EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO EM C&T: O ALGORITMO APRIORI

Wesley Romão, Carlos A. P. Niederauer, Alejandro Martins, Aran Tcholakian, Roberto C. S. Pacheco e Ricardo M. Barcia

Programa de Pós-Graduação em Engenharia de Produção - PPGEP
Universidade Federal de Santa Catarina, Centro Tecnológico
C. P. 476 - CEP 88040-900 - Florianópolis, SC, Brasil
E-mail: romao@eps.ufsc.br

ABSTRACT: this paper considers the problem of discovering association rules between items in the Brazilian Research Groups Database. The Apriori algorithm was used to mining this database searching for association rules. The aim of this work is show how Apriori can be a real alternative to traditional decision making tools. In conclusion, is showed that Apriori discovered interesting association rules, although new tests with some adjust and improvement will be necessary in the futures works.

KEYWORDS: data mining, association rules, Apriori algorithm.

RESUMO: o trabalho utiliza uma técnica de *Data Mining*, o algoritmo *Apriori*, para descobrir regras de associação no banco de dados do Diretório dos Grupos de Pesquisa no Brasil, versão 3.0, recentemente disponibilizado pelo CNPq. O objetivo é demonstrar a viabilidade do algoritmo como uma ferramenta alternativa na avaliação de C&T. Utilizou-se o pesquisador como unidade de análise. Os resultados conduziram à descoberta de regras consistentes e interessantes, confirmando o potencial da ferramenta como instrumento de gestão de C&T.

1. INTRODUÇÃO

A grande quantidade de dados acumulados nos bancos de dados informatizados das organizações pode esconder conhecimentos valiosos e úteis para a tomada de decisão. Há uma relação inversa entre o volume de dados existentes e a necessidade de conhecimento estratégico, ou seja, apesar das

informações resumidas e significativas para tomada de decisão serem de volume menor, geralmente ela não está disponível e exige a sua extração a partir de grandes quantidades de dados que crescem com o tamanho e a idade das empresas. Uma seqüência natural neste processo seria: Dados → Informação → Conhecimento → Decisão. Neste contexto, o desafio que se apresenta para as organizações pode ser encarado como a resolução de duas questões básicas:

1. Como organizar os dados?
2. Como extrair conhecimento dos dados organizados?

A primeira questão pode ser equacionada através da construção de um *Data Warehouse*, tecnologia que permite armazenar as informações, anteriormente dispersas, através da identificação, compreensão, integração e agregação dos dados, de forma a posicioná-los nos locais mais apropriados visando a atender à estratégia organizacional das empresas (Brackett,1996).

Entretanto, em resposta à segunda questão, para extrair conhecimento de um sistema de *Data Warehouse*, são necessárias ferramentas de exploração, hoje conhecidas como *Data Mining* (Mineração de Dados). O *Data Mining* reúne uma série de técnicas, com destaque para as estatísticas, probabilísticas e de inteligência artificial, capazes de fornecer respostas a várias questões ou mesmo descobrir novas informações em grandes bancos de dados. O *Data Mining* é especialmente útil em casos nos quais não se conhece a pergunta, mas, mesmo assim, existe a necessidade de respostas. Isto o distingue, por exemplo, de um *Executive Information System* (Nigro, 1997).

Considere, como exemplo, um banco de dados contendo registros de clientes e mercadorias vendidas. Uma consulta ao banco de dados para a extração da informação poderia ser: "Quantos computadores foram vendidos para o cliente X na data dd/mm/aa?". Esta seria uma operação comum da baixa administração da empresa. Entretanto, as técnicas de *Data Mining* visam atender às aplicações de níveis administrativos mais elevados, tais como: marketing (mala direta direcionada), planejamento de estoque, abertura de novas filiais e outras decisões estratégicas. Uma técnica de *Data Mining* poderia extrair conhecimento do tipo “SE (idade = '[25 a 35] anos') E (profissão = 'advogado') ENTÃO (compra = 'computador')” com uma freqüência, por exemplo, de 90%. A informação obtida poderia ser usada para responder a uma provável pergunta do setor de marketing: Quais os clientes que têm alta probabilidade de comprar computadores?

Existem diversas técnicas de *Data Mining* disponíveis na literatura (Chen *et alli*, 1996; Cheung *et alli*, 1996). Uma das técnicas mais atraentes é a Mineração de Regras de Associação, que tem como

destaque o algoritmo *Apriori*. Ele pode trabalhar com um número grande de atributos, gerando várias alternativas combinatórias entre eles. O algoritmo Apriori realiza buscas sucessivas em toda a base de dados, mantendo um ótimo desempenho em termos de tempo de processamento (Agrawal & Srikant, 1994).

O algoritmo *Apriori* foi aplicado à base de dados do Diretório dos Grupos de Pesquisa no Brasil, versão 3.0, o qual foi recentemente disponibilizado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para o público em geral. O Diretório se constitui numa fonte estratégica de informações sobre a pesquisa brasileira, inventariando os pesquisadores e sua produção intelectual. Os dados coletados pelo Diretório dizem respeito ao triênio 1995-97.

Uma das metas do CNPq é construir ferramentas para a aquisição e análise de dados mais aprofundados relativos ao sistema brasileiro de Ciência e Tecnologia (C&T). Esta expectativa motivou a aplicação do *Data Mining* à atual versão do Diretório.

O trabalho objetiva demonstrar a viabilidade do emprego do algoritmo *Apriori* aplicado à analise de pesquisadores e grupos de pesquisa. Esta abordagem pode ser um instrumento valioso no auxílio à gestão e à política de C&T. Além do mais, como objetivo secundário, este trabalho tem um caráter didático, pois o algoritmo foi implementado em um computador pessoal, podendo ser reproduzido e testado sem que seja necessário recorrer a computadores de maior porte, ambiente em que o *Apriori* geralmente é executado.

O artigo está organizado em seis seções, incluindo esta Introdução. Na seção seguinte, explora-se o tema de mineração de regras de associação, em especial o algoritmo *Apriori*. Na seção 3, apresenta-se o Diretório, traçando um perfil do mesmo. A seção 4 traz a metodologia adotada. Na seção 5, são apresentados os resultados e, na seção 6, as conclusões do trabalho.

2. MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO

Freqüentemente, em grandes bancos de dados que armazenam milhares de itens, como aqueles existentes em redes de supermercados, deseja-se descobrir associações importantes entre os itens comercializados, tal que a presença de alguns deles em uma transação (compra e venda) implique na presença de outros na mesma transação. O objetivo, então, é encontrar todas as regras de associação relevantes entre os itens, do tipo X (antecedente) $\Rightarrow$ Y (conseqüente). Para tratar desta questão, Agrawal *et alli* (1993) propuseram um modelo matemático, onde as regras de associação geradas devem atender a um *suporte* e *confiança* mínimos especificados pelo decisor. O *suporte*

corresponde à freqüência com que ocorrem os padrões em toda a base; é a percentagem dos pesquisadores da base que possuem X e Y. Enquanto que a *confiança* é uma medida da força das regras, vale dizer, a percentagem de pesquisadores de X que possuem Y.

O suporte mínimo (*minsup*) é a fração das transações que satisfaz a união dos itens do conseqüente com os do antecedente, de forma que estejam presentes em pelo menos *s%* das transações no banco de dados. A confiança mínima (*minconf*) garante que ao menos *c%* das transações que satisfaçam o antecedente das regras também satisfaçam o conseqüente das regras.

2.1 DECLARAÇÃO FORMAL DO PROBLEMA

Seja um banco de dados contendo informações sobre os pesquisadores atuantes no Brasil. Almeja-se descobrir associações importantes entre esses dados, onde:

- $I = \{i_1, i_2, ..., i_m\}$ é um conjunto de literais, denominados itens. São as características e atributos dos pesquisadores. Por exemplo, I = {idade, sexo, ..., área de atuação, artigos publicados};
- T é um conjunto de certos itens de um pesquisador, tal que $T \subseteq I$;
- D é uma tabela representando todas as características e atributos de todos os pesquisadores; e
- X, Y são conjuntos de itens específicos dos pesquisadores, tal que $X \subseteq T$ e $Y \subseteq T$.

Uma regra de associação é uma implicação da forma $X \Rightarrow Y$, onde $X \subset I$, $Y \subset I$ e $X \bullet Y = \varnothing$. A regra $X \Rightarrow Y$ pertence a D com confiança c se $c\%$ dos registros em D que contém X também contém Y. A regra $X \Rightarrow Y$ tem suporte s em D se $s\%$ dos registros em D contém $X \cup Y$. (Agrawal, 1994). Então, dado uma tabela D, o objetivo é descobrir as regras de associação interessantes.

O problema de descobrir todas as regras de associação, tal como formulado por Agrawal *et alli* (*op. cit.*), pode ser decomposto em duas etapas:

1. encontrar todos os conjuntos de itens (*itemsets*) que apresentam suporte maior que o suporte mínimo estabelecido pelo decisor. Os *itemsets* que atendem a este quesito são denominados *itemsets freqüentes*; e
2. utilizar os *itemsets freqüentes* obtidos para gerar as regras de associação do banco de dados.

O desempenho geral da mineração de regras de associação é determinado pela primeira etapa, a qual exige sucessivas buscas na base de dados. Em se encontrando o conjunto dos *itemsets freqüentes*, as regras de associação correspondentes podem ser diretamente identificadas.

Algoritmos que realizem a contagem eficiente dos grandes *itemsets* são a chave para o sucesso dos métodos de mineração em grandes bancos de dados (Chen *et alli*, 1996).

2.2. O ALGORITMO *APRIORI*

O algoritmo *Apriori* é um dos algoritmos mais conhecidos quando o assunto é mineração de regras de associação em grandes bancos de dados centralizados. Ele encontra todos os conjuntos de itens freqüentes, denominados *itemsets freqüentes* (L_k).

O algoritmo principal (Apriori) faz uso de duas funções: a função Apriori_gen, para gerar os candidatos e eliminar aqueles que não são freqüentes, e a função Genrules, utilizada para extrair as regras de associação. Os Anexos I, II e II apresentam, respectivamente, os códigos do algoritmo Apriori, da função Apriori_gen e da função Genrules. O algoritmo e as duas funções foram implementadas através do ambiente de simulação Matlab®.

O primeiro passo do algoritmo Apriori (Anexo I) é realizar a contagem de ocorrências dos itens para determinar os *itemsets freqüentes* de tamanho unitário (1-*itemsets freqüentes*). Os passos posteriores, *k*, consistem de duas fases. Primeiro, os *itemsets freqüentes* L_{k-1}, encontrados no passo anterior (k-1) são utilizados para gerar os conjuntos de itens potencialmente freqüentes, os *itemsets candidatos* (C_k). O procedimento para geração de candidatos é descrito no parágrafo seguinte. Na seqüência, é realizada uma nova busca no banco de dados, contando-se o suporte de cada candidato em C_k.

A geração dos *itemsets candidatos*, de antemão, toma como argumento L_{k-1}, o conjunto de todos (k-1)-*itemsets freqüentes*. Para tal, utiliza-se a função *Apriori_gen* (Anexo II), que retorna um superconjunto de todos os *k-itemsets freqüentes.* A intuição por trás desse procedimento é que se um *itemset* X tem suporte mínimo, todos os seus subconjuntos também terão (Agrawal & Shafer, 1996). A função, em um primeiro estágio, une L_{k-1} com L_{k-1}. No estágio seguinte, são eliminados os *itemsets* $c_k \in C_k$, desde que um dado (k-1)-*subset* de c_k não pertença a L_{k-1}.

Para viabilizar a implementação, no Matlab®, do algoritmo A*priori* e da função *Apriori-gen*, extraídos de Agrawal & Srikant (1996), efetuou-se uma adaptação da estrutura de dados substituindo a estrutura original, em forma de *hash tree*, por matrizes, mantendo a lógica e seqüência originais do algoritmo.

O último passo é a descoberta das regras de associação, obtida através da função *Genrules* (Anexo III). A geração de regras, para qualquer *itemset freqüente*, significa encontrar todos os *subsets* não

vazios de *l*. Assim, para todo e qualquer *subset a*, produz-se uma regra $a \Rightarrow (l - a)$ somente se a razão (suporte (*l*)/suporte(*a*)) é ao menos igual a confiança mínima estabelecida pelo usuário.

Para gerar regras com múltiplos conseqüentes, são considerados todos os *subsets*. Por exemplo, dado um *itemset ABCD*, considera-se primeiro o *subset ABC*, seguido de *AB*, etc. Se $ABC \Rightarrow D$ não atinge uma confiança suficiente (confiança < *minconf*), não é necessário verificar se $AB \Rightarrow CD$.

2.3. UM EXEMPLO DA EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO

O exemplo a seguir, adaptado de Chen *et alli* (1996) e de Agrawal & Srikant (1994), demonstra como funciona a extração de regras de associação através do *Apriori*.

Suponha um banco de dados formado somente por um grupo de pesquisa, GP. Suponha, também, que este grupo seja composto por cinco pesquisadores, associando-se a cada um deles cinco itens categóricos, conforme mostrado na Tabela 1.

Registro	Itens Categóricos				
(pesquisador)	*A*	*B*	*C*	*D*	*E*
501	1	0	0	1	0
502	0	1	1	0	1
503	1	0	1	0	1
504	0	1	1	0	1
505	0	1	1	0	0
freqüência	2	3	4	1	3

Legenda:

A = pesquisador estrangeiro; *B* = pesquisador brasileiro;
C = sexo feminino; *D* = sexo masculino; e
E = pesquisador doutor.
0 = ausência do atributo; e 1 = presença do atributo.

Tabela 1 - Dados do fictício grupo de pesquisa GP.

Seja C_k o conjunto de *k-itemsets candidatos*, onde $k = 5$. Cada membro c_k deste conjunto tem dois campos: *itemset* e *contador de suporte*, representados, respectivamente, por *its* e *cs* na Figura 1. Seja L_k o conjunto dos *k-itemsets freqüentes*. De modo análogo, cada membro deste conjunto também possui *its* e *cs*.

O primeiro passo do algoritmo conta a freqüência com que os itens ocorrem para determinar os 1-*itemsets freqüentes* (última linha da Tabela 1). Posteriormente, obtém-se o conjunto de candidatos 1-*itemsets*, C_1, mostrado na Figura 1. Assumindo um suporte mínimo igual a dois, ou seja, *minsup* = 40%, L_1 é composto pelos elementos de C_1 com suporte igual ou superior a 40%. No exemplo,

somente o *itemset D* não atendeu a esta condição, ficando L_1 composto por {*A*}, {*B*}, {*C*} e {*E*}.

Para descobrir o conjunto dos 2-*itemsets freqüentes*, de modo a continuar satisfazendo ao suporte mínimo, o *Apriori* usa a concatenação $L_1 * L_1$ para gerar o conjunto candidato C_2, que consiste de 2-*itemsets*. Por exemplo, {*C*} e {*E*} geram {*CE*}. Mais uma vez, cada ocorrência é computada. No caso {*CE*} ocorre três vezes em GP (registros 502, 503 e 504). L_2 é determinado com base no suporte de cada candidato de C_2. Agora são excluídos {*AB*}, {*AC*} e {*AE*}, pois têm suporte inferior ao mínimo estabelecido.

C1	
its	*cs*
{*A*}	2
{*B*}	3
{*C*}	4
{*D*}	1
{*E*}	3

L1	
its	*cs*
{*A*}	2
{*B*}	3
{*C*}	4
{*E*}	3

C2	
its	*cs*
{*AB*}	0
{*AC*}	1
{*AE*}	1
{*BC*}	3
{*BE*}	2
{*CE*}	3

L2	
its	*cs*
{*BC*}	3
{*BE*}	2
{*CE*}	3

C3		**L3**	
is	*cs*	*its*	*cs*
{*BCE*}	2	{*BCE*}	2

Figura 1 - Geração dos *itemsets candidatos* (*C*) e dos *itemsets freqüentes*(*L*).

A geração de C_3 é obtida a partir de L_2 de uma maneira distinta. Os futuros *itemsets* candidatos devem manter uma ordem lexicográfica, tal que quando a concatenação $L_2 * L_2$ for realizada, o primeiro item de um *itemset* seja idêntico ao primeiro item do outro *itemset* e assim sucessivamente. Porém, o último item do *itemset* deve ser menor, lexicograficamente, que o último item do outro *itemset.* Esta regra pode ser representada como

$$itemset\_p = \{p_1, p_2, ..., p_n\},\ itemset\_q = \{q_1, q_2, ..., q_n\} \qquad (1)$$

sendo necessário que

$$p_1 = q_1, p_2 = q_2, ..., p_n < q_n \qquad (2)$$

Na Figura 1, o *itemset* candidato {*BCE*}, em C_3, foi formado concatenando {*BC*} com {*BE*}, pois *B* = *B* e *C* < *E*. Este foi o único conjunto que pôde ser formado, pois não há outra concatenação que satisfaça (2). A concatenação {*BC*}*{*CE*}, por exemplo, não satisfaz (2), pois, lexicograficamente, $p_1 = B$ é menor que $q_1 = C$.

O passo seguinte é descobrir as regras de associação. No caso do fictício grupo de pesquisa GP, supondo uma confiança mínima de 60% e mantendo o suporte mínimo em 40%, uma regra provável seria $BC \Rightarrow E$. Para ela, a confiança é igual suporte(*BCE*)/suporte(*BC*), cujo resultado é $\frac{2}{3}$, ou 66%,

satisfazendo a condição imposta (ver Tabela 2).

Para a regra $BC \Rightarrow E$ ou (pesquisador brasileiro; sexo feminino) $\Rightarrow$ (pesquisador doutor), seu suporte seria o percentual de ocorrências de *BCE* com relação ao total de pesquisadores do grupo, que resulta em 40% ($\frac{2}{5}$). Então, esta é uma regra válida. Isto eqüivale a dizer que, das pesquisadoras brasileiras, 66% têm doutorado, muito embora estas brasileiras portadoras do título de doutor correspondam a apenas 40% dos indivíduos do grupo.

Outra provável regra seria $B \Rightarrow CE$. Para esta situação, o valor da confiança seria idêntico, pois a razão suporte (*BCE*)/suporte (*B*) também é igual a $\frac{2}{3}$.

Registro (pesquisador)	Itens Categóricos		
	B	*C*	*E*
501	0	0	0
502	1	1	1
503	0	1	1
504	1	1	1
505	1	1	0
Freqüência	3	4	3

Tabela 2 - Ocorrências dos conjuntos de atributos

3. O DIRETÓRIO DOS GRUPOS DE PESQUISA NO BRASIL

O Diretório dos Grupos de Pesquisa no Brasil, coordenado pelo CNPq, é uma base de dados implementada desde 1992. Originou-se em 1991, a partir de uma proposta de elaboração de um Almanaque de Pesquisa no CNPq, bem como no levantamento de grupos de pesquisa realizado pelo Fórum Nacional de Pró-Reitores de Pesquisa. O intuito era organizar o Programa de Laboratórios Associados encomendado em 1990 pela então Secretaria de Ciência e Tecnologia (atual Ministério de Ciência e Tecnologia) (Guimarães, 1994). O objetivo básico do projeto era oferecer um suporte informacional atualizado sobre as atividades de pesquisa, pretendendo obter, periodicamente, a configuração dos recursos humanos e a organização da produção científica e tecnológica brasileiras.

Hoje, o Diretório tem o claro objetivo de ser uma plataforma de informação básica sobre o parque científico e tecnológico brasileiro (CNPq, 1999). O esforço empreendido pelo Brasil após a Segunda Grande Guerra, gerou o maior parque de C&T da América Latina. Entretanto, ainda há carência de informação organizada a respeito, o que enfraquece e dificulta a tomada de decisão sobre os desígnios da C&T nacional. Tal fato transforma o Diretório em instrumento essencial para

a gestão de C&T (Martins & Galvão, 1994).

O Diretório possui três finalidades importantes: a) fortalecer o intercâmbio entre pesquisadores brasileiros, bem como entre estes e pesquisadores estrangeiros; b) preservar a memória da atividade de pesquisa; e c) ferramenta estratégica para as atividades de planejamento do CNPq. Esta terceira finalidade é de vital importância em processos de avaliação e acompanhamento (A&A). Por sua vez, os procedimentos de A&A são fundamentais para a tomada de decisão no âmbito do CNPq, quer em nível estratégico, quer no âmbito gerencial, como por exemplo, na formulação de políticas de investimentos em C&T (CNPq, 1998).

A primeira versão do Diretório teve seu trabalho de campo realizado no segundo semestre de 1993, englobando a produção de C&T do triênio 1990-92. Seus resultados foram publicados em novembro de 1994. As informações para a segunda versão foram colhidas em 1995, cobrindo o biênio 1993-94. A atual versão, a terceira, teve seu trabalho de campo realizado no final de 1997 e foi disponibilizada ao público em geral no segundo trimestre de 1998. Esta última versão pode ser acessada através da home page do CNPq (http://www.cnpq.br/gpesq3).

3.1 ORGANIZAÇÃO DO DIRETÓRIO

O Diretório está organizado em quatro bases de dados relacionais independentes, mas que interagem: 1) grupos de pesquisa; 2) pesquisadores; 3) linhas de pesquisa e; 4) produção científica, tecnológica e artística. Para este trabalho, foram utilizados dados das bases (2) e (4). Vale ressaltar que a base (2) traz toda informação gerada e disseminada pelos grupos no período censitário.

A base de dados da versão 3.0 do Diretório contém informações sobre 33.675 pesquisadores distribuídos em 8.544 grupos de pesquisa pertencentes a 181 instituições. Esses pesquisadores geraram 339.568 produtos, tais como artigos em periódicos, comunicações em eventos, livros, formação de recursos humanos, etc.

O Diretório tem como unidade básica de análise o *grupo de pesquisa*. Um grupo de pesquisa é caracterizado por um ou dois pesquisadores líderes, podendo ter ou não outros pesquisadores, técnicos, estudantes de graduação ou pós-graduação e estagiários. Todos devem trabalhar em uma ou mais linhas de pesquisa, compartilhando instalações, equipamentos e demais recursos. Devido às peculiaridades de cada área do conhecimento, não há uma estrutura rigorosa. Assim, um grupo pode contar apenas com um pesquisador, trabalhando individualmente com seus estudantes. Também não é obrigatório que os integrantes de um grupo pertençam a uma única instituição. Do mesmo modo, a

definição de "pesquisador" também é flexível. O único requisito é que o integrante do grupo tenha pelo menos concluído a graduação.

4. METODOLOGIA DA PESQUISA

Calcado no objetivo de demonstrar o potencial do algoritmo *Apriori* na extração de regras para apoio à gestão de C&T, foi utilizado todo o universo de 33.675 pesquisadores cadastrados no Diretório, os quais foram escolhidos como unidades de análise.

A metodologia empregada possui três grandes blocos: 1) pré-processamento dos dados; 2) aplicação do algoritmo *Apriori*; e 3) análise e refinamento dos resultados.

O primeiro passo do pré-processamento consistiu em realizar a seleção de alguns itens, dos pesquisadores, considerados relevantes. Esses itens são os atributos que caracterizam um pesquisador, tais como sua idade e sexo, e os indicadores de produção de C&T, como a quantidade de artigos publicados no período 1995-97. Esta seleção permitiu testar o algoritmo em um computador pessoal, através do *software* Matlab, o qual se ajusta muito bem às aplicações acadêmicas. Todos os itens foram extraídos dos módulos Recursos Humanos e Produção Científica e Tecnológica do Diretório. Foram escolhidos cinco itens categóricos e três quantitativos, conforme relação apresentada na Figura 2.

ITENS CATEGÓRICOS

Nacionalidade (brasileira, estrangeira);
Idade ([>24], [25..29], [30..34], [35..39], [40..44], [45..49], [50..54], [55..59], [60..64], [• 65]);
Sexo (masculino, feminino);
Titulação Máxima (graduação, especialização/aperfeiçoamento, mestrado, doutorado);
Grande Área do Conhecimento (Ciências Agrárias, Ciências Biológicas, Ciências das Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias e Ciência da Computação, Lingüística, Letras e Artes);

ITENS QUANTITATIVOS

Artigos Publicados em Periódicos (0, [1..2], [3..5], [6..10], [11..20], [21..30], [31..40], [• 41]);
Dissertações Orientadas ([0], [1], [2], [3], [4..7], [8..12], [• 13]); e
Teses Orientadas ([0], [1], [2], [3], [4..5], [6..7], [8..9], [• 10]).

Figura 2 – Itens selecionados para a aplicação do algoritmo *Apriori*.

Em um segundo estágio, os itens selecionados foram classificados conforme sua natureza: categóricos ou quantitativos. Os itens categóricos, tais como Nacionalidade e Sexo, foram representados na forma booleana. Por exemplo: O item Sexo foi representado pelos atributos booleanos Sexo_M e Sexo_F. A alternativa de atributos booleanos é uma forma de representar dados categóricos para economizar memória e reduzir o tempo de processamento.

Os itens quantitativos foram estratificados em faixas. Exemplificando, o item Idade foi dividido em faixas com amplitude de 5 anos e cada faixa foi considerada como um atributo categórico. O mesmo ocorreu com os itens Artigos Publicados em Periódicos, Dissertações Orientadas e Teses Orientadas.

Este pré-processamento gerou um arquivo contendo 33.675 registros, cada um com 50 campos, onde cada registro representa um pesquisador.

Esta abordagem de itens categóricos e quantitativos permitiu gerar uma tabela de dados onde a ausência de um item é representada pelo valor '0' e a presença pelo valor '1'. Assim, se um pesquisador é mulher, o item Sexo_F recebe o valor '1' e o item Sexo_M recebe o valor '0'.

Feito isto, a etapa seguinte é aplicar o *Apriori* para encontrar, entre os itens categóricos e quantitativos, as regras de associação que sejam relevantes para o decisor. Isto é obtido através de delimitações (restrições) impostas pelo decisor na execução do algoritmo, através da estipulação de suporte e confiança mínimos. Esta parte da metodologia pode ser resumida nos seguintes passos:

1) selecionar os itens quantitativos e categóricos;

2) transformar os itens quantitativos em categóricos, particionando-os em intervalos iguais;

3) extrair os itens selecionados, através da linguagem SQL, gerando uma tabela booleana;

4) calcular o suporte para cada item;

5) encontrar todos os itens cujo suporte seja pelo menos igual ao suporte mínimo, obtendo o conjunto de *itemsets freqüentes*;

6) a partir do conjunto de *itemsets freqüentes*, gerar as regras de associação que apresentem confiança pelo menos igual à confiança mínima; e

7) escolher as regras interessantes.

5. RESULTADOS

No estudo, duas abordagens foram utilizadas. A primeira abordagem realizou várias simulações utilizando diversos valores de suporte e confiança mínimos, em busca dos itens mais freqüentes, conforme a proposta original do algoritmo. A segunda abordagem tratou dos itens menos freqüentes, acrescentando ao algoritmo um limite máximo para o suporte.

5.1 RESULTADOS PRELIMINARES

A Tabela 3 resume o número de regras encontradas com a primeira abordagem, em função de suporte e confiança mínimos estabelecidos e considerando a busca por itens mais freqüentes. A primeira constatação é que o número de regras geradas é inversamente proporcional aos valores determinados para *minconf* e *minsup*, pois, à medida que estes últimos decrescem, aumenta o número de regras.

Simulação	Suporte mínimo (*minsup*)	Confiança mínima (*minconf*)	Número de regras geradas
S1	70%	80%	5
S2	60%	50%	8
S3	50%	40%	29
S4	50%	20%	29
S5	25%	20%	67

Tabela 3 –Regras geradas em função de suporte e confiança mínimos.

Para a simulação S1, o número reduzido de regras foi acompanhado de alguma redundância. Por exemplo, observe-se as regras a seguir:

- R1: (brasileiro; não titulou doutor) $\Rightarrow$ (não titulou mestre), c = 85%, s = 76,4%.
- R2: (brasileiro) $\Rightarrow$ (não titulou doutor; não titulou mestre), c = 80%, s = 76,4%

Com a redução de *minconf* e *minsup*, o aumento de regras minerou outros itens. Com *minsup* = 60% e *minconf* = 50% (simulação S2), além de quatro regras idênticas às da simulação S1, outras regras selecionaram somente homens, ou somente doutores. Isto é coerente e indica que o algoritmo foi implementado corretamente, pois a maioria dos pesquisadores inventariados pelo Diretório são homens (58%) e são doutores (55%) (CNPq, 1998).

Para as simulações S3, S4 e S5, o aumento no número de regras foi acompanhado do aumento de redundância. Porém, novas regras foram descobertas. Agora, as regras passaram a considerar mulheres (42% dos pesquisadores do Diretório são mulheres (CNPq, *op. cit.*)).

Mas, em se tratando de pesquisadores, é bastante provável que o decisor esteja interessado em descobrir o comportamento das minorias, posto que elas podem revelar pesquisadores com alta performance ou, por outro lado, com baixo rendimento. Este é o enfoque a seguir.

5.2 RESULTADOS APÓS A MODIFICAÇÃO DO ALGORITMO

Para buscar regras que possam revelar as exceções do Diretório, introduziu-se uma restrição adicional ao suporte, que passou a ter um limite máximo. A Tabela 4 mostra o número de regras geradas para cinco simulações, com a introdução de limite máximo ao suporte.

Simulação	Suporte mínimo (*minsup*)	**Suporte máximo (*maxsup*)**	Confiança mínima (*minconf)*	Número de regras geradas
S6	1%	**10%**	15%	3
S7	1%	**20%**	25%	4
S8	3%	**49%**	20%	19
S9	1%	**50%**	1%	243
S10	0.1%	**5%**	1%	39

Tabela 4 – Resultados impondo-se um suporte máximo.

As simulações S6 e S7, geraram regras até certo ponto óbvias, apesar de destacar “minorias”. Eis duas das regras obtidas:

- Regra 1, S7: (título = graduação) $\Rightarrow$ (idade < 24 anos), s = 1,05%, c = 15,85%.
- Regra 3, S8: ([6..10] artigos produzidos) $\Rightarrow$ (titulou 1 mestre), s = 1,07%, c = 16,38%.

A interpretação de *s* em Regra 1 da simulação S7 é: 1,05% dos pesquisadores são graduados e possuem menos de 24 anos. A interpretação de *c* é: 15,85% dos graduados têm menos de 24 anos. Esta informação implica que 84,15% desses pesquisadores têm 24 anos ou mais. A conclusão é que há um pequeno contingente muito jovem envolvido com pesquisa. Porém, o decisor poderia ser levado a perguntar, por exemplo, qual é a faixa etária dos 84,15% restantes? Há pessoas nesta situação em idade avançada? Porque não realizaram um mestrado ou doutorado? Será que não estaria havendo uma certa confusão em registrar pessoal técnico como pesquisador? Ou seja, as duas regras, mais do que constatações normais, induzem a explorar em maior profundidade os dados.

Já as regras geradas pelas simulação S8, S9 e S10 revelaram nichos distintos de pesquisadores. Na simulação S8 foi possível encontrar algumas características dos pesquisadores do sexo feminino e dos mestres entre 30 e 35 anos, entre outras.

Na simulação S9, apesar do excesso de regras, novamente apareceu um agrupamento referente às mulheres. Agora, descobriu-se que 24,84% delas são estrangeiras e que 30,98% estão concentradas nas Ciências Exatas. Além do mais, 24,5% das pesquisadoras com graduação possuem entre 25 e 29 anos e que 34,34% delas atuam em grupos das Ciências da Saúde.

Uma outra constatação é a formação de *clusters* de regras por Idade, Sexo, Titulação, Grande Área, etc. Veja, na Tabela 5, um *cluster* correspondente à distribuição dos mestres entres as grandes áreas. Observe que a área de Ciências Sociais e a área de Lingüistica, Letras e Artes são as que possuem o menor contingente de pesquisadores com o título de mestre.

Regra	Titulação	Grande área	Confiança
154.	mestre	Ciências Exatas	10.68
155.	"	Ciências Biológicas	12.35
156.	"	Engenharias e C. Computação	15.66
157.	"	Ciências da Saúde	16.49
158.	"	Ciências Agrárias	15.04
159.	"	Ciências Sociais	**8.21**
160.	"	Ciências Humanas	16.47
161.	"	Lingüística, Letras e Artes	**5.10**

Tabela 5 - *Cluster* de Grandes Áreas (simulação S9).

Analisando um *cluster* referente a publicação de artigos, observou-se que os pesquisadores das Ciências Exatas, Biológicas e Saúde publicaram até sete artigos no período; os das Ciências Agrárias e Humanas publicaram até cinco artigos, ao passo que os das Engenharias, Ciências Sociais e Lingüística, Letras e Artes publicaram no máximo dois artigos. Isto pode estar revelando o perfil dessas áreas no tocante à disseminação do conhecimento em periódicos científicos.

Por último, na simulação S10, foram utilizados baixos valores para o suporte e a confiança. Surgiram, então, os pesquisadores estrangeiros, onde 3,18% produziram entre 11 e 20 artigos, 2,85% formaram entre 4 e 7 mestres e 4,88% têm mais de 64 anos. Ou seja, são grupos seletos (os estrangeiros são minoria) e de alta produtividade.

Em resumo, diferentemente da proposta original do *Apriori*, concebido para extrair regras onde a alta freqüência é essencial, no contexto dos grupos de pesquisa a confiança da regra é mais significativa do que o suporte, mesmo que ambos sejam baixos. Importa saber, além da organização da maioria, o perfil e o desempenho de alguns poucos pesquisadores.

Uma constatação final diz respeito aos limites impostos ao suporte. Eles provocaram a eliminação de itens com suporte fora destes limites. No entanto, o suporte final das regras depende da

freqüência dos itens resultantes combinados, que deve ser menor ou igual a freqüência dos itens individuais. Logo, dado *minsup* e *maxsup*, e considerando *s* como sendo o suporte da regra, tem-se: *minsup* <= *s* <= *maxsup*, apesar de que, na prática, observou-se que em geral *minsup* < *s* < *maxsup*.

6. CONCLUSÕES

As simulações realizadas sobre alguns atributos e indicadores dos pesquisadores inventariados pelo Diretório confirmaram o potencial do algoritmo *Apriori*. Regras de associação consistentes foram geradas e nichos específicos foram descobertos. Se, por um lado, algumas delas são previsíveis, por outro lado o algoritmo estimula o aprofundamento do conhecimento por parte do decisor, onde a sua sensibilidade e experiência são fundamentais. Ele pode intervir estipulando limites para o suporte e a confiança.

Neste trabalho promoveu-se pequenas modificações no algoritmo *Apriori*, em especial a introdução de um limite superior para o suporte. Com isto, foi possível explorar com maior profundidade as informações de grupos seletos de pesquisadores.

Entretanto, ficou evidente a necessidade de ajustes de modo a eliminar problemas como a redundância de regras. Em trabalhos seguintes, pretende-se explorar este campo e focalizar a questão da mistura de dados quantitativos com categóricos, proposta por Srikant & Agrawal (1996).

Futuramente, pretende-se utilizar todos os itens do Diretório e explorar o desempenho do algoritmo, quer implementando novas funções ao *Apriori*, como as propostas em Agrawal & Srikant (1994), quer refinando e inserindo restrições ao algoritmo.

Tenciona-se, também, aplicar o algoritmo ao Banco de Currículos do CNPq e outras bases de C&T, como a da CAPES, sobre a pós-graduação brasileira.

A crise financeira que se abate sobre o setor de C&T, em especial no apoio à pesquisa, torna imperativo que os processos de tomada de decisão sejam executados com base em conhecimento estratégico obtido à partir de procedimentos de acompanhamento e avaliação realizados sobre uma base íntegra de dados. Bancos de dados consistentes são, portanto, o primeiro passo. Além disso, conhecer como se organiza o parque científico e tecnológico nacional é fundamental para uma gestão segura do setor, permitindo, inclusive, traçar diretrizes políticas realistas. Por isso, ferramentas de apoio à tomada de decisão devem ser estimuladas. O presente trabalho é uma alternativa viável, merecendo ser explorada em maior profundidade.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Mining association rules between sets of items in large databases. **Proc. of the ACM SIGMOD Conference**. Washington, DC, USA, p. 207-216, maio 1993. Disponível na Internet. http://www.almaden.ibm.com/u/ragrawal/pubs.html. 3 junho 1999.

AGRAWAL, R., SHAFER, J. C. Parallel mining of association rules. **IEEE Transactions on Knowledge and Data Engineering,** vol. 8, NO. 6, December 1996.

AGRAWAL, R. & SRIKANT, R. Fast algorithms for mining association rules. **Proc. of the 20th Int'l Conference on Very Large Databases**. Santiago, Chile, set. 1994. Disponível na Internet. http://www.almaden.ibm.com/u/ragrawal/pubs.html. 3 junho 1999.

AGRAWAL, R. & SRIKANT, R. Mining generalized association rules in large relational tables. **Proc. of 21st Int'l Conference on Very Large Databases**. Zurique, Suiça, set/1995. Disponível na Internet. http://www.almaden.ibm.com/u/ragrawal/pubs.html. 3 junho 1999.

BRACKETT, M. H. **The data warehouse challange: taming data chaos**. New York: John Wiley & Sons, 1996.

CHEN, M-S., HAN, J. E YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering,** v. 8, n. 6, p.886-883, 1996.

CHEUNG, D. W., NG , V. T. & FU, A. W. Efficient mining of association rules in distributed databases. **IEEE Transactions on Knowledge and Data Engineering**, v.8, n. 6, p. 911-922, 1996.

CNPq. **Construindo o futuro: propostas e realizações da gestão 95-98.** Brasília: CNPq, 1998.

CNPq. **Diretório dos Grupos de Pesquisa no Brasil. Versão 3.0**. Disponível na Internet. http//:www.cnpq.br/gpesq3. 31 mar.1999.

GUIMARÃES, R. **Avaliação e fomento de C&T no Brasil: propostas para os anos 90.** Brasília: MCT/CNPq, 1994.

HANSELMAN, D., LITTLEFIELD, B. **Matlab: versão do estudante: guia do usuário.** São Paulo: Makron Books, 1997.

MACHADO, C. Como dar o tiro certo na hora de decidir. **Informática Exame**, março de 1996. Disponível na Internet. http://www2.ulo.com.br/info/arquivo/ie120/capa.html. 19 junho 1998.

MARTINS, G. M., GALVÃO, G. O diretório dos grupos de pesquisa no Brasil: perspectivas de fomento e avaliação. **Educação Brasileira**, v.16, n. 33, p. 11-29, 1994.

NIGRO, M. O melhor caminho até seu cliente. **Byte Brasil**, p. 44-66, janeiro de 1997.

SRIKANT, R. & AGRAWAL, R. Mining quantitative association rules in large relational tables. **Proc. of ACM SIGMOD Conf. on Management of Data.** Montreal, Canadá, jun. 1996.Disponível na Internet. Disponível na Internet. http://www.almaden.ibm.com/u/ragrawal/pubs.html. 3 junho 1999.

StatSoft Portugal Ltda. **Data mining com o statistica**. Disponível na Internet. http://www.statsoftinc.com/portugal/datamin.html. 28 junho 1998.

ANEXO I

ALGORITMO APRIORI

```
%INICIALIZAÇÃO
arquivo = 'dados4.txt'
minsup = 0.0; maxsup = 100;          minconf = 0.0%
k = 1;  Natributos = 50;        NL = 0;
%LEITURA DO ARQUIVO
arq = fopen(arquivo,'rt');
i = 1;
linha = str2num(fgetl(arq));
for j = 1:Natributos,
   Sup(j) = linha(j);
```

```
end;
while  ~feof(arq)
  i = i + 1;
  linha = str2num(fgetl(arq));
  for j = 1:Natributos,
    Sup(j) = Sup(j) + linha(j);  %Cálculo do Suporte Geral
  end;
end;
st = fclose(arq);
Npesquisadores = i
fator = 100/Npesquisadores;
Sup = Sup*fator;
%Passo (1)--> L = {1-itemset}
%==================================
Nlarges = 0;
for col=1:Natributos
  if Sup(col) >= minsup & Sup(col) < maxsup,
    Nlarges = Nlarges + 1;
    L(Nlarges,1) = col;
    L(Nlarges,2) = Sup(col);
  end; %if
end; %for col
clear Sup

%Passo (2) --> Laço geral para obter L
%==================================
while Nlarges(k) > 1,
  k = k + 1;
%Passo 3 --> Chamar a função apriori_gen para obter c
%====================================
  apriori_gen;
        for i = 1:Ncandid
    soma = 0;
    %LEITURA DO ARQUIVO
    arq = fopen(arquivo,'rt');
    for j = 1:Npesquisadores
      linha = str2num(fgetl(arq));
          for w = 1:k
          if linha(c(i,w)) == 1
          adic = 1;
              else
              adic = 0;
            break;
        end; %if
            end; %for w
          soma = soma + adic;
    end; %for j
    st = fclose(arq);
```

```
    c(i,k+1) = soma*fator; %c recebe a coluna de Suporte
  end; %for i
  %Passo 9 - OBTER L(k)
  %====================
  j = 0;
  clear large
  for i = 1:Ncandid
    if c(i,k+1)>= minsup & c(i,k+1)< maxsup
        j = j + 1;
            large(j,:) = c(i,:);
        end; %if
  end; %for i
  Nlarges(k) = j;
  if Nlarges(k) > 0
    L(1:Nlarges(k),1:k+1,k) = large;
    NL = k;
  end;
end; %while
%Chamar a função genrules para extrair as regras genrules
'FIM DO ALGORITMO APRIORI'
```

ANEXO II

FUNÇÃO APRIORI_GEN

```
%OBTER OS CANDIDATOS E PODA (join step e prune step)
        Ncandid = 0;
        if k==2,
          for fixo = 1:Nlarges-1
          for varia = (fixo+1):Nlarges
          ss = 1;
        item(ss) = L(fixo,1);
              ss = ss + 1;
            item(ss) = L(varia,1);
          Ncandid = Ncandid + 1;
        c(Ncandid,:) = item;
            end; %for varia
    end; %for fixo
  else %para k>2
    for i=1:(k-1)
      flag(i) = 1;
    end; %for i
        for fixo = 1:(Nlarges(k-1)-1)
        for ss = 1:(k-1)
        item_ant(ss) = L(fixo,ss,k-1);
      end; %for ss
```

```
        if item_ant(ss) < Natributos
                    for i = 2:(k-1)
                              test(i-1) = item_ant(i);
                    end; %for i
                    for varia = (item_ant(ss)+1):Natributos
                    test(i) = varia;
                  pert = 0;
            for j = (fixo+1):Nlarges(k-1) %Procura em L
                    for w=1:(k-1)
                          if test(w) == L(j,w,k-1)
                    pert(w) = 1;
                       else
                       pert(w) = 0;
                  end; %if
              end; %for w
                    if pert == flag
                              break;
                          end; %if
                       end; %for j
                    if pert == flag  %item pertence a L
                    item = item_ant;
                    item(k) = varia; %novo componente
                  Ncandid = Ncandid + 1;
                              c(Ncandid,:) = item;
                end; %if
          end; %for varia
        end; %if
            end; %for fixo
  end; %else
'FIM DA FUNÇÃO APRIORI_GEN'
```

ANEXO III

FUNÇÃO GENRULES

```
%ESTRAÇÃO DE REGRAS  A --> B
if NL < 2,
  NL
  'Não há Regras.'
else
  'Extraindo Regras...'
  Nregras = 0;
  for  k = NL:-1:2
    for i = 1:Nlarges(k)
      itemset = L(i,:,k);
      m = k - 1;
      while m > 0
```

```
          a = itemset(1:m);
          b = itemset(m+1:k);
          tamb = 0;
          for w = m+1:k
            tamb = tamb + 1;
          end;
          for j = 1:Nlarges(m)
            if L(j,1:m,m) == a
              break;
            end;
          end; %for j
          sup_a = L(j,m+1,m);
          conf = (itemset(k+1)/sup_a)*100;
          if conf >= minconf
            Nregras = Nregras + 1;
            A(Nregras,1:m) = a;        %antecedente da regra
            B(Nregras,1:tamb) = b;     %consequente da regra
            S(Nregras) = itemset(k+1); %suporte da regra
            C(Nregras) = conf;         %confiança da regra
            m = m - 1;                 %próximo subset
          else
            break; %não precisa avaliar subsets
          end; %if conf
        end; %while
      end; %for i
    end; %for k
    clear L
    Nregras
    if Nregras > 0,
      A
      B
      format bank;
      'Suporte = ',S'
      'Confiança = ',C'
    end; %if Nregras
  end; %if NL
'FIM DA FUNÇÃO GENRULES'
```