



## MÉTODO ROCK APLICADO NA DETECÇÃO AUTOMÁTICA DE RESPOSTAS EM UM SISTEMA EDUCACIONAL

PEDROTTI, Tales<sup>1</sup>

CHICON, Patricia Mariotto Mozzaquatro<sup>2</sup>

**Resumo:** Atualmente, a expansão e utilização de plataformas *Web* voltadas à aprendizagem educacional online, como os cursos de ensino à distância e o Ambiente Virtual de Aprendizagem (AVA), vêm gerando grandes volumes de dados. Esses dados geralmente contêm informações que, ao serem utilizadas no contexto pedagógico, poderão beneficiar tanto alunos quanto professores. Contudo, devido à complexidade e ao volume, torna-se inviável o processo manual de extração dessas informações, sendo necessária a utilização de ferramentas específicas para sua extração. Neste contexto, o presente trabalho propõe um estudo sobre o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*), composto por diversas etapas para extração de informações em dados brutos. Uma das etapas deste processo é a Mineração de Dados (*Data Mining*), utilizada para extrair padrões dos dados. Entre as técnicas utilizadas na mineração foi escolhido o método de Agrupamento de Dados (*Clustering*), utilizando o algoritmo ROCK a ser utilizado em um sistema educacional, com base em um estudo de caso aplicando o mesmo na detecção automática de respostas.

**Palavras-chave:** ROCK. Sistema educacional. Mineração de dados.

**Abstract:** Nowadays the expansion and use of Web platforms aimed at online educational learning, such as distance learning courses and the Virtual Learning Environment (VLE), have generated large volumes of data. These data often contain information which, when used in the educational context, can benefit both students and teachers. However, due to the complexity and volume, it is not feasible the manual process of extracting this information, requiring the use of specific tools for its extraction. In this context, this article proposes a study on the process of Knowledge Discovery in Databases (*Knowledge Discovery in Databases*), composed of several steps for extracting information in raw data. One of the steps in this process is the Data Mining (*Data Mining*), used to extract data standards. Among the techniques used in data mining was chosen the Clustering technique using the ROCK algorithm to be used in an educational system based on a case study applying the same in the automatic detection of answers.

**Keywords:** ROCK. Educational system. Data Mining.

### 1. INTRODUÇÃO

Atualmente com a expansão da *internet*, houve o crescimento da utilização de serviços e plataformas *Web*, principalmente para aquelas voltadas a aprendizagem educacional *online*. Entre essas plataformas destacam-se os cursos de Educação a Distância

<sup>1</sup> Acadêmico do Curso de Ciência da Computação. E-mail: tales.pedrotti@gmail.com.

<sup>2</sup> Professora do curso de Ciência da Computação. E-mail: patriciamozzaquatro@gmail.com.



(EaD) e o Ambiente Virtual de Aprendizagem (AVA), cuja distribuição vem crescendo gradualmente assim como o número de usuários, gerando grandes quantidades de dados (COSTA *et al.*, 2012). Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2008) que realiza os censos educacionais para o Ministério da Educação (MEC) o Brasil tem 727.961 alunos em cursos à distância, apenas no nível de graduação. Os dados produzidos pelos usuários são geralmente armazenados em repositórios de dados, como bancos de dados e *data warehouses*, de maneira centralizada ou distribuída, existentes em diversos domínios como o educacional, financeiro, médico, comercial e científico (REZENDE, 2005). No domínio educacional, mais especificamente em instituições de ensino como escolas e universidades, é possível explorar as bases de dados em busca de informações que poderão trazer benefícios tanto aos alunos quanto aos professores. No entanto essas informações são raramente obtidas de forma direta por falta de ferramentas apropriadas para sua extração (MITRA *et al.*, 2002). A área que busca extrair informações a partir de grandes volumes de dados utilizando métodos e ferramentas com o fim de gerar conhecimento é conhecida como Mineração de Dados (do inglês, *Data Mining*).

No contexto educacional, com a expansão dos cursos à distância, pesquisadores da área da Informática na Educação têm mostrado interesse em utilizar mineração de dados para investigar perguntas científicas na área da educação como: “Quais são os fatores que afetam a aprendizagem?” ou “Como desenvolver sistemas educacionais mais eficazes?”. Diante desses fatores surgiu uma nova área de pesquisa conhecida como Mineração de Dados Educacionais (do inglês, *Educational Data Mining*, EDM). A EDM é uma área de pesquisa específica da mineração de dados, cujo objetivo é desenvolver ou adaptar métodos e algoritmos de mineração já existentes a fim de explorar e analisar conjuntos de dados coletados em ambientes educacionais (BAKER *et al.*, 2011).

Ao analisar as informações armazenadas nas bases de dados, observou-se que grande parte desses dados (cerca de oitenta por cento) está em formato não estruturado, na qual uma parte significativa desses dados é formada por textos (KUECHLER, 2007). Esses textos constituem um grande repositório de dados organizacional, envolvendo documentos, e-mails, projetos e o próprio conhecimento adquirido (HAN; KAMBER, 2006). Neste contexto, surgiu à área de exploração conhecida como Mineração de Texto (do inglês, *Text Mining*, TM) também conhecida como Mineração de Dados Textuais (do inglês, *Text Data Mining*, TDM) ou Descoberta de Conhecimento em Bases Textuais (do inglês, *Knowledge-Discovery in Text*, KDT). A mineração de texto pode ser definida como um conjunto de técnicas e processos



utilizados na extração descoberta de conhecimento a partir de documentos em textos não estruturados (LOPES, 2004).

Considerando os fatos mencionados, o presente trabalho teve como objetivo a elaboração de um sistema educacional de perguntas e respostas, detectando respostas corretas de forma automática. O desafio encontra-se na detecção automática em um banco de dados textual não estruturado utilizando o método de agrupamento hierárquico aglomerativo ROCK. Uma base de dados é considerada “não estruturada” quando os grupos e categorias dos dados armazenados são desconhecidos (BAKER *et al.*, 2011). Neste caso, conforme Metz (2006) são utilizados algoritmos para descobrir padrões nos dados a partir de alguma caracterização de regularidade, sendo esses padrões denominados *clusters*. Neste contexto, foi realizada uma análise de *clusters*, criando-se uma padronização dos dados, classificando-os em diferentes grupos e categorias. Para a detecção desses grupos e categorias, foi utilizada a técnica de Agrupamento de Dados (*Clustering*), por meio da aplicação do algoritmo de agrupamento hierárquico aglomerativo ROCK.

O estudo buscou medir o desempenho do método ROCK aplicado na detecção automática de respostas em um sistema educacional.

## **2. REVISÃO DA LITERATURA**

As subseções a seguir irão abordar: o *learning* analítico, no que diz respeito à utilização de ferramentas educacionais *online* e a importância dos dados gerados por estas; a mineração de dados, tarefa do processo de descoberta de informações em bases de dados que tem como objetivo descobrir informações úteis a partir de conjuntos de dados; a técnica de agrupamento de dados, em específico a abordagem hierárquica aglomerativa, com o objetivo de classificar agrupamentos de dados por meio de um índice de similaridade e o método ROCK, algoritmo aplicado na técnica de agrupamento hierárquico aglomerativo.

### **2.1 Learning analítico**

Com a expansão e a crescente utilização de serviços e plataformas *Web* como os ambientes educacionais – entre eles o curso de Educação a Distância (EaD) e o Ambiente Virtual de Aprendizagem (AVA) – grandes volumes de dados são gerados, sendo que grande parte desses dados podem conter informações que, ao serem exploradas, poderão trazer benefícios tanto a estudantes e professores quanto às instituições de ensino. Segundo Baker *et al.* (2011) a exploração de dados coletados em ambientes educacionais permite compreender



de forma mais eficaz e adequada o comportamento dos alunos, como eles aprendem e o papel do contexto na qual a aprendizagem ocorre, por exemplo, identificar em que situação um tipo de abordagem instrucional (aprendizagem individual ou colaborativa) proporciona melhores benefícios educacionais aos alunos. Neste contexto, surge o *learning* analítico (*Learning Analytics*, LA), uma nova área de pesquisa envolvendo um processo de medida, coleta, análise e relato dos dados produzidos por alunos e seus contextos, com o objetivo de entender e aperfeiçoar o aprendizado e o ambiente em que este ocorre (LAK, 2011).

Ao ser aplicado em instituições de ensino, disciplinas e cursos, o processo de *learning* analítico fornece informações valiosas sobre o que realmente está acontecendo e sugere maneiras através das quais os educadores podem melhorar o processo de ensino-aprendizagem, além de indicar quais alunos podem desistir do curso ou aqueles que precisam de ajuda adicional para melhorar seu desempenho (SIEMENS; LONG, 2011).

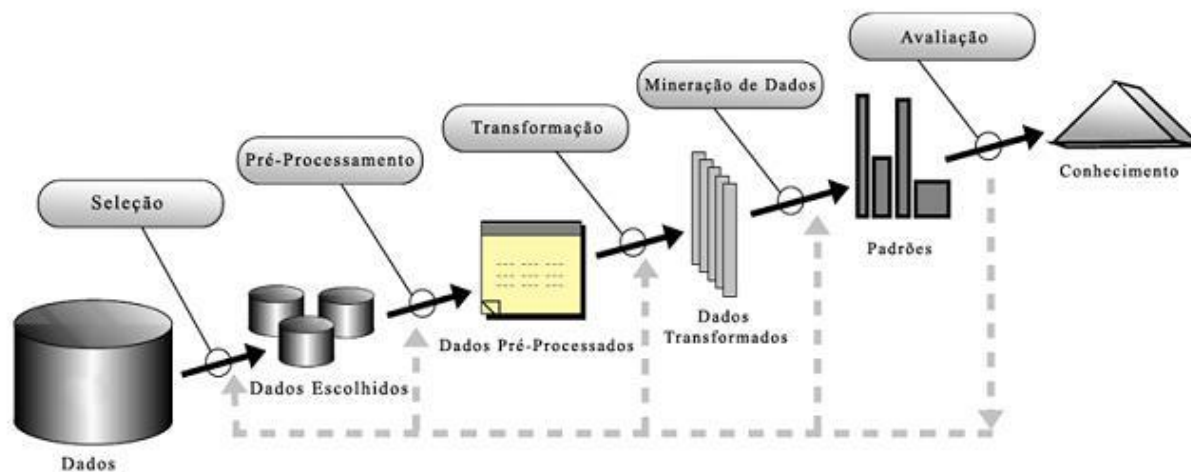
Segundo Chatti *et al.* (2012) o processo de *learning* analítico aplica diferentes métodos para detectar padrões interessantes ocultos em conjuntos de dados educacionais. Um destes métodos é a Mineração de Dados.

## **2.2 Mineração de Dados**

A Mineração de Dados (*Data Mining*, MD) é a etapa de um processo mais amplo conhecido como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*, KDD) que consiste na aplicação de métodos para extração de padrões interessantes a partir dos dados, de forma iterativa e repetitiva; enquanto que o KDD é definido como “...um processo não trivial, interativo e iterativo, para identificar em dados padrões válidos, novos, potencialmente úteis e compreensíveis” (FAYYAD *et al.*, 1996). A Figura 1 ilustra as etapas do processo de KDD.



Figura 1. Processo de KDD



Fonte: Fayyad *et al.* (1996).

As etapas do processo de KDD descritas seguem o modelo proposto por Fayyad *et al.* (1996) no qual o processo é composto por nove etapas:

- Entendimento do Domínio da Aplicação: definição do tipo de conhecimento a ser descoberto, identificando o objetivo do processo de KDD do ponto de vista do usuário;
- Seleção: seleção de um conjunto de dados ou um subconjunto de variáveis e amostra de dados, onde a descoberta deverá ser realizada;
- Pré-processamento: inclui operações básicas como remoção de ruídos, coleta de informações necessárias para modelar ou estimar ruído e escolha de estratégias para manipular campos de dados ausentes;
- Transformação: localização de características úteis para representar os dados dependendo do objetivo da tarefa;
- Definição do Objetivo da Descoberta de Conhecimento: escolha do objetivo ou tarefa para a Mineração de Dados;
- Análise Exploratória: escolha dos algoritmos de mineração e seleção de métodos a serem utilizados para a busca de padrões de dados;
- Mineração de Dados: aplicação da mineração de dados em busca de padrões de interesse em forma de representação ou um conjunto de representações, incluindo regras de classificação ou árvores, regressão e agrupamento (*clustering*);
- Avaliação: interpretação dos padrões minerados, com possível retorno às etapas 1-7 para posterior iteração;



- Conhecimento Descoberto: incorporação do conhecimento descoberto em outro sistema ou documentá-lo e reportá-lo às partes interessadas.

O processo de KDD pode também ser visualizado em três grandes grupos: pré-processamento (preparação dos dados para a tarefa de mineração, envolvendo as tarefas de seleção, limpeza e transformação dos dados), mineração de dados (o processo de extração de padrões dos dados) e pós-processamento (análise do conhecimento extraído).

### 2.2.1 Técnica de Agrupamento de Dados

A técnica de Agrupamento de Dados ou Clusterização (*Clustering*) é uma tarefa da mineração de dados que tem por finalidade a classificação de conjuntos finitos e similares de dados por meio da formação de agrupamentos para descrever os dados (JAIN *et al.*, 1999). Um agrupamento (ou *cluster*) é uma coleção de registros (ou objetos) similares entre si, porém distintos dos outros registros nos demais agrupamentos (CAMILO; SILVA, 2009). Conforme Calil *et al.* (2008) no processo de agrupamento, os objetos são agrupados de tal forma que as similaridades *intraclusters* (dentro de um mesmo *cluster*) sejam maximizadas e as similaridades *interclusters* (entre *clusters* diferentes) sejam minimizadas.

Segundo Jain *et al.* (1999) os algoritmos de agrupamento podem ser classificados de acordo com o método adotado para definição dos *clusters*. Para Prass (2012), a escolha do método a ser utilizado na análise depende do tipo de variável e do propósito da aplicação. Os métodos utilizados para o agrupamento podem ser classificados em: hierárquicos, partição (métodos mais tradicionais), baseados em modelos, baseados em grades e baseados em densidade (métodos mais modernos).

Na técnica de agrupamento hierárquico, dado um conjunto de dados inicial, são construídos agrupamentos de objetos de dados, de modo que objetos pertencentes a um mesmo *cluster* possuem alta similaridade e objetos pertencentes a *clusters* diferentes possuem baixa similaridade, resultando em uma hierarquia que descreve um particionamento diferente a cada nível analisado (METZ, 2006).

Segundo Metz (2006) a técnica de agrupamento hierárquico pode ser realizada de duas formas: por aglomeração ou por divisão. Na aglomeração cada objeto é considerado um *cluster* unitário. Em seguida pares desses *clusters* são iterativamente agrupados de acordo com um índice de similaridade, até que todos os objetos pertençam à apenas um *cluster*. Na divisão inicia-se com apenas um agrupamento contendo todos os objetos e procede dividindo o conjunto de objetos em *clusters* cada vez menores até que cada objeto pertença



exclusivamente a um *cluster* ou até que se alcance o critério de parada, frequentemente o número de *clusters* desejados.

Embora haja duas estratégias para o agrupamento hierárquico, geralmente a abordagem aglomerativa é referenciada em trabalhos relacionados ao agrupamento hierárquico, sem muito interesse nos métodos divisivos. Um fator predominante para essa preferência é a complexidade computacional desses métodos (METZ, 2006).

### 2.2.1.1 Método ROCK

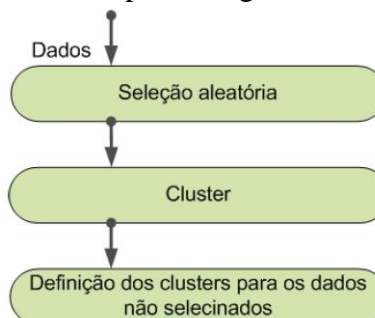
O método ROCK (*RObust Clustering using LinKs*, Agrupamento Robusto utilizando *Links*) é um algoritmo hierárquico aglomerativo proposto por Guha *et al.* (1999) voltado ao agrupamento de atributos categóricos (nominais) e booleanos (binários).

O algoritmo se baseia principalmente no uso de medidas de proximidade para agrupar os elementos, em substituição às medidas de distância abordadas nos métodos de *cluster* para dados numéricos. A cada passo unem-se os elementos, utilizando uma medida de proximidade ou vizinhança, diferente da abordagem utilizada pelos algoritmos clássicos de agrupamento (SOUSA, 2012).

O processo do algoritmo ROCK inicia-se com a construção de um grafo esparso de uma dada matriz de similaridade de dados usando um limite de similaridade e o conceito de vizinhos compartilhados, aplicando um algoritmo de agrupamento hierárquico sobre o grafo. Neste grafo, os elementos são os nós do grafo e a aresta entre dois elementos indica que eles fazem parte de um mesmo *cluster* (METZ, 2006).

As etapas envolvidas no processo de agrupamento do algoritmo ROCK são demonstradas na Figura 2.

Figura 2. Etapas do algoritmo ROCK



Fonte: Metz (2006).



Conforme a Figura 2, uma amostra de elementos é selecionada aleatoriamente a partir dos dados originais. Na etapa seguinte, aplica-se um algoritmo de agrupamento hierárquico utilizando o conceito de ligações sobre essa amostra. Por último, os *clusters* identificados são utilizados para determinar o grupo ao qual pertence cada elemento não selecionado na primeira etapa, os quais são atribuídos ao *cluster* que maximiza o número de ligações (METZ, 2006).

A utilização do método ROCK no agrupamento hierárquico aglomerativo possui suas vantagens e limitações. Para Guha *et al.* (1999) o algoritmo ROCK gera *clusters* de melhor qualidade do que os algoritmos tradicionais e também exibe boa escalabilidade, porém possui um pior caso de complexidade de tempo com relação a computação das listas de vizinhos e ligações.

### 2.3 Trabalhos correlatos

O presente trabalho visando um embasamento bibliográfico baseou-se nas pesquisas descritas a seguir:

O trabalho “Interpretação de *clusters* gerados por algoritmos hierárquicos” de Metz (2006) propôs um projeto envolvendo a implementação e integração de um módulo de aprendizado não supervisionado baseado na abordagem de *clustering* hierárquico integrado ao ambiente computacional “DISCOVER”, um ambiente desenvolvido como uma ferramenta de apoio à pesquisa em mineração de dados, composto por diversas ferramentas de pré-processamento de dados, extração de padrões e pós-processamento de conhecimento com finalidade de servir de uso e extensão por pesquisadores de aprendizado de máquina, mineração de dados e textos. Outro objetivo do trabalho foi à avaliação da utilização de metodologias de interpretação de *clusters* para facilitar o trabalho do especialista de domínio na realização da etapa de interpretação dos padrões extraídos pelos algoritmos de *clustering* hierárquicos, principalmente em relação ao método ROCK. Os conceitos e definições sobre as etapas do processo de *clustering*, bem como a abordagem hierárquica e os algoritmos utilizados, contribuíram como uma importante fonte de pesquisa e um melhor entendimento para a realização deste trabalho.

O trabalho “Análise de *Clusters* via Algoritmo ROCK aplicada ao problema da mineração de dados sobre criminalidade em Belo Horizonte” de Sousa (2012) propôs um estudo de um banco de dados gerado a partir de boletins de ocorrência registrados na Secretaria de Segurança Pública de Minas Gerais em Belo Horizonte no ano de 2010. O





estudo proposto teve como o objetivo de aplicar técnicas específicas para o tratamento e modelagem de grandes volumes de dados, com baixo custo computacional, ao investigar a formação de possíveis padrões de semelhança entre variáveis provenientes de informações sobre criminalidade. Como método para a pesquisa e a detecção de *clusters* foi desenvolvido uma versão adaptada do algoritmo ROCK utilizando o *software* MATLAB. A utilização desta ferramenta obteve eficiência e eficácia, conseguindo em um tempo computacional razoável agrupar indivíduos que realmente possuíam alguma similaridade, além de separar os demais, garantindo a homogeneidade dentro do *cluster* e a heterogeneidade fora dele. Uma das contribuições científicas interessantes deste trabalho foi a eficiência do algoritmo ROCK em relação ao tempo computacional mesmo para uma quantidade elevada de observações.

### 3 METODOLOGIA

Este trabalho caracterizou-se por adotar uma abordagem do tipo quali-quantitativa, pois uma pesquisa qualitativa é aquela que “não se preocupa com a quantidade numérica, mas sim com o aprofundamento da compreensão de um grupo social.” (GOLDENBERG, 1999).

O trabalho foi dividido da seguinte forma:

Etapa 1 – Estudo Teórico: Definição da área de pesquisa – Engenharia de *Software* voltada à mineração de dados; Pesquisa bibliográfica; Análise do problema – Extração e agrupamento de dados de ambientes educacionais utilizando mineração de dados; Pesquisa sobre mineração de dados; Pesquisa sobre a técnica de agrupamento de dados; Análise e descrição sobre o método de agrupamento hierárquico aglomerativo ROCK.

Etapa 2 – Desenvolvimento prático: Aplicação do método de agrupamento hierárquico aglomerativo ROCK.

Etapa 3 – Validação e apresentação dos resultados.

### 4. RESULTADOS E DISCUSSÕES

Foi realizado um estudo de caso do método de agrupamento hierárquico aglomerativo ROCK utilizando uma base de dados.

Para o agrupamento de dados pelo método ROCK foram utilizada amostras de uma base de dados de um questionário educacional, composta por registros de perguntas e respostas dos usuários. A Figura 3 ilustra a tela inicial do sistema educacional.



Figura 3. Tela inicial do sistema educacional



Fonte: Elaborado pelo autor.

Para o processo de agrupamento foi utilizada a ferramenta GEPHI, uma plataforma de código aberto para manipulação e visualização de grafos dinâmicos e hierárquicos, de todos os tipos de redes à sistemas complexos (GEPHI, 2015).

As bases de dados foram geradas no formato de arquivo CSV (*Comma Separated Values*, Valores Separados por Vírgula) e importadas pelo GEPHI.

Para o processo de agrupamento de dados pelo método ROCK foi utilizada uma base de dados geral (total), com a criação de duas outras bases de dados contendo seleções dos registros (amostras) da base de dados original (etapa de teste e de treinamento). A base de dados e as amostras foram utilizadas no processo de agrupamento são descritas a seguir:

- Amostra total: correspondente à base de dados original (10 (dez) perguntas e 500 (quinhentos) usuários);
- Amostra 50% (cinquenta por cento) mista: correspondente à metade do total dos registros da base de dados original (5 (cinco) perguntas e 250 (duzentos e cinquenta) usuários);
- Amostra 30% (trinta por cento) mista: correspondente a trinta por cento do total dos registros da base de dados original (3 (três) perguntas e 150 (cento e cinquenta) usuários).

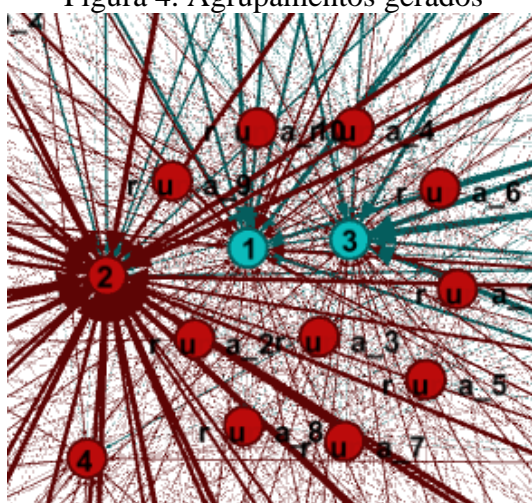
A base de dados e as amostras são compostas por uma classe “usuario” que, por sua vez, é composta por dois atributos: “usuario” e “resultado”. O atributo “usuario” refere-se ao



registro do usuário cadastrado no sistema, enquanto o atributo “resultado” refere-se ao registro das respostas.

Para o agrupamento dos dados foi aplicado o algoritmo “Force Atlas” a base de dados, a fim de gerar os *clusters*. O algoritmo “Force Atlas” agrupa os dados pelo método ROCK e tem como objetivo representar espacialmente as ligações entre os nós pertencentes a *clusters* distintos, de forma a criar um mapa de suas interações. A Figura 4 ilustra os agrupamentos identificados e gerados pelo algoritmo.

Figura 4. Agrupamentos gerados

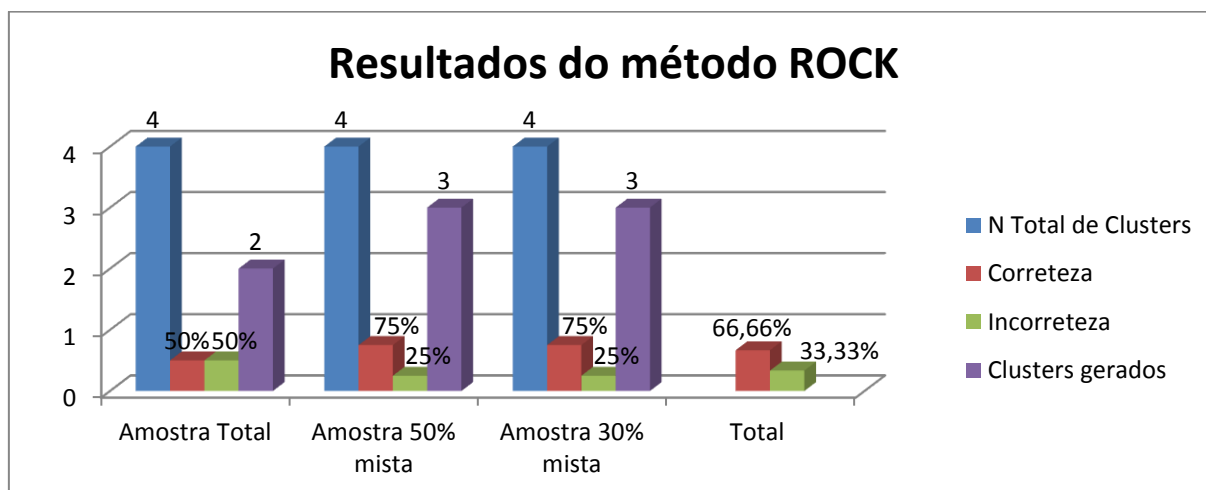


Fonte: Elaborado pelo autor.

Ao término do processamento do algoritmo, verificou-se que os *clusters* 1 (um) e 3 (três) possuem relacionamento em comum.

A Figura 5 ilustra o resultado do método ROCK.

Figura 5. Resultados do método ROCK



Fonte: Elaborado pelo autor.



Conforme o gráfico da Figura 5, foram analisadas três amostras da base de dados. A análise da “Amostra Total” resultou em índices de correteza e incorreteza de 50%, gerando 2 *clusters* em um total de 4 *clusters* analisados. A “Amostra 50% mista” resultou em um índice de correteza de 75% e de incorreteza 25%, gerando 3 *clusters* em um total de 4 . A “Amostra 30% mista” foi equivalente a “Amostra 50% mista”. O índice total de correteza e incorreteza das amostras analisadas pelo método ROCK foi de 66,66% e 33,33% respectivamente.

## 5. CONSIDERAÇÕES FINAIS

Ao término do estudo realizado em relação ao desempenho do método ROCK na detecção automática de respostas em um sistema educacional, foi possível afirmar que o processo de agrupamento hierárquico aglomerativo de dados pelo método ROCK aplicado às amostras da base de dados analisadas foi realizado com sucesso. A aplicação do método ROCK gerou os resultados esperados como a detecção dos agrupamentos e as similaridades das respostas.

Analisando os resultados, observou-se que o método ROCK obteve um desempenho positivo em contraposição à complexidade de tempo, detectando os agrupamentos de forma correta e rápida.

Foi possível constatar a similaridade entre perguntas e respostas corretas, de modo que a aplicação do método ROCK tornou possível a detecção de agrupamentos e a elaboração dos índices de correteza e incorreteza das respostas.

Para a aplicação do método ROCK foi necessária a instalação da ferramenta GEPHI, visto que a ferramenta WEKA não oferece suporte nativo ao algoritmo (necessita instalação de *plugin*).

## REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, R. S. J. de; ISOTANI, S.; CARVALHO, A. M. J. B. de. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, Ano XIX, n. 2, Agosto de 2011.

CALIL, L. A. de Almeida.; CARVALHO, D. R.; SANTOS, C. B. dos. VAZ, M. S. M. G.; **Mineração de Dados e Pós-processamento em Padrões Descobertos**. Ponta Grossa: UEPG.



Publicatio UEPG – Ciências Exatas e da Terra, Agrárias e Engenharias, v. 14, n. 3, p. 212, 2008.

CHATTI, M. A.; DYCKOFF, A. L.; SCHROEDER, U.; THÜS, H. *A reference model for learning analytics. International Journal of Technology Enhanced Learning*, v. 4, n. 5, p. 318-331, 2012.

COSTA, E.; BAKER, R. S. de; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. **Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações**. Rio de Janeiro: Jornada de Atualização em Informática na Educação – JAIE, 2012.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. *The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM*, v. 39, n. 11, p. 29-30, 1996.

\_\_\_\_\_. *From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence – AAAI*, v. 17, n. 3, p. 41, 1996.

GEPHI. *The Open Graph Viz Platform*. Disponível em: <<http://www.gephi.github.io>>. Acesso em: 25 abril. 2015.

GOLDENBERG, Mirian. **A Arte de Pesquisar: Como fazer Pesquisa Qualitativa em Ciências Sociais**. 3ª ed. Rio de Janeiro: Record, 1999.

GUHA, S.; RASTOGI, R.; SHIM, K. **ROCK: A Robust Clustering Algorithm for Categorical Attributes. Proceedings of the 15th International Conference on Data Engineering**, p. 512-521, 1999.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 2ª ed. USA: Morgan Kaufmann Publishers Inc, 2006.

INEP. **Instituto Nacional de Estudos e Pesquisas**, 2008. Disponível em: <[http://www.acheseucurso.com.br/pdf/Relatorio\\_observa\\_2.pdf](http://www.acheseucurso.com.br/pdf/Relatorio_observa_2.pdf)>. Acesso em: 15 ago. 2014.



JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. *Data clustering: a review*. *ACM Computing Surveys*, v. 31, n. 3, p. 264-323, 1999.

KUECHLER, L. W. *Business Applications of Unstructured Text*. *Communications of the ACM*, vol. 50, n. 10, p. 86-93, 2007.

LAK. *1st International Conference on Learning Analytics and Knowledge 2011*, 27 fev-1 mai, 2011. Disponível em: <<http://www.tekri.athabascau.ca/analytics>>. Acesso em: 21 mai. 2015.

LOPES, Maria Célia Santos. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**. Tese de pós-graduação. Rio de Janeiro: COPPE-UFRJ, 2004.

METZ, Jean. **Interpretação de clusters gerados por algoritmos de clustering hierárquico**. Dissertação de mestrado, São Carlos: USP, 2006.

MITRA, S.; PAL, S. K.; MITRA, P. K. *Data mining in soft computing framework: A survey*. *IEEE Transactions on Neural Networks*, vol.XIII, n.1, p. 3-14, 2002.

OCHI, L. S.; DIAS, C. R.; SOARES, S. S. F. **Clusterização em Mineração de Dados**. Programa de Pós-Graduação em Computação, Rio de Janeiro: UFF, 2004.

PRASS, F. S.; **KDD – Uma Visão Geral do Processo**. FP2 Tecnologia, 2012. Disponível em: <<http://www.fp2.com.br/blog/index.php/2012/artigo-kdd-uma-visao-geral-do-processo>>. Acesso em: 17 abr. 2014.

REZENDE, S. O. **Mineração de Dados**. XXV Congresso da Sociedade Brasileira de Computação, São Leopoldo: UNISINOS, 2005.

SIEMENS, G.; LONG, P. *Penetrating the fog: Analytics in learning and education*. *Educause Review*, v. 46, n. 5, p. 30-32, 2011.



# XVII

Seminário Internacional  
de Educação no MERCOSUL



[www.unicruz.edu.br/mercosul](http://www.unicruz.edu.br/mercosul)

SOUSA, C. R. dos Santos. **Análise de Clusters via Algoritmo ROCK Aplicada ao Problema da Mineração de Dados sobre Criminalidade em Belo Horizonte.** Dissertação de Mestrado, Minas Gerais: CEFET-MG, 2012.