10º FEPEG FÓRUM ENSINO · PESQUISA EXTENSÃO · GESTÃO
RESPONSABILIDADE SOCIAL: INDISSOCIABILIDADE ENSINO, PESQUISA E EXTENSÃO UNIVERSITÁRIA
Realização: Unimontes MINAS GERAIS
Apoio: FAPEMIG FADENOR
ISSN 1806-549 X

Autor(es): RENÊ RODRIGUES VELOSO, LUIS PAULO TOLENTINO FERNANDES

Mineração de Dados aplicada ao Observatório do Trabalho no Norte de Minas

Introdução

Atualmente, a tarefa de extrair informações úteis de grandes volumes de dados é cada vez mais comum, sendo que, geralmente, necessita-se de um especialista em análise de dados para a realização dessa tarefa, que é árdua e demorada. Contudo, a área de Computação tem avançado bastante em fornecer métodos automáticos (algoritmos) para extração e análise desses grandes volumes de dados. Tais algoritmos podem, assim, ser facilmente entendidos e aplicados, permitindo uma maior rapidez e escala no volume de dados a serem processados. Para demonstrar isso, apresentamos neste trabalho um estudo de caso utilizando os microdados da RAIS (MINISTÉRIO DO TRABALHO, 2015) e os algoritmos Apriori e FP-Growth. O objetivo do estudo é a obtenção das chamadas regras de associação no contexto do mercado de trabalho formal do Norte de Minas. Acredita-se que as informações e os métodos utilizados podem ser relevantes para o Observatório do Trabalho no Norte Minas, projeto permanente do Grupo de Estudos e Pesquisas em Administração (GEPAD) da Universidade Estadual de Montes Claros (Unimontes).

Material e métodos

No método proposto por Tan et al. (2009) a mineração de dados é uma das etapas do processo de Knowledge Discovery in Databases (KDD) que é simplificado em três grandes etapas: pré-processamento dos dados, mineração e pós-processamento. A Figura 1 apresenta as etapas do processo de KDD, onde o objetivo da etapa do pré-processamento é preparar os dados brutos para a aplicação dos algoritmos na etapa seguinte, a etapa de mineração de dados. Nessa etapa, toda informação relevante é extraída por métodos automáticos (algoritmos), como, por exemplo, relações de causa-efeito, agrupamento de dados com características semelhantes, análises de tendências, classificação e previsão. A etapa de pós-processamento, por sua vez, garante que os resultados sejam filtrados e se tornem úteis em sistemas de apoio a decisões.

A etapa de mineração foi realizada neste trabalho utilizando os algoritmos Apriori e FP-Growth (Frequent Pattern-Growth), ambos algoritmos de geração de regras de associação. As regras de associação têm como premissa encontrar relações ou padrões frequentes entre conjuntos de dados. Através dessa análise, pode-se descobrir relações úteis na forma $\alpha \Rightarrow \beta$ que muitas vezes não são visíveis devido ao volume de dados. . Por exemplo, uma relação encontrada nos dados da RAIS é a seguinte: {Salário Mínimo, Feminino} → {Adulto}, a qual indica que trabalhadores do norte de Minas Gerais do sexo feminino que recebem salário mínimo têm idade adulta (entre 18 e 59 anos) (VASCONCELOS, 2004).

Os algoritmos Apriori e FP-Growth, baseiam-se no seguinte esquema geral: se qualquer padrão de comprimento k não é frequente na base de dados, seu comprimento (k + 1) não será frequente. A aplicação é feita por meio de um processo iterativo, visando gerar o conjunto de candidatos de comprimento (k +1) a partir do conjunto de padrões frequentes de comprimento k (para $k \geq 1$), e verificar suas frequências de ocorrência na base de dados. No entanto, a geração de candidatos pode se tornar um processo extenso quando há um grande número de padrões ou seja, quando a base de dados tem um número expressivo de itens. Pensando nisso, o FP-Growth foi criado com o objetivo de superar essas limitações utilizando um conceito diferente do Apriori, codificando o conjunto de dados em uma estrutura compacta em forma de árvore chamada Frequent Pattern tree (FP-tree) e extrai os conjuntos de itens frequentes diretamente desta estrutura (TAN, 2009).

Resultados e discussão

Nos experimentos, utilizou-se a Relação Anual de Informações Sociais (RAIS) do Ano base de 2015. A RAIS é disponibilizada pelo Ministério de Trabalho e Emprego. Entre outras funções da RAIS, está o provimento de dados para a elaboração de estatísticas do trabalho e a disponibilização de informações do mercado de trabalho aos gestores públicos.

Os dados da RAIS são disponibilizados em forma de texto, o que dificulta o seu processamento, os dados do ano base de 2015 por exemplo, têm 3.4 Gigabytes e um total de 7.355.176 registros. Assim, para melhorar a manipulação dos dados e iniciar a extração de informações, foi realizada uma etapa de conversão dos dados brutos para o banco de dados MYSQL (MYSQL, 2016), que é um sistema de gerenciamento de banco de dados muito difundido e que utiliza linguagem SQL (*Structured Query Language*, ou Linguagem de Consulta Estruturada) (ENGENHOFER, 1996) para a manipulação dos registros. Após a conversão dos dados, utilizou-se apenas as informações das cidades do norte de Minas Gerais. Em seguida, extraiu-se os seguintes campos:

ISSN 1806-549 X

- Sexo do trabalhador;
- Município de trabalho;
- Remuneração média;
- Idade;
- Subsetor de trabalho segundo o IBGE;
- Tempo de Emprego.

Os dados filtrados geraram 324.868 registros e, em seguida, uma etapa de discretização de alguns campos foi realizada. O campo *Idade* foi separada em "jovem", "adulto" e "idoso", o campo *Remuneração média* em "salário mínimo", "salário baixo", "salário médio" e "salário alto", e também atribuímos campos textuais aos campos *Sexo do trabalhador* e *Subsetor de trabalho segundo o IBGE*, pois estes campos eram numéricos (contínuos) e dificultaria a análise das regras geradas pelos algoritmos Apriori e FP-Growth. Terminada a fase de pré-processamento iniciou-se a fase de Mineração de Dados. Nesta fase, a primeira tarefa foi ajustar os valores de suporte e confiança do algoritmo Apriori, permitindo a geração de regras mais consistentes. A fase de pós-processamento foi simplificada, uma vez que os próprios algoritmos já geram as regras, bastando apenas interpretá-las. Os dados mostrados na Tabela 1 foram gerados pelos algoritmos Apriori e FP-Growth usando um suporte mínimo de 15% e confiança de 70%. Na tabela, são mostrados os itens frequentes, as regras de associação geradas e o conhecimento, isto é, a interpretação que pode-se obter de cada regra. Embora não apresentado neste trabalho, o algoritmo FP-Growth obteve um tempo de processamento menor que o algoritmo Apriori. Contudo, as regras geradas por ambos foram as mesmas.

Conclusões

Os resultados preliminares mostram que a técnica utilizada é consistente e permite a extração de regras de associação em grandes bases de dados. Experimentos demonstraram que as estatísticas extraídas da base RAIS estão em conformidade com aquelas que seriam geradas por especialistas humanos (com grau de certeza de mais 89%), sendo que, devido à quantidade de informações a serem analisadas, representa ainda ganhos significativos de tempo e esforço na tarefa de análise dos dados. Por meio desta técnica, é possível a construção de ferramentas automatizadas que auxiliem o Observatório do Trabalho no Norte de Minas no fornecimento de informações úteis aos tomadores de decisão da região.

Referências bibliográficas

EGENHOFER, M. Spatial, **SQL: A Query and Presentation Language**. IEEE Transactions on Knowledge and Data Engineering, 6:86-95, 1994.

MINISTÉRIO DO TRABALHO. Rais - **Relação Anual de Informações Sociais**, 2015. Disponível em: <http://www.rais.gov.br/sitio/sobre.jsf>. Acesso em: 02 nov. 2016.

MYSQL. The world's most popular open source database, 2016 Disponível em: <https://www.oracle.com/br/mysql/index.html>. Acesso em: 02 nov. 2016.

TAN, P., Steinbach, M., Kumar, V. **Introdução ao Data Mining – Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009.

VASCONCELOS, L.M.R; Carvalho, C. L. **Aplicação de Regras de Associação para Mineração de Dados na Web**, 2004. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf>. Acesso em: 02 nov. 2016.

Tabela 1. Regras de conhecimentos extraídas da base de dados da RAIS ano base 2015 com base apenas no Norte de Minas Gerais. A coluna Itens Frequentes mostra a porcentagem de vezes que os itens aparecem juntos nos registros da base de dados. A coluna Regras Geradas mostra a regra extraída e o grau de confiança da regra gerada.

Itens Frequentes	Regras Geradas	Conhecimento Extraído

ISSN 1806-549 X

Adulto, Feminino = 41%	Fem => Adulto Grau de certeza = 93%	41% dos adultos (entre 18 e 59 anos) são do sexo feminino e 59% são do sexo masculino com grau de certeza de 93%.
Adulto, Baixo = 46%	Baixo => Adulto Grau de certeza = 93%	46% dos adultos recebem salário entre R$880 e R$1700 com grau de certeza de 93%.
Adulto, 314330 = 37%	314330 => Adulto Grau de certeza = 90%	37% dos trabalhadores adultos estão trabalhando na cidade de Montes Claros (314330) com grau de certeza de 90%.
Masc, Mínimo, Adulto = 26%	Masc, Mínimo => Adulto Grau de certeza = 88%	26% dos adultos trabalhadores do norte de Minas Gerais e do sexo masculino recebem salário mínimo com grau de certeza de 88%.
Masc, Baixo, Adulto = 26%	Masc, Baixo => Adulto Grau de certeza = 92%	26% dos trabalhadores do sexo masculino que recebem salário entre R$880 e R$1700 tem idade adulta. Grau de certeza de 92%.
Mínimo, Fem, Adulto = 22%	Fem, Mínimo => Adulto Grau de certeza = 90%	22% dos trabalhadores do sexo feminino, recebem salário mínimo e tem idade adulta. Grau de certeza da regra de 90%.
Adulto, 314330, Masc = 20%	Masc, 314330 => Adulto Grau de certeza = 89%	20% dos adultos, trabalhadores do sexo masculino, estão trabalhando em Montes Claros.Grau de certeza de 89%.
Adm. Pública Autárquica, Adulto = 23%	Adm. Pública Autárquica => Adulto Grau de certeza = 94%	23% dos adultos trabalhadores trabalham no ramo da Administração Pública Autárquica. Grau de certeza de 94%.
Com. Varejista, Adulto = 17%	Com. Varejista =>Adulto Grau de certeza = 89%	17% dos adultos trabalhadores trabalham no ramo do Comércio Varejista. Grau de certeza de 94%.
Salário Médio, Adulto = 15%	Adulto => Salário Médio Grau de certeza = 96%	15% dos adultos trabalhadores recebem salário médio (de R$1700 à R$4000). Grau de certeza de 96%.

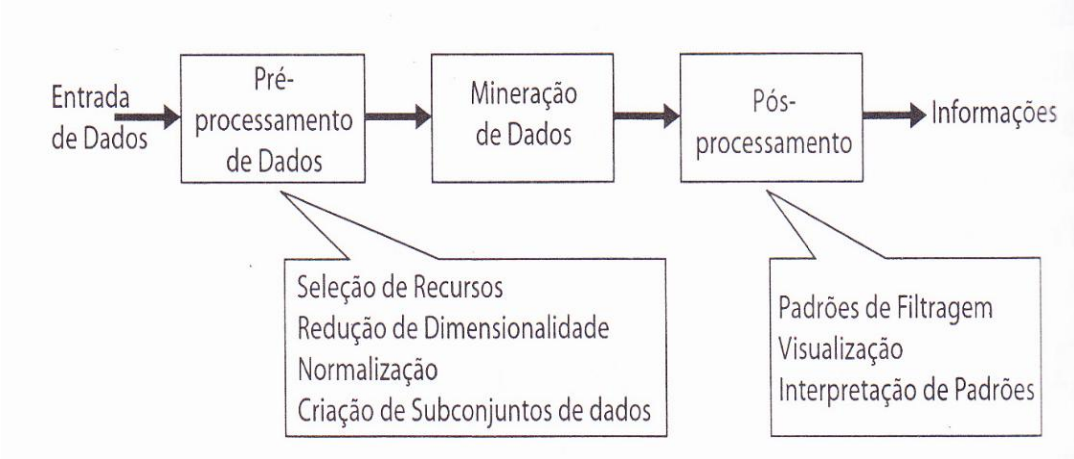

Figura 1. Etapas do processo de *Knowledge Discovery in Databases* (KDD).