Mineração de Dados a partir do Currículo Lattes com a Ferramenta WEKA

Un Hee Schiefelbein[1], Igor Camargo Moiano[1], Tairone Livinalli[1], Milene Santos Teixeira[1], Matheus Ricalde de Souza[1], Juçara Salete Gubiani[1]

[1]Universidade Federal de Santa Maria (UFSM) ï 97.105-900 ï Santa Maria ï RS ï Brasil

{peace.unhee, milene.tsi, matheusricaldee}@gmail.com, {igor.moiano, taironelivinalli}@hotmail.com, jucara@ufsm.br

Abstract. *Data mining aims to mine or extract knowledge in large volumes of data using tools and techniques that, by the use of learning algorithms, are able to explore a data set extracting or highlighting patterns to assist in knowledge discovery. This paper makes a review on data mining and shows the result of the processing of the algorithm J48, which is implemented in the WEKA tool. The data were extracted from the base of the Lattes curriculum of a set of professors from the Federal University of Santa Maria.*

Resumo. *A mineração de dados tem como objetivo minerar ou extrair conhecimento em grandes volumes de dados usando ferramentas e técnicas, que por meio de algoritmos de aprendizagem, são capazes de explorar um conjunto de dados, extraindo ou evidenciando padrões para auxiliar na descoberta de conhecimento. O presente trabalho faz uma revisão sobre mineração de dados e mostra o resultado do processamento do algoritmo J48 implementado na ferramenta WEKA. Os dados foram extraídos da base do currículo Lattes de um conjunto de professores da Universidade Federal de Santa Maria.*

1. Introdução

Segundo Tan (2009), ao longo dos últimos anos com avanço da tecnologia, organizações começam a armazenar uma grande quantidade de dados, estes que já não possuem a mesma dimensionalidade e complexidade de décadas anteriores, dificultando a filtragem de informações úteis. Surge então a mineração de dados, uma tecnologia que combina métodos tradicionais de análises de dados com algoritmos sofisticados para processamento de grandes volumes de dados.

Para Cios et al. (2007) desde o surgimento de sistemas computacionais, o principal objetivo das organizações tem sido o armazenamento de dados. Nas últimas décadas essa tendência ficou ainda mais evidente com a queda nos custos e a facilidade da aquisição de *hardware*, tornando assim, possível armazenar quantidades cada vez maiores de dados. Estruturas de armazenamento novas e mais complexas foram desenvolvidas, tais como: banco de dados, *data warehouses*, bibliotecas virtuais, *web* e outras.

O artigo inicia com a metodologia usada para o desenvolvimento do trabalho na

seção 2, seguindo na seção 3 com a fundamentação teórica da Mineração de Dados. Na seção 4 é apresentada a ferramenta WEKA, assim como o seu funcionamento, seguida pela seção 5 onde é apresentado o uso do Lattes com a Mineração de Dados. Após, tem-se a seção 6 com os resultados e discussões sobre o trabalho e por fim, na seção 7 apresentam-se as conclusões sobre o estudo.

2. Metodologia

O trabalho faz um estudo bibliográfico sobre mineração de dados e a partir deste é realizado um estudo de caso utilizando a ferramenta WEKA. Por fim, são feitas algumas considerações.

3. Mineração de Dados

De acordo com Fayyad et al. (1996), a definição de mineração é dada da perspectiva do aprendizado de máquina: "Mineração de Dados é um passo no processo de descoberta de conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados."

A mineração de dados é uma parte integral da descoberta de conhecimento em banco de dados (KDD- *Knowledge Discovery in Databases*), que é processo geral de conversão de dados brutos em informações úteis. O processo consiste em vários passos de transformação: pré-processamento, mineração e o pós-processamento dos resultados da mineração de dados. Fayyad et al. (1996).

4. Software para Mineração de Dados: WEKA

A ferramenta *Waikato Environment for Knowledge Analysis* – WEKA foi desenvolvida pela Universidade de Waikato na Nova Zelândia, utilizando linguagem de programação Java e possuindo código aberto emitido sob a GNU *General Public License*. WEKA possui um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados onde, os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamado a partir de seu próprio código Java. WEKA contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação, e visualização. É também bem adequada para o desenvolvimento de novos sistemas de aprendizagem máquina. (Hall, et al. 2009).

5. Mineração de Dados a Partir dos Resumos do Currículo Lattes (CNPq)

O trabalho foi desenvolvido com base nas etapas processamento dos dados, pré-processamento, mineração e pós-processamento dos dados.

Para o pré-processamento dos dados, ocorre a extração de somente o resumo do currículo Lattes de 67 professores de uma universidade, o qual foi obtido automaticamente por meio do uso de um *script* desenvolvido para este fim. Após isto, as informações foram estruturadas no padrão do WEKA (.arff). O nome do conjunto de dados é especificado através da marcação @RELATION areadoconhecimento, @atribute resumo string e @atribute area {computação, outraArea} para os atributos e os dados definidos por meio da marcação @data, como apresentados na figura 1.

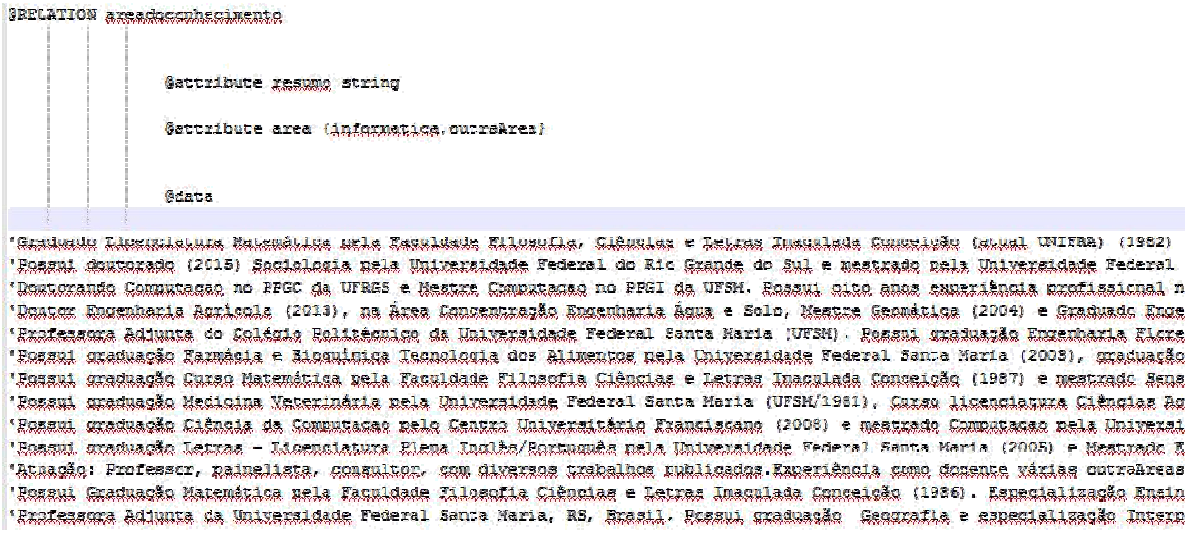

Figura 1. Dados no padrão WEKA (.arff)

Com o arquivo já no formato é possível fazer uma a limpeza dos dados transformando as *Strings* em vetores, podendo então remover palavras que não contribuirão com mineração, como demonstrado na figura 1.

Para a mineração dos dados optou-se pela escolha do algoritmo J48, que permite gerar uma árvore de decisão, sendo assim, possível entender a relação entre as palavras do resumo dos atributos e a classe "área" procurada. A ferramenta WEKA extrai os classificadores (ou modelo de classificação) para identificar a classe à qual pertence uma determinada observação de uma base de dados, a partir de suas características (seus atributos).

O atributo "área" representa o atributo classe, ele é utilizado para indicar se a área do professor pertence a "informática" ou "outraArea", enquanto o atributo "resumo" é preditivo, seus valores serão analisados para que seja descoberto o modo como eles se relacionam com o atributo classe.

6. Análise e Discussão dos Resultados

A árvore de decisão gerada por meio do algoritmo J48 apresenta a palavra **Computação** como nó raiz (figura 2), pois foi considerado pelo algoritmo classificador como o atributo mais importante para determinar se o professor pertencia a área de informática ou outra.

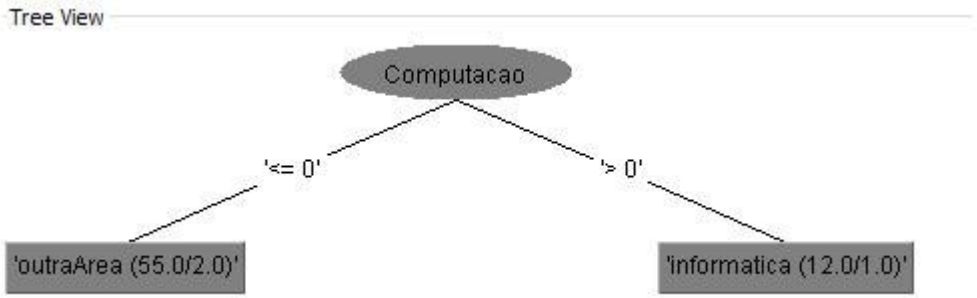

Figura 2. Árvore de Decisão formada pelo algoritmo J48

Os resumos analisados pelo algoritmo que continham então a palavra **Computação** foram selecionados como '> 0', determinado assim qual classe eles pertencem, no caso informática, como exemplificado na figura 3.

```
J48 pruned tree
------------------

Computacao <= 0: outraArea (55.0/2.0)
Computacao > 0: informatica (12.0/1.0)
```

Figura 3. Determinação das classes por meio da palavra selecionada pelo algoritmo J48

A árvore de decisão determinou palavras chaves relacionando as mesmas aos professores da área de informática. O algoritmo realizou uma seleção semelhante ao mundo real através da mineração de dados realizada. Portanto através do estudo de caso realizado é possível implementar a mineração de dados através da ferramenta WEKA para realizar uma otimização na busca de professores da área de informática através do currículo lattes.

7. Conclusões

O algoritmo selecionou a palavra Computação, pois esta é a palavra que melhor pode identificar o resumo como sendo da área da informática. Esta análise tornou evidente a importância do uso de palavras objetivas na descrição dos resumos do currículo Lattes, para que os mecanismos de buscas encontrem cada vez mais resultados com precisão. O trabalho proposto demonstrou ser possível realizar a mineração de dados do currículo lattes e pretende se futuramente realizar a implementação de um sistema utilizando a mineração de dados proposta.

Referências

Cios, K. J., Pedrycz, W., Swiniarski, R. W. e Kurgan, L. A. (2007). Data Mining - A Knowledge Discovery Approach. Springer.

CNPq. Plaforma Lattes. Disponível em: http://lattes.cnpq.br/

Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.

Han, J. e Kamber, M. (2006). Data Mining: Concepts and Techniques. Elsevier.

Larose, D. T. (2005) Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons, Inc.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009).

The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Tan, P.N., Michael, S. e Vipin, K. (2009). "Introdução ao datamining: mineração de dados". Ciência Moderna.