EXTRAÇÃO DE CONHECIMENTO DA PLATAFORMA LATTES UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS: ESTUDO DE CASO POLI/UPE

Trabalho de Conclusão de Curso

Engenharia da Computação

Aluno: Bruno Carlos Sales de Morais

Orientador: Prof. Dr. Carmelo Jose Albanez Bastos Filho

UNIVERSIDADE
DE PERNAMBUCO

Bruno Carlos Sales de Morais

Extração de conhecimento da Plataforma Lattes utilizando técnicas de Mineração de Dados: estudo de caso POLI/UPE

Monografia apresentada para obtenção do Grau de Bacharel em Engenharia da Computação pela Universidade de Pernambuco

Orientador:
Prof. Dr. Carmelo Jose Albanez Bastos Filho

Departamento de Sistemas e Computação
Escola Politécnica de Pernambuco
Universidade de Pernambuco

Recife - PE, Brasil

dezembro de 2010

Resumo

Atualmente, existe uma grande dificuldade na aquisição de dados sobre a produção científica do corpo docente dentro das universidades brasileiras. Entretanto, a maior parte destas informações pode ser encontrada na plataforma Lattes - um sistema desenvolvido pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para auxiliar na gestão de ciência, tecnologia e inovação no Brasil. Mineração de Dados é um conjunto de técnicas que podem ser utilizadas para extração de novas informações em uma base de dados. Desta forma, a utilização de técnicas de Mineração de Dados pode ajudar na extração de informações importantes na plataforma Lattes. O presente trabalho apresenta uma ferramenta capaz de extrair dados automaticamente de currículos da plataforma Lattes. Além disso, o principal objetivo deste trabalho é aplicar técnicas de Mineração de Dados nas informações extraídas dos currículos Lattes dos professores da Escola Politécnica de Pernambuco (POLI). Desta maneira, as novas informações podem ser utilizadas para auxiliar na tomada de decisões dos gestores da Universidade de Pernambuco com relação a investimentos nos cursos e nos docentes da instituição.

Abstract

Nowadays, the Brazilian universities have a difficult task that is to collect the scientific production of their researchers. However, this information is widely available on the Internet by the Lattes platform - a web-based system developed by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). The Lattes platform aims to help the management of science, technology and innovation in Brazil. Data mining is a set of techniques that can be used to extract patterns from a database. Thus, some techniques from Data Mining may be used to extract useful information inside the Lattes platform. The goal of this research is to present a tool to automatically extracts data from Lattes curriculums. Moreover, we used some Data Mining techniques to extract hidden information from Lattes curriculums of the professors of Polytechnic School of Pernambuco (POLI). This new information can be used to guide decision making of investiments in the institution.

Agradecimentos

Primeiramente agradeço a Deus por me dar discernimento e forças para realizar este trabalho, pois acredito que sem Ele não conseguiria superar as dificuldades enfrentadas ao longo deste curso.

Aos meus pais Edilson e Nadja que sempre me apoiaram e me encorajaram a concluir esta graduação. Reconheço todo o esforço que eles fizeram para que eu pudesse estudar sem a preocupação de ter que trabalhar. Um agradecimento especial ao meu irmão Dyego Carlos por todo o apoio, companheirismo e ajuda na realização deste trabalho, que por muitas vezes abdicou de dormir para me ajudar.

Aos meus avós maternos, Elza e Gabriel, e paternos, Dulcelene e Morais, pelo apoio incondicional e sábios conselhos durante toda a minha vida.

Aos professores do curso de Engenharia da Computação, que sem dúvida contribuíram para o que sou hoje como profissional e ser humano. Em especial agradeço ao professor Carmelo, que mesmo com tantas ocupações sempre esteve disponível para me oferecer excelente orientação durante todo este trabalho.

E por fim, agradeço aos meus colegas de turma, em especial, Carlos Eduardo Buarque, Francisco Marinho e Marcos Antonio, pelas muitas madrugadas que dividimos para que os inesquecíveis projetos desenvolvidos ao longo do curso pudessem ser concluídos.

Sumário

Lista de Figuras

Lista de Tabelas

Tabela de Símbolos e Siglas

ARFF - *Attribute-Relation File Format*

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

DOI - *Digital Object Identifier*

FACEPE - Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco

FINEP - Financiadora de Estudos e Projetos

HTML - *HyperText Markup Language*

IC - Iniciação Científica

ISSN - *International Standard Serial Number*

JCR - *Journal Citation Reports*

KDD - *Knowledge Discovery in Databases*

LHS - *Left Hand Side*

MYSQL - Banco de dados relacional que utiliza a Linguagem de Consulta Estruturada (SQL)

PHP - *Hypertext Preprocessor*

POLI - Escola Politécnica de Pernambuco

RHS - *Right Hand Side*

TCC - Trabalho de Conclusão de Curso

URL - *Uniform Resource Locator*

WEKA - *Waikato Environment for Knowledge Analysis*

1 Introdução

Atualmente, com a facilidade de armazenar dados, as empresas e organizações possuem registros de suas atividades cada vez mais completos. Porém, poucas são as organizações que utilizam esses dados para transformá-los em conhecimento útil para a sua gestão. Ou seja, as empresas e organizações possuem grande quantidade de dados, mas não fazem processamentos e análises desses dados, de forma que sejam produzidas informações que possam auxiliar nas suas tomadas de decisões.

A informação está se tornando cada vez mais a principal matéria-prima de grandes organizações, por isso faz-se necessário a aplicação de processos que acelerem a extração de informações de grandes bases de dados. Neste contexto, o processo KKD (*Knowledge Discovery in Databases*) pode ser utilizado para auxiliar a descoberta de conhecimento útil em grandes bases de dados [1].

Uma das principais etapas do processo KDD, a Mineração de Dados, consiste na aplicação de algoritmos com a finalidade de extrair padrões de comportamento em uma base de dados [2]. Algumas das mais importantes tarefas descritivas da Mineração de Dados são clusterização e regras de associação. Dada essa natureza, essas tarefas foram selecionadas para serem aplicadas neste trabalho. As regras de associação demonstram o quanto a ocorrência de um conjunto de itens implica na ocorrência de algum outro conjunto distinto de itens nos registros de uma mesma base de dados [3]. Já a clusterização pode ser utilizada para agrupar os dados de acordo com uma medida de similaridade pré-definida [4], [5], [6].

1.1 Formulação do Problema

Atualmente, existe uma dificuldade enorme nas universidades brasileiras para aquisição de dados sobre a produção científica do seu corpo docente. Entretanto, a maior parte destes dados pode ser encontrada na plataforma Lattes, que é um sistema de informação desenvolvido pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para auxiliar a gestão

de ciência, tecnologia e inovação no Brasil [7]. Sendo, portanto, uma rica fonte de informações sobre a produção científica, tecnológica e bibliográfica dos pesquisadores do Brasil.

A utilização de técnicas de Mineração de Dados [8] pode auxiliar na extração de informações importantes desta plataforma e, consequentemente, guiar melhores investimentos na área de Ciência & Tecnologia das universidades. Porém, a determinação de padrões que são realmente úteis, ainda requer uma grande interação com analistas humanos, o que torna o processo de extração do conhecimento uma tarefa não trivial.

Este trabalho visou estudar os conceitos e técnicas da Mineração de Dados e aplicar o conhecimento adquirido para descobrir informações não explícitas nos currículos Lattes dos professores da Escola Politécnica de Pernambuco - POLI. Este trabalho propõe a utilização do WEKA (*Waikato Environment for Knowledge Analysis*), que é um *software* livre com implementações de diversas técnicas de Mineração de Dados, para auxiliar na extração do conhecimento [9]. Especificamente, este trabalho se concentra em:

- Selecionar dados disponíveis na plataforma Lattes referentes à produção científica de professores da POLI.
- Projetar e desenvolver uma ferramenta para extrair os dados selecionados da plataforma Lattes automaticamente;
- Selecionar técnicas de Mineração de Dados para serem aplicadas aos dados obtidos;
- Utilizar o WEKA para aplicar as técnicas selecionadas;
- Analisar e organizar as informações obtidas, de forma que o conhecimento gerado possa ser utilizado por gestores da instituição.

1.2 Estrutura da monografia

No capítulo 2, é apresentada uma revisão bibliográfica que contempla os principais conceitos necessários à compreensão deste trabalho.

No capítulo 3, são apresentadas a plataforma Lattes, a ferramenta desenvolvida para extração de dados a partir da plataforma Lattes e a ferramenta de Mineração de Dados (WEKA).

No capítulo 4, são descritos o estudo de caso foi realizado e os resultados da aplicação dos algoritmos de mineração de dados.

O capítulo 5 contém as conclusões e dificuldades encontradas. Além disso, neste capítulo são sugeridos trabalhos futuros.

2 Revisão bibliográfica

Neste capítulo são apresentados conceitos e definições que são necessários à compreensão deste trabalho. Na seção 2.1, são apresentadas as etapas do processo KDD. A seguir, na seção 2.2, é apresentado o conceito de Mineração de Dados, bem como suas principais tarefas. Na seção 2.3, são apresentadas as principais fases da clusterização, como pré-processar os diferentes tipos de dados para a clusterização e o algoritmo Kmeans. Por fim, na seção 2.4, são apresentados os principais conceitos acerca de regras de associação e o algoritmo Apriori.

2.1 Processo KDD

Historicamente, encontrar padrões em dados é conhecido por diferentes nomes em diferentes comunidades (por exemplo, extração de conhecimento, descoberta de informação, arqueologia de dados, processamento de padrões de dados e Mineração de Dados). O termo Mineração de Dados é muitas vezes utilizado em comunidades de estatísticos, pesquisadores de banco de dados, e mais recentemente por gerentes de sistemas de informação. Mas segundo [2]), o termo KDD (*knowledge-discovery in databases*) referencia o processo global de descobrir conhecimento útil em grandes bases de dados, sendo a Mineração de Dados um passo particular desse processo que consiste na execução de algoritmos de reconhecimento padrões em base de dados.

KDD pode ser definido como [1]: O processo não trivial de identificação de padrões válidos, inovadores, potencialmente úteis e principalmente compreensíveis em bases de dados.

O processo KDD é interativo e iterativo, envolvendo numerosos passos com muitas decisões tomadas pelo analista. Nas próximas seções, são apresentados os passos básicos que compõem esse processo, conforme fluxograma ilustrado na figura 1.

Figura 1: Etapas do Processo KDD.

2.1.1 Identificação do domínio da aplicação

Nesse passo é realizado um estudo do domínio da aplicação para definir os objetivos e metas a serem alcançados com o processo KDD. Esse passo é importante, pois para que seja extraída informação útil dos dados, as pessoas envolvidas no processo KDD devem possuir algum grau de conhecimento sobre a área da aplicação.

2.1.2 Seleção dos dados

Os sistemas, normalmente, armazenam vários atributos de um objeto, mas nem sempre todos estes atributos são relevantes para a análise em questão. Assim, faz-se necessário uma criteriosa avaliação de quais atributos realmente agregam informações ao conjunto de dados para que esses venham a ser utilizados no processo KDD.

A qualidade dos dados armazenados é muito importante, pois ela determina a qualidade dos resultados obtidos, de forma diretamente proporcional. Porém, o processo KDD ainda é muito dependente da avaliação de analistas humanos e do seu conhecimento sobre a base de dados pesquisada, pois caso um atributo que contenha informações importantes seja desprezado nesta fase o resultado poderá não ser satisfatório.

2.1.3 Limpeza e Pré-processamento dos dados

No mundo real os dados tendem a ser incompletos, ruidosos e inconsistentes, o que torna necessária a aplicação de técnicas para corrigir essas falhas. A seguir são apresentadas algumas das atividades que podem ser realizadas na etapa de limpeza dos dados [10]:

- *Valores ausentes*. Por vezes, os valores de alguns atributos não estão presentes nas bases, para lidar com esse problema pode-se: ignorar os registros que possuem valores ausentes; preencher os valores ausentes manualmente; usar um valor constante, usar o valor médio do atributo ou usar valores estatísticos para preencher os valores ausentes.

- *Valores fora do padrão (outliers)*. Para resolver esse tipo de problema pode-se utilizar uma das seguintes técnicas: agrupamento, que consiste no agrupamento dos valores similares, facilitando a identificação e exclusão de valores fora do padrão; inspeção humana e computador, que consiste em uma inspeção feita por pessoas e computadores para identificar e excluir os valores fora do padrão; e regressão, na qual os dados podem ser ajustados por meio de funções de regressão.

- *Dados inconsistentes*. Esses erros ocorrem normalmente porque o usuário entra com um dado incorreto, mas pode ocorrer também redundância de dados, ou seja, dados que possuem o mesmo valor semântico, mas que foram inseridos com nomes diferentes. Uma forma de resolver esse problema é por meio da análise de correlação, que consiste na medida de relacionamento entre dois atributos.

2.1.4 Transformação dos dados

Por vezes é necessário realizar transformações nos dados para que os algoritmos de mineração possam ser executados, dentre as mais comuns estão:

- *Normalização de atributos*. Forma de harmonizar as escalas dos atributos em um pequeno intervalo especificado.

- *Padronização de atributos*. Atributos redundantes devem ser eliminados, utilizando a análise de correlação para mapear múltiplos atributos para uma simples entidade, por exemplo.

- *Redução dos dados*.

  Agregação. Agrega e sumariza os dados. Por exemplo, os dados de vendas diárias podem ser agregados de forma a calcular os montantes mensais e anuais.

  Compressão de dados. Mecanismos de codificação são usados para reduzir o tamanho do conjunto de dados.

  Redução da dimensionalidade. Eliminação de atributos irrelevantes à técnica de mineração aplicada.

  Redução numérica. Diminuir o número de instâncias, por exemplo.

- *Projeção de dados*. Criação de novos atributos julgados relevantes a partir de outros atributos existentes.

2.1.5 Mineração de dados

Esse é o passo onde os padrões são descobertos por meio de algoritmos específicos que são aplicados várias vezes até que sejam extraídas informações úteis dos dados. Por ser um dos passos mais importantes do processo KDD, este passo é descrito mais detalhadamente na seção 2.2.

2.1.6 Interpretação e Avaliação dos padrões

Os resultados obtidos pelos algoritmos de mineração devem ser avaliados por analistas para que sejam julgados como úteis ou não. Os padrões que forem considerados úteis devem ser interpretados de forma que se tornem compreensíveis para os usuários finais do sistema.

O processo KDD é considerado iterativo, pois pode ser necessária sua execução por várias vezes até que, por meio da repetição de qualquer um dos passos anteriores, se obtenha resultados satisfatórios [1].

2.1.7 Utilização do conhecimento

Nesse passo, o conhecimento extraído é consolidado sendo incorporado a um sistema, utilizado diretamente pelo usuário final ou, simplesmente, documentado e relatado às pessoas interessadas. Os resultados dos passos anteriores devem ainda ser analisados para que possíveis conflitos entre o conhecimento existente e o conhecimento adquirido sejam solucionados [1].

2.2 Mineração de Dados

Mineração de Dados é um dos principais passos do processo KDD, que consiste na aplicação de algoritmos com a finalidade de extrair padrões de comportamento em uma base de dados [1]. Essa é uma área de pesquisa multidisciplinar, que inclui inteligência computacional, matemática (estatística) e banco de dados. Um dos grandes desafios da Mineração de Dados é propor algoritmos que sejam capazes de lidar com escalabilidade e alta dimensionalidade dos dados, ou seja, trabalhar com grandes quantidades de dados.

Segundo [1], o processo de Mineração de Dados possui dois objetivos principais: predição, onde a ideia é prever o comportamento futuro de algumas variáveis da base de dados; e descrição, onde a ideia é identificar padrões que representem a distribuição dos itens de tal forma que esses padrões sejam passíveis de interpretação.

É importante diferenciar o que é uma tarefa e o que é uma técnica de Mineração de Dados. Uma tarefa está relacionada a *o que* se pretende buscar nos dados, ou seja, que tipo de padrões deseja-se encontrar. Já uma técnica, está relacionada a *como* encontrar os padrões de interesse. Por exemplo, para o caso de um gasto no cartão de crédito de um cliente acima do normal, para este caso pode-se definir uma tarefa, que seria detecção de desvios, e uma técnica para resolver o problema, que seria redes neurais, por exemplo [8]. A seguir, são apresentadas as principais tarefas da Mineração de Dados, de acordo com o proposto em [1]:

- Classificação - uma tarefa preditiva, que consiste em determinar uma função que mapeie cada item de uma base dados a uma das classes previamente definidas. Um exemplo de classificação é identificação de objetos de interesse em grandes bases de imagens.
- Regressão - é uma tarefa preditiva, cujo objetivo é estimar o valor de uma variável com base nos valores de outras variáveis. Alguns dos exemplos de regressão são: prever o PIB de um país; estimar a probabilidade de um paciente sobreviver, dados os resultados de um conjunto de exames; e prever séries temporais.
- Detecção de desvios/anomalias - é uma tarefa preditiva, cujo objetivo é detectar itens que possuam características significativamente diferentes do comportamento normal do restante dos dados, como por exemplo detecção de fraudes em cartões de crédito.
- Sumarização - é uma tarefa descritiva, que consiste em definir um conjunto mínimo de características que seja capaz de identificar um subconjunto de objetos. As técnicas de sumarização são comumente aplicadas para análise exploratória de dados e geração de relatórios automatizados.
- Modelo de Dependência - descreve dependências significativas entre os atributos. Esses modelos existem em dois níveis: estruturado, que especifica (geralmente em forma de gráfico) quais variáveis são localmente dependentes; e quantitativo, que especifica o grau de dependência usando alguma escala numérica.
- Análise de Séries Temporais - modela características sequenciais, como dados que possuem dependências no tempo. O objetivo é modelar os estados do processo de geração da sequência, extraindo e relatando os desvios e tendências no tempo.

As tarefas de clusterização e regra de associação, por serem as escolhidas para aplicação no estudo de caso deste trabalho, são descritas mais detalhadamente nas seções 2.3 e 2.4, respectivamente.

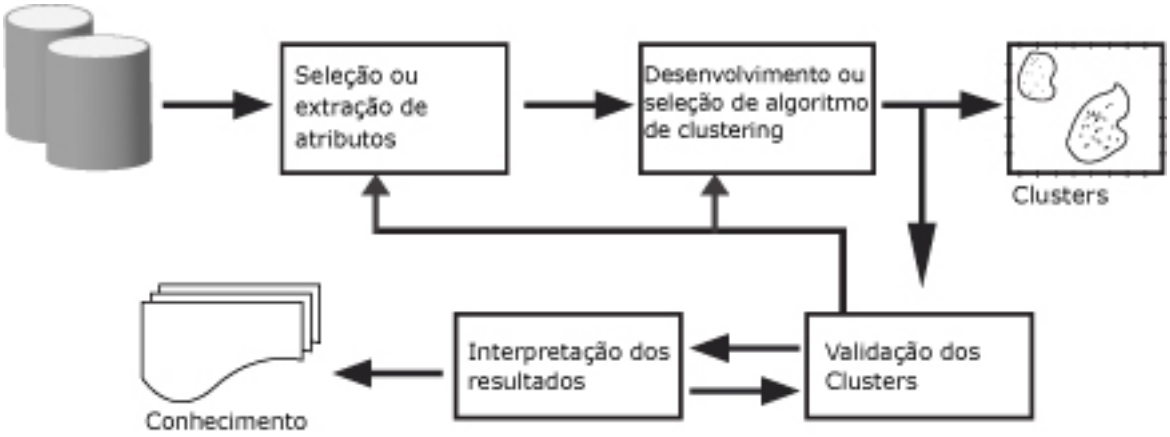

Figura 2: Principais fases da clusterização.

2.3 Clusterização

A **clusterização**, também conhecida como análise de *cluster*, é uma tarefa descritiva, cujo objetivo é encontrar grupos de objetos, tais que, objetos pertencentes a um grupo sejam similares entre si e diferentes de objetos que pertencem a outro grupo [4], [5], [6].

A tarefa de classificação pode ser dividida em classificação supervisionada e classificação não supervisionada. Em classificação supervisionada, dada uma coleção de objetos e suas classes, o problema é classificar um novo objeto para o qual a classe não é conhecida. Normalmente, os objetos classificados são divididos em um conjunto que é usado para aprender a descrição da classe (treinamento) e outro que é usado para testar a informação aprendida. No caso da clusterização (classificação não supervisionada), as classes não são previamente conhecidas, de forma que, o problema se torna classificar objetos por meio de alguma medida de similaridade de forma que os grupos tenham um significado relevante.

2.3.1 Principais fases da clusterização

De acordo com [11], a atividade de clusterização envolve os seguintes passos (figura 2):

1. *Seleção ou extração de atributos.* Seleção de atributos é o processo de identificação de atributos distintos de um conjunto de atributos candidatos, enquanto a extração de atributos por meio de transformações atributos gera novos atributos a partir dos originais.

2. *Medida de similaridade.* A medida de similaridade determina o quanto um objeto é similar a outro, possibilitando assim, a identificação de qual *cluster* o objeto deve pertencer. Essa medida deve ser específica para o domínio sob o qual está sendo a aplicada a

clusterização. Na seção 2.3.3 são apresentadas algumas das medidas que são comumente usadas.

3. *Algoritmo de clusterização.* Atualmente existe uma grande variedade de algoritmos de clusterização, mas não há um algoritmo que possa ser usado para resolver todo e qualquer problema. Dessa forma, é fundamental analisar o problema para selecionar ou desenvolver um algoritmo adequado. Na seção 2.3.2, são apresentados os tipos de algoritmos de clusterização.

4. *Validação dos* clusters. Todo algoritmo de clusterização consegue gerar partições, independente de essas possuírem significado semântico. E, além disso, algoritmos diferentes normalmente agrupam os objetos de forma diferente. Por isso, é necessário que os resultados obtidos pelos algoritmos sejam validados por meio de critérios de avaliação eficientes. Há três tipos de critérios de validação: os índices externos, nos quais, os resultados encontrados pelos algoritmos são comparados com uma estrutura pré-estabelecida; os índices internos, nos quais, os resultados obtidos pelos algoritmos são examinados a fim de determinar se eles são intrinsecamente apropriados para os dados de entrada; e os índices relativos, que comparam os resultados obtidos por algoritmos diferentes para decidir qual melhor representa as características dos dados.

5. *Interpretação dos resultados.* Por fim, o resultado dos algoritmos de clusterização deve ser interpretado por especialistas para que sejam atribuídos significados aos *clusters* de forma que o usuário possa compreendê-los.

Como visto na figura 2, a análise de *cluster* apresenta um fluxo para realimentação, pois por vezes são necessárias execuções com algoritmos de clusterização diferentes até que se obtenham bons resultados.

2.3.2 Classificação das técnicas de clusterização

De acordo com [4], [10], os principais tipos de técnicas de clusterização podem ser classificados em:

- *Técnicas Particionais*: o conjunto de *clusters* não possui intersecção de modo que cada objeto pertence a exatamente um *cluster*.

- *Técnicas Hierárquicas*: o conjunto de *clusters* está organizado e aninhado como uma árvore.

- *Técnicas baseadas em densidade*: são técnicas baseadas na noção de densidade, ou seja, um objeto pertence a uma região densa se na sua vizinhança existem pelo menos L objetos, onde L é um limiar definido pelo usuário. Nessas técnicas, diferentemente das outras, o número de *clusters* não precisa ser especificado.

2.3.3 Tipos de dados e medidas de similaridade

Nesta seção, são apresentados os tipos de dados que normalmente ocorrem em clusterização, como pré-processá-los para análise e algumas das medidas de similaridade mais usadas [10].

Supondo que um conjunto de dados para ser clusterizado contenha n objetos, os quais podem representar pessoas, casas, documentos, países e outros. Algoritmos de clusterização tipicamente operam sobre as seguintes estruturas de dados:

- Matriz de dados (ou objetos por atributos) - esta estrutura representa n objetos, tais como pessoas, com p atributos, tais como altura, peso, idade, sexo e outros. A estrutura está na forma de uma tabela relacional, ou matriz $n \times p$ (n objetos $\times$ p atributos) como mostrado na figura 3.

$$\begin{pmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{pmatrix}$$

Figura 3: Matriz de dados.

- Matriz de dissimilaridade (ou objetos por objetos) - esta estrutura armazena uma coleção de proximidades que são avaliadas para todos os pares de n objetos. Ela é muitas vezes representada por uma matriz $n \times n$, conforme pode ser observado na figura 4, na qual $d(i, j)$ é a diferença ou dissimilaridade medida entre os objetos i e j.

.

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{pmatrix}$$

Figura 4: Matriz de dissimilaridade.

2.3.3.1 Atributos Contínuos

Atributos contínuos são atributos que podem ser relacionados a funções matemáticas contínuas. Exemplos desse tipo de atributo incluem peso, altura, temperatura ambiente e outros. As medidas de dissimilaridade mais usadas para esse tipo de atributo são as distâncias Euclidiana, Manhattan e Minkowski [10].

A unidade de medida usada pode afetar a clusterização, para que isso não ocorra os dados devem ser normalizados, ou seja, dar a todos os atributos igual peso. Porém, há casos em que pode ser útil dar mais importância a um determinado conjunto de dados, por exemplo, quando clusterizando candidatos a jogadores de basquete, pode-se dar mais peso ao atributo altura.

Uma forma de normalizar medidas é convertendo as medidas originais em atributos sem unidade. Dados os valores medidos para um atributo f, a normalização pode ser feita da seguinte maneira:

- Calcular o desvio absoluto médio, sf:

$$sf = \frac{1}{n}(\mid x_{1f} - m_f \mid + \mid x_{2f} - m_f \mid + \cdots + \mid x_{nf} - m_f \mid), \tag{2.1}$$

  no qual $x_{1f}, \cdots, x_{nf}$ são n medidas de f, e m_f é o valor médio de f, ou seja,

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf}). \tag{2.2}$$

- Calcular a medida normalizada (escore-z):

$$z_{if} = \frac{x_{if} - m_f}{s_f}. \tag{2.3}$$

Há outras medidas de dispersão robustas, tal como desvio absoluto mediano, mas a vantagem de usar desvio absoluto médio é a facilidade de detectar *outliers*, pois os escores-z de

outliers não se tornam muito pequenos.

Após a normalização (caso seja necessária), deve ser estabelecida uma medida de dissimilaridade. Uma das medidas mais usadas é a distância Euclidiana, que pode ser definida como:

$$d(i,j) = \sqrt{(\mid x_{i1} - x_{j1} \mid^2 + \mid x_{i2} - x_{j2} \mid^2 + \cdots + \mid x_{ip} - x_{jp} \mid^2)}, \tag{2.4}$$

na qual $i = (x_{i1}, x_{i2}, ..., x_{ip})$ e $j = (x_{j1}, x_{j2}, ..., x_{jp})$ são dois objetos de dados p-dimensional.

Outra métrica bem conhecida é a distancia Manhattan, que pode ser defina como:

$$d(i,j) = \mid x_{i1} - x_{j1} \mid + \mid x_{i2} - x_{j2} \mid + \cdots + \mid x_{ip} - x_{jp} \mid \ . \tag{2.5}$$

A distância Minkowski é uma generalização das distâncias Euclidiana e Manhattan:

$$d(i,j) = (\mid x_{i1} - x_{j1} \mid^q + \mid x_{i2} - x_{j2} \mid^q + \cdots + \mid x_{ip} - x_{jp} \mid^q)^{1/q}, \tag{2.6}$$

na qual q é um inteiro positivo. A distância Minkowski representa a distância Manhattan quando $q = 1$ e a distância Euclidiana quando $q = 2$.

Se a cada atributo é atribuído um peso de acordo com sua importância, a distância Euclidiana ponderada, por exemplo, pode ser calculada da seguinte maneira:

$$d(i,j) = \sqrt{w_1 \mid x_{i1} - x_{j1} \mid^q + w_2 \mid x_{i2} - x_{j2} \mid^q + \cdots + \mid x_{ip} - x_{jp} \mid^2}. \tag{2.7}$$

O cálculo de medidas levando em consideração pesos pode ser realizada utilizando as demais distâncias também, de forma semelhante a mostrada para distancia Euclidiana na equação 2.7.

2.3.3.2 Atributos Binários

Uma abordagem para medir a dissimilaridade entre dois atributos binários é calcular a matriz de dissimilaridade dos dados. Se todos os atributos binários possuem os mesmos pesos, tem-se a tabela 1, na qual a é o número de atributos iguais a 1 para ambos os objetos i e j, b é o número de atributos que é igual a 1 para i e igual 0 para j, c é o número de atributos que é igual 0 para i e igual 1 para j, d é o número de atributos iguais a 0 para ambos os objetos i e j, e p é o número total de atributos $(p = a + b + c + d)$.

Um atributo binário pode ser simétrico ou assimétrico. Ele é simétrico quando os dois estados possuem o mesmo valor e peso, ou seja, não existe preferência para atribuir 0 ou 1

Tabela 1: Tabela de contingência para atributos binários.

		objeto j		
		1	0	Soma
objeto i	1	a	b	$a+b$
	0	c	d	$c+d$
	Soma	$a+c$	$b+d$	p

ao atributo. Um exemplo é o atributo sexo. A similaridade baseada em atributos binários simétricos é chamada similaridade invariante, pois o resultado não muda quando alguns ou todos os atributos binários possuem valores diferentes. Para similaridades invariantes, o mais conhecido coeficiente é o coeficiente de casamento simples (*simple matching coefficient*), que pode ser definido como:

$$d(i,j) = \frac{b+c}{a+b+c+d}. \tag{2.8}$$

Um atributo binário é assimétrico se os estados não são igualmente importantes, tal como o resultado positivo ou negativo de um teste para determinar a presença de uma dada doença. Dados dois atributos binários assimétricos, a ocorrência de dois 1s (casamento positivo) é considerado mais importante que dois 0s (casamento negativo). A similaridade baseada em atributos binários assimétricos é chamada similaridade não invariante. Para similaridades não invariantes, o mais conhecido coeficiente é o Coeficiente de Jaccard, no qual o número de casamentos negativos, d, não é considerado importante, portanto pode ser ignorado no cálculo do coeficiente. O Coeficiente de Jaccard pode ser calculado da seguinte maneira:

$$d(i,j) = \frac{b+c}{a+b+c}. \tag{2.9}$$

2.3.3.3 Atributos Nominais

Um atributo nominal é uma generalização de atributos binários, na qual ele pode possuir mais do que dois estados. Por exemplo, um atributo que mapeia a cor dos olhos de uma pessoa pode assumir os valores: verde, castanho e azul.

Os estados um atributo nominal podem ser denotados por letras, símbolos ou um conjunto de inteiros, mas os inteiros não representam nenhuma relação de ordem, apenas servem para manipular os dados.

A dissimilaridade entre dois objetos i e j pode ser calculada usando a abordagem de casamento simples (*simple matching*) como mostrado a seguir:

$$d(i,j) = \frac{p-m}{p}, \tag{2.10}$$

na qual m é o número de casamentos (ou seja, número de atributos nos quais i e j são iguais) e p é o número total de atributos.

Atributos nominais podem ser codificados como atributos binários assimétricos por meio da criação de um atributo binário para cada um dos estados que o atributo nominal possua. Para cada atributo nominal do objeto que pertença a um estado, o atributo binário correspondente a este estado é mapeado para 1 enquanto o restante dos atributos binários são mapeados para 0. Desta maneira os coeficientes de cálculo de dissimilaridade apresentados para atributos binários podem ser utilizados.

2.3.3.4 Atributos Ordinais

Um atributo ordinal discreto lembra um atributo nominal, exceto pelo fato que os M estados de um atributo ordinal estão ordenados numa sequência significativa. Um exemplo de atributo ordinal é um atributo cujos valores estão na escala Likert, que é uma escala muito usada em pesquisas de opinião, ou seja, os atributos podem assumir valores como: muito bom, bom, razoável, ruim e muito ruim. Atributos ordinais podem ser obtidos pela discretização de valores de atributos contínuos por meio da divisão da faixa de valores em um número finito de classes.

Supondo que f é um conjunto de atributos ordinais descrevendo n objetos. O cálculo da dissimilaridade de f envolve os seguintes passos:

- O valor de f para o i-ésimo objeto é x_{if}, f possui M_f estados ordenados, representando o ranking $1, ..., M_f$. Substitui-se cada x_{if} por seu elemento correspondente no ranking, r_{if} pertence $1, ..., M_f$.
- Como cada atributo ordinal pode ter um número diferente de estados, é necessário mapear a faixa de cada atributo em $[0 - 1]$ para que cada atributo possua o mesmo peso. Isso pode ser feito substituindo o valor de r_{if} do i-ésimo objeto no f-ésimo atributo por:
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}. \tag{2.11}$$
- A dissimilaridade pode ser calculada usando qualquer uma das formas apresentadas para atributos contínuos, usando z_{if} para representar o valor do atributo f para o objeto i.

2.3.3.5 Atributos em escala não linear

Atributos em escala não linear são atributos que expressam uma medida em uma escala não linear, como uma escala exponencial, por exemplo. Há três maneiras de calcular dissimilaridade

para objetos que possuem esse tipo de atributo:

- Tratar atributos em escala não linear como atributos em escala linear. Porém usualmente essa não é um boa escolha uma vez que a escala pode ser distorcida.
- Aplicar transformações logarítmicas a um atributo em escala não linear f tendo valor x_{if} para o objeto i por meio da fórmula $y_{if} = \log(x_{if})$. Os valores de y_{if} podem ser tratados como atributos em escala linear.
- Tratar x_{if} como atributo ordinal e sua posição no raking como atributo contínuo.

2.3.3.6 Objetos formados por atributos de vários tipos

É muito comum que objetos sejam descritos por atributos de vários tipos. Surge, portanto, a necessidade de uma forma de medir dissimilaridade para tais objetos.

Uma abordagem é juntar todos os atributos de um determinado tipo e realizar a clusterização para cada tipo de atributo. Outra abordagem é mapear todos os atributos para o intervalo $[0; 1]$ e usar medidas como distância Euclidiana. Há várias outras abordagens, porém uma das mais poderosas é [12]:

$$d(i,j) = \frac{\sum_{f=1}^{p}, \delta_{ij}^{f} d_{ij}^{f}}{\sum_{f=1}^{p} \delta_{ij}^{f}}, \tag{2.12}$$

na qual o indicador $\delta_{ij}^{f} = 0$ se (1) x_{if} ou x_{jf} está faltando (quando o valor do atributo f está faltando para o objeto i ou j, por exemplo), ou (2) $x_{if} = x_{jf} = 0$ e o atributo f é binário assimétrico; caso contrário, $\delta_{ij}^{f} = 1$. A contribuição do atributo f para dissimilaridade entre os objetos i e j, d_{ij}^{f}, é calculada d seguinte maneira:

1. Se f é binário ou nominal: $d_{ij}^{f} = 0$, se $x_{if} = x_{fi}$; caso contrário, $d_{ij}^{f} = 1$.
2. Se f é contínuo: $d_{ij}^{f} = \frac{|x_{if} - x_{jf}|}{\max_h(x_{hf}) - \min_h(x_{hf})}$, no qual h executa sobre todos os objetos que possuem valor para o atributo f.
3. Se f é ordinal ou não escalar: calcula-se o ranking r_{if} e $z_{if} = \frac{r_{if}-1}{M_f-1}$, e trata z_{if} como atributo contínuo.

2.3.4 Algoritmo *K-means*

O algoritmo ***k-means*** é um algoritmo particional (os *clusters* são disjuntos), exclusivo (cada objeto pertence a um único *cluster*) e baseado em protótipos (cada objeto do *cluster* está

mais próximo ao protótipo que define o *cluster* do que dos protótipos de quaisquer outros *clusters*). O seu funcionamento pode ser descrito pelos seguintes passos [11]:

1. *Atribuir valores iniciais para os centróides (protótipos).* Nesse passo, são escolhidos os k objetos dentro do banco de dados que serão utilizados como centros dos *clusters* (centróides). Essa escolha pode ser feita de diversas maneiras, dentre elas: selecionar as k primeiras entradas; ou selecionar k entradas aleatoriamente.

2. *Associar objetos aos centróides.* Nesse passo, cada objeto é associado, de acordo com a medida de similaridade, ao centróide mais próximo.

3. *Recalcular centróides.* Nesse passo, para cada *cluster* é recalculado o calor do centróide a partir da média dos objetos pertencentes ao *cluster*.

4. *Iteração.* O algoritmo repete os passos 2 e 3 até que não haja mudança nos centróides ou até que relativamente ocorram poucas mudanças nos centróides.

Para avaliação dos *clusters* criados pelo algoritmo k-*means* a medida mais comumente usada é a soma dos erros quadrados (*Sum of Square Error - SSE*), que pode ser calculada de acordo com a seguinte equação:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x), \tag{2.13}$$

na qual, x representa um objeto pertencente ao *cluster* C_i, m_i representa o centróide do *cluster* i e k é o número de *clusters*. Uma forma de reduzir o SSE é aumentando o número de *clusters*.

2.4 Regras de Associação

A técnica de regras de associação é uma das tarefas descritivas da Mineração de Dados, que demonstra o quanto a ocorrência de um conjunto de itens implica na ocorrência de algum outro conjunto distinto de itens nos registros de uma mesma base de dados [3]. Desse modo, o objetivo das regras de associação é encontrar itens que ocorrem de forma simultânea e frequente em transações de grandes bases de dados, facilitando a compreensão do comportamento dos dados.

A aplicação mais comum de regras de associação é a análise de transações de compras, por isso, ao longo dessa seção será utilizado um exemplo de uma pequena base de dados

que armazena compras realizadas por clientes de um supermercado (figura 5). Como pode ser observado na figura 5, cada transação da base de dados armazena a relação de produtos adquiridos por um cliente específico.

Transações	Produtos Comprados
1	biscoito, manteiga, café, arroz
2	manteiga, feijão, ovos, pão, queijo
3	café, leite, ovos, pão
4	leite, café, manteiga, feijão, pão, arroz
5	leite, café, pão, queijo

Figura 5: Base de dados com transações de clientes.

2.4.1 Conceitos e Definições

Nessa seção são apresentados conceitos e definições, que são necessários à compreensão do processo de mineração de regras de associação [13].

O pré-requisito para que a técnica de regras de associação possa ser aplicada é que a base de dados esteja no formato de uma **tabela booleana de itens-transações** (figura 6) ou a uma **tabela de itens-transações** (figura 7). A tabela de itens-transações é um caso particular da tabela booleana de itens-transações, onde apenas os itens que possuem valor um na tabela booleana de itens-transações aparecem na tabela de itens-transações. Dessa forma, quando os dados não estão nos formatos apresentados, deve ser realizado o pré-processamento dos dados.

	a_1	a_2	$\cdots$	a_m
t_1	1	1	$\cdots$	1
t_2	0	1	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_n	1	0	$\ldots$	1

Figura 6: Tabela booleana de itens-transações.

t_1	a_1	a_2	$\cdots$	a_m
t_2		a_2	$\cdots$	a_m
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_n	a_1		$\ldots$	a_m

Figura 7: Tabela de itens-transações.

Uma regra de associação pode ser representada como uma implicação na forma $\boldsymbol{LHS \Rightarrow RHS}$, onde LHS e RHS são conjuntos disjuntos de itens que representam respectivamente,

o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra. Um exemplo de regra que poderia ser extraída da base de dados da figura 5 é {café, leite} ⇒ {pão}, cujo significado é que clientes que compraram café e leite tendem a comprar pão também.

Ao conjunto de atributos ou itens ordenados lexicograficamente dá-se o nome de ***itemset***. Um *itemset* com k elementos costuma ser referenciado como ***k-itemset***. Um exemplo de 2-*itemset* é {café, leite}. O **suporte** de um *itemset* Z, $sup(Z)$, indica a porcentagem de transações da base de dados que contêm os itens de Z, ou seja,

$$sup(Z) = \frac{n(Z)}{N} \cdot 100, \tag{2.14}$$

onde $n(Z)$ é o número de transações nas quais Z ocorre e N é o número total de transações da base de dados. Por exemplo, o suporte do *itemset* {café,leite} é de 60% para a base de dados mostrada na figura 5.

Já o suporte de uma regra $LHS \Rightarrow RHS$ indica a frequência com que $LHS \cup RHS$ ocorre no conjunto de dados, ou seja,

$$sup(LHS \Rightarrow RHS) = sup(LHS \cup RHS) = \frac{n(LHS \cup RHS)}{N} \cdot 100, \tag{2.15}$$

onde $n(LHS \cup RHS)$ é o número de transações nas quais LHS e RHS ocorrem juntas e N é o número total de transações da base de dados. No exemplo 1, é mostrado como calcular o suporte da regra {café, leite} ⇒ {pão} para a base de dados apresentada na figura 5.

Exemplo 1 $sup(\{café, leite\} \Rightarrow \{pão\}) = \frac{n(\{café, leite\} \cup \{pão\})}{5} \cdot 100 = \frac{3}{5} \cdot 100 = 60\%.$

A **confiança** de uma regra $LHS \Rightarrow RHS$, $conf(LHS \Rightarrow RHS)$, representa, dentre as transações que contêm LHS, a porcentagem de transações que também contêm RHS, ou seja,

$$conf(LHS \Rightarrow RHS) = sup(LHS \cup RHS)/sup(LHS) \cdot 100. \tag{2.16}$$

No exemplo 2, é calculada a confiança da regra {café, leite} ⇒ {pão} para a base apresentada na figura 5.

Exemplo 2 $conf(café, leite \Rightarrow pão) = \frac{sup(café, leite \Rightarrow pão)}{sup(café, leite)} \cdot 100 = \frac{60}{60} \cdot 100 = 100\%.$

Um k-*itemset* é dito **frequente** quando o seu suporte é maior ou igual ao suporte mínimo definido pelo usuário.

De acordo com [3], a técnica de regras de associação pode ser descrita formalmente da seguinte maneira:

Seja $I = \{i_1, i_2, ..., i_n\}$ o conjunto de itens que compõem uma base de dados D e T o conjunto de transações da mesma base de dados, cada transação $t_i \in T$ é composta por um conjunto de itens tal que $t_i \subseteq I$. A regra de associação é uma implicação na forma $LHS \Rightarrow RHS$, onde $LHS \subset I$, $RHS \subset I$ e $LHS \cap RHS = \emptyset$.

Dessa forma, o processo de obtenção de regras de associação pode ser dividido em duas etapas:

1. Determinar todos os k-*itemsets* frequentes.
2. Para cada k-*itemset* encontrado na etapa anterior, com $k >= 2$, gerar regras de associação (permutações de subconjuntos) que possuam confiança maior ou igual à confiança mínima especificada pelo usuário.

Para determinar os k-*itemsets* freqüentes o algoritmo comumente utilizado é o *Apriori*, que será apresentado na próxima seção.

2.4.2 Algoritmo *Apriori*

O algoritmo *Apriori*, desenvolvido por [3], é capaz de encontrar todos os itemsets frequentes em uma base de dados. Esse algoritmo considera a seguinte propriedade para diminuir o espaço de busca:

Propriedade 1 *Se um* itemset Z *não é frequente então para todo* itemset A, $Z \cup A$ *não será frequente.*

Inicialmente o algoritmo conta a ocorrência dos itens, determinando o 1-*itemsets* frequentes que são armazenados em L_1, onde L_1 representa o conjunto de 1-*itemsets* frequentes. Depois de forma iterativa, para encontrar o L_k, o algoritmo constrói um conjunto de k-*itemsets* candidatos, C_k, através de um *join* entre os elementos do L_{k-1}. A seguir, ele poda C_k usando a propriedade 1 e calcula o suporte dos candidatos que não foram podados. E por fim, são identificados os k-*itemsets* frequentes, que são armazenados em L_k. O conjunto que contém todos os *itemsets* frequentes é formado pela união dos conjuntos L_k de k-*itemsets* frequentes.

A seguir é apresentada, no exemplo 3, a utilização do algoritmo *Apriori* para obtenção de regras de associação.

Exemplo 3 *Aplicação do algoritmo Apriori para obtenção de regras de associação na base de dados da figura 5, considerando o suporte mínimo 50% e a confiança mínima 90%.*

1. Determinando todos os k-itensets frequentes;

   - Gera os candidatos a 1-itensets frequentes, C_1:

$\boldsymbol{sup(\{leite\}) = \frac{3}{5} \cdot 100 = 60\%}$
$\boldsymbol{sup(\{café\}) = \frac{4}{5} \cdot 100 = 80\%}$
$\boldsymbol{sup(\{pão\}) = \frac{4}{5} \cdot 100 = 80\%}$
$sup(\{biscoito\}) = \frac{1}{5} \cdot 100 = 20\%$
$sup(\{manteiga\}) = \frac{2}{5} \cdot 100 = 40\%$
$sup(\{queijo\}) = \frac{2}{5} \cdot 100 = 40\%$
$sup(\{ovos\}) = \frac{2}{5} \cdot 100 = 40\%$
$sup(\{arroz\}) = \frac{2}{5} \cdot 100 = 40\%$
$sup(\{feijão\}) = \frac{2}{5} \cdot 100 = 40\%$

   Logo, $L_1 = \{\{café\}, \{leite\}, \{pão\}\}$.

   - Gera os candidatos a 2-itensets frequentes, C_2:

$\boldsymbol{sup(\{café, leite\}) = \frac{3}{5} \cdot 100 = 60\%}$
$\boldsymbol{sup(\{café, pão\}) = \frac{3}{5} \cdot 100 = 60\%}$
$\boldsymbol{sup(\{leite, pão\}) = \frac{3}{5} \cdot 100 = 60\%}$

   Logo, $L_2 = \{\{café, leite\}, \{café, pão\}, \{leite, pão\}\}$.

   - Gera os candidatos a 3-itensets frequentes, C_3:

$\boldsymbol{sup(\{café, leite, pão\}) = \frac{3}{5} \cdot 100 = 60\%}$

   Logo, $L_3 = \{\{café, leite, pão\}\}$.

   Como L_3 só possui um itemset o algoritmo para de iterar, pois não é possível gerar candidatos a 4-itemsets.

2. Gerando regras de associação

   Para gerar as regras deve-se permutar os k-itemsets frequentes (k>=2) e selecionar as regras que possuem confiança maior ou igual a confiança mínima especificada pelo usuário. As regras de associação que obedecem as especificações são as que estão em negrito na tabela da figura 8.

REGRA DE ASSOCIAÇÃO	SUPORTE	CONFIANÇA
{leite} => {café}	**60%**	**100%**
{café} => {leite}	60%	75%
{leite} => {pão}	**60%**	**100%**
{pão} = {leite}	60%	75%
{café} => {pão}	60%	75%
{pão} => {café}	60%	75%
{café, leite} => {pão}	**60%**	**100%**
{leite, pão} => {café}	**60%**	**100%**
{café, pão} => {leite}	**60%**	**100%**

Figura 8: Regra de Associação.

3 Plataforma Lattes e Ferramentas

Neste capítulo, são apresentadas a plataforma Lattes e algumas ferramentas que foram utilizadas na realização deste trabalho. A seção 3.1 apresenta a plataforma Lattes e seus componentes. A seguir, na seção 3.2, é apresentada a estrutura do sistema de currículos Lattes. Na seção 3.3, é apresentada a ferramenta utilizada apra realizar as tarefas de Mineração de Dados. Por fim, na seção 3.4, é apresentada uma ferramenta de extração e estruturação dos dados obtidos a partir da plataforma Lattes.

3.1 Plataforma Lattes

A Plataforma Lattes é um sistema de informação desenvolvido pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para auxiliar a gestão de ciência, tecnologia e inovação no Brasil [14]. Essa foi lançada em 16 de agosto de 1999, com a versão inicial do sistema de currículos Lattes.

A plataforma é composta pela integração de quatro sistemas distintos: currículo Lattes, que é um sistema de informação responsável por registrar a vida curricular pregressa e atual dos pesquisadores; diretório de grupos de pesquisa, que é um sistema responsável por manter informações sobre os grupos de pesquisa existentes no país; diretório de instituições, cujo objetivo é armazenar informações sobre os institutos de pesquisa, universidades e outros, que demandam fomento ao CNPq; e sistema gerencial de fomento, cujo objetivo é aumentar a qualidade das atividades de fomento do CNPq.

3.2 Estrutura do currículo Lattes

O currículo Lattes está estruturado de forma hierárquica, os níveis mais altos da hierarquia são:

- **Apresentação:** módulo inicial do sistema, é composto por um resumo do currículo do

usuário e a data da última atualização do currículo;

- **Dados Gerais:** este módulo agrupa dados de identificação, endereços, formação acadêmica e complementar, atuação profissional, áreas de atuação e outros;
- **Produção bibliográfica:** concentra toda a produção bibliográfica tais como artigos completos, livros, textos em jornais e outros;
- **Produção técnica:** agrupa informações sobre a produção técnica do pesquisador tais como softwares, produtos, trabalhos técnicos e outros;
- **Orientações:** módulo destinado a todas as orientações ou supervisões (concluídas ou em andamento);
- **Projetos:** neste módulo são encontrados os projetos do pesquisador;
- **Eventos:** contém informações relacionadas a eventos que o pesquisador organizou ou paticipou;
- **Bancas:** contém informações relacionadas a bancas e comissões julgadoras;
- **Citações:** reune indicadores de referências de outros pesquisadores aos trabalhos publicados do pesquisador.

3.3 Ferramenta de Mineração de Dados

Para realização das tarefas de Mineração de Dados neste trabalho foi utilizada uma ferramenta de código aberto chamada WEKA (*Waikato Environment for Knowledge Analysis*). Essa ferramenta foi desenvolvida na Universidade de Waikato na Nova Zelândia e possui implementações de vários algoritmos de Mineração de Dados [9].

A interface gráfica do WEKA é amigável, porém é importante destacar os principais elementos. A Figura 9 apresenta a tela de pré-processamento da ferramenta destacando os principais elementos:

- (a) Esse botão permite a seleção de bases de dados no formato ARFF (*Attribute-Relation File Format*);
- (b) Nessa área podem ser selecionados algoritmos para pré-processar os dados (discretizar atributos, por exemplo);

- (c) Nessa área são apresentados os atributos da base de dados;
- (d) Apresenta informações quantitativas e estatísticas sobre o atributo selecionado na área *Atribute*.

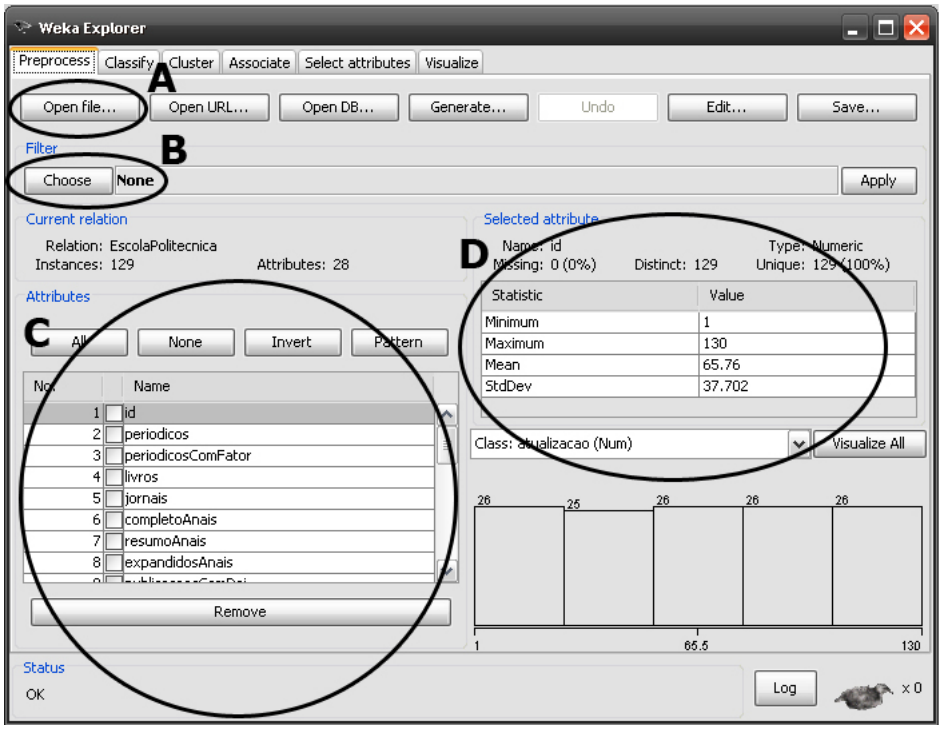

Figura 9: Tela de pré-processamento da ferramenta WEKA.

A ferramenta WEKA recebe como entrada arquivos no formato ARFF, que são compostos por 3 elementos (Figura 10): (a) *relation*, que define um nome para a relação estudada; (b) *atribute*, onde são especificados os atributos que compõem a base de dados; e (c) *data*, que contempla os dados separados por vírgulas [15].

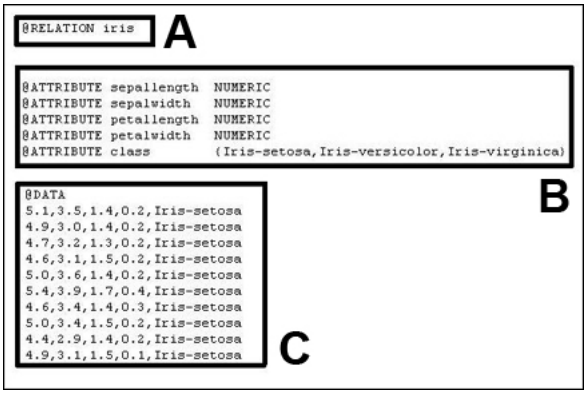

Figura 10: Exemplo de arquivo ARFF.

Neste trabalho são estudados clusterização e regras de associação, por isso é necessário

compreender como o WEKA apresenta os resultados dessas tarefas. A Figura 11 apresenta um exemplo de resultados do WEKA para clusterização, destacando-se os principais elementos:

- (a) Apresenta os centróides de cada *cluster* e dos dados completos;
- (b) Apresenta a distribuição dos objetos entre os *clusters* de acordo com o atributo classe escolhido.

Cluster centroids:

Attribute	Full Data (129)	Cluster# 0 (115)	1 (14)
periodicos	2.2403	1.3304	9.7143
periodicosComFator	0.7984	0.5043	3.2143
livros	0.5581	0.3478	2.2857
jornais	0.6899	0.2957	3.9286
completoAnais	15.3566	8.5043	71.6429
resumoAnais	3.6822	1.4696	21.8571
expandidosAnais	0.6589	0.3565	3.1429
publicacoesComDoi	1.0775	0.4435	6.2857
publicacoes	23.186	12.3043	112.5714
oriMSC	1.7907	0.3478	13.6429
oriPHD	0.1938	0.0261	1.5714
oriESP	1.6202	0.713	9.0714
oriTCC	4.5736	2.5652	21.0714
oriIC	4.0543	1.6348	23.9286
oriOutras	2.4419	1.3565	11.3571
orientacoes	14.6744	6.6435	80.6429
bancaEX	10.6744	5.5304	52.9286
bancaJUL	1.7287	0.9478	8.1429
ProjetoCnpq	0.3178	0.1739	1.5
ProjetoFacepe	0.155	0.0609	0.9286
ProjetoFinep	0.0853	0.0087	0.7143
atvDirecao	1.6357	1.3043	4.3571
atvEnsino	4.4186	3.4783	12.1429
atvProjeto	2.0388	1.1217	9.5714
atvPesquisa	1.124	0.7826	3.9286
atualizacao	758.8217	841.2087	82.0714

A

B

Class attribute: departamento
Classes to Clusters:

0	1	<-- assigned to cluster
31	3	basico
11	2	computacao
22	8	civil
30	1	eletrica
21	0	mecanica

Figura 11: Exemplo de resultados de clusterização.

A Figura 12 apresenta um exemplo de resultados gerados pelo WEKA quando a tarefa de regras de associação é realizada, destacando-se os principais elementos:

- (a) As regras são ordenadas pela confiança.
- (b) Os valores depois de antecedentes e consequentes das regras representam o número de suas respectivas ocorrências.

3.4 Ferramenta de extração de dados da plataforma Lattes

As técnicas de Mineração de Dados selecionadas para serem estudadas neste trabalho devem ser aplicadas sobre dados estruturados, como na internet os dados estão na forma textual

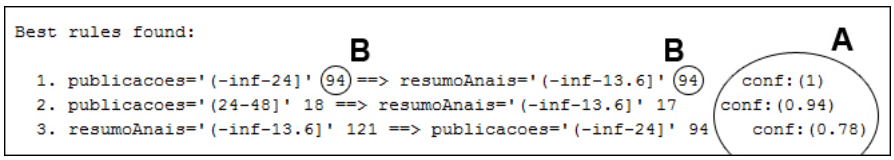

Figura 12: Exemplo de resultados de regras de associação.

houve a necessidade da construção de uma ferramenta para extração e estruturação dos dados da Plataforma Lattes. Portanto, foi desenvolvida uma ferramenta que obtém os dados dos currículos e transforma-os em um banco de dados. Nesta seção é descrito o funcionamento dessa ferramenta e a forma como os dados foram organizados após a extração.

A ferramenta foi desenvolvida utilizando PHP (*Hypertext Preprocessor*) [16], que é uma linguagem de programação muito usada em páginas dinâmicas da web, e o sistema de gerenciamento de banco de dados relacional MYSQL [17]. Os principais dados selecionados dos currículos dos pesquisadores são:

- *id* - Atributo que identifica individualmente cada professor;
- *periódicos* - Número de artigos completos publicados em periódicos;
- *periodicosComFator* - número de artigos completos publicados em periódicos que possuem fator de impacto. O fator de impacto é uma avaliação feita pela JCR (*Journal Citation Reports*) [18] para medir o desempenho de um jornal com relação a outros da mesma área;
- *livros* - número de capítulos de livro publicados;
- *jornais* - número de textos em jornais de noticiais/revistas;
- *completoAnais* - número de trabalhos completos publicados em anais de congressos;
- *resumoAnais* - número de resumos publicados em anais de congresso;
- *expandidosAnais* - número de resumos expandidos publicados em anais de congresso;
- *publicaçõesComDoi* - número de publicações que possuem DOI (*Digital Object Identifier*), que permite localizar e acessar materiais na web - especialmente, publicações em periódicos e obras protegidas por copyright;
- *publicações* - atributo projetado a partir da soma de *periódicos*, *livros*, *jornais*, *completosAnais*, *resumoAnais* e *expandidosAnais*;

- *oriMSC* - número de orientações de mestrado;
- *oriPHD* - número de orientações de doutorado;
- *oriESP* - número de orientações de especialização;
- *oriTCC* - número de orientações de trabalho de conclusão de curso;
- *oriIC* - número de orientações de iniciação científica;
- *oriOutras* - número de outras orientações;
- *orientacoes* - atributo projetado a partir da soma de *oriMSC*, *oriPHD*, *oriESP*, *oriTCC*, *oriIC* e *oriOutras*;
- *bancaEX* - número de participações em bancas examinadoras;
- *bancaJUL* - número de participações em bancas julgadoras;
- *projetoCnpq* - número de projetos financiados pelo CNPQ;
- *projetoFacepe* - número de projetos financiados pela FACEPE(Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco);
- *projetoFinep* - número de projetos financiados pelo FINEP(Financiadora de Estudos e Projetos);
- *atvDirecao* - número de atividades de direção;
- *atvEnsino* - número de atividades de ensino;
- *atvProjeto* - número de atividades de projeto;
- *atvPesquisa* - número de atividades de pesquisa;
- *departamento* - departamento ao qual o professor pertence (básico, computação, civil, elétrica, mecânica);
- *atualização* - número de dias decorridos desde a última atualização.

A arquitetura da ferramenta é composta por 3 componentes (figura 13):

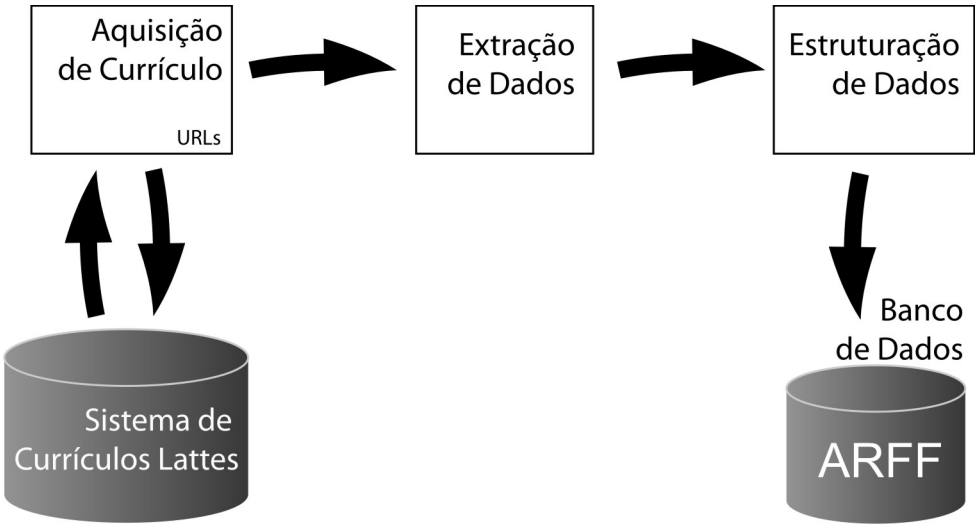

Figura 13: Arquitetura da ferramenta de extração.

- *Aquisição de currículo*. A tarefa inicial realizada pela ferramenta é a obtenção do conteúdo dos currículos Lattes dos professores pesquisadores da Escola Politécnica de Pernambuco no formato HTML (*HyperText Markup Language*) utilizando o endereço eletrônico (URL - *Uniform Resource Locator*) dos currículos.

- *Extração de dados*. As informações necessárias à criação da base de dados são extraídas do texto HTML, por meio de funções PHP que reconhecem expressões regulares (padrões) em *strings*. O texto semi-estruturado em HTML facilitou o estabelecimento dos padrões, uma vez que as *tags* puderam ser utilizadas como delimitadores para identificação dos dados de interesse.

- *Estruturação dos dados*. Por fim, os dados extraídos são armazenados em um arquivo ARFF (figura 14).

```
@relation EscolaPolitecnica
@attribute id numeric
@attribute periodicos numeric
@attribute periodicosComFator numeric
@attribute livros numeric
@attribute jornais numeric
@attribute completoAnais numeric
@attribute resumoAnais numeric
@attribute expandidosAnais numeric
@attribute publicacoesComDoi numeric
@attribute publicacoes numeric
@attribute oriMSC numeric
@attribute oriPHD numeric
@attribute oriESP numeric
@attribute oriTCC numeric
@attribute oriIC numeric
@attribute oriOutras numeric
@attribute orientacoes numeric
@attribute bancaEX numeric
@attribute bancaJUL numeric
@attribute ProjetoCnpq numeric
@attribute ProjetoFacepe numeric
@attribute ProjetoFinep numeric
@attribute atvDirecao numeric
@attribute atvEnsino numeric
@attribute atvProjeto numeric
@attribute atvPesquisa numeric
@attribute departamento {basico,computacao,civil,eletrica,mecanica}
@attribute atualizacao numeric
@data

6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,civil,819
1,0,0,0,0,0,1,0,0,1,0,0,3,0,0,0,3,0,0,0,0,0,1,2,0,0,civil,5
2,12,0,2,25,71,5,0,0,115,5,0,9,9,10,6,39,30,8,0,0,0,6,7,11,6,civil,226
3,6,0,5,3,114,1,0,0,129,17,1,37,13,14,0,82,22,4,0,3,0,5,4,7,2,civil,451
```

Figura 14: Base de dados no formato ARFF.

4 Estudo de Caso

Neste capítulo é apresentado como foram realizadas as tarefas de clusterização e regras de associação, bem como a análise dos resultados provenientes dessas tarefas. Na seção 4.1, são apresentadas características gerais sobre o estudo. A seguir, na seção 4.2, são apresentados os experimentos de clusterização. E por fim, na seção 4.3, são apresentados os experimentos realizados usando regras de associação.

4.1 Características gerais do estudo

Os experimentos foram realizados com 129 professores, do quadro efetivo da Escola Politécnica de Pernambuco (POLI) e estão divididos em 5 grupos: básico, mecânica, elétrica, civil e computação. Os professores estão distribuídos da seguinte maneira: 34 no básico, 21 em mecânica, 31 em elétrica, 30 em civil e 13 em computação.

Os nomes e cursos dos professores foram obtidos junto ao setor de recursos humanos da POLI, de forma que a partir dos nomes dos professores, os endereços eletrônicos dos seus currículos Lattes puderam ser adquiridos e utilizados como entrada para a ferramenta de extração apresentada no capítulo 3.

Os dados dos currículos Lattes dos professores foram extraídos no dia 28 de outubro de 2010. Para realizar a Mineração de Dados foi utilizada uma ferramenta WEKA apresentada no capítulo 4.

4.2 Clusterização

A tarefa de clusterização foi executada com várias configurações, mas em todas foi utilizado o algoritmo *K-means* juntamente com a distância Euclidiana, para medir a similaridade entre os objetos.

O primeiro experimento foi realizado com 2 *clusters*. Nesse caso, o algoritmo separou os

professores que possuem boa produção científica (*cluster* 1 - 14 professores) dos que possuem pouca produção científica (*cluster* 0 - 115 professores), conforme pode ser observado na figura 15. Uma característica que pode ser observada é que os professores pertencentes ao *cluster* 1 em média atualizaram seus currículos há menos de 3 meses, já os professores pertencentes ao *cluster* 0 em média atualizaram seus currículos há mais de 2 anos. Outra característica que pode ser observada é que embora o *cluster* 1 tenha agrupado professores com número de publicações elevadas (em média 113 publicações), poucas publicações possuem DOI (em média 6 publicações).

Cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(129)	(115)	(14)
periodicos	2.2403	1.3304	9.7143
periodicosComFator	0.7984	0.5043	3.2143
livros	0.5581	0.3478	2.2857
jornais	0.6899	0.2957	3.9286
completoAnais	15.3566	8.5043	71.6429
resumoAnais	3.6822	1.4696	21.8571
expandidosAnais	0.6589	0.3565	3.1429
publicacoesComDoi	1.0775	0.4435	6.2857
publicacoes	23.186	12.3043	112.5714
oriMSC	1.7907	0.3478	13.6429
oriPHD	0.1938	0.0261	1.5714
oriESP	1.6202	0.713	9.0714
oriTCC	4.5736	2.5652	21.0714
oriIC	4.0543	1.6348	23.9286
oriOutras	2.4419	1.3565	11.3571
orientacoes	14.6744	6.6435	80.6429
bancaEX	10.6744	5.5304	52.9286
bancaJUL	1.7287	0.9478	8.1429
ProjetoCnpq	0.3178	0.1739	1.5
ProjetoFacepe	0.155	0.0609	0.9286
ProjetoFinep	0.0853	0.0087	0.7143
atvDirecao	1.6357	1.3043	4.3571
atvEnsino	4.4186	3.4783	12.1429
atvProjeto	2.0388	1.1217	9.5714
atvPesquisa	1.124	0.7826	3.9286
atualizacao	758.8217	841.2087	82.0714

Figura 15: Experimento com 2 agrupamentos.

O segundo experimento que foi realizado visou separar os professores em 3 *clusters*. Nesse caso o algoritmo separou professores com pouca (*cluster* 0 - 24 professores), razoável (*cluster* 2 - 91 professores) e boa (*cluster* 1 - 14 professores) produção científica, conforme pode ser observado na figura 16. Os professores pertencentes ao *cluster* 1 são os mesmos que pertenciam ao *cluster* 1 no experimento anterior, de forma que com o aumento de uma unidade no número de *clusters*, o algoritmo separou melhor o grupo de professores que possuíam pouca produção científica. Desta forma, o *cluster* 0 agrupa os professores que possuem pouca produção científica, orientações, projetos e atividades. Já o *cluster* 2 agrupa os professores que possuem algumas publicações em anais de congresso (em média 12 publicações) e orientações de TCC e IC. Além disso, os professores do *cluster* 2 que possuem atividades de ensino cadastradas,

já participaram de algumas bancas examinadoras (em média 6,6 bancas) e atualizaram seus currículos há menos de 1 ano.

Cluster centroids:

Attribute	Full Data (129)	Cluster# 0 (24)	1 (14)	2 (91)
periodicos	2.2403	0.5417	9.7143	1.5385
periodicosComFator	0.7984	0.1667	3.2143	0.5934
livros	0.5581	0.0417	2.2857	0.4286
jornais	0.6899	0.0417	3.9286	0.3626
completoAnais	15.3566	3.9167	71.6429	9.7143
resumoAnais	3.6822	0.4583	21.8571	1.7363
expandidosAnais	0.6589	0.0833	3.1429	0.4286
publicacoesComDoi	1.0775	0	6.2857	0.5604
publicacoes	23.186	5.0833	112.5714	14.2088
oriMSC	1.7907	0.0417	13.6429	0.4286
oriPHD	0.1938	0	1.5714	0.033
oriESP	1.6202	0.8333	9.0714	0.6813
oriTCC	4.5736	1.7917	21.0714	2.7692
oriIC	4.0543	0.2083	23.9286	2.011
oriOutras	2.4419	0.0417	11.3571	1.7033
orientacoes	14.6744	2.9167	80.6429	7.6264
bancaEX	10.6744	1.375	52.9286	6.6264
bancaJUL	1.7287	0	8.1429	1.1978
ProjetoCnpq	0.3178	0	1.5	0.2198
ProjetoFacepe	0.155	0	0.9286	0.0769
ProjetoFinep	0.0853	0	0.7143	0.011
atvDirecao	1.6357	1.0833	4.3571	1.3626
atvEnsino	4.4186	1.4583	12.1429	4.011
atvProjeto	2.0388	0.0833	9.5714	1.3956
atvPesquisa	1.124	0.125	3.9286	0.956
atualizacao	758.8217	2745.375	82.0714	339.011

Figura 16: Experimento com 3 agrupamentos.

O terceiro experimento foi realizado com 4 *clusters*. Nesse caso o algoritmo agrupou professores com muita produção no *cluster* 3 (12 professores), os que produzem razoavelmente no *cluster* 1 (23 professores), os que produzem pouco no *cluster* 2 (71 professores) e os que produzem muito pouco no *cluster* 0 (23 professores), como pode ser observado na figura 17. Juntando os professores dos *clusters* 3 e 1, pode-se notar que aproximadamente 25 professores possuem produção relevante e os demais possuem produção inexpressiva.

O quarto experimento foi realizado com 5 *clusters*. Nesse caso o algoritmo estratificou os dados da seguinte maneira: *cluster* 0, professores que não possuem, a princípio, o viés de pesquisa; *cluster* 1, professores que possuem algumas pesquisas (em média 23 publicações), orientam (em média 27 orientações) e participaram de algumas bancas examinadoras (em média 18 bancas) e julgadoras (em média 3,4 bancas); *cluster* 2, onde estão agrupados a maioria dos professores (57% do total), e é caracterizado por professores que possuem poucas publicações (em média 11 publicações); *cluster* 3, que possui professores com perfil parecido com os do *cluster* 1, com a diferença de que os pertencentes ao *cluster* 3 possuem mais publicações (em média 42 publicações) e menos orientações (em média 21); *cluster* 4, professores que produzem bastante em pesquisa (em média 123 publicações), orientam muitos alunos (em

Cluster centroids:

Attribute	Full Data (129)	Cluster# 0 (23)	1 (23)	2 (71)	3 (12)
periodicos	2.2403	0.5652	4.3043	1.0141	8.75
periodicosComFator	0.7984	0.1739	2.4348	0.3099	1.75
livros	0.5581	0.0435	1.3043	0.1408	2.5833
jornais	0.6899	0.0435	0.8696	0.1831	4.5833
completoAnais	15.3566	4.087	18	7.9718	75.5833
resumoAnais	3.6822	0.4783	3.6522	1.4085	23.3333
expandidosAnais	0.6589	0.087	1.5217	0.1268	3.25
publicacoesComDoi	1.0775	0	2.0435	0.2254	6.3333
publicacoes	23.186	5.3043	29.6522	10.8451	118.0833
oriMSC	1.7907	0.0435	2.087	0.0986	14.5833
oriPHD	0.1938	0	0.3913	0.0282	1.1667
oriESP	1.6202	0.8696	0.3478	0.7606	10.5833
oriTCC	4.5736	1.8696	5.6522	1.7324	24.5
oriIC	4.0543	0.2174	6.6957	0.8028	25.5833
oriOutras	2.4419	0.0435	6.3913	0.169	12.9167
orientacoes	14.6744	3.0435	21.5652	3.5915	89.3333
bancaEX	10.6744	1.4348	14.6087	4.0845	59.8333
bancaJUL	1.7287	0	3.2609	0.6056	8.75
ProjetoCnpq	0.3178	0	0.9565	0.0141	1.5
ProjetoFacepe	0.155	0	0.1739	0.0423	1.0833
ProjetoFinep	0.0853	0	0	0.0141	0.8333
atvDirecao	1.6357	1.1304	1.5217	1.2676	5
atvEnsino	4.4186	1.4783	6.4348	3.6901	10.5
atvProjeto	2.0388	0.087	3.9565	0.6056	10.5833
atvPesquisa	1.124	0.1304	1.8261	0.7183	4.0833
atualizacao	758.8217	2803.2609	169.5217	401.0423	86.6667

Figura 17: Experimento com 4 agrupamentos.

média 91 orientações), participam de muitas bancas examinadoras (em média 61,6 bancas) e julgadoras (em média 8,5 bancas) e participam de muitas atividades de ensino, pesquisa, projeto e direção (figura 18). Embora os professores pertencentes ao *cluster* 4 possuam um elevado número de publicações, na média eles possuem menos publicações com fator de impacto do que os professores pertencentes ao *cluster* 3 e também menos projetos financiados pelo CNPQ.

Conforme pode ser visto na figura 19, os departamentos de elétrica, mecânica e básico possuem poucos professores com perfil de pesquisador. Já nos departamentos de civil e computação, pode-se notar uma maior distribuição dos professores entre os diferentes perfis. Além disso, pode-se destacar que o departamento de mecânica praticamente não possui professores com o perfil de pesquisador, e o departamento de civil possui aproximadamente 12 professores com esse perfil (levando em consideração os professores de civil dos *clusters* 1, 3 e 4 do experimento 4). Embora os professores do curso de civil possuam o maior número absoluto de professores com perfil de pesquisador, esse curso também possui um grande número de professores que aparentemente não apresentam viés de pesquisa (18 professores levando em consideração os *clusters* 0 e 2 do experimento 4). Outra característica observada nos experimentos 2, 3 e 4 é que o departamento de computação não possui professores sem viés de pesquisa e, além disso, mais de metade dos seus professores possuem boa produção científica

Cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(129)	(23)	(12)	(73)	(10)	(11)
periodicos	2.2403	0.5652	2.5833	1.0411	6.8	9.1818
periodicosComFator	0.7984	0.1739	0.75	0.3425	4.7	1.6364
livros	0.5581	0.0435	2	0.1507	0.7	2.6364
jornais	0.6899	0.0435	1.5833	0.1781	0.2	4.9091
completoAnais	15.3566	4.087	14.3333	8.1233	26.1	78.2727
resumoAnais	3.6822	0.4783	1.9167	1.3973	6.6	24.8182
expandidosAnais	0.6589	0.087	1.3333	0.1781	1.5	3.5455
publicacoesComDoi	1.0775	0	1	0.2466	5	5.3636
publicacoes	23.186	5.3043	23.75	11.0685	41.9	123.3636
oriMSC	1.7907	0.0435	1.6667	0.0959	4	14.8182
oriPHD	0.1938	0	0	0.0274	1	1.1818
oriESP	1.6202	0.8696	0.6667	0.7397	0	11.5455
oriTCC	4.5736	1.8696	8.6667	1.8082	3.9	24.7273
oriIC	4.0543	0.2174	6.0833	0.8767	10.6	25
oriOutras	2.4419	0.0435	10.75	0.2329	1.3	14.0909
orientacoes	14.6744	3.0435	27.8333	3.7808	20.8	91.3636
bancaEX	10.6744	1.4348	18	4.3836	13	61.6364
bancaJUL	1.7287	0	3.4167	0.6438	4.1	8.5455
ProjetoCnpq	0.3178	0	0.0833	0.0548	2.1	1.3636
ProjetoFacepe	0.155	0	0.25	0.0548	0	1.1818
ProjetoFinep	0.0853	0	0	0.0137	0	0.9091
atvDirecao	1.6357	1.1304	3.75	0.9589	1.2	5.2727
atvEnsino	4.4186	1.4783	9.0833	2.9041	10	10.4545
atvProjeto	2.0388	0.087	3.8333	0.6849	4.8	10.6364
atvPesquisa	1.124	0.1304	2.25	0.6712	1.8	4.3636
atualizacao	758.8217	2803.2609	315.25	380.0411	89.4	90.2727

Figura 18: Experimento com 5 agrupamentos.

(levando em consideração os professores de computação dos *clusters* 1, 3 e 4 do experimento 4 e dos *clusters* 1 e 3 do experimento 3). O experimento 1 mostra que o número de professores com muitas pesquisas relevantes é igual a 14, uma vez que ele separou professores com elevado número de publicações (*cluster* 1) dos demais professores (*cluster* 0).

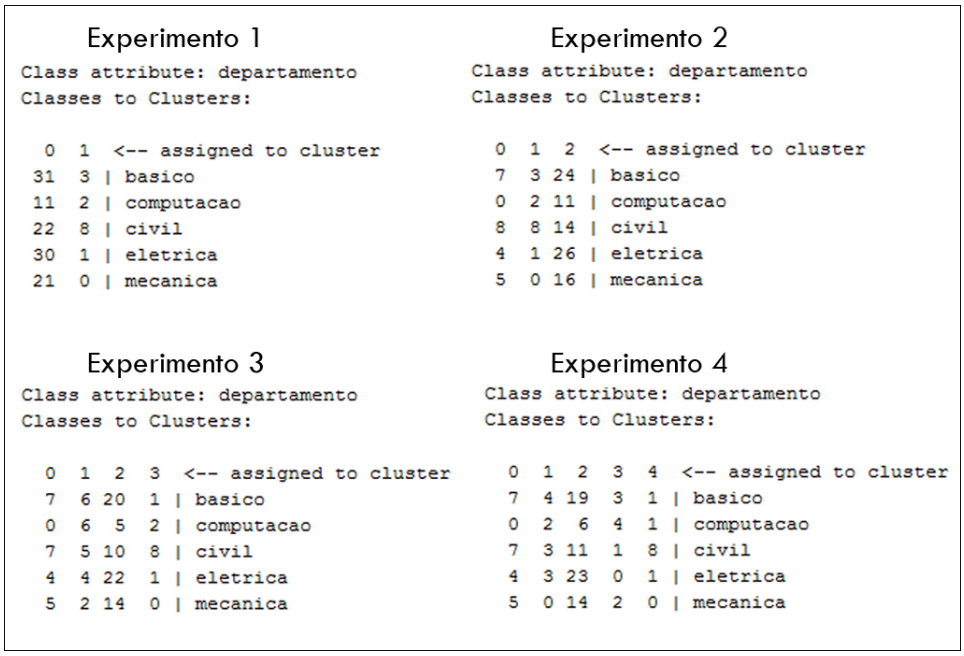

Figura 19: Distribuição dos *clusters* de acordo com os departamentos.

4.3 Regras de associação

A tarefa de regras de associação também foi analisada neste trabalho, com o objetivo de obter relacionamentos entre os atributos dos professores. Para realização desta tarefa foi utilizado o algoritmo *Apriori*. Porém, para que esse algoritmo pudesse ser aplicado fizeram-se necessários alguns pré-processamentos dos dados, uma vez que a maioria dos atributos selecionados neste trabalho é numérica e o algoritmo só funciona com dados nominais.

A ferramenta *WEKA* possui vários algoritmos para pré-processamento dos dados, de forma que para este trabalho foram utilizados dois algoritmos não-supervisionados: discretize, para discretizar os dados em intervalos; e *numeric-to-nominal*, para transformar os dados de numéricos para nominais.

Após a realização de vários testes alterando os valores do suporte mínimo e da confiança do algoritmo e os atributos de entrada, os resultados apresentaram algumas regras interessantes. Os experimentos que apresentaram melhores resultados foram executados com o parâmetro confiança mínima igual 0,7, sendo esses os experimentos que são descritos a seguir.

Um primeiro experimento com os atributos publicações e orientações produziu as regras apresentadas na figura 20, que indicam um forte relacionamento entre o atributo publicações e o atributo orientações, como pode ser visto se um professor possui poucas publicações (inferior a 24 na regra 1 e entre 24 e 48 na regra 2) implica que possui poucas orientações também (com uma confiança de 0,99 na regra 1 e 0,83 na regra 2).

```
Best rules found:

  1. publicacoes='(-inf-24]' 94 ==> orientacoes='(-inf-24.7]' 93    conf:(0.99)
  2. publicacoes='(24-48]' 18 ==> orientacoes='(-inf-24.7]' 15    conf:(0.83)
  3. orientacoes='(-inf-24.7]' 112 ==> publicacoes='(-inf-24]' 93    conf:(0.83)
```

Figura 20: Regras de associação que relacionam os atributos *publicações* e *orientações*.

O segundo e terceiro experimentos foram realizados selecionando os atributos departamento e orientações (figura 21) e departamento e publicações (figura 22), respectivamente. Estas regras extraídas mostram um relacionamento entre os departamentos e as publicações e orientações de um professor. Assim como nos resultados apresentados na tarefa de clusterização, estas regras indicam que os professores dos departamentos de elétrica, mecânica e básico possuem poucas publicações e orientações, pois estas regras com confiança maior que 0,8 mostram que se o departamento ao qual o professor pertence é mecânica, básico ou elétrica implica que seu número de publicações é inferior a 24,7 e seu número de orientações é inferior a 24.

```
Best rules found:

  1. departamento=mecanica 21 ==> orientacoes='(-inf-24.7]' 21    conf:(1)
  2. departamento=basico 34 ==> orientacoes='(-inf-24.7]' 32    conf:(0.94)
  3. departamento=eletrica 31 ==> orientacoes='(-inf-24.7]' 29    conf:(0.94)
```

Figura 21: Regras de associação que relacionam os atributos *departamento* e *orientações*.

```
Best rules found:

  1. departamento=mecanica 21 ==> publicacoes='(-inf-24]' 19    conf:(0.9)
  2. departamento=basico 34 ==> publicacoes='(-inf-24]' 28    conf:(0.82)
  3. departamento=eletrica 31 ==> publicacoes='(-inf-24]' 25    conf:(0.81)
```

Figura 22: Regras de associação que relacionam os atributos *departamento* e *publicações*.

O quarto experimento foi realizado com os atributos departamento e publicacoesComDoi (figura 23). Dentre as regras extraídas pode-se destacar a primeira, que relaciona os professores do departamento de civil com poucas publicações com DOI (inferior a 4,2 publicações), ou seja, embora possa ser constatado nos resultados da tarefa de clusterização que tais professores possuem muitas publicações, poucas possuem DOI.

```
Best rules found:

  1. departamento=civil 30 ==> publicacoesComDoi='(-inf-4.2]' 30    conf:(1)
  2. departamento=mecanica 21 ==> publicacoesComDoi='(-inf-4.2]' 21    conf:(1)
  3. departamento=eletrica 31 ==> publicacoesComDoi='(-inf-4.2]' 30    conf:(0.97)
  4. departamento=basico 34 ==> publicacoesComDoi='(-inf-4.2]' 32    conf:(0.94)
```

Figura 23: Regras de associação que relacionam os atributos *departamento* e *publicações-ComDoi*.

O quinto experimento foi realizado com os atributos departamento e publicacoesComFator (figura 24). Assim como no terceiro experimento, pode-se destacar a segunda regra, que relaciona os professores do departamento de civil com poucos periódicos com fator de impacto (inferior a 2 periódicos).

```
Best rules found:

  1. departamento=eletrica 31 ==> periodicosComFator='(-inf-2]' 29    conf:(0.94)
  2. departamento=civil 30 ==> periodicosComFator='(-inf-2]' 28    conf:(0.93)
  3. departamento=basico 34 ==> periodicosComFator='(-inf-2]' 31    conf:(0.91)
  4. departamento=mecanica 21 ==> periodicosComFator='(-inf-2]' 19    conf:(0.9)
```

Figura 24: Regras de associação que relacionam os atributos *departamento* e *publicações-ComFator*.

O sexto, sétimo e oitavo experimentos foram realizados utilizando os atributos publicações e periódicos (figura 25), publicações e completosAnais (figura 26) e publicações e resumoAnais (figura 27), respectivamente. As regras extraídas por estes experimentos mostram que artigos publicados em periódicos, trabalhos completos e resumos publicados em anais de congressos influenciam muito o número de publicações gerais de um professor, pois com regras com

confiança acima de 0,78 pode ser observado que quando o número de publicações é inferior a 24 o número de artigos publicados em periódicos é inferior a 2,6, o número de trabalhos completos publicados em anais de congresso é inferior a 12,9 e o número de resumos publicados em anais de congresso é inferior a 13,6.

```
Best rules found:

  1. periodicos='(-inf-2.6]' 90 ==> publicacoes='(-inf-24]' 82    conf:(0.91)
  2. publicacoes='(-inf-24]' 94 ==> periodicos='(-inf-2.6]' 82    conf:(0.87)
```

Figura 25: Regras de associação que relacionam os atributos *publicações* e *periódicos*.

```
Best rules found:

  1. completoAnais='(-inf-12.9]' 86 ==> publicacoes='(-inf-24]' 84    conf:(0.98)
  2. publicacoes='(-inf-24]' 94 ==> completoAnais='(-inf-12.9]' 84    conf:(0.89)
```

Figura 26: Regras de associação que relacionam os atributos *publicações* e *completoAnais*.

```
Best rules found:

  1. publicacoes='(-inf-24]' 94 ==> resumoAnais='(-inf-13.6]' 94    conf:(1)
  2. publicacoes='(24-48]' 18 ==> resumoAnais='(-inf-13.6]' 17    conf:(0.94)
  3. resumoAnais='(-inf-13.6]' 121 ==> publicacoes='(-inf-24]' 94    conf:(0.78)
```

Figura 27: Regras de associação que relacionam os atributos *publicações* e *resumoAnais*.

5 Conclusão

Este trabalho apresentou uma ferramenta, desenvolvida na linguagem PHP, cujo objetivo básico é extrair dados automaticamente de currículos da plataforma Lattes. Ainda neste trabalho, são aplicadas técnicas de Mineração de Dados aos dados extraídos por essa ferramenta, produzindo informações úteis à coordenação de pesquisa da Escola Politécnica de Pernambuco.

O trabalho apresenta importantes análises da produção cientifica dos professores da POLI, por meio da aplicação de algoritmos Mineração de Dados implementados pelo WEKA, clusterização e regras de associação. Os experimentos apresentam uma importante contribuição em termos de quais aspectos são característicos a perfis tanto de professores com pesquisas relevantes, quanto a perfis de professores aparentemente sem viés de pesquisa.

Uma característica que pode ser observada é que os professores que possuem viés de pesquisa procuram manter seus currículos atualizados (em média atualizaram seus currículos há menos de 3 meses), já os professores sem aparente viés de pesquisa em média atualizaram seus currículos há mais de 2 anos.

Além disso, pode-se concluir que os cursos de elétrica, mecânica e básico possuem poucos professores com perfil de pesquisador. Já nos cursos de civil e computação, pode-se notar uma maior distribuição dos professores entre os diferentes perfis. Também pode-se destacar que o curso de mecânica praticamente não possui professores com o perfil de pesquisador e o curso de civil possui aproximadamente 12 professores com esse perfil. No entanto, os professores do curso de civil possuem poucas publicações com doi e poucos periódicos com fator de impacto. Outra característica importante que pode ser extraída é que o curso de computação não possui professores no perfil que agrupa professores sem viés de pesquisa.

Outra informação importante extraída indica um forte relacionamento entre o atributo publicações e o atributo orientações, isso pode ocorrer porque publicações normalmente são produzidas em conjunto com alunos orientados. Ou seja, o incentivo a orientações, como bolsas de iniciação científica e mestrado, podem resultar em mais publicações relevantes para instituição.

Dentre as dificuldades encontradas durante o desenvolvimento deste trabalho algumas devem ser destacadas. O desenvolvimento da ferramenta de extração concentrou as maiores dificuldades encontradas, devido a falta de padronização dos dados na plataforma Lattes, uma vez que os dados podem ser inseridos nos currículos Lattes de forma subjetiva, ou seja, cada pesquisador pode colocar um mesma informação de várias formas diferentes. Outra dificuldade foi encontrada para acessar o fator de impacto das publicações cadastradas nos currículos Lattes, pois é necessário passar por um processo de autenticação, o que impossibilitaria a automação da ferramenta de extração desenvolvida. Para resolver este problema, o acesso a esta informação foi feito colocando o ISSN (*International Standard Serial Number* - utilizado para individualizar o título de uma publicação seriada) na URL da plataforma Lattes que realiza a busca desta informação no portal ISI Web of Knowledge, pois como os servidores da plataforma Lattes possuem cadastro neste portal, desta forma não se faz necessária a autenticação no sistema.

Como trabalho futuro sugere-se a realização de outras tarefas de Mineração de Dados, tais como aplicações de redes neurais para fazer previsões da produção científica da POLI de acordo com informações de anos anteriores. Outro trabalho futuro sugerido é ampliar o escopo da pesquisa e realizar um estudo com os professores de todas as unidades que compõem a Universidade de Pernambuco.

Referências

[1] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, p. 37–54, 1996.

[2] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, v. 39, p. 27–34, 1996.

[3] AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. *Proceedings 20th International Conference Very Large Data Bases*, VLDB, p. 487–499, 1994.

[4] BERKHIN, P. *Survey Of Clustering Data Mining Techniques*. San Jose, CA, 2002. Disponível em: <http://www.ee.ucr.edu/ barth/EE242/clustering_survey.pdf>. Acesso em: 18 de setembro de 2010.

[5] XU, R.; WUNSCH, D. Survey of clustering algorithms. IEEE Transactions on neural networks, v. 16, p. 645–678, 2005.

[6] JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice Hall, 1988.

[7] LATTES. Plataforma Lattes. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 15 de outubro de 2010.

[8] AMO, S. D. *Técnicas de Mineração de Dados*. [S.l.], 2004.

[9] HALL, M. et al. The weka data mining software: An update. SIGKDD Explorations, v. 11, 2009.

[10] HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Morgan Kaufmann, 2006.

[11] JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM computing surveys (CSUR)*, v. 31, 1999.

[12] GOWER, J. C. A general coefficient of similarity and some of its properties. Biometrics, v. 27, p. 857–871, 1971.

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

[14] CNPQ. Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em: <http://www.cnpq.br/>. Acesso em: 15 de outubro de 2010.

[15] ARFF. Attribute Relation File Format. Disponível em: <http://www.cs.waikato.ac.nz/ ml/weka/arff.html>. Acesso em: 15 de outubro de 2010.

[16] PHP. Disponível em: <www.php.net>. Acesso em: 18 de setembro de 2010.

[17] MYSQL. Disponível em: <www.mysql.com>. Acesso em: 18 de setembro de 2010.

[18] JCR. Journal Citation Reports. Disponível em: <http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports>. Acesso em: 15 de outubro de 2010.