Revista da Universidade Vale do Rio Verde
ISSN: 1517-0276 / EISSN: 2236-5362
v. 17 | n. 1 | Ano 2019

ESTUDO SOBRE OS ALGORITMOS DE CLUSTERIZAÇÃO *HIERARCHICAL CLUSTERER* E *SIMPLE K-MEANS* APLICADOS NO AGRUPAMENTO DE PADRÕES SIMILARES

Flávio de Brum
Universidade de Cruz Alta - UNICRUZ
flaviodebrum@gmail.com

Patricia Mariotto Mozzaquatro
Universidade de Cruz Alta - UNICRUZ
patriciamozzaquatro@gmail.com

Jocias Maier Zanatta
Universidade Federal de Santa Maria - UFSM
josk85@hotmail.com

RESUMO

O quantitativo de informações disponíveis na internet faz com que os usuários busquem na tecnologia da informação ferramentas para auxiliar na execução de suas tarefas, com isso, emerge a técnica de mineração de dados. O presente estudo tem o objetivo de implementar e comparar dois algoritmos de mineração de dados *Simple K-Means* e *Hierarchical Clusterer*, medindo sua eficiência na identificação de padrões similares entre sub- área, palavras-chave e artigos acadêmicos, gerando assim clusters baseando-se na similaridade de interesses entre usuários e obras consultadas. Busca-se ainda elucidar a seguinte problemática: De que forma a Mineração de Dados pode contribuir para a identificação de padrões similares entre grande área, área, sub-área, palavras-chave e artigos acadêmicos. Conclui-se com o estudo que o algoritmo de clusterização *Hierarchical Clusterer* apresentou maior eficiência no agrupamento de dados similares em uma base de dados.

Palavras-chave: Mineração de Dados; Técnica de Clusterização; *Hierarchical Clusterer*; *Simple K-means*.

STUDY ON CLUSTERIZATION ALGORITHMS HIERARCHICAL CLUSTERER AND SIMPLE K-MEANS APPLIED IN THE GROUPING OF SIMILAR PATTERNS

ABSTRACT

The amount of information available on the Internet makes users seek in information technology tools to assist in the execution of their tasks, thus, the data mining technique emerges. The present study aims to implement and compare two Simple K-Means and Hierarchical Clusterer data mining algorithms, measuring their efficiency in identifying similar patterns between sub-area, keywords and academic articles, thus generating clusters based on in the similarity of interests between users and consulted works. It also seeks to elucidate the following problem: How Data Mining can contribute to the identification of similar patterns between large area, area, sub-area, keywords and academic articles. It is concluded with the study that the clustering algorithm Hierarchical Clusterer presented greater efficiency in the grouping of similar data in a database.

Keywords: Data Mining; Clustering Technique; Hierarchical Clusterer; Simple K-means.

Recebido em: 29/07/2018 - Aprovado em: 19/03/2019 - Disponibilizado em: 15/07/2019

INTRODUÇÃO

Atualmente, com a crescente quantidade de informações disponíveis na web observa-se que os usuários buscam por ferramentas que possam auxiliá-los na realização de suas tarefas e, para isso recorrem geralmente às tecnologias de informação.

A Internet passa a ser uma das maiores tecnologias que permitem a busca por informação, porém mesmo com ferramentas de busca o usuário tende a despender muito tempo nesta procura. Visando facilitar este acesso pela informação surgem as técnicas de Mineração de Dados (MD), uma das etapas do processo de *Knowledge Discovery in Database* (KDD), situado entre a preparação de dados e a interpretação dos resultados, objetivando buscar relacionamentos implícitos entre dados (MOZZAQUATRO, 2006).

O KDD pode ser definido como um processo que tem a capacidade de extrair de Bancos de Dados informações desconhecidas, válidas e utilizáveis a fim de auxiliar em uma tomada de decisão identificando e validando padrões novos, potencialmente úteis e compreensíveis em dados.

A mineração de dados atua sobre um grande volume de dados, utilizando uma enorme variedade de técnicas que de maneira automática fazem a exploração destes dados à procura de padrões e tendências, buscando relacionamentos implícitos entre as áreas de conhecimento (SOMMERVILLE, 2007).

Com a chegada da web 2.0, há uma grande demanda de informações e falta de mecanismos personalizáveis que se ajustem partir do perfil do usuário. Informações boas o suficiente para suprir necessidade do conhecimento por assuntos específicos, são difíceis de encontrar pelo simples fato de os motores de busca nem sempre responderem com o grau de qualidade esperado. O problema da falta de acesso à informação foi substituído pela necessidade de filtrá-la e apresentá-la de acordo com as necessidades dos usuários.

Neste contexto, expõe-se o problema de pesquisa: De que forma a Mineração de Dados pode contribuir para a identificação de padrões similares entre grande área, área, sub- área, palavras-chave e artigos acadêmicos.

As técnicas de Mineração de dados apresentam-se como solução ao problema abordado, ou seja, as mesmas filtram informações de acordo com o perfil de interesses dos usuários e assim recomendam itens que atendam as expectativas e necessidades dos mesmos (CAZELLA; REATEGUI, 2005).

Assim, o artigo tem o objetivo de implementar e comparar dois algoritmos de mineração de dados *Simple K-Means* e *Hierarchical Clusterer*, medindo sua eficiência na identificação de padrões similares entre sub-área, palavras-chave e artigos acadêmicos, gerando assim clusters baseando-se na similaridade de interesses entre usuários e obras consultadas.

TÉCNICA DE MINERAÇÃO DE DADOS CLUSTERIZAÇÃO

A técnica de Clusterização tem a função de unir dados em grupos com conteúdo semelhante e dados diferentes em grupos distintos gerando uma diferenciação de conteúdo em cada grupo (CASTANHEIRAS, 2008). Esta técnica é um tipo de padrão descritivo que agrupa objetos físicos ou abstratos em categorias ou grupos de objetos baseados em algum critério de similaridade, identificando aglomerações que descrevam os dados (MOZZAQUATRO, 2006).

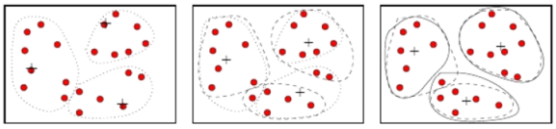
O processo é executado de forma automática por algoritmos de clusterização que tem a função de identificar características distinguíveis de um conjunto de dados. Neste contexto, pode-se citar as principais formas analisadas pelo processo: busca de similaridades, formação de conjuntos por diferenciação de classes e subclasses, como também a formação de diferentes estruturas.

A técnica de clusterização é dita um processo de mineração de dados que utiliza o aprendizado não supervisionado, pois os elementos que fazem parte da base de entrada não possuem o seu grupo definido. Muitas vezes nem mesmo o número de grupos é previamente definido.

Dentro dessa estrutura de informações que precisam de um aprendizado para gerar resultados pode-se citar algumas formas de condução destes dados como os algoritmos de agrupamento (*Simple k-Means*) e *Hierarchical Clusterer*.

O algoritmo *Simple K-Means* é um algoritmo que tem por função principal agrupamento de dados em k conjuntos diferenciados entre as especificações do grupo de dados. Funciona como uma localização por distância verificando caminhos viáveis minimizados. O algoritmo sugere a utilização de um dado como referência no espaço do conjunto para busca dos padrões. A partir deste momento começa a medição das diferenças de localizações de k conjuntos dos dados perante seu centro. Para localização é escolhido um representante chamado de centroide que tem por função servir de referência para o agrupamento. (MUNZ, 2007), (DUTRA, 2008). Após começa a formação dos conjuntos entre os espaçamentos minimizados e longos, quanto mais próximo estiver o dado, o mesmo se tornará parte daquele conjunto. Já os que estiveram mais distantes em relação ao centro formam outros k conjuntos. A Figura 1 ilustra as etapas do algoritmo *Simple K-Means*.

Figura 1 – Etapas do Algoritmo *Simple K-Means*



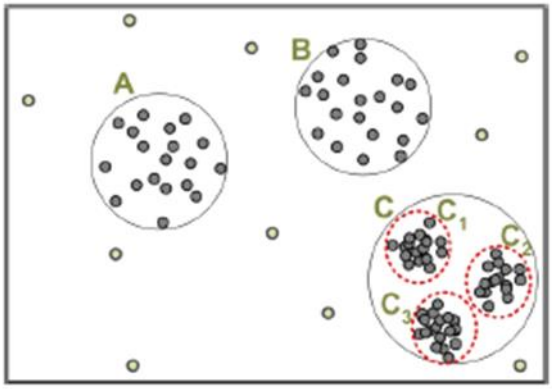
Fonte: Adaptado pelos autores.

O Algoritmo *Hierarchical Clusterer* possui uma abordagem que constrói os agrupamentos de modo que exemplos pertencentes ao mesmo cluster possuem alta similaridade e exemplos pertencentes a cluster diferentes possuem baixa similaridade. O resultado obtido difere do algoritmo *Simple K-Means*, pois não é constituído apenas de uma partição do conjunto de dados inicial, mas sim de uma hierarquia que descreve um particionamento diferentes à cada nível analisado (METZ, 2006).

Um conjunto de dados, geralmente, contém diversos clusters e esses clusters, por sua vez, são compostos de sub-clusters. Os sub clusters podem ainda ser formados a partir do agrupamento de outros clusters menores (sub-sub clusters), e assim sucessivamente.

Duas estratégias podem ser utilizadas para a implementação de algoritmos de clusterização hierárquicos: Aglomerativa (*botton-up*) e Divisiva (*top – down*). Na primeira, cada exemplo é considerado um cluster unitário. Em seguida, pares desses clusters são iterativamente agrupados de acordo com um índice de similaridade, até que todos os exemplos pertençam a apenas um cluster. Por outro lado, a abordagem decisiva é iniciada com apenas um agrupamento contendo todos os exemplos e procede dividindo o conjunto de exemplos em cluster cada vez menores, até que cada exemplo pertença exclusivamente a um cluster ou até que se alcance o critério de parada, frequentemente o número de clusters desejados (MURTAGH, 1983). A Figura 2 apresenta clusters com diferentes densidades.

Figura 2 – Algoritmo Hierarchical Clusterer: Clusters com diferentes densidades



Fonte: Adaptado pelos autores.

IMPLEMENTAÇÃO DO ALGORITMO *HIERARCHICAL CLUSTERER*

Neste algoritmo, os agrupamentos são obtidos pela representação dos clusters em uma estrutura conhecida como dendograma que consiste de um tipo especial de árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos. Dessa maneira, um agrupamento hierárquico agrupa os dados de modo que se dois exemplos são agrupados em algum nível, nos níveis mais acima eles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de clusters. Essa técnica permite analisar os clusters em diferentes n´níveis de granularidade, pois cada nível do dendograma descreve um conjunto diferente de agrupamentos.

Inicialmente, para instanciar os 49 (quarenta e nove) artigos científicos a base de dados foi dividida aplicando a regra use training set. Foi selecionado um conjunto de treinamento e dividido em duas partes: cerca de 70% e 30% dos dados utilizados para criar o modelo. Após, para testar a exatidão do algoritmo foi aplicada a regra *supplied test set* com os dados restantes, colocando-os em um conjunto de testes. Esta etapa da amostra de teste é fundamental devido ao problema do super ajuste, ou seja, caso sejam fornecidos dados em excesso na criação do modelo, o mesmo será, na verdade, perfeitamente criado, mas somente para esses dados. Caso se necessite prever valores futuros desconhecidos deve-se criar um conjunto de testes. A Tabela 1 apresenta a classificação do algoritmo apresentando respectivamente K=2, 5 e 10.

Tabela 1 – Agrupamento centroides (2, 5, 10)

	K= 2			K=5			K=10		
Agrupamento	M.T	1. Teste	Média	M. T	Modelo Teste	Média	M. T	Modelo Teste	Média
Instancias classificadas corretamente	17,3333%	14,8148%	16,07%	20%	22,2222%	21,1111%	22,67%	33,33%	28%
Instancias classificadas incorretamente	82,6667%	85,1852%	83,9259%	80%	77,7778%	78,8889%	77,33%	66,6667%	71,99%
Número total de instâncias	34	15		34	15		34	15	

Fonte: Elaborado pelos autores.

Conforme ilustra a Tabela 1, o algoritmo *Hierachical clusterer* agrupou com k=2 16,07% de instancias corretamente e 83,92% incorretamente. Após foi validada a amostra com alteração no centroide k=5 apresentando um agrupamento de 21,11% de instancias corretamente e 78,88 incorretamente. Finalmente foi validado o k=10 obtendo-se um agrupamento de 28% dos dados agrupados corretamente e 71,99% agrupados incorretamente.

A Tabela 2 apresenta a média geral do algoritmo *Hierarchical clusterer.*

Tabela 2 – Média Geral do algoritmo *Hierarchical clusterer*

Algoritmo	Instancias classificadas corretamente	Instancias classificadas incorretamente
Hierarchical clusterer	21,7037%	78,2696%

Fonte: Elaborado pelos autores.

Conforme a Tabela 2 observa-se que o algoritmo *Hierarchical Clusterer* apresenta um percentual de 21,7037% das instancias classificadas corretamente e 78,2696% classificadas incorretamente.

IMPLEMENTAÇÃO DO ALGORITMO *K-MEANS*

A técnica de clusterizacão permite ao usuário separar os dados em diversos grupos. Com a clusterização é possível dividir os conjuntos de dados em diversos clusters agrupando-os conforme seu grau de similaridade.

Para definir quão similar é um dado de outro, deve-se utilizar uma "medida de similaridade. Tal medida pode ser simplesmente definida como a distância Euclidiana entre os valores dos atributos de cada dado. Ou seja, quanto menor a distância entre tais valores, mais similares são esses dados (METZ, 2006).

Na etapa 1 cada cluster é representado por um centróide que representa o ponto central do cluster. O número k de clusters deve ser informado previamente pelo usuário, que definirá os elementos de cada agrupamento, comparando cada dado ao centroide de cada cluster. Para definição dessa similaridade de cada dado a cada centroide, são utilizadas funções de similaridade. Os dados foram agrupados utilizando o software Weka, o qual

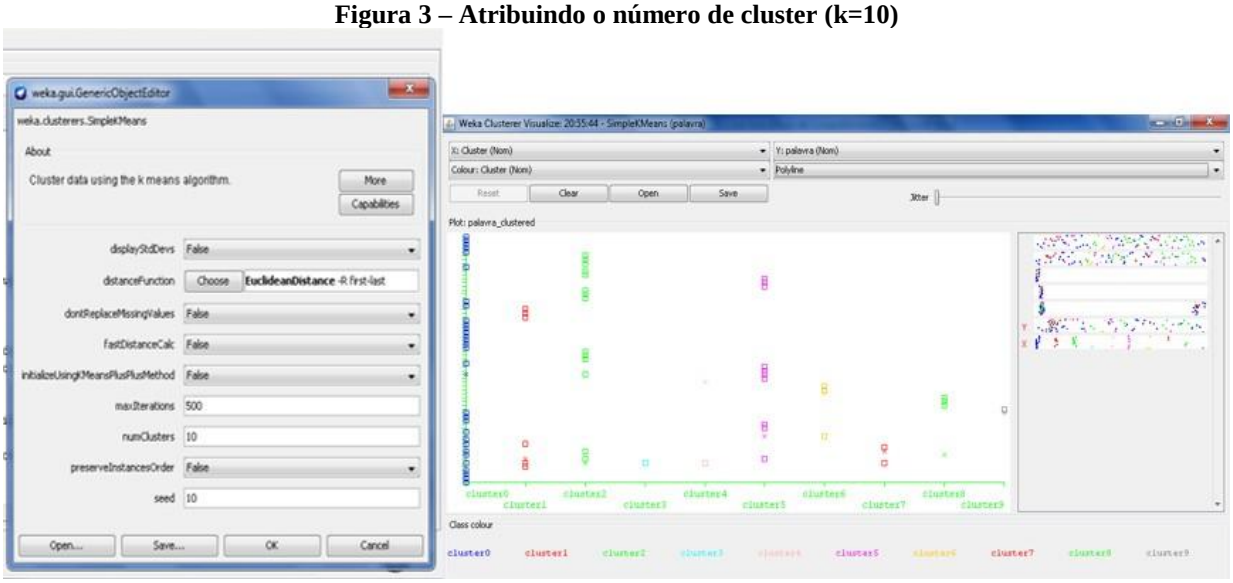
integra o algoritmo de clusterização *Simple K-means* na base de dados já pré-processada.

O usuário deve informar o valor de k, ou seja, o número de cluster a serem agrupados, o algoritmo define k centroides iniciais, que representarão os grupos que receberão os dados. Cada cluster, então, deve possuir um centroide que servirá como referência para que sejam definidos os clusters de cada dado, através do cálculo de similaridade com cada centróide. O software Weka recebe arquivos de entrada, contendo os dados (já formatados) a serem analisados, e aplica o algoritmo solicitado sobre tais dados.

Inicialmente, foi criada a base de dados na ferramenta Weka, ordenando que se clusterizassem todas as instâncias analisadas. Obteve-se 49 (quarenta e nove) artigos científicos, sendo processados e analisados pelo software Weka.

A similaridade tem que ser media para cada nova entrada de dados. A similaridade é calculada pela regra da distância Euclidiana que tem por objetivo formar clusters pela sua proximidade em relação a cada centroide pré-definido. Atribuiu-se o valor inicial de 10 (dez) para o k, ou seja, 10 centroides iniciais (10 clusters). Observou-se o que segue na Figura 3.

Figura 3 – Atribuindo o número de cluster (k=10)



Fonte: Elaborado pelos autores.

Conforme ilustra a Figura 3, O cluster oº cluster foi o que mais reuniu dados dentre todos os clusters. Isso aconteceu devido à similaridade da maioria das palavras – chave e sub-áreas integrantes de cada artigo científico, apresentando diferentes nuances que foram dispostas nos clusters restantes.

Foram analisadas 49 instancias relacionadas a artigos acadêmicos, palavras – chave e sub-áreas, ou seja, quais artigos apresentam palavras –chaves e subáreas

similares. Constatou-se que 82 instancias (80,3922%) foram classificadas incorretamente.

Observa-se que no cluster 0 (44 = 43%) dos artigos integram a palavra –chave Educação; no cluster 1 (10 = 10%) dos artigos integram a palavra Computação Móvel; cluster 2 (20 = 20%) integra a palavra chave AVA; cluster 3 (1= 1%) não foi agrupado; cluster 4 (3 = 3%) Web, cluster 5 (11= 11%) Informática na Educação, cluster 6 (4= 4%) Sistemas Operacionais, cluster 7 (3= 3%) Matemática, cluster 8 (4= 4%) Aprendizagem e cluster 9 (2= 2%) Objeto de Aprendizagem.

Importante frisar que o maior cluster do primeiro experimento (com 10 centróides iniciais) agrupou 44 dados (43%), enquanto que o maior cluster do segundo experimento (com 5 centróides) agrupou 68 dados (67%). Com um K = 2 agrupou-se 92 dados (90%). Isso mostra a efetividade do algoritmo, agrupando um número maior de cluster com um reduzido valor de K (centroides). A Tabela 3 apresenta o agrupamento dos dados no algoritmo *Simple k-means*.

Tabela 3 – Agrupamento dos dados no algoritmo *Simple K-Means*

	K= 2	K=5	K=10
Agrupamento			
Instancias classificadas corretamente	19,6078%	15,6863%	19,6078%
Instancias classificadas incorretamente	80,3922%	84,3137%	80,3922%
Número total de instâncias	49	49	49

Fonte: Elaborado pelos autores.

A Tabela 4 apresenta a média geral do algoritmo *Simple K-means.*

Tabela 4 – Média Geral do algoritmo *Simple K-means*

Algoritmo	Instancias classificadas corretamente	Instancias classificadas incorretamente
K-means	18,3021%	81,8679%

Fonte: Elaborado pelos autores.

Conforme a Tabela 4 observa-se que o algoritmo *Simple k-means* agrupou corretamente 18,3021% dos dados e 81,8679% incorretamente em uma amostra total de 49 instâncias.

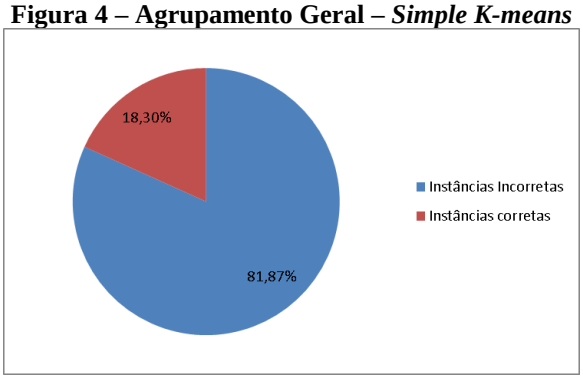
ESTUDO COMPARATIVO DOS ALGORITMOS *HIERARCHICAL CLUSTER* E *K-MEANS* APLICADOS NO AGRUPAMENTO DE INFORMAÇÕES SIMILARES

Esta seção é dedicada a apresentar um estudo comparativo entre os algoritmos de clusterização *Simple K-means* e *Hierarchical clusterer* aplicado no agrupamento de dados similares.

A pesquisa desenvolvida buscou aplicar os dois algoritmos citados no agrupamento da seguinte situação: relações entre sub-áreas,

artigos acadêmicos e palavras-chave. A Figura 4 ilustra o agrupamento geral obtido com a aplicação do algoritmo *Simple K-means* nas três diferenciações de centroides (2, 5, 10).

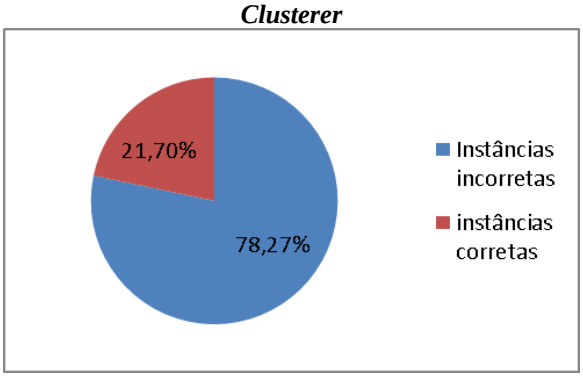
Figura 4 – Agrupamento Geral – *Simple K-means*



Fonte: Elaborado pelos autores.

Após testou-se o experimento com o algoritmo *Hierarchical clusterer* delimitando também o número de centroides de 2, 5 e 10, obteve-se o resultado apresentado na Figura 5.

Figura 5 – Agrupamento Geral – *Hierarchical Clusterer*



Fonte: Elaborado pelos autores.

Portanto, pode-se constatar que o algoritmo de clusterização *Hierarchical clusterer* aplicado no agrupamento de informações similares apresentou melhor desempenho comparado ao algoritmo *Simple K-means*. A Tabela 5 apresenta a média geral dos dois algoritmos validados.

Tabela 5 – Média Geral: *Hierarchical Clusteterer* e *Simple K-means*

Algoritmos	Instancias classificadas corretamente	Instancias classificadas incorretamente
K-means	18,3021	81,8679%
Hierarchical clusterer	21,7037%	78,2696%

Fonte: Elaborado pelos autores.

Conforme ilustra a Tabela 5 (em uma porcentagem de 100%) pode-se observar que o algoritmo *Simple K-means* agrupou corretamente 18,3021% dos dados, enquanto o algoritmo *Hierarchical Clusterer* agrupou corretamente 21,7037% dos dados corretamente. Observa-se um diferencial entre ambas, em média de 3,4016%. Assim comprova-se na prática que o algoritmo *Hierarchical Clusterer* se apresenta mais eficiente no agrupamento de dados similares (minerando texto) comparado ao *Simple K- Means*.

CONSIDERAÇÕES FINAIS

Este artigo apresentou um estudo sobre a técnica de mineração de dados clusterização implementando os algoritmos *Hierarchical Clusterer* e *Simple K-means*.

A partir da análise visual dos resultados fornecidos pelos citados, gerada pela ferramenta WEKA, foi possível visualizar e entender os dados constatando-se assim que o algoritmo de clusterização *Hierarchical Clusterer* apresentou maior eficiência no agrupamento de dados similares em uma base de dados. Tal resultado justifica-se por não ser constituído apenas de uma partição do conjunto de dados inicial, mas sim de uma hierarquia que descreve um

particionamento diferentes à cada nível analisado (METZ, 2006).

Acredita-se que com os resultados obtidos nesta pesquisa abrem-se novos campos de estudos relacionados à área.

REFERÊNCIAS

CASTANHEIRAS, G.LUCIANA. **Aplicação de Técnicas de Mineração de dados em Problemas de Classificação de Padrões.** Departamento de Engenharia Elétrica,Universidade Federal de Minas Gerais,2008.

CAZELLA, S. C., Reategui, E., **Mini-course: Recommender Systems**. In: Encontro Nacional de Inteligência Artificial (ENIA), 2005, São Leopoldo, RG, Brazil.

DUTRA,Vieira Luciano. **Seleção de Candidatos: Uma Estratégia para Incorporação da Distância de Mahalanobis no Algoritmo K-médias.7°** Brazilian Conference on Dynamics, Control and Applications-FTC-Unesp at Presidente Prudente,SP, Brasil, 2008.

METZ, Jean. **Interpretação de Clusters gerados por algoritmos de clustering Hierárquico.** Instituto de Ciências Matemática e de Computação. Dissertação de Mestrado, São Paulo, 2006.

MOZZAQUATRO, Patrícia Mariotto. **Estudo da Aquisição e Modelos de Usuários da Biblioteca Digital Acadêmica**. Trabalho de Conclusão de Curso em Sistemas de Informação. Universidade Luterana do Brasil - ULBRA, 2006.

MUNZ, Gerhard; LI, Sa; CARLE, Georg. **Traffic anomaly detection using K-means clustering.** In: Proceedings of Leistungs-,Zuverlassigkeits-Undverlasslichkeitsbewertung Von Kommunikationsnetzen Und Verteilten systemem, 4.GI/ITG-Worshop MMBnet 2007, Hamburg, Germany, September 2007.

MURTAGH, F. **A survey of recent advances in hierarchical clustering algorithms**. The Computer Journal, 26, 1983.

SOMMERVILLE, Ian. **Engenharia de Software**. São Paulo – Pearson Addison Wesley, 2007.

Flávio de Brum
Especialista em Segurança da Informação e Gestão de Riscos pela Faculdade Meridional – IMED, Graduado em Ciência da Computação pela Universidade de Cruz Alta – UNICRUZ.

Jocias Maier Zanatta
Doutorando em Administração pela Universidade Federal de Santa Maria – UFSM, Mestre em Desenvolvimento pela Universidade Regional do Noroeste do Estado do Rio Grande do Sul - UNIJUI, Especialização em Gestão Financeira, Controladoria e Auditoria e Graduação em Administração pela Sociedade Educacional Três de Maio - SETREM.

Patricia Mariotto Mozzaquatro
Doutoranda em Modelagem Matemática pela Universidade Regional do Noroeste do Estado do Rio Grande do Sul – UNIJUI, Mestre em Ciência da Computação e Especialista em Tecnologias da Informação e Comunicação aplicadas a Educação pela Universidade Federal de Santa Maria – UFSM, e Graduada em Sistemas de Informação pela Universidade Luterana do Brasil – ULBRA. Professora da Universidade de Cruz Alta – UNICRUZ.