5ª MOSTRA ACADÊMICA UNIMEP
23 a 25 de outubro de 2007
UNIMEP
Universidade Metodista de Piracicaba
Tema:
EDUCAÇÃO BRASILEIRA:
Extinção ou Sustentabilidade na Universidade

5º Congresso de Pós-Graduação

MINERAÇÃO DE DADOS USANDO REGRAS DE ASSOCIAÇÃO MULTI-RELACIONAL QUANTITATIVA

Autor(es)

EDERSON GARCIA

Co-Autor(es)

MARINA TERESA PIRES VIEIRA

Orientador(es)

MARINA TERESA PIRES VIEIRA

1. Introdução

A Mineração de Dados ou Data Mining, segundo Han e Kamber (2001), é a principal etapa do processo de descoberta de conhecimento em banco de dados (KDD - Knowledge Discovery in Databases) e tem como objetivo encontrar padrões em dados armazenados em um banco de dados. Existem diversos tipos de tarefas de mineração, sendo as mais usuais as tarefas de associação, classificação e agrupamento. A tarefa de mineração determina a técnica de mineração a ser usada para buscar padrões. A tarefa de associação (HAN; KAMBER, 2001) adotada neste trabalho, tem por objetivo encontrar padrões em um conjunto de dados que contêm itens que estão relacionados à ocorrência de outros itens. Essa tarefa gera regras que são representadas na forma de uma implicação X⇒ Y, onde X e Y representam um conjunto de itens, e a regra gerada representa a implicação de que onde ocorre o conjunto de itens X também ocorre o conjunto de itens Y; com os dados de uma padaria poderíamos ter a regra, conforme apresentado no exemplo 1: Exemplo 1: Pão ⇒ Leite Essa regra tem como interpretação: “quem compra pão tende a comprar leite”. As tarefas de mineração de dados utilizam medidas de interesse, sendo que, no caso das regras de associação, as mais comuns são as medidas de suporte e confiança descritas em (AGRAWAL; IMIELINSKI; SWAMI,1993). A medida de suporte demonstra a freqüência com que os itens ocorrem em relação ao total de dados analisados e a medida de confiança representa a freqüência com que os itens X ocorrem em relação à ocorrência dos itens Y.

As medidas de suporte e confiança podem ser representadas pelas formulas que estão representadas na figura formula.jpg Os primeiros algoritmos de mineração de dados usando tarefa de associação foram destinados à extração de regras envolvendo dados categóricos (AGRAWAL; IMIELINSKI; SWAMI, 1993; AGRAWAL; SRIKANT, 1994; HAN; PEI; YIN, 2000). Depois, surgiram os algoritmos que tratam dados quantitativos (SRIKANT; AGRAWAL, 1996; HONG; KUO; CHI, 1999; AUMANN; LINDELL, 1999; FUKUDA et

al, 1996; MILLER; YANG, 1997; LENT; SWAMI; WIDOM, 1997; POSSAS et al; 2000). A abordagem adotada por Aumann e Lindell (1999) é de particular interesse do presente trabalho e se baseia na distribuição de valores dos atributos quantitativos, usando, para isso, medidas estatísticas como a média e a variância. Os autores separam a regra gerada em dois lados, o primeiro contendo um subconjunto da população e o outro lado contendo o comportamento interessante do subconjunto, formando assim uma regra conforme apresentada no exemplo 2: Exemplo 2: sexo = feminino ⇒ salário: médio = $7.90/h (salário médio geral = $9.02/h) que tem como interpretação: As pessoas do sexo feminino ganham em média $7.90 dólares por hora e a média geral é de $9.02 dólares por hora. Com isso é relevante afirmar que as pessoas do sexo feminino ganham abaixo da média geral do salário que é $9.02 dólares por hora de acordo com a regra gerada. Uma característica dos algoritmos citados é que eles manipulam os dados que estão contidos em uma única tabela. Algumas pesquisas recentes têm focado no processo de mineração de dados envolvendo múltiplas tabelas, chamada mineração multi-relacional (DŽEROSKI, 2003; NESTOROV; JUKIC, 2003; RIBEIRO, 2004; RIBEIRO; VIEIRA, 2004; RIBEIRO; VIEIRA; TRAINA, 2005; PIZZI, 2006). Essas pesquisas estão concentradas na mineração de dados categóricos, não levando em consideração os dados quantitativos. Os trabalhos de Ribeiro (2004) e Pizzi (2006), que também fazem parte da abordagem adotada neste trabalho, propõem técnicas de mineração envolvendo múltiplas tabelas que não estão relacionadas entre si. As autoras desenvolveram algoritmos, baseados em algoritmo existente de mineração de regras de associação, nos quais são introduzidos novos conceitos, como o conceito de Bloco, Segmento e Peso de um item; além disso, as medidas de interesse como suporte e confiança são alterados com relação aos conceitos definidos por Agrawal, Imielinski e Swami (1993), para que os padrões gerados possam representar melhor a verdadeira relação entre os itens das múltiplas relações. O trabalho dessas autoras foi motivado pelo fato que a analise conjunta de múltiplas tabelas permite relacionar múltiplos assuntos; com essa análise conjunta geram-se regras do tipo: X⇒ Y, onde X e Y são itemsets (conjuntos de itens) de tabelas distintas, isto é, X pertence a uma tabela e Y pertence a uma outra tabela. Para exemplificar o tipo de regras geradas, considere as tabelas apresentadas no Quadro 1: Trabalho_realizado (nroaluno, nroTrabalho, VlNota, ConceitoTrab) e Prova_realizada(nroaluno, nroProva, VlNota, ConceitoProva) Uma das possíveis regras geradas através dos conceitos de Ribeiro (2004) é apresentada no exemplo 3 Exemplo 3: ConceitoTrab=A ⇒ ConceitoProva=A. suporte(ConceitoTrab=A⇒ ConceitoProva=A) = 40% confiança(ConceitoTrab=A⇒ ConceitoProva=A) = 100% Peso(ConceitoTrab=A) = 100% Peso(ConceitoProva=A) = 80% Que tem o significado: “O aluno que obtém o conceito A em algum trabalho, tende a obter conceito A em alguma prova”

No quadro 1 são apresentados exemplos de blocos e segmentos definidos por Ribeiro (2004). Um bloco é a unidade de analise do processo de mineração multifatos, sendo este um conjunto de transações de tabela que contém o mesmo valor de um atributo. Nas tabelas do Quadro 1 esse atributo é nroAluno e os blocos estão destacados com cores alternadas. Um segmento é formado por um conjunto de blocos de tabelas distintas que possuem o mesmo valor para o atributo comum, isto é, os blocos das tabelas distintas relacionados por um atributo comum. Os vários conceitos introduzidos de mineração multi-relacional, blocos e segmentos e de mineração quantitativa estão sendo considerados para compor uma nova forma de composição das regras geradas, visando obter maior expressividade nas regras geradas.

2. Objetivos

O objetivo deste trabalho é combinar a abordagem de mineração de dados multi-relacional adotada por Ribeiro (2004) com uma abordagem de mineração de dados quantitativos (AUMANN; LINDELL, 1999) usando medidas estatísticas, para propor uma estratégia de mineração de dados multi-relacional visando obter maior expressividade nas regras geradas. Com essa proposta pretende-se melhorar as regras de associação apresentadas por Ribeiro (2004) adicionando informações quantitativas às regras geradas, permitindo desse modo uma melhor expressividade dessas regras. Gerando regras conforme apresentada no exemplo 4: Exemplo 4: ConceitoTrab=A (média trabalho = 9.6 (media geral = 6.03)) ⇒ ConceitoProva=A (média Prova =9.1 (média geral = 6.37)). suporte(ConceitoTrab=A⇒

ConceitoProva=A) = 40% confiança(ConceitoTrab=A⇒ ConceitoProva=A) = 100% Peso(ConceitoTrab=A⇒ ConceitoProva=A) = 38% Que tem o seguinte significado: Os alunos que tiram em média nota 9.6 no trabalho que é 59% acima da média geral de todos os indivíduos e tendem tiram em média nota 9.1 na prova que é 43% acima da média geral de todos os indivíduos. Comparado a regra do exemplo 3 com a regra do exemplo 4, percebe-se que a regra do exemplo 4 possui uma expressividade maior, pois apresenta uma comparação quantitativa da regra gerada em relação ao total da população.

3. Desenvolvimento

Para se alcançar os objetivos do trabalho estão sendo realizados estudos sobre a funcionalidade do algoritmo existente proposto por Ribeiro (2004) e também do algoritmo proposto por Aumann e Lindell (1999) e, paralelamente, está sendo feita a revisão bibliográfica baseada em artigos relacionados à mineração de dados multi-relacional e à mineração de dados quantitativos com o objetivo de dar suporte teórico sobre os conceitos a serem aplicados. A etapa seguinte será a especificação e implementação da estratégia para a geração das regras de associação em dados multi-relacional e quantitativos. Objetivando com isso criar um algoritmo baseado no Connection desenvolvido em Java para gerar as regras apresentadas em 4.

4. Resultados

Nas duas abordagens apresentadas neste trabalho, tanto a abordagem de Ribeiro (2004), quanto a de Pizzi (2006), fazem uso de um peso para manter a real relação entre os itens e seu comportamento nas tabelas de origem. Desse modo, quando é feita a geração da regra, esse peso é considerado e com isso afeta a definição da regra forte. Portanto o peso é um ponto de destaque desses trabalhos para tornar confiável a mineração multi-relacional. Contudo, se um segmento possuir três tuplas com diferentes valores para X na tabela t1, e que apontam para o mesmo segmento que possui duas tuplas de Y com valores diferentes na tabela t2, as medidas de suporte, confiança ou peso não representam a diferença entre esse segmento e um outro segmento que possui apenas uma tupla para X com uma tupla para Y. Isso faz com isso que os dois segmentos tenham a mesma relevância para gerar as regras. No trabalho proposta, é evidenciado o relacionamento entre os itemsets das diferentes tabelas com sua freqüência dentro do segmento.

5. Considerações Finais

Neste trabalho pretende-se explorar as técnicas adotadas por Ribeiro (2004) e Pizzi (2006) combinando-as com as técnicas de mineração de dados quantitativos, buscando com isso melhorar os resultados das regras geradas pelo processo de mineração. Na literatura são raros os trabalhos tratando a mineração de dados multi-relacional considerando dados de tabelas não relacionadas diretamente e não há trabalhos que tratam de dados quantitativos envolvendo tabelas não relacionadas diretamente. Espera-se com este trabalho obter maior expressividade paras as regras geradas dando assim maior clareza às regras e possibilitando com isso a extração de mais conhecimento da base de dados. Esse novo algoritmo pode ser aplicado em bases de dados de diferentes áreas de aplicação, possibilitando novas descobertas de conhecimento nos dados. Ainda, pode-se estender a mineração multi-relacional para outras tarefas de mineração como classificação ou agrupamento.

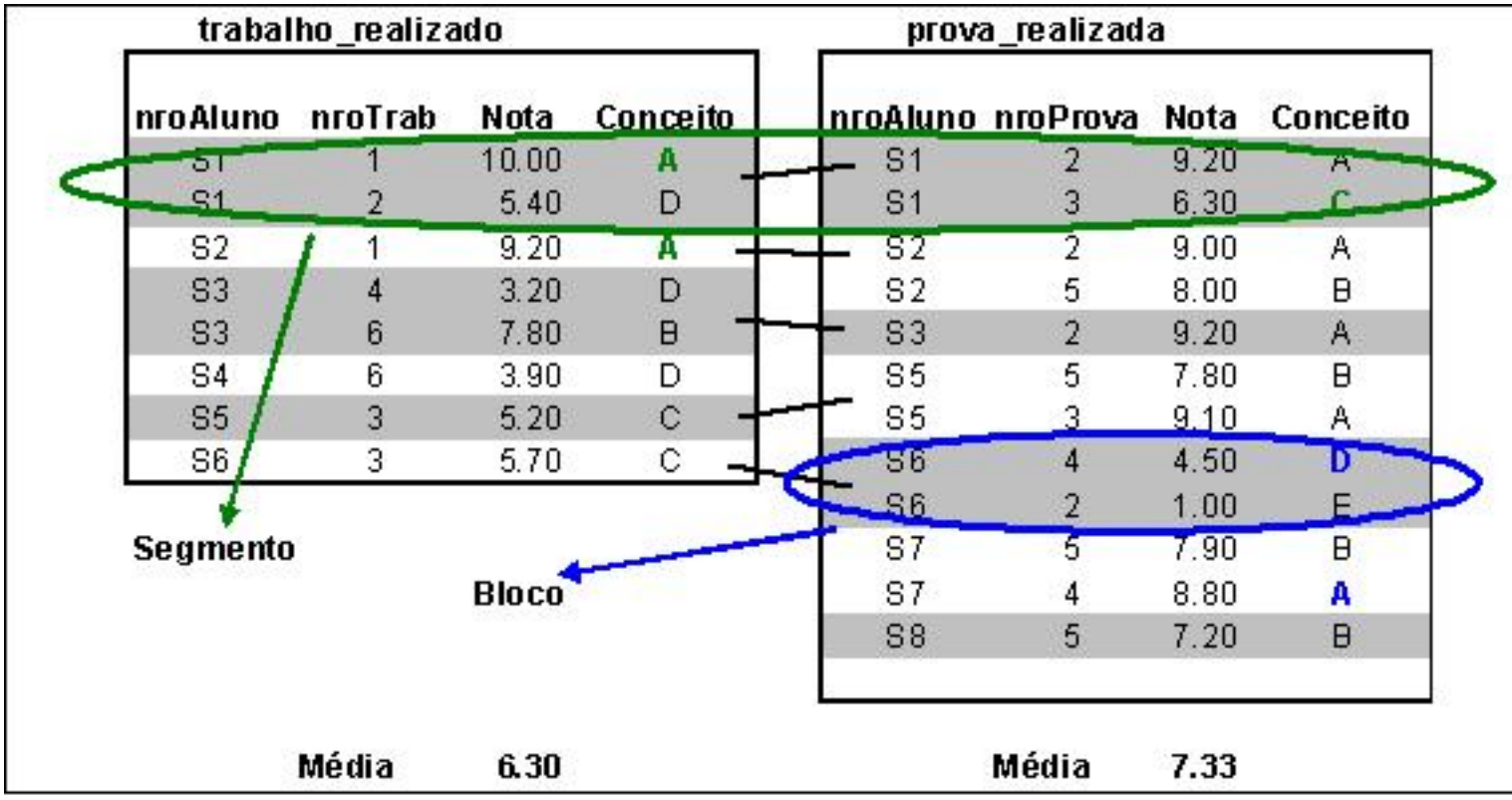
Referências Bibliográficas

AGRAWAL, R.; IMIELINSKI, T; SWAMI, A. ***Mining association rules between sets of items in large databases****, in Proc. of the ACM SIGMOD Intl Conf. on Management of Data, Washington, D.C., USA, 1993, pp. 207-216.*

AGRAWAL, R.; SRIKANT, R. ***Fast algorithms for mining association rules****. In: Proc. of the Intl Conf. on*

Very Large Databases, Santiago de Chile, Chile, 1994.
AUMANN,Y.; LINDELL, Y. ***A statistical theory for quantitative association rules.*** *In: FIFTH ACM SIGKDD INT. CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING, Aug. 1999. p.261-270.*
CALDERS, T.; GOETHALS, B.; JAROSZEWICZ, S. ***Mining Rank-Correlated Sets of Numerical Attributes.*** *In Proc. ACM SIGKDD, August 20–23, 2006, Philadelphia, Pennsylvania, USA. p. 96-105*
DžEROSKI, S. O. ***Multi-relational data mining: an introduction.*** *ACM SIGKDD Explorations Newsletter, Volume 5, Issue 1, 2003. p. 1 - 16.*
FUKUDA, T.; MORIMOTO, Y,; MORISHITA, S.; TOKUYAMA, T. ***Data minig using two-dimensional optimized association rules: Scheme, algoritms, and visualization.*** *In: PROC. OF THE 1996 ACM SIGMOD INT. CONF. MANAGEMENT OF DATA, Montreal, Canada, June, 1996. p. 13-23.*
HAN, J.; KAMBER, M. ***Data Mining - Concepts and Techniques.*** *1a edição. Nova York: Morgan Kaufmann, 2001.*
HAN, J.; PEI, J.; YIN, Y. ***Mining frequent patterns without candidate generation.*** *In: Proc. of the ACM SIGMOD Intl Conf. on Management of Data, Dallas, Texas, USA, 2000.*
HONG, T., KUO, C., CHI, S. ***Minining Association Rules from Quantitative Data.*** *The Eighth International Fuzzy Systems Association World Congress, 1999. Departament of Information Management. I-Shou University. Taiwan*
LENT, B. A.; SWAMI, A.; WIDOM, J. ***Clustering association rules.*** *In: PROC. 1997 INT. CONF. DATA ENGINEERING, Birmingham, England, Apr. 1997. p. 220-231.*
MATA, J; ALVAREZ, J, L; RIQUELME, J, C. ***An Evolutionary Algorithm to Discover Numeric Association Rules.*** *In: PROC. OF THE 2002 ACM SIGMOD INT. CONF. MANAGEMENT OF DATA, Madrid, Spain, 2002. p. 590-594.*
MILLER, R.;YANG, Y. ***Association rules over interval data.*** *In: 1997 ACM SIGMOD INT. CONF. MANANGENT OF DATA, Tucson, Arizona, 1997. p. 452-461.*
NESTOROV, S.; JUKIC, N. ***Ad-Hoc Association-Rule Mining within the Data Warehouse.*** *In: Proc. of 36th Annual Hawaii International Conference on System Sciences (HICSS3), Big Island, Hawaii, 2003. p.232a.*
PIZZI, L. ***Mineração de Dados em Múltiplas Tabelas.*** *88 f. Dissertação (Dissertação de Mestrado) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos, SP, 2006.*
POSSAS, B; MEIRA,W; RESENDE, R. ***Geração de regras de associação quantitativas****. In 14th Simpósio Brasileiro de Banco de Dados., Outubro 1999.*
PÔSSAS, B.; MEIRA JR, W.; CARVALHO, M.; RESENDE, R. ***Using quantitative information for efficient association rule generation.*** *In: ACM SIGMOD RECORD, v.29, n. 4, Dec. 2000. p. 19-25.*
RIBEIRO, M. ***Mineração de Dados em Múltiplas Tabelas Fato de Data Warehouse.*** *131 f. Dissertação (Dissertação de Mestrado) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos, SP, 2004.*
RIBEIRO, M. X.; VIEIRA, M. T. P. ***A New Approach for Mining Association Rules in Data Warehouses.*** *In: 6th International Conference On Flexible Query Answering Systems, Lyon, France, 2004.*
RIBEIRO, M. X.; VIEIRA, M. T. P.; TRAINA, A. J. M. ***Mineração de Regras de Associação Usando Agrupamentos.*** *In: I Workshop sobre Algoritmos de Mineração de Dados (WAMD2005), Uberlândia, MG, Brasil, 2005.*
SRIKANT, R; AGRAWAL, R. ***Mining quantitative association rules in large relational tables.*** *Technical report, IBM Almaden Research Center, San Jose, CA, 1996.*
WANG, W; YANG, J; YU, P. ***Efficient mining of weighted association rules (WAR).*** *In Proc. ACM SIGKDD 2000, Boston, MA USA. p. 270-274*

Anexos



$$Suporte = \frac{\text{Ocorrências de X}}{\text{Total de Ocorrências do BD}}$$

$$Confiança = \frac{\text{suporte}(X \cup Y)}{\text{suporte}(X)}$$