AMBIENTE PARA APLICAÇÃO DAS TÉCNICAS DE MINERAÇÃO DE DADOS - *SURVEYS* DE GALÁXIAS

Thiago Crestani[1]; Fábio José Rodrigues Pinheiro[2]; Fábio Rafael Herpich[3];Fernando José Braz[4]; Marcelo Massocco Cendron[5]; Vinícius Barreto Klein[6]; Leila Lisiane Rossi[7]

RESUMO

O artigo descreve os passos realizados para a obtenção de um ambiente para aplicação das técnicas de mineração de dados. Inicialmente foi realizado o download dos *Surveys* (pesquisas) das bases relacionadas ao projeto como parte da etapa de *Matching* (agrupamento) de banco de Dados. Estes *Surveys* não envolveram os dados do tipo imagem, mas apenas os dados que descrevem os atributos das galáxias em formato texto ou numérico. Em seguida foram criados novos campos a partir do banco de dados já existente, a serem usados futuramente para a criação de gráficos. Tendo os dados coletados e implantados, foi realizada uma pesquisa com o objetivo de identificar qual a melhor linguagem de programação para minerar o grande volumes de informações. Utilizando a linguagem de programação *Python* foram gerados gráficos que serviram para mostrar características das galáxias. Na sequência foram criadas divisões de classes nesses gráficos para facilitar a visualização de atributos em grupos de galáxias. Na etapa final foi implementada a técnica de divisão de dados *Watershed* com o software *Matlab* e com ele juntamente com as linguagens *Python* e *C* foi gerada a técnica de divisão de dados usando clusters.

Palavras-chave: Galáxias, Mineração de Dados, Banco de Dados

1 INTRODUÇÃO

Como resultado da evolução tecnológica dos tempos atuais, a Astronomia passou a ser uma ciência que se utiliza de volumes de dados gigantescos, obtidos através de projetos de varredura do céu em grande escala. Para otimizar o processamento de informação obtida através deste sistema, surgiu a necessidade de uma nova forma de pesquisa na área *de* Astrofísica Observacional, a mineração de dados. Com isso tornou-se imprescindível o desenvolvimento de ferramentas de armazenamento, organização, integração, análise e exploração de dados, introduzindo um novo conceito de ciência observacional chamado de Observatório Virtual (VO).

[1] *Thiago Crestani, Ciência da Computação, 5º semestre, thiagocrestani@gmail.com*

[2] *Professor Orientador Fábio José Rodrigues Pinheiro, fabio@ifc-videira.edu.br*

[3] *Co- Orientador Fábio Rafael Herpich, fabiorafaelh@gmail.com*

[4, 5, 6, 7] *Professores Colaboradores do Projeto,fernando.braz@ifc-araquari.edu.br;marcelo.cendron,vinicius, leila.rossi@ifc-videira.edu.br*

Para cada *survey* são desenvolvidas as ferramentas necessárias para a obtenção da informação nele contida da forma mais eficiente. Porém, não há nenhuma forma de correlação já desenvolvida quando trata-se da comparação entre bancos distintos. Fazendo-se o sistema de *cross matching* entre dois bancos distintos, cada qual com sua função específica, pode-se obter um novo e mais amplo banco de dados, com um maior número de informações para cada objeto. E para que as pesquisas destes bancos de dados aconteçam de forma eficiente, é importante o uso das tecnologias de Banco de Dados e Mineração de Dados resultando em novos *surveys* com detalhes revelados, para o mesmo objeto, por *surveys* distintos. E ainda com a aplicação de algoritmos otimizados é possível executar este processo de *matching* com taxas razoáveis de tempo de resposta. O presente trabalho tem como objetivo descrever as etapas realizadas para a criação do ambiente para a aplicação das técnicas de mineração de dados de galáxias e está organizado conforme segue: O Capítulo 2 apresenta os Procedimentos Metodológicos. No Capítulo 3 são apresentados os Resultados e Discussões. O Capítulo 4 apresenta a Conclusão e finalmente no Capítulo 5 são apresentadas as Referências.

2 PROCEDIMENTOS METODOLÓGICOS

Os dados utilizados no projeto são provenientes das bases SDSS e WISE, ambos com vários objetos astronômicos e disponibilizadas para download.

SDSS - O Sloan Digital Sky Survey (SDSS), (YORK et.al.2000), (http://www.sdss.org/) é um dos maiores projetos da história da astronomia. Em 8 (oito) anos de operação (2000 – 2008), obteve imagens do céu profundo e multicoloridas de mais de um quarto do céu, criando mapas tridimensionais contendo mais de 930000 galáxias e mais de 120000 quasares. Em geral, os dados são disponibilizados à comunidade científica com atualizações anuais. Os dados deste período estão disponíveis no Data Release 7 (DR7, http://www.sdss.org/dr7/). Não obstante, o projeto ainda continua com o Third Sloan Digital Sky Survey (SDSS-III, http://www.sdss3.org/), iniciado em Julho de 2008 e tem previsão de durar até 2014. A primeira liberação de dados do SDSS-III foi através do Data Release 8 (DR8, http://www.sdss3.org/dr8) em

Janeiro de 2011, com uma segunda atualização em Agosto de 2012 com o Data Release 9 (DR9, http://www.sdss3.org/dr9).

O SDSS utiliza um telescópio dedicado com espelho primário de 2.5 metros de diâmetro, situado no Apache Point Observatory, no Novo México nos EUA e está acoplado a dois poderosos instrumentos: uma câmera CCD com 120 megapixels e 1.5 graus quadrados de área de céu, que equivale a cerca de oito vezes a área de uma Lua Cheia, e um par de espectrógrafos equipados com fibras óticas, capazes de obter a medida espectral de mais de 600 galáxias e quasares em uma simples observação.

WISE - O Wide-field Infrared Survey Explorer (WISE, Wright et al. 2010, http://adsabs.harvard.edu/abs/2010AJ....140.1868W, http://wise.ssl.berkeley.edu) é uma missão astronômica observacional idealizada e executada pela NASA (http://www.nasa.gov), com o objetivo de prover um vasto volume de dados de alta qualidade, potencializando a bagagem de conhecimento adquirido até os dias atuais. Isto proporciona a ampliação significativa do conhecimento sobre o Sistema Solar, Via Láctea e, não obstante, sobre o Universo. O WISE é um telescópio espacial que utiliza-se da parte espectral do infravermelho para imagear todo o céu.

Em 14 de Março de 2012, foi liberado o WISE All-Sky Data Release (http://wise2.ipac.caltech.edu/docs/release/allsky/) com dados para o céu inteiro nos comprimentos de onda 3.4, 4.6, 12 e 22 μm (W1, W2, W3 e W4, respectivamente), com uma resolução angular de 6.1”, 6.4”, 5.5” e 12.0” nas quatro bandas. A sensibilidade até 5σ medida em cada banda é melhor que 0.08, 0.11, 1 e 6 mJy para as quatro respectivas bandas em regiões fora da eclíptica, onde a luz zodiacal densa prejudica estas condições. No total, foram disponibilizados mais de 560 milhões de objetos detectados nas imagens do WISE.

Com o grande volume de dados obtidos nas duas frentes de trabalho (os grandes bancos de dados SDSS, WISE e o Observatório Municipal Domingos Forlin), o processamento e armazenamento se tornaram limitados. Para suprir essas necessidades, foi criado um ambiente de banco de dados, no qual é possível realizar consultas otimizadas, como por exemplo, de grupos bem definidos de galáxias obtidos a partir da aplicação de técnicas de mineração de

dados. As ferramentas da construção dessa infraestrutura foram totalmente desenvolvidas na instituição, proporcionando o contato da pesquisa com a prática aos alunos envolvidos no projeto.

As grandes bases de dados como a usada no projeto dificultam a descoberta de conhecimento quando são usadas as tecnologias tradicionais, pois estas não possuem a capacidade de revelar tendências ou padrões de relacionamentos entre as ocorrências de itens de dados. Uma possível solução é a aplicação da Mineração de Dados, a qual pode ser executada sobre vários tipos de dados como os bancos de dados relacionais, data *warehouses*, bancos de dados multimídia, bancos de dados espaciais, entre outros. Com isso é possível descobrir padrões nos dados até então desconhecidos. Para tal finalidade podem ser aplicadas técnicas de mineração de dados como as descritas a seguir:

[**Classificação**] Esta tarefa mapeia dados dentro de classes pré-definidas. Isto é um tipo de aprendizado supervisionado, tendo em vista que as classes são determinadas antes da etapa de exame dos dados. Algoritmos de classificação necessitam que as classes sejam definidas baseadas em valores de atributos de dados. Um padrão de entrada é classificado dentro de uma ou mais classes baseado na similaridade com as classes pré-definidas.

[**Regressão**] O objetivo nesta tarefa é mapear um item de dados para uma variável de previsão. Isto pode ser feito através do aprendizado da função que executa este mapeamento.

[**Análise de Séries Temporais**] Esta tarefa permite examinar o valor de um atributo ao longo do tempo, enquanto que diferentes séries dos valores de atributos são obtidos. Utilizando estas séries temporais é possível, por exemplo, investigar o nível de similaridade entre diferentes séries.

[**Agrupamento**] Esta tarefa tem alguma similaridade com a tarefa de classificação. A principal diferença é que, nesta tarefa, os agrupamentos (classes), não são previamente conhecidas. Na tarefa de agrupamento os grupos são definidos pelos próprios dados. Isto acontece através da avaliação da similaridade entre os dados em atributos previamente definidos.

[**Regras de Associação**] Regras de associação são utilizadas para mostrar o relacionamento entre itens de dados. Esta é uma das tarefas mais populares na área de Mineração de Dados. O objetivo principal da regra de associação é

identificar padrões de co-ocorrências dos itens de dados (AGRAWAL; IMIELINSKI; SWAMI, 1993), (FAYYAD et al., 1996), (GANTI; GEHRKE; RAMAKRISHNAN, 1999), (AGRAWAL; SRIKANT, 1995). A compra de um produto quando outro produto é adquirido previamente, este é um exemplo de uma regra de associação.

[**Descoberta de Sequência**] O objetivo desta tarefa é encontrar padrões sequenciais em dados. Os padrões são baseados em uma sequencia temporal de ações. Em um ambiente de compras online, por exemplo, o cliente pode efetuar suas compras seguindo uma determinada sequencia temporal. Estas sequências podem identificar padrões de comportamento de compras.

No presente projeto foi aplicada principalmente a técnica de agrupamento permitindo assim a geração de *clusters* de galáxias conforme as suas características.

3 RESULTADOS E DISCUSSÕES

Os principais resultados obtidos no projeto foram a criação do ambiente para aplicação das Técnicas de Mineração de Dados em Pesquisa de *Matching* entre *Surveys* de Galáxias e a aplicação das técnicas de Mineração de Dados *Watershed* (linha divisora de águas) (Figura 1) a qual consiste em inundar a imagem, que nesse caso é um gráfico, e em seguida remover as partes inundadas, sobrando apenas as linhas com os pontos mais altos *e Simple-K-Means (*agrupamento). Além da criação de softwares para a conversão de dados e também gráficos do tipo *WHAN* (Figura 2) e *BPT* que são usados para classificar galáxias.

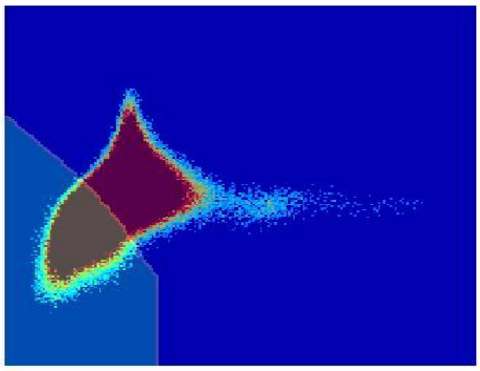

Figura 1 - Aplicação da técnica de Watershed em um gráfico Whan com os parâmetros W1_W4 – U_R

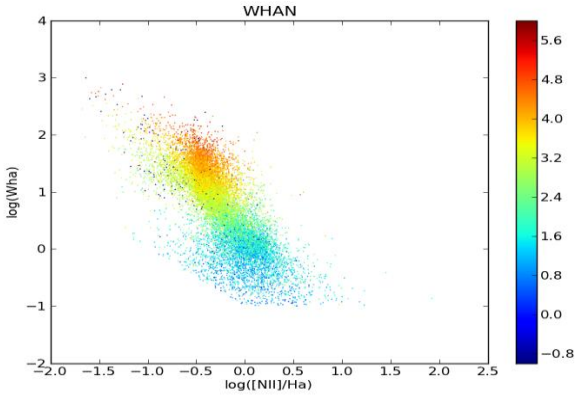

Figura 2 - Gráfico do tipo Whan

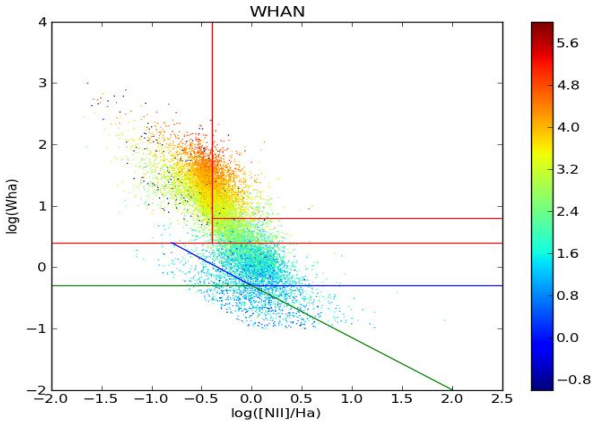

Figura 3 - Gráfico do tipo Whan com divisões de classes

Após a obtenção das bases foi iniciado um trabalho de estudo sobre qual ferramenta de banco de dados seria a melhor para armazenar as informações. Escolheu-se o software *PostgreSQL* devido a sua robustez e bom desempenho se tratando de grandes volumes de dados.Então foi desenvolvido um software na linguagem de programação *Java* capaz de realizar a conversão dos dados no formato original *Fits,* para o *SQL.* Em seguida foi feita a inserção dos dados na base, através de linha de comando.

Com os dados armazenados foi realizada uma limpeza na base, isto é, foram retirados valores que se tratavam de erros na captura, valores fora dos padrões, que deveriam ser removidos para não interferirem nos resultados finais. Então, através de comandos *SQL* foi feita a remoção desses dados. Com a base limpa efetuou-se a criação de novos campos de dados, esses campos seriam utilizados futuramente para a criação de gráficos, bem como a comparação entre valores. Eles foram gerados a partir de comando *SQL*. Os

parâmetros para gerá-los saíram de campos pré-existentes, que foram subtraídos entre si, gerando as combinações.

Após a base estar devidamente pronta para a mineração era necessário a interação dela com uma linguagem de programação capaz de gerar gráficos com os pontos no banco de dados. Realizou-se uma pesquisa bibliográfica e também testes, e constatou-se que a melhor linguagem para essas tarefas seria a linguagem de programação *Python* devido a sua simplicidade a alta produtividade. Então, foi implementado um software que a partir do banco de dados gera gráficos do tipo *WHAN* e *BPT* usados para mostrar características de galáxias. Com a melhoria desse mesmo software, gráficos do tipo *WHAN* podem ser gerados com divisões de classes (Figura 3), permitindo a melhor visualização de diferentes tipos de galáxias.

Tornou-se necessária traçar uma linha divisória nos gráficos do tipo *WHAN* para a visualização de galáxias jovens e antigas. Para este problema foi utilizado a *Watershed*. Neste caso foi implementada uma variação dessa técnica, que consiste em remover os pontos mais altos da gráfico, sobrando apenas uma linha que é traçado pelos pontos mais baixos do gráfico. A técnica foi desenvolvida em duas etapas. Inicialmente os gráficos foram preparados, utilizando a linguagem *python.* Foram criados gráficos em forma de matrizes, que continham informações de quantos pontos apareciam em cada parte do gráfico. Em seguida com o software matemático *matlab* os gráficos foram transformados em imagens, e em seguida submetidos a uma serie de técnicas de tratamento de imagens para a implementação da técnica *Watershed.* Ao final com o próprio *matlab* a técnica foi implantada e então como resultado foi obtida a linha de separação entre galáxias jovens e antigas. (Figura 1)

4 CONCLUSÃO

Com o uso da informática, a interpretação de dados astronômicos torna-se mais rápida e produz melhores resultados. Com um ambiente especifico para a mineração de dados criado e ferramentas adequadas é possível produzir diversos resultados de enorme importância para a comunidade científica. A implantação de técnicas de mineração de dados permite obter várias informações e padrões presentes em grupos de galáxias. Contudo, devido à

complexidade do projeto e os resultados obtidos até o momento, considera-se necessária a continuidade do mesmo através da aplicação de outras técnicas de mineração de dados, permitindo assim compará-las e escolher as melhores para a referida base de dados. Espera-se também obter uma melhor separação entre as classes de galáxias, bem como o estudo das propriedades do meio interestelar, bem como a efetiva interferência de gás e poeira sobre a luz produzida pelas estrelas componentes. Estes dados serão disponibilizados no Observatório Virtual, o qual poderá ser acessado tanto pela comunidade acadêmica quanto pelo público externo.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. N. Mining association rules between sets of items in large databases. Proceedings of the ACM International Conference on Management of Data, p. 207–216, 1993.

AGRAWAL, R.; SRIKANT, R. Mining sequential patterns. In: YU, P. S.; CHEN, A. S. P. (Ed.). Eleventh International Conference on Data Engineering. Taipei, Taiwan: IEEE Computer Society Press, 1995. p.3–14. Disponível em: <citeseer.ist.psu.edu/agrawal95mining.html>.

DUNHAM, M. H. Data mining introductory and advanced topics. [S.l.]: Prentice Hall, 2003.

FAYYAD, U. et al. (Ed.). Advances in knowledge discovery and data mining. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. ISBN 0-262-56097-6.7

GANTI, V.; GEHRKE, J.; RAMAKRISHNAN, R. Mining very large databases. Computer, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 32, n. 8, p. 38–45, 1999. ISSN 0018-9162.

SDSS - Disponível em: <http://www.sdss.org/> - Visto em 30 de Abril de 2013

Wright, E. L. et al. The Wide-field Infrared *Survey* Explorer (WISE): Mission Description and Initial On-orbit Performance. v. 140, p. 1868–1881, dez. 2010.

YORK, D. G. et al. The Sloan Digital Sky *Survey*: Technical Summary. The Astronomical Journal, v. 120, n. 3, p. 1579–1587, set. 2000. ISSN 00046256. Disponível em: <http://stacks.iop.org/1538-3881/120/i=3/a=1579>