Centro: Tecnologia da Informação

Curso: Sistemas de Informação

Titulo: **Aplicação da mineração de dados na melhoria da qualidade de dados em coleções científicas botânicas**

Autor(es) Jefferson da Silva Lyrio; Jorge José de Freitas Júnior; Juliana Miranda da Silva; Luís Alexandre Estevão da Silva; Mayara Luana Pereira Matias Ventura

E-mail para contato: laes.estacio@gmail.com IES: UNESA

Palavra(s) Chave(s): qualidade de dados, mineração de dados, coleções científicas, banco de dados

RESUMO

Dados de coleções científicas botânicas são de vital importância para o estudo da biodiversidade. As amostras de plantas mantidas em herbários e seus respectivos registros armazenados nos bancos de dados servem para diversas pesquisas biológicas e são testemunhos da ocorrência das espécies nas regiões de coleta. Apesar do aumento constante do volume de dados das coleções científicas, a qualidade dos dados não é a ideal, exigindo um esforço considerável dos pesquisadores no processo de limpeza dos dados, de forma que as pesquisas são efetivamente realizadas somente após essa etapa. O problema apresentado é a motivação para o presente trabalho que propõe como tema a definição de um processo de avaliação e melhora da qualidade dos dados de coleções científicas botânicas com a utilização da mineração de dados, tanto em relação aos erros taxonômicos quanto em relação aos erros encontrados nos dados de coletas, como por exemplo, a validação dos dados de ocorrência das espécies com a distribuição geográfica das espécies, que pode ser obtida na literatura. A hipótese para a formulação da proposta é a de que a análise de dados realizada com os algoritmos mineração de dados pode ser usada para identificar erros e outliers facilitando uma possível correção. Para a aplicação da proposta é definido um workflow científico, no qual a qualidade de dados é testada nas fases do ciclo de geração de conhecimento, desde a coleta dos dados e análise exploratória, a posterior aplicação dos algoritmos da mineração de dados, geração e visualização dos resultados. A metodologia de pesquisa do trabalho consiste nas seguintes etapas: 1) Estudo dos dados das coleções científicas, com a análise dos atributos mais utilizados e a identificação dos principais tipos de erros existentes; 2) Limpeza e padronização de dados; 3) Estudo para a determinação dos tipos de algoritmos da mineração de dados mais adequados para cada um dos tipos de erros identificados na Etapa 1, como os algoritmos da análise de associação e de agrupamentos; 4) A definição da sequência adequada de aplicação desses algoritmos; 5) O desenvolvimento das etapas do workflow científico; 6) Geração do conhecimento. Até o presente momento o trabalho encontra-se desenvolvido até a Etapa 3. Para as primeiras análises foram usados dados de coletas pertencentes ao Instituto de Pesquisas Jardim Botânico do Rio de Janeiro em um total aproximado de 852.736 registros. Resultados: um dos principais atributos das coletas é o nome do coletor, porém, campos livres de críticas de entrada de dados permitiram uma grande inconsistência nos valores, acarretando em uma variedade de 79.556 nomes de coletores, onde essa variedade proporciona casos como onde um mesmo coletor apresenta mais de 60 formas distintas de escrita de seu nome. Justificada a necessidade da limpeza dos dados, 30 rotinas desenvolvidas em SQL e automatizadas no R, permitiram alcançar uma redução de inconsistências de 23.58% (18.764) registros nos nomes de coletores. Após a limpeza, uma nova etapa de avaliação da qualidade dos dados foi iniciada, utilizando o algoritmo de regras de associação Apriori. O objetivo foi analisar possíveis associações entre o nome do coletor e o nome da família de plantas ao qual o pesquisador tem especialidade. Os padrões esperados foram obtidos, tanto em relação a associação positiva (comprobatórios) quanto em relação a associação negativa (suspeitos). Todo o desenvolvimento do projeto vem sendo feito com software livre, com o pacote estatístico R e o sistema gerenciador de banco de dados PostgreSQL. Como conclusão obtida através dos resultados iniciais comprovou-se que a mineração de dados pode ser usada em dados de coleções científicas botânicas no processo de qualidade de dados auxiliando os administradores das bases de dados na busca da melhora do processo de geração do conhecimento.