encontram-se diferentes usuários nas diferentes camadas. Um administrador, por exemplo, em nível operacional, trabalha primariamente com informações diárias e operações de rotina, do tipo ***o que***, encontradas em arquivos e base de dados, na base da pirâmide informacional. Esses criam dados. Enquanto analistas de negócios e executivos, responsáveis por indicarem direções, formularem estratégias e táticas e supervisionando a sua execução, necessitam de informações de maior fôlego. Preocupam-se com tendências, padrões, fraquezas, ameaças, pontos fortes e oportunidades, informações de mercado e mudanças tecnológicas. Necessitam de informações do tipo ***por quê***

[2] Aplicação de ferramentas de reformatação a fim de gerar a base de dados de trabalho, ou seja, transformar a base de dados bruta. Neste caso específico, foi utilizado o *software* de reformatação *Infotrans*.

[3] Processo que consiste em transformar dados sem qualquer significado em informação útil. O processo de agregação de valor é composto de organização, análise, síntese e julgamento (Taylor, 1986).

e do tipo ***e se***. Necessitam de informações internas e externas. São os criadores e os que demandam dados analisados com alto valor agregado, as do topo da pirâmide.

Uma visão geral das etapas envolvidas no *DM* é mostrada na figura 3. O processo inicia-se com a definição clara do problema – 1ª etapa –, seguida da 2ª etapa, que é a seleção a fim de identificar todas as fontes internas e externas de informação e selecionar o subconjunto de dados necessário para aplicação de *DM*, que contemple o problema. A 3ª etapa corresponde à preparação dos dados, que inclui o pré-processamento, sendo a que exige maior esforço. Está dividida em ferramentas de visualização e ferramentas de reformatação dos dados, o que corresponde a 60% do trabalho de *DM*, situação ilustrada na figura 4. A preparação é crucial para a qualidade final dos resultados, por isso as ferramentas utilizadas são tão importantes. Os *softwares* dedicados a esta etapa devem estar prontos para muitos processos: agregar valor, efetuar conversão, filtrar variáveis, possuir formato de exportação de dados, trabalhar com base de dados relacionais, mapear variáveis de entrada, entre outros. Essas etapas assemelham-se em tudo com as etapas do ciclo informacional ou o processo de gestão da informação realizado dentro do domínio temático da ciência da informação, em especial no processo de recuperação de informações (*information retrieval)*[4].

Passamos agora à quarta etapa de análise dos resultados obtidos do processo de *DM,* que tem dois aspectos fundamentais a considerar: informar as novas descobertas e apresentá-las de maneira que possam ser exploradas potencialmente. Nesta fase, recomenda-se a participação de um especialista da área de que trata a base de dados, a fim de solucionar questões técnicas específicas que possam influir na análise. Gerentes de negócios e executivos podem ser envolvidos nesta fase.

Podemos obter com a aplicação de *Data Mining* vários tipos de descoberta de conhecimento. Dentre eles, descoberta de associações, descoberta de agrupamentos, descoberta de classificações, descoberta de regras de previsão, hierarquias de classificação, descoberta de padrões seqüenciais, descoberta de padrões em séries temporais, categorização e segmentação, que estão em Alvares (2000).

[4] O termo recuperação da informação "engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação" (Mooers, 1951).

FIGURA 3
Etapas do processo de Data Mining

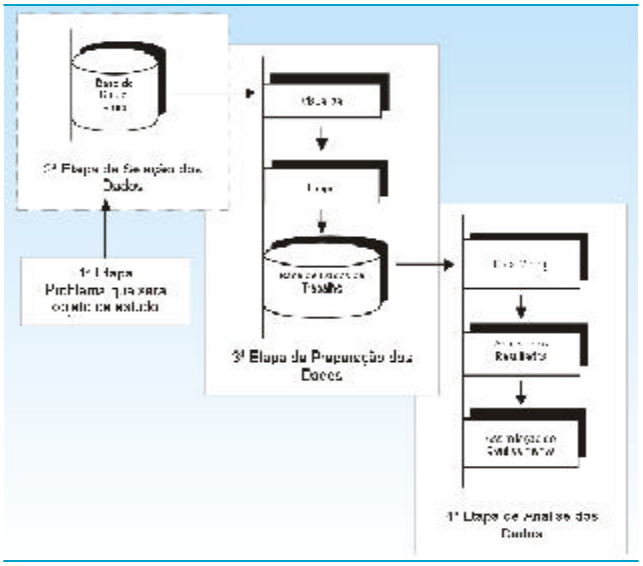
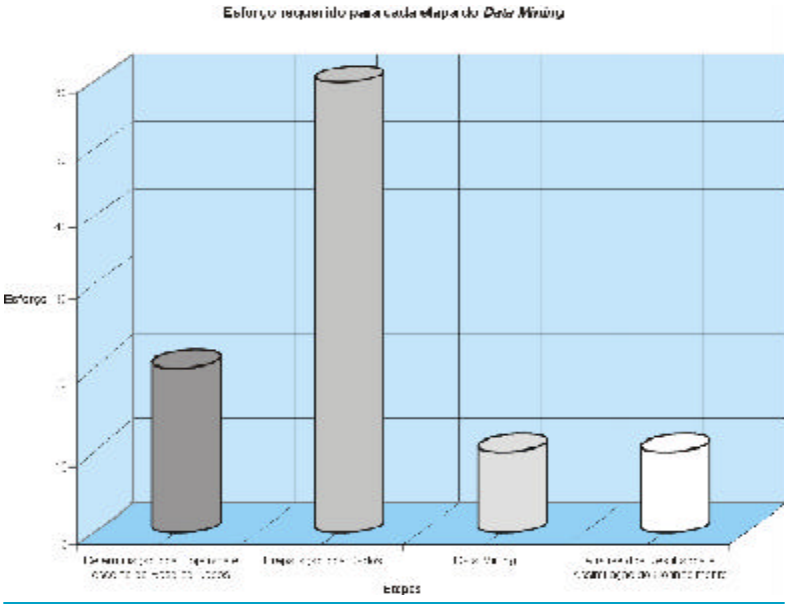


FIGURA 4
Esforço típico requerido para cada etapa do Data Mining



Fonte: Cabena *et alii*, 1997.

ESTUDO DE CASO PARA A APLICAÇÃO DO DATA MINING

Base de Dados DocThèses

A base de dados selecionada para o estudo de *Data Mining* foi a *DocThèses*, que é o catálogo das teses defendidas nas universidades francesas. Este catálogo é de responsabilidade da *Agence Bibliographique de l'Enseignement Supérieur (ABES)*, ligada ao Ministério da Educação

Nacional, da Pesquisa e da Tecnologia da França, e tem como missão alimentar o Sistema Universitário de Documentação, localizar e cadastrar os fundos documentários das bibliotecas de ensino superior, bem como monitorar a normalização da catalogação e da indexação dos trabalhos.

A base *DocThèses* é disponibilizada em *CD-ROM,* e, para o presente estudo, utilizou-se a versão de 2000. A extração para o estudo de caso correspondeu às teses que tiveram o Brasil como assunto da pesquisa. O total da amostra foi de 1.355 teses (registros bibliográficos), na qual também estavam incluídas todas as teses de brasileiros defendidas na França de 1969 até 1999.

O formato de cada referência bibliográfica (ocorrência) segue a seguinte estrutura:

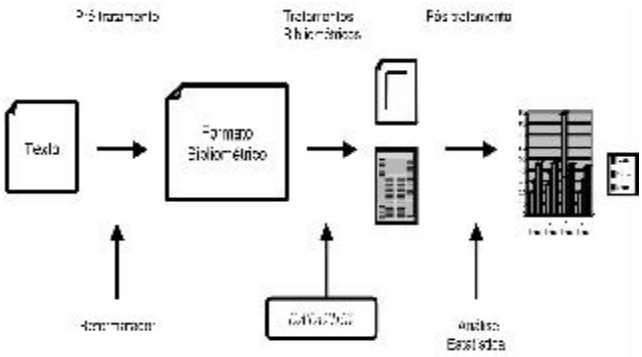
- autor;
- título;
- orientador;
- disciplina (área do conhecimento);
- palavras-chave;
- ano de defesa;
- universidade ou estabelecimento de defesa;
- texto integral.

Optamos por estudar várias tendências de comportamento, geradas e selecionadas a partir da aplicação do *software* bibliométrico [5] *Dataview*[6], que nas seções seguintes serão alvo de comentários e análises.

Metodologia simplificada

Após a etapa de preparação dos dados na qual foi empregado o *software Infotrans* versão 4.0[7] e pronta a base de dados de trabalho, passamos ao *Data Mining* empregando o *Dataview*, *software* bibliométrico de extração de indicadores de tendência elaborado pelo Centre de Recherche Rétrospective de Marseille (CRRM) da Universidade Aix-Marseille III, Centre de St. Jérôme, Marselha – França.

FIGURA5
Posição do Dataview em um estudo bibliométrico



Fonte: Rostaing, 2000.

O *Dataview* está assentado em métodos bibliométricos com o objetivo fundamental de transformar dados em inteligência para a tomada de decisão visando a gerar elementos para a análise estatística. Nesta tarefa, a reformatação dos dados é uma condição básica para o tratamento bibliométrico. Após a análise estatística, as informações extraídas deverão influir decisivamente na geração de conhecimento e inteligência, na qual dois aspectos deverão ser considerados: valor da informação e validade da informação para o processo decisório (Tarapanoff; Araújo Jr. & Cormier, 2000).

Tanto o valor quanto a validade da informação vão influir decisivamente na prospecção de conhecimento em bases de dados (*KDD*). Esta é a filosofia que deve nortear qualquer estudo referente à mineração de dados, bem como a geração de conhecimento. Apresentamos, na figura 5, a posição do *Dataview* em um estudo bibliométrico.

Outra importante característica do *software Dataview* em questão coincide com a característica de mensuração da bibliometria, estabelecida em bases numéricas, que, por sua vez, é realizada por meio da ocorrência. Assim sendo, para cada unidade ou elemento bibliográfico, deve-se considerar a ocorrência de três maneiras, em a) estado primário – simples localização das ocorrências, presença ou ausência dos elementos da referência; b) estado condensado – desdobramento destas ocorrências, ou freqüências; c) co-ocorrência, que representa a combinação do estado primário e do estado condensado. Desta maneira, serão geradas listas – freqüência de ocorrência e co-ocorrência e quadros –, matrizes de presença e ausências (Rostaing, 2000).

[5] Bibliometria – Estudo de aspectos da produção, distribuição e uso da informação registrada a partir de modelos matemáticos para o processo de tomada de decisão (Tarapanoff; Miranda & Araújo Jr., 1995). Aplicação de métodos estatísticos ou matemáticos sobre um conjunto de referências bibliográficas (Rostaing, 1996).

[6] Protótipo desenvolvido no *CRRM – Université de Aix-Marseille III* – França, cedido para a realização desta pesquisa.

[7] *Software* da *IuK Rieth GmbH* de reformatação empregado na etapa de preparação dos dados para o *Data Mining.*

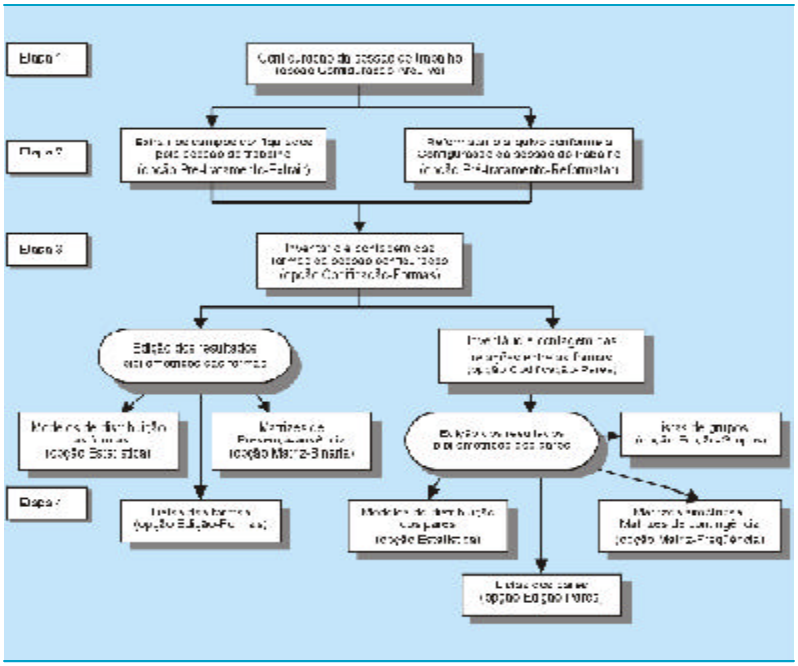
Na figura 6, apresentamos uma visão esquemática das etapas de uma sessão de trabalho no *Dataview.*

Importante para a entendimento dos dados é conhecer as três leis básicas da bibliometria:

1) Lei de Bradford (ou Lei da Dispersão): concentra sua descrição no comportamento repetitivo das ocorrências em um determinado campo do saber. Bradford escolheu o periódico para a sua análise, devido às suas características de incidência de assuntos e tendências, e observou que poucos periódicos produzem muitos artigos e muitos periódicos produzem poucos artigos.

2) Lei de Lotka: analisa a produção científica dos autores, ou seja, determina a contribuição de cada um deles para o progresso da ciência. A Lei de Lotka tem o seguinte enunciado: o número de autores que produzem ***n*** trabalhos é proporcional a ***1/1*** elevado a ***n²*** dos autores que produzem apenas um trabalho;

FIGURA 6
Etapas de uma sessão de trabalho no Dataview



Fonte: Rostaing, 2000.

3) Lei de Zipf: é a chamada lei quantitativa fundamental da atividade humana. Subdivide-se na Primeira Lei de Zipf, que corresponde à freqüência das palavras que aparecem em um texto (número de ocorrência das palavras). É regida pela seguinte expressão matemática:

$$K = R X F$$

Onde: ***K*** = constante; ***R*** = ordem das palavras; ***F*** = freqüência das palavras.

A Segunda Lei de Zipf estabelece as palavras de baixa freqüência que ocorrem de modo que várias palavras acabam por apresentar a mesma freqüência (Tarapanoff, Miranda & Araújo Jr., 1995).

Para este trabalho, consideraremos a curva Zipf conforme a figura 7.

Segundo Quoniam (1992), na curva de Zipf temos:

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. É aí que as transferências de tecnologia relacionadas aos novos temas devem ser consideradas;

FIGURA 7
Curva de Zipf

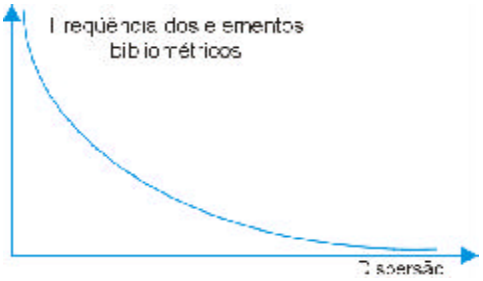
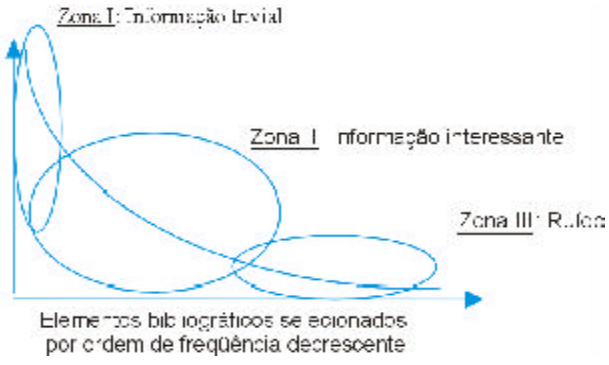


FIGURA 8
Zonas de distribuição



Fonte: baseado em Quoniam, 1992.

Zona III – Ruído: tem como característica possuir conceitos ainda não emergentes, onde é impossível afirmar se eles serão emergentes ou se são apenas ruído estatístico;

As Zonas I, II e III estão representadas na curva de Zipf, conforme a figura 8 (ver p. 24).

A partir deste referencial, optamos por apresentar os resultados do *Data Mining* aplicado na Base *DocThèse*, considerando apenas as Zonas I e II pelos seus aspectos de definição dos temas centrais da análise bibliométrica e da informação potencialmente inovadora, respectivamente. Os resultados são apresentados no gráfico 1.

As áreas de economia, sociologia e ciências tecnológicas correspondem a um terço do total de teses que tinham o Brasil como país de origem do pesquisador ou como tema da pesquisa, seguidas de perto com 101 e 98 teses pelas áreas de geografia e biologia respectivamente, conforme podemos verificar no gráfico 1, e que correspondem à Zona I – Informação trivial.

Com efeito, a França, por possuir, junto com a Alemanha, uma das mais importantes e tradicionais escolas de sociologia, torna-se, por conseqüência, atraente para o desenvolvimento de trabalhos acadêmicos nesta área, fato comprovado pelo gráfico 1. O mesmo acontece com a economia, em que podemos somar também forte interesse por assuntos latino-americanos. Assuntos de procura e interesse regular pelos alunos que pesquisam sobre o Brasil.

A área tecnológica, por sua vez, tem na França um dos países líderes mundiais no desenvolvimento tecnológico, abrigado por um eficiente sistema de inovação tecnológica que justifica sua posição no *ranking* desta pesquisa. Nessas áreas, dos orientadores presentes, nota-se pela tabela 1 que a produção se concentra em torno dos profissionais que juntos respondem por 20%, 18,8% e 7,1% do total das teses defendidas.

Cabe ressaltar que as áreas analisadas foram as responsáveis pelo crescimento e pela presença de brasileiros na França no período compreendido até 1994, tendo a partir de então entrado em declínio sua procura.

A Zona II – Informação interessante, por sua vez, representa aquelas áreas emergentes, o que se comprova através das áreas de pedagogia, ciências médicas, estudos ibero-americanos e história, que estão no período que vem desde 1995 em ascensão. Alguns dos fatos que influenciam o interesse por estas áreas do conhecimento residem no peso do novo paradigma científico e tecnológico, como é

ANÁLISE DOS RESULTADOS

Ocorrência de teses com a recuperação do termo "Brasil" por áreas do conhecimento

GRÁFICO 1
Ocorrência por Áreas do Conhecimento

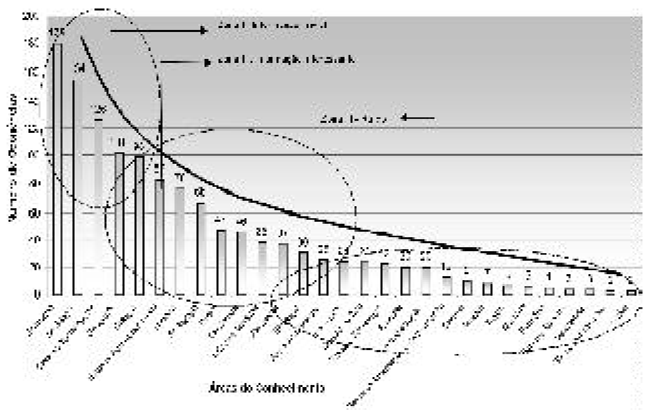


TABELA 1
Orientadores que se destacaram por área do conhecimento

Área do conhecimento	1º grupo de orientadores que se destacaram	% do total de teses orientadas	2º grupo de orientadores que se destacaram	% do total de teses orientadas
Zona I – Informação Trivial				
Economia	I. Sachs	[illegible]	P. Salama	[illegible]
Sociologia	[illegible]	[illegible]	[illegible]	[illegible]
Ciências Tecnológicas	[illegible]	[illegible]	[illegible]	[illegible]
Geografia	M. Rochefort	[illegible]	[illegible]	[illegible]
Zona II – Informação Interessante				
Biologia	[illegible]	[illegible]		
Estudos Ibero-americanos	[illegible]	[illegible]		
História	F. Mauro	25,0%	K. Mattoso	[illegible]
Pedagogia	[illegible]	[illegible]	[illegible]	[illegible]
Direito	[illegible]	[illegible]		
Linguística	[illegible]	[illegible]		
Ciências Médicas	[illegible]	7,9%		
Psicologia	Sem concentração significativa			
Etnologia	[illegible]	[illegible]	M. Godelier	10,0%

o caso das áreas de educação e medicina, permeadas constantemente por novas descobertas e tecnologias que tendem a avançá-las no conhecimento. No caso da história, por estarmos vivendo um período de forte transição neste tipo de sociedade, obriga releituras constantes e busca de explicações incessantes sobre a sua nova condição.

Destaca-se ainda na área da história Fré5deric Mauro, que dentre todos os orientadores, foi o professor que mais orientou teses no período de 1969 a 1999, com 25% do total relativo ao primeiro grupo de orientadores (Zona II – Informação interessante), conforme o gráfico 2, a seguir, pode demonstrar. Este desempenho se deve

fundamentalmente a uma influência preponderante da historiografia francesa na academia brasileira. Na década de 30, uma missão francesa composta por vários professores de diferentes áreas trouxe ao Brasil o eminente professor Fernand Braudel, um dos artífices da primeira geração fundadora da Escola de Análise francesa, que reunia ainda importantes historiadores, tais como March Bloch e Lucien Fébvre. Na época e por conta da missão francesa, é fundado o Departamento de História da Universidade de São Paulo, inaugurando a decisiva influência da historiografia francesa no Brasil. No caso específico do professor Fréderic Mauro, a sua importância é grande nesta escola historiográfica, junto com Georges Duby e Jacques Le Goff entre outros, pertence à segunda geração.

Como conseqüência dos fatos relatados, podemos afirmar que não só a incidência de teses orientadas por Mauro respondem pela importância numérica da área de história verificada, mas que, sem dúvida nenhuma, deve-se também ao fato de que a historiografia francesa é a principal catalisadora do interesse dos historiadores brasileiros que buscam formação no exterior.

Concentração de teses defendidas por cidades

Ao traçar a trajetória dos pesquisadores na França, obtivemos o gráfico 3, cujo resultado demonstra que 62% das teses defendidas concentram-se em Paris. Das 40% restantes, apenas em Montpellier, Toulouse, Marseille, Grenoble e Bordeaux se concentram 50%. Os demais 50% aproximadamente estão em outras 30 cidades francesas[8].

Período de defesa das teses entre 1969 e 1999

Por meio da análise da tabela 2, verificamos que somente a área do Direito, no período de 1974 a 1978, alcançou altos índices de interesse, a maior concentração observada em relação a todas as demais áreas. Esta situação chama a atenção e pode ser explicada, em parte, pelas circunstâncias políticas vividas no Brasil na década de 70.

A coincidência da alta concentração de teses defendidas na França com o auge da ditadura militar implantada no Brasil a partir de 1967 elevou o interesse pela compreensão do estado de direito implantado, sobretudo relativo aos direitos e garantias fundamentais do cidadão.

Sobre a área da lingüística, nota-se que o ápice ocorreu no período de 1980 a 1984, tendendo a partir de 1995 a recuperar o interesse.

[8] As instituições de ensino superior francesas têm o mesmo nome das cidades que as abrigam. Desta forma, quando falamos na concentração de teses defendidas em *Marseille*, estamos nos referindo a *Université de Aix-Marseille* e assim por diante.

GRÁFICO 2
Orientadores que se destacaram por área do conhecimento por grupos

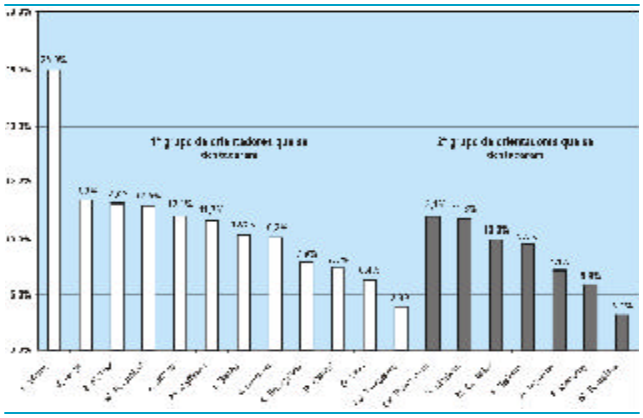


GRÁFICO 3
Concentração de teses defendidas por cidade

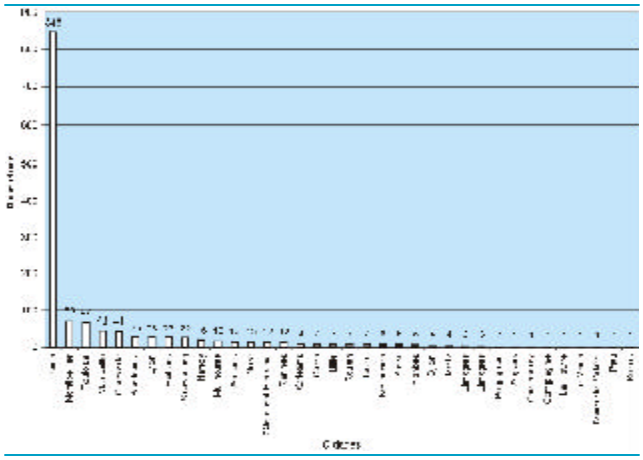


TABELA 2
Incidência de áreas do conhecimento por períodos de anos (1969 – 1999)

Áreas do conhecimento	1995-1999	1990-1994	1985-1989	1980-1984	1975-1979	1974-1978	1969-1973
Arte e Arqueologia	18	4	4	8	0	0	1
Biologia	38	50	19	10	2	*	1
Ciência das Religiões	1	0	1	0	0	0	0
Ciência e Tecnologia	5	14	3	0	2	*	0
Economia	12	45	39	61	23	15	3
Ciência Política	10	2	5	3	0	*	2
Pedagogia	17	17	12	15	5	3	0
Ciências Médicas	14	10	12	0	0	0	0
Cinema	3	2	3	0	0	0	0
Direito	5	14	7	9	10	70	0
Estudos Ibero-americanos	28	18	16	17	5	7	0
Filologia	9	13	3	4	1	*	0
Filosofia	2	2	1	0	0	0	0
Filosofia	0	9	3	4	4	2	0
Física	1	1	0	0	0	0	0
Geografia	20	28	19	16	9	6	7
Administração	1	6	1	6	0	0	0
História	20	19	17	9	7	7	4
Ciência da Informação e Comunicação	1	2	1	5	3	2	0
Lingüística	10	9	7	14	4	3	2
Literatura	3	5	4	4	0	0	1
Música	3	3	0	0	1	*	0
Odontologia	0	0	2	0	0	0	0
Psicologia	5	18	5	5	2	3	2
Química	2	1	2	0	0	0	0
Semiologia	0	0	1	0	1	0	0
Sociologia	42	50	21	10	18	13	6
Teatro	2	0	4	0	0	0	0
Ciências Tecnológicas	52	59	25	7	5	3	1
Veterinária	2	1	0	0	0	0	0

Grosso modo, em relação ao número de teses defendidas ao longo do período, constata-se que, desde 1996, o número vem decaindo velozmente, conforme podemos visualizar no gráfico 4. O motivo para tal ocorrência, talvez, resida no fato de que desde 1999 há contingenciamento de bolsas de estudos para o exterior nas áreas de humanas e sociais, o que faz com que a área tecnológica sozinha não alcance altos índices como no conjunto.

É interessante verificarmos que no período de relativo equilíbrio da curva, que oscilou entre 36 e 58 teses defendidas entre 1980 e 1990, houve uma média de aproximadamente 47 teses defendidas a cada ano, sobressaindo-se neste período a área de economia.

Na área de ciência da informação, foram 12 teses defendidas entre 1974 e 1999. O período áureo deu-se entre 1980 e 1984, com um total de cinco trabalhos. Dentre os orientadores, destacam-se F. Ballet seguido de J. Meyriat. Os outros cinco, cada qual responsável por uma tese, são P. Albert, M. Menou, M.Mouillard, J. Perriault e G. Thibault, sendo Bordeux a base desse último e o único fora da cidade de Paris. M. Menou[9], por sua vez, tem estado a serviço da ciência da informação como consultor internacional, no Canadá, onde desenvolve vários trabalhos na linha de impacto da informação para o desenvolvimento. No Brasil, desenvolveu extensa consultoria junto ao Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), ligado ao Ministério da Ciência e Tecnologia.

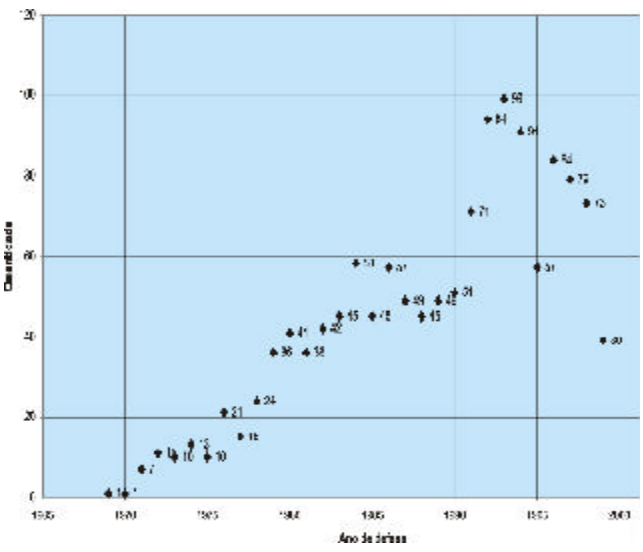
Com relação a Zona III – a chamada zona de ruído, apesar de sua característica de não possuir ainda conceitos emergentes , consagrados e ser pouco conclusiva, deve ser monitorada sistematicamente, pois pode revelar, ou até mesmo permitir, na análise de sinais fracos, a inferência de interesses futuros de aperfeiçoamento e pesquisa. Desta maneira, não devemos desprezá-la *a priori*. Nessa zona aparecem as áreas de arte e arqueologia, literatura, ciência política, ciência e tecnologia, filosofia, administração, ciência da informação e comunicação entre outros.

CONCLUSÃO

A análise da Base de Dados *DocThèses* em relação à recuperação da palavra Brasil por meio do *Data Mining* foi reveladora em relação às áreas do conhecimento escolhidas, orientadores que se destacaram, período cronológico de maior concentração de teses defendidas e cidades escolhidas.

[9] Michel Menou é autor de livro básico para a área de ciência da informação, em conjunto com Claire Guinchat: *Introdução geral às técnicas da informação e documentação.* Tradução de Míriam Vieira da Cunha. Brasília : IBICT, 1994.

GRÁFICO 4
Incidência de defesa de teses por períodos de anos



A descoberta de conhecimento foi ocorrendo gradativamente, à medida que o processo de *Data Mining* foi se concretizando. Na primeira etapa – definição do problema –, optou-se por explorar a base no que se refere ao Brasil, tanto em palavras-chave como por origem do orientador. A segunda etapa de limpeza dos dados levou ao primeiro contato com os dados, extraindo apenas aqueles potencialmente de interesse para a descoberta de algum padrão. Na terceira etapa, a realização do *DM* propriamente dito, optou-se pela utilização do *software Dataview*, que já traz embutido no sistema regras de estatística e de visualização de dados para a descoberta do conhecimento. As primeiras análises e constatações advêm desta fase, de acordo com o objeto da pesquisa. Na quarta etapa, análise dos dados, por fim, novas associações foram realizadas e o conhecimento emergiu.

Os resultados obtidos são uma amostra de como órgãos nacionais de fomento à pesquisa e de formação de recursos humanos de alto nível, como aFundação Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (Capes) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), podem direcionar seus investimentos em áreas do conhecimento consideradas relevantes, por intermédio do conhecimento descoberto nas bases de dados.

Os indicadores obtidos e sua aplicação reforçam o objetivo do processo de *Data Mining* no sentido de transformar dados em informação a ser empregada no processo decisório das organizações que tomam decisão relacionada à preservação e à inovação do conhecimento. As áreas de economia, sociologia e história, embora não sejam prioritárias como áreas de ponta para o desenvolvimento

no Brasil, são essenciais para entender os processos econômico, social e histórico brasileiros, que têm forte presença e influência francesa do ponto de vista de orientação teórica e cultural. Essas áreas merecem continuar a receber investimentos. O impacto desta influência se fez presente na Mostra do Redescobrimento do Brasil (2000), na qual muitos documentos de viajantes franceses atestaram a sua presença e influência no país. Outras áreas, como as das ciências tecnológicas, mereceriam reflexão sobre a sua retração e ainda outras áreas de interesse emergente mereceriam reflexão para a composição de acordos de cooperação bilateral técnico, cultural e econômico.

Impossível esgotar todas as inferências de acordos políticos, técnicos, econômicos e culturais que podem ser obtidas graças a análises de bases de dados do tipo estudado. Outras bases, de diversas procedências, de outros países trariam basicamente possibilidades de inferências diferenciadas.

O presente artigo pode ser o início de uma série abordando a trajetória do interesse de pesquisa pelo Brasil em vários países, buscando paralelismos e descobrindo conhecimento entre os resultados encontrados, aplicando o *Data Mining* como uma efetiva ferramenta gerencial.

REFERÊNCIAS BIBLIOGRÁFICAS

ACM special interest group on knowledge discovery in data and data mining. Disponível em: < http://www.acm.org/sigkdd/ >

ALVARES, Lillian. *Aplicação de data mining em bases de dados especializadas em ciência da informação para obtenção de informações sobre a estrutura de pesquisa e desenvolvimento em ciência da informação no Brasil*. Brasília, 2000. Monografia (Especialização) UFRJ/ECO, MCT/INT/IBICT.

CABENA, Peter *et al. Discovering data mining:* from concept to implementation. New Jersey : Prentice Hall, 1997.

DATAMATION magazine. Disponível em: < http://www.datamation.com>

FAYYAD, U.M. *et al. Advances in knowledge discovery and dataming.* Cambridge, Ma : AAAI Press, 1996.

INFORMATION discovery. Disponível em: < http://www.datamining.com>

INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING. Disponível em: < http://www.digimine.com/usama/datamine/kdd99/ >

MOOERS, C. N. Zatacoding applied to mechanical organization of knowledge. *American Documentation*, v. 2, p. 20-32, 1951.

NORTON, M. Jay. Knowledge discovery in databases. *Library Trends*, v. 48, n. 1, p. 9-21, Summer 1999.

QUONIAM, Luc. Bibliométrie sur des référence bibliographiques: methodologie. *In*: DESVALS H.; DOU, H. (Org.). *La veille technologique.* Paris : Dunod, 1992. p. 244 – 262.

QUONIAM, Luc. *Les productions scientifiques en bibliométrie et dossier de travaux*. Marseille : Université de Droit d'Economie et des Sciences d'Aix-Marseille III, 1996.

ROSTAING, Hervé. *La bibliometrie et ses tecniques*. Toulouse : Sciences de la Societé, 1996.

ROSTAING, Hervé. *Guide d'utilisation de dataview:* logiciel bibliométrique d'aide à l'élaboration d'indicateurs de tendances. Marseille : CRRM, [2000].

TARAPANOFF, Kira.; ARAUJO JÚNIOR., Rogério Henrique de; CORMIER, Patricia Marie Jeanne. Sociedade da informação e inteligência em unidades de informação. *Ciência da Informação*, Brasília, v. 29, n. 3, p. 91-100, set./dez. 2000.

TARAPANOFF, Kira.; MIRANDA, Denir Mendes; ARAUJO JÚNIOR., Rogério Henrique de. *Técnicas para tomada de decisão nos sistemas de informação*. Brasília : Thesaurus, 1995.

TARAPANOFF, Kira. (Org.). *Inteligência organizacional e competitiva.* Brasília : Editora Universidade de Brasília, 2001.

TYSON, K. W. M. *The complete guide to competitive intelligence*. Chicago: Kirk Tyson International, 1998.

ZANASI, Alessandro. Competitive intelligence trough data mining public sources. *Competitive Intelligence Review*, v. 9, n. 1, p. 44-54, 1998.

Agradecimentos

Os autores agradecem os esclarecimentos referentes à análise dos resultados obtidos neste trabalho aos seguintes especialistas: Patricia Marie Jeanne Cormier (especialista em inteligência competitiva), Carlos Henrique Araújo (mestre em sociologia) e Nildo Luzio (mestre em história).