TECNO-LÓGICA
Revista do depto. de Química e Física, do depto. de Engenharia, Arquitetura e Ciências Agrárias e do Mestrado em Tecnologia Ambiental
http://online.unisc.br/seer/index.php/tecnologica

SISTEMA DE RECOMENDAÇÃO DE PRODUTOS UTILIZANDO MINERAÇÃO DE DADOS

Cássio Alan Garcia[1*], *Rejane Frozza*[1,2]

[1]*Departamento de Informática - Universidade de Santa Cruz do Sul, Av. Independência, 2293, Santa Cruz do Sul, Rio Grande do Sul, Brasil*

[2]*Programa de Pós-Graduação em Sistemas e Processos Industriais (Mestrado) - Universidade de Santa Cruz do Sul, Av. Independência, 2293, Santa Cruz do Sul, Rio Grande do Sul, Brasil*

**E-mail: cassioalangarcia@gmail.com*

Recebido em 24 /12/ 2012
Aceito em 28/02/ 2013

RESUMO

Este artigo apresenta o estudo da técnica de mineração de dados aplicada aos sistemas de recomendação e o desenvolvimento de um sistema computacional, para indicação de produtos a usuários conforme suas preferências e características. Com a utilização de técnicas de Mineração de Dados, Clusterização e Regras de Associação, advindas da área de Inteligência Artificial, o sistema desenvolvido visa integrar essas duas técnicas, gerando um sistema híbrido que fornece informações para a geração de recomendações.

Palavras-chave: Mineração de dados, Sistemas de Recomendação, *e-commerce*.

1 Introdução

Sistemas de Recomendação vêm sendo de grande valia em ambientes como, por exemplo, os e-commerces (comércio eletrônico), tornando possível conhecer melhor os interesses e hábitos do consumidor, personalizando o relacionamento entre cliente e empresa [1]. A recomendação de um produto pode ser um fator importante para a atração ou então perda de um cliente.

Em ambientes de busca, devido ao grande volume de informações na Internet, os Sistemas de Recomendação podem ajudar a refinar os resultados e minimizar o tempo de busca, fornecendo um retorno de pesquisa de forma individualizada e ágil, apresentando informações realmente relevantes ao usuário [2].

A Mineração de Dados, através de suas técnicas de filtragem e descoberta de padrões, auxilia na localização de usuários com interesses semelhantes. Isso ocorre através das descobertas de padrões nos dados, que dão suporte a Filtragens Colaborativas, por exemplo. Além de auxiliar, ela torna-se uma necessidade, devido ao grande volume de informações.

O objetivo principal deste trabalho foi desenvolver um sistema que se fundamenta em Sistemas de Recomendação, utilizando-se de Mineração de Dados. As tarefas de Data Mining estudadas mais profundamente foram as de Clustering e Regras de Associação, para o desenvolvimento de algoritmos que gerem dados de entrada para um Sistema de Recomendação. Para isso, foram estudados assuntos relacionados à data mining, e sistemas de recomendação, bem como pesquisados trabalhos relacionados a essas áreas.

Este trabalho traz contribuições científicas para a área de Inteligência Artificial, no sentido de mostrar a possibilidade da integração das tarefas de Clusterização e Regras de Associação na área de Mineração de Dados, bem como a possibilidade da criação de um sistema híbrido de Inteligência Artificial – integrando Data Mining e um Sistema de Recomendação.

Em busca de suprir essas necessidades foi desenvolvido um sistema que recomenda produtos a clientes, com base no seu histórico de compras que é comparado com compras de outros clientes, buscando, dessa forma, prever o que o usuário poderia vir a se interessar e que é apresentado neste artigo.

2 Mineração de Dados e Sistema de Recomendação

Esta seção apresenta um embasamento sobre as áreas envolvidas nesta pesquisa, tais como conceitos e técnicas de Inteligência Artificial, a qual abrange Mineração de Dados e Sistemas de Recomendações. Em Mineração de Dados, o estudo está focado em duas tarefas: Clusterização (Clustering) e Regras de Associação. A Figura 1 ilustra estes níveis conceituais.

Figura 1 - Sistema Híbrido

2.1 Minerações de Dados

A Mineração de Dados é, muitas vezes, considerada como sinônimo de KDD (Knowledge Discovery in Databases), Descoberta (ou Extração) de Conhecimento em Bases de Dados, mas representa uma etapa deste grande processo.

"Extração de Conhecimento em Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados". Para isso, são utilizadas técnicas de diversas áreas do conhecimento, como estatística, matemática, bancos de dados, inteligência artificial, visualização de dados e reconhecimento de padrões [3][4].

Os principais termos envolvidos nesse processo são [5]:

• Dados: Conjunto de fatos ou casos em uma base de dados.
• Padrões: Abstrações de um conjunto de dados em uma linguagem descritiva de conceitos.
• Processo: Busca de padrões e avaliação do conhecimento, sendo dividida em várias etapas.
• Válidos: Resultados devem satisfazer restrições/testes pré-definidos.
• Novos: Os padrões encontrados devem ter certo grau de novidade.
• Compreensíveis: Deve haver uma forma de análise mais profunda dos dados e padrões descobertos.
• Conhecimento: Fortemente relacionado com medidas de utilidade, originalidade e compreensão.

Ao ser explorada a área de Descoberta de Conhecimento, remete-se aos níveis hierárquicos da informação que são descritos a seguir [6]:

• Dados: Classe mais baixa da informação, constituindo representações de fotos, textos, gráficos imagens, sons, entre outros. Ou seja, dados são sinais que não foram processados, integrados, avaliados ou interpretados, constituindo assim a matéria-prima para a produção de informação.
• Informação: A informação são os dados que passaram por um processamento e que podem, dessa forma, ser compreensíveis às pessoas. Podem ser classificados como informação a exibição de um arquivo textual ou gráfico, uma fotografia revelada, entre outros. Através da interpretação e integração de vários dados e informações, obtém-se o conhecimento.
• Conhecimento: São informações analisadas e avaliadas, sendo confiáveis, relevantes e importantes. É fruto da combinação de informações, sendo por meio deste que pessoas responsáveis por tomadas de decisões buscam uma compreensão mais efetiva da situação do problema.
• Inteligência: É o conhecimento que foi sintetizado e aplicado a uma determinada situação para entendê-la melhor, ou seja, a informação sintetizada e relevante ao contexto do problema.

Considerando tais níveis da informação, a mineração de dados atua sobre os dados armazenados nas bases, buscando por padrões úteis e compreensíveis para que seja gerado o conhecimento.

O processo de KDD é composto pelas seguintes etapas [5]: Identificação do problema, Pré-Processamento, Extração de Padrões, Pós-Processamento, conforme Figura 2.

Figura 2 - Etapas do processo de Descoberta de Conhecimento em Bases de Dados [5]

Na Identificação do Problema, são determinados os objetivos e metas a serem alcançados no processo de Data Mining através do estudo do domínio da aplicação, mostrando-se necessário o pleno conhecimento deste [5].

Na segunda etapa, chamada Pré-Processamento, trata-se da limpeza dos dados e redução de volume, para que seja reduzido o tempo de processamento e a utilização de memória feita pelos algoritmos de mineração [5].

A Extração de Padrões é a fase de Mineração de Dados propriamente dita. Nesta fase é escolhida a tarefa e definido o algoritmo a ser utilizado, podendo ser executado mais de uma vez, já que esta etapa é um processo iterativo, para que haja a extração de padrões [5].

O Pós-Processamento é o momento em que os conhecimentos são interpretados e utilizados em processos de tomada de decisão ou em Sistemas Inteligentes, sendo possível retornar a qualquer etapa anterior.

As medidas de desempenho (precisão, velocidades, entre outras) também são executadas nesta fase, podendo, caso necessário, ajustar parâmetros e voltar a alguma etapa anterior para ser executada novamente.

2.1.1 Tarefa de Mineração de Dados

Em função do objetivo a ser alcançado é feita a escolha da tarefa, que pode ser classificada em preditiva ou descritiva [5], para aplicação sobre a base de dados.

Nas tarefas preditivas, a abordagem é botton-up, ou seja, a pesquisa é feita de forma a encontrar padrões frequentes, tendências e generalizações, a fim de encontrar informações implícitas nos dados [7].

As tarefas descritivas possuem abordagem top-down, onde existem hipóteses previamente formuladas que são testadas para a verificação da sua veracidade. Esta abordagem busca encontrar respostas que confirmem ou neguem as hipóteses, enquanto que a anterior revela informações que não haviam sido imaginadas, gerando as próprias hipóteses [7].

A escolha da(s) tarefas(s) a ser(em) utilizada(s) na etapa de Mineração de Dados é feita em função dos objetivos a serem alcançados, por isso não há como definir uma tarefa que seja mais eficiente em qualquer situação. Uma vez definida a tarefa, escolhe-se a técnica a ser empregada.

A seguir, são descritas algumas tarefas comumente utilizadas:

- Classificação (tarefa de predição)

A tarefa de Classificação mapeia dados de entrada em um número finito de classes, de forma a criar uma relação de cada exemplo com certa classe. Utiliza-se destas relações para predizer a classe de um novo e desconhecido exemplo [5].

É exemplo a classificação de clientes de um banco, classificando-os em possíveis bons pagadores ou maus pagadores, podendo, com isso, determinar se deve conceder crédito ao cliente ou não [3].

- Regressão (tarefa de predição)

Esta tarefa é bastante semelhante à anterior, diferenciando-se apenas no fato de que o atributo a ser predito é contínuo em vez de discreto [5].

O objetivo da tarefa de regressão que é "encontrar a relação entre um conjunto de atributos de entrada e um atributo-meta contínuo". Ainda cita o seguinte exemplo: sendo o atributo de entrada X = {x1, x2, ..., xn}, e y o atributo-meta, a tarefa de regressão procura buscar um mapeamento y = f(x1, x2, ..., xn} [4].

- Clustering ou Agrupamento (tarefa de descrição)

Busca detectar a existência de diferentes grupos, ou *clusters*, dentro de um determinado conjunto de dados, baseando-se em medidas de similaridade ou modelos probabilísticos, determinando quais são estes grupos, caso existam, dividindo grupos heterogêneos em sub-grupos homogêneos. Dados com atributos (ou um subconjunto de atributos escolhidos) parecidos são agrupados/segmentados no mesmo *cluster*, podendo ainda um dado estar classificado em mais de um *cluster* [5][8].

Clustering pode ser considerada a mais importante técnica de aprendizagem não supervisionada dentre as demais técnicas desse tipo, trata-se de encontrar uma estrutura em uma coleção de dados não classificados. Agrupamento é o processo de organização de objetos em grupos cujos membros são semelhantes de alguma forma [9].

- Regras de associação (tarefa de descrição)

Esta tarefa faz o levantamento de quanto um conjunto de atributos contribui para a presença de outro conjunto. Na área de marketing, é também conhecida como "análise de cestas de venda", havendo um estudo de como os itens estão relacionados. Pode ser aplicada em estudos de preferências, tentando descobrir afinidades entre itens, para, por exemplo, criar pacotes de vendas para os consumidores. Essa atividade serve como exemplo para a teoria de que a presença de um item em uma transação implica na presença de outro, sendo que o banco de dados é visto como uma coleção de transações em que cada uma envolve um conjunto de itens [10].

A regra de associação possui dois lados, o direito e o esquerdo (X ➔ Y) que significa que se X existe em alguma transação, há uma determinada possibilidade de Y existir também [5].

- Sumarização (tarefa de descrição)

Segundo Rezende (2003), "a Sumarização envolve métodos para encontrar uma descrição compacta para um subconjunto de dados". Tal descrição identifica e apresenta de forma concisa e compreensível as principais características dos dados em um conjunto de dados. A técnica de visualização é uma função de Sumarização que é necessária para se obter um entendimento intuitivo do conjunto de dados, fazendo uso de diagramas, baseados em proporção e dispersão, histogramas, entre outros [5][8].

Com tarefas como as supracitadas, os Sistemas de Recomendação fornecem uma alternativa às interfaces das tecnologias de filtragem e recuperação de informações, diferenciando-se pela predição dos conteúdos interessantes e úteis ao usuário [11].

2.2 Sistema de Recomendação

Um Sistema de Recomendação (SR) busca criar um ambiente personalizado para cada usuário. O caso mais conhecido é o de comércio eletrônico (e-commerce), com base nas informações absorvidas pelo sistema de forma explícita e/ou implícita, sendo que a primeira se dá através de, por exemplo, um cadastro de usuário, em que ele define seus gostos e preferências. Já a segunda forma se dá através de análise de comportamento do usuário, como produtos adquiridos anteriormente, caminhos (links) percorridos pelo usuário dentro do site (logs), *rating* de produtos, entre outros. Com o uso desses dados, SR podem recomendar produtos, informações, serviços ou pessoas.

Os SR podem ser classificados, quanto a sua forma, em dois grupos: Baseados em Conteúdo, que recomendam itens semelhantes àqueles relacionados com um usuário alvo, recomendando itens individualmente e partindo do princípio de que usuários tendem a se interessar por itens similares aos que demonstraram interesse anteriormente; e Colaborativos que se relacionam com usuários que possuem interesses em comum com o usuário em questão, efetuando a recomendação a grupos de usuários semelhantes, levando em conta que é comum pessoas recomendarem ou pedir recomendação de itens de qualquer natureza, ficando, assim, a encargo do SR fazer o processo "boca-a-boca" [12][13].

2.2.1 Coleta de dados implícita e explícita

A busca de informações sobre o usuário ocorre de forma explícita e/ou implícita.

Na primeira forma, o usuário preenche um formulário ao se cadastrar, referente a dados pessoais, preferências, interesses, ou de alguma forma indica espontaneamente o que lhe interessa. Esse processo demanda tempo do usuário, sem falar na inexistência de confiança no sistema por parte do usuário para que forneça informações pessoais.

Já na coleta de informações implícita, é interpretado, por exemplo, o comportamento (navegação) do usuário dentro do sistema, obtendo-se assim informações sobre suas necessidades e preferências. Isto sem a interferência do usuário e sem a implicação de ter de estar preenchendo formulários para que sejam feitas as recomendações [1].

2.2.2 Tipos de Sistemas de Recomendação (ou Técnicas de Filtragem)

Sistemas de Recomendação, na área comercial, buscam a fidelidade dos clientes para assim aumentar a lucratividade das empresas. Sendo assim, criaram-se estratégias de recomendação para se alcançar esses objetivos. De forma ampla, podem-se ser citadas as estratégias de listas de recomendação, em que são criadas listas de itens sem a análise mais profunda dos dados, recomendadas listas de itens mais vendidos, por exemplo, tendo como principal vantagem a simplicidade de implementação e por outro lado a desvantagem de que são listas fixas para todos os clientes; as avaliações de usuário são outro tipo de estratégia em que os usuários avaliam e comentam itens; outra forma de recomendação é a criação de uma página exclusiva para isto [14].

A filtragem baseada em conteúdo tem esse nome pelo fato de os sistemas que a utilizam realizam uma filtragem baseada em análises dos conteúdos dos itens que podem ser recomendados com base no perfil do usuário [11]. O perfil dos itens é composto por alguns atributos que o descrevem e é utilizado para ser aplicada uma função de similaridade e, com base nisso, recomendar conteúdo ao usuário.

Essa metodologia é amplamente aplicada nas áreas de recomendação de textos, sendo que são geradas de forma automática descrições dos itens para serem comparados com os interesses do usuário, a fim de verificar a relevância deste item.

A filtragem baseada em conteúdo é mais indicada para a recomendação de textos (artigos, páginas da web), pois é possível verificar a similaridade com os interesses do usuário ao identificar termos comuns entre o texto e estes interesses. Já a aplicação desta filtragem na recomendação de produtos já se torna mais difícil, pois se deve avaliar atributos (características como cor, peso, preço, marca) destes produtos para serem recomendados [1]. Isto tudo, considerando-se que os usuários tendem a se interessar por itens semelhantes aos que já procuraram anteriormente.

A Filtragem Colaborativa foi desenvolvida para suprimir os pontos fracos da filtragem baseada em conteúdo, por não exigir nenhum tipo de descrição dos itens, mas sim se baseando na troca de experiências entre as pessoas que possuem interesse em comum, sendo os itens filtrados de acordo com avaliações dos demais usuários [1].

O principal fator que difere um sistema de recomendação colaborativo de outro é a forma como é calculada a similaridade entre usuários. Como a Filtragem Colaborativa se dá através de avaliações explícitas de itens feitas pelos usuários, os que avaliam de forma semelhante os mesmos conteúdos são considerados usuários com preferências similares, ficando claro então que o conteúdo que um usuário do grupo X gostou, será também do gosto dos demais usuários pertencentes a este grupo. Neste caso, trata-se de um sistema personalizado [16].

A Tabela 1 faz uma breve comparação entre as filtragens estudadas, relacionando suas vantagens e desvantagens.

Tabela 1: Vantagens e desvantagens dos dois principais tipos de Filtragens

SISTEMAS DE RECOMENDAÇÃO		
FILTRAGEM	VANTAGENS	DESVANTAGENS
Baseada em Conteúdo	- Bons resultados para usuários incomuns - Independe do número de usuários para haver boa recomendação	- Baixo desempenho devido à falta de informações no momento inicial do sistema - Não há relacionamento entre usuários
	- A qualidade das recomendações melhora com o tempo	- Dificuldade para mapear arquivos multimídia, bom como texto com sinônimos
Colaborativa	- Relacionamento entre usuários - Recomendação de itens com base no histórico de outros usuários relacionados	- Baixo desempenho devido à falta de informações no momento inicial do sistema - Baixo desempenho se o usuário não tiver uma quantidade considerável de relacionamentos
		- Quando um item é adicionado ao sistema e ainda não foi classificado por nenhum usuário, não é recomendado

3 Trabalhos relacionados

Trabalhos relacionados aos temas de Sistemas de Recomendação e de Mineração de Dados foram pesquisados e estudados, a fim de se construir uma tabela comparativa, a partir de critérios definidos.

Os critérios julgados mais relevantes são quanto à forma de coleta de informações dos usuários (implícita/explícita), tipo de filtragem utilizado no sistema (colaborativa, baseada em conteúdo, outra), estratégia de recomendação, técnica utilizada (relacionada às tarefas de mineração de dados: regras de associação, agrupamentos, classificação, outra).

Os trabalhos analisados relacionados a Sistemas de Recomendação e Mineração de Dados são listados na Tabela 2 e Tabela 3, respectivamente [15].

Avaliando-se tais informações, pode-se destacar que, na área de sistemas de recomendação, o tipo de filtragem mais adotado foi o Colaborativo. Nos casos em que a filtragem Baseada em Conteúdo foi utilizada, também se utilizou a Colaborativa, resultando em uma filtragem Híbrida. Quanto às coletas de informações, aplica-se geralmente a Implícita, sendo acompanhada muitas vezes da Explícita, como suporte ao *startup* do sistema.

No que se refere aos trabalhos relacionados à Mineração de Dados, a abordagem foi em sua totalidade a preditiva; a tarefa, em sua grande maioria, foi a de Classificação de objetos alvos e predominou o uso da técnica de árvores de decisão e algoritmo A priori [15].

4 Sistema CPM e Resultados

Para o desenvolvimento do sistema CPM (Customer Preferences Mining) tomou-se como base uma arquitetura híbrida fundamentada em técnicas de Inteligência Artificial – Sistema de Recomendação e Mineração de Dados.

O problema a ser abordado está no fato de que em um ambiente de vendas, não basta apenas oferecer produtos de forma organizada, separados por categorias, de fácil acesso ao consumidor. Faz-se necessário oferecer produtos que possam vir a interessá-lo, buscando-se, assim, um aumento da lucratividade. Para realizar tais indicações de produtos, avalia-se o perfil deste consumidor.

A forma escolhida para resolver esse problema foi gerar a recomendação de itens com base no seu histórico de compras, comparando com compras realizadas por outros usuários, podendo ser avaliados todos os clientes, ou apenas clientes que pertençam a um mesmo cluster.

Para a elaboração desta solução, em um primeiro momento, foi realizado o estudo referente à fundamentação teórica com base em uma ampla pesquisa bibliográfica, buscando fundamentar conceitos e técnicas de Mineração de Dados e Sistemas de Recomendação, bem como um levantamento de trabalhos e sistemas já existentes nestas áreas. Após, escolheu-se as tecnologias a serem adotadas para a implementação do protótipo proposto, partindo-se para o desenvolvimento do sistema e análise dos resultados dos experimentos realizados.

A figura 3 representa os procedimentos metodológicos da pesquisa e desenvolvimento. Cada uma destas etapas é detalhada a seguir.

Tabela 2: Comparativo entre os Trabalhos Relacionados a Sistemas de Recomendação

Projeto	**Coleta de Dados**	**Tipo de Filtragem**	**Estratégia de recomendação**
GroupLeans	Explícita	Colaborativa	Correlação entre usuário com mesmas atribuições de notas
Ringo	Explícita	Colaborativa	Correlação entre usuário com mesmas atribuições de notas
Fab	Explícita (notas) Implícita (itens lidos)	Híbrida (Colab + Cont) e Agentes	Várias etapas de filtragens
MovieLens	Explícita	Colaborativa	Lista dos cinco melhores filmes
TeamWorks	Implícita (no fluxo de informações)	Colaborativa	Filtragem de documentos irrelevantes
QuickStep	Implícita	Híbrida (Colab + Cont)	Ontologia para capturar preferências de usuários
Amazon.com	Explícita e Implícita	Híbrida (Colab + Cont)	Lista de Recomendação Avaliações e Comentários Itens Semelhantes Associação por conteúdo Email
CDNow.com	Explícita e Implícita	Híbrida (Colab + Cont)	Várias formas de recomendação
RecDoc	Explícita e Implícita	Híbrida (Colab + Cont)	Recomendação baseada em conteúdo no momento da consulta; Recomendação colaborativa em off-line.
Saraiva	Explícita	Híbrida (Colab + Cont)	Lista de Recomendação Avaliações e Comentários Itens Semelhantes Associação por conteúdo
Submarino	Explícita e Implícita	Híbrida (Colab + Cont)	Lista de Recomendação Avaliações e Comentários Itens Semelhantes Associação por conteúdo Email

Tabela 3: Comparativo entre os Trabalhos Relacionados à Mineração de Dados

Projeto	Abordagem	Tarefa Utilizada	Técnica	Algoritmo
Azarias (2009)	Preditiva	Classificação	Árvores de Decisão	-
Queiroga (2005)	Preditiva	Classificação	Árvores de Decisão	-
Adeodato et al (2005)	Preditiva	Classificação	Árvores de Decisão	A Priori + RNA Backpropagation
ZEE-MG	Preditiva	Regra de Associação	-	A Priori

(-) não localizado

A pesquisa bibliográfica é o ponto inicial deste projeto, passando-se então para a definição das tecnologias adotadas para o desenvolvimento da solução, sendo que a plataforma de desenvolvimento escolhida foi o. Net Framework 4.0 na linguagem C Sharp (C#), utilizando o IDE Microsoft Visual Studio 2010.

Esse sistema está dividido em três módulos: o motor de recomendação; interfaces usuário-sistema para solicitação e visualização das recomendações a serem geradas pelo motor e configurações; e base de dados.

A Figura 4 apresenta o fluxo de informações entre os módulos e o processamento do sistema de recomendação proposto.

O motor de recomendação é responsável pela aplicação das técnicas de Mineração de Dados (Clustering e Regras de Associação) sobre a base de dados e geração das recomendações; no ambiente de configuração o administrador do sistema informa os parâmetros para a realização da Clusterização (melhor explicada a seguir), geração das regras de associação e das recomendações; a interface principal é o meio de entrada de pedidos e saída de recomendação.

Para ser gerada a recomendação final, são definidas quatro grandes etapas no sistema, que são ilustradas na Figura 5.

A primeira etapa é a de configuração do sistema, em que o administrador define os parâmetros para a aplicação dos algoritmos. Dentre essas configurações, está o número máximo de recomendações a serem geradas e o índice mínimo de Suporte e Confiança, necessários para a geração das Regras de Associação. A Figura 6 mostra a interface de configurações do protótipo. Conforme descrito na seção 2.1.4 – Regras de Associação – os coeficientes de Suporte e Confiança referem-se à quantidade de vezes em que o lado esquerdo da regra precisa aparecer dentre as transações (Suporte), e a quantidade de vezes que o lado direito da regra deve ocorrer nas transações que o lado esquerdo da regra atende (Confiança).

Figura 3 – Procedimento de execução do trabalho

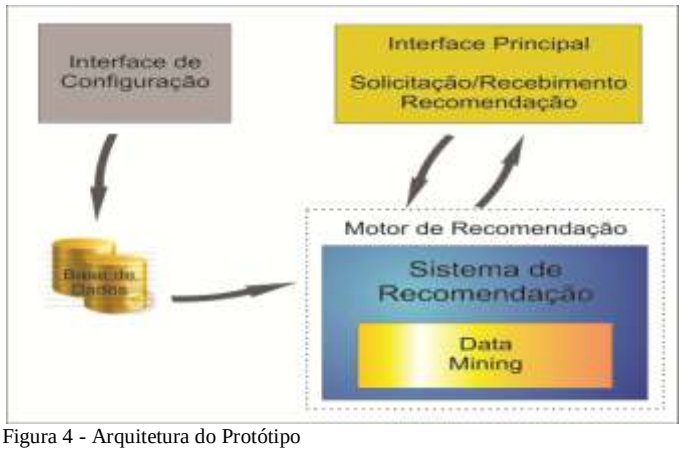

Figura 4 - Arquitetura do Protótipo

Figura 5 – Ordem de Execução para geração da recomendação

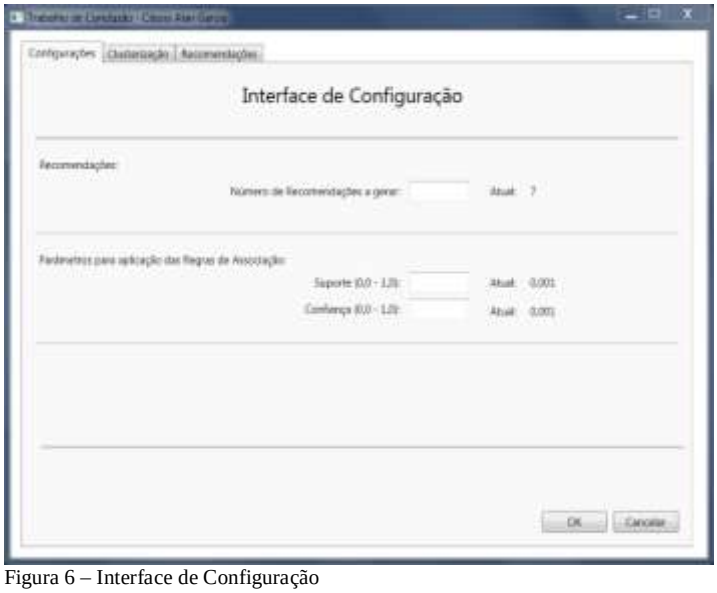

Figura 6 – Interface de Configuração

O sistema cria grupos de usuários semelhantes, através da tarefa de *Clustering* da Mineração de Dados. Como parâmetro de entrada, é informado, na tela de Clusterização (Figura 7), o número de *Clusters* a serem gerados. Esse número precisa ser avaliado pelo administrador, sendo que quanto mais *clusters* forem criados, maior será a especialização de cada um, ou seja, mais restritas serão as características dos clientes em cada *cluster*. Uma desvantagem de um número elevado de *clusters* está no menor número de clientes em cada *cluster*, podendo, assim, obter-se um restrito número de regras geradas ao se considerar apenas clientes de um mesmo *cluster*. Porém tendo a vantagem de tais regras serem mais específicas ao cliente que estará recebendo as recomendações.

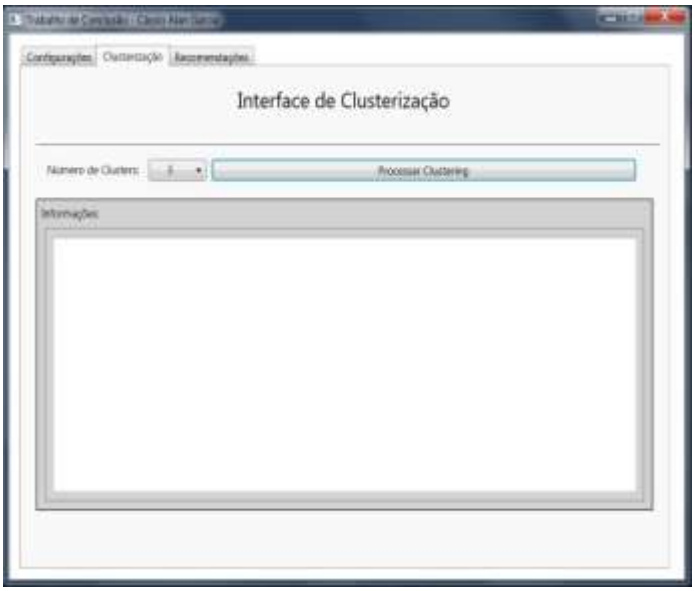

Figura 7 – Interface de Clusterização

Nesta etapa são criadas as Regras de Associação advinda da Mineração de Dados, com base no histórico de compras, levando em consideração apenas os usuários do mesmo *cluster*, criado na etapa anterior, ou considerando todos os clientes – conforme o solicitado na interface de Recomendações (Figura 8).

Mediante solicitação, o sistema recomenda itens ao usuário escolhido, tendo como base as Regras de Associação geradas. De acordo com a seção 3.4 – Tipos de Sistema de Recomendação, tais abordagens resultam em uma Filtragem Colaborativa, já que são utilizadas informações de outros clientes na tentativa de prever o que seria de interesse ao cliente em questão. Na geração de recomendações, mais de uma Regra de Associação pode indicar a recomendação de um mesmo produto. Caso o produto já tenha sido recomendado, tais regras são desconsideradas, visto que as regras são avaliadas seguindo o índice de Confiança do melhor para o pior (maior índice para o menor).

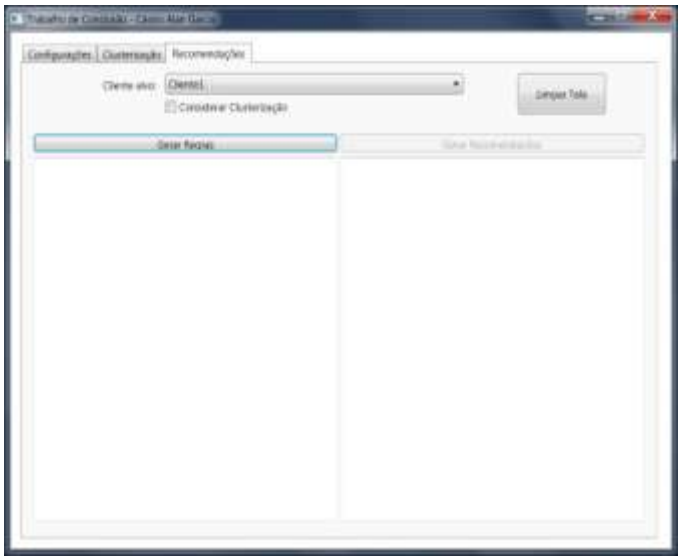

Figura 8 – Interface de Recomendações

Um Sistema de Recomendação funciona como um servidor, capaz de recuperar preferências do usuário, seja de forma implícita ou explícita, para fornecer conteúdo relacionado a itens nos quais já demonstrou interesse [14].

Nesse protótipo, a coleta de dados é feita de forma implícita, com base no histórico de transações dos usuários.

Como o foco do desenvolvimento está no motor de recomendação, não há implementação de interface para os clientes definirem explicitamente seus interesses (coleta de dados explícita).

Para a execução do protótipo desenvolvido, foi modelada uma base de dados que supre as necessidades dos algoritmos para fins de geração de recomendações, dispensando-se informações não úteis à resolução do problema.

Para a realização da tarefa de Clusterização, é utilizado o algoritmo K-means, implementado da seguinte forma:

Foi desenvolvida uma Stored Procedure que realiza esta tarefa, seguindo o princípio do algoritmo K-means, tendo como parâmetro de entrada uma variável K, que define o número de *clusters* que serão criados. Esse número precisa, necessariamente, ser inferior ao número de clientes, para que não exista *cluster* vazio.

Para isto, a Stored Procedure utiliza os K primeiros clientes da base de dados como semente para a realização do processo. Então para cada cliente na base de dados, encontra-se a semente mais próxima, para que seja o grupo (*cluster*) deste cliente.

Quando todos os clientes são associados a um *cluster*, determinam-se novos centroides – que é o ponto médio de cada grupo, substituindo as sementes geradas no início do processo. Com esses novos centroides, o processo se repete, associando novamente os clientes ao grupo do centroide mais próximo de si e calculando novos centroides.

Esse processo finaliza quando nenhum cliente trocar de *cluster*, ou seja, os clusters atingem um formato estável, chegando-se a uma clusterização final.

Em auxílio a esta Stored Procedure é utilizada uma função SQL que calcula a distância de um ponto a outro. Essa função é desenvolvida para encontrar o centroide que possua a menor distância até o cliente.

Para definir a distância entre um cliente e outro, são avaliados dois atributos: o valor total em compras realizadas pelo cliente e a média de valores dos itens comprados, gerando-se assim um gráfico bidimensional. Vale ressaltar que podem ser utilizado quantos atributos julgar-se necessário. Para fins de ilustração, foram adotadas duas dimensões.

No desenvolvimento da geração de Regras de Associação, inicia-se buscando todos os produtos, calculando o suporte destes e selecionando apenas aqueles que atendam o suporte mínimo definido na tela de configurações do protótipo.

Depois de realizado este filtro, são geradas combinações de itens, cruzando todos os itens, calculando seu suporte e aplicando o mesmo filtro, permitindo apenas combinações que atendam o suporte mínimo definido. Nessa etapa, são gerados todos os grupos possíveis de produtos. Por exemplo, em uma base que contenha 3 produtos {a, b, c}, são geradas as combinações {a}, {a, b}, {a, c}, {b, c} e {a, b, c} e calculado o índice de suporte destes grupos, sendo que neste momento, a ordem dos itens não é relevante.

Após, para as combinações que atendem o suporte mínimo exigido, são geradas sequências de produtos, em que, por sua vez, a ordem dos itens é importante para o próximo passo. Exemplificando, ao considerar o grupo de produtos {a, b, c} listado acima, são geradas as seguintes sequências: {a, b, c}, {a, c, b}, {b, a, c}, {b, c, a}, {c, a, b} e {c, b, a}.

No próximo passo, são efetivamente criadas as Regras de Associação. Sendo que para cada sequência criada na etapa anterior, são geradas regras avaliando-se compras realizadas pelos clientes do mesmo *cluster* ao qual pertence o cliente em questão, ou considerando-se todos os clientes da base, seguindo o que é especificado no momento da execução.

Tendo como exemplo a sequência {c, a, b}, são geradas as regras {c -» a, b}, {c, a -» b}. Para cada regra é verificado o índice mínimo de Confiança configurado.

Considerando que os itens {c, a} estejam presentes em 60% das compras, e o item {b} esteja em 90% destas ocorrências, diz-se que a regra {c, a -» b} possui um suporte de 60% e confiança de 90%.

Geradas as regras, o sistema está apto a gerar recomendações para o cliente em questão, sendo que a geração das recomendações é realizada avaliando-se cada regra. Caso o(s) produto(s) do lado esquerdo da regra estiver(em), em sua totalidade, dentre os produtos adquiridos pelo cliente, recomenda-se os produtos do lado direito da regra.

Utilizando a base de dados gerada pelo software auxiliar desenvolvido (Figura 10), foram realizados experimentos para fins de verificação de resultados e para comparações entre a geração de regras de associação fazendo-se uso ou não de Clusterização.

A base de dados é populada com dois mil clientes, oito mil produtos e três mil vendas, sendo que para venda é sorteado um cliente e produtos que serão associados àquela venda. Essas condições foram adotadas na geração da base, pois é uma situação que pode acontecer em um ambiente real.

A configuração utilizada para os índices mínimos de suporte e confiança no experimento foi de 0,1% em ambos. Tais índices foram configurados com valores extremamente pequenos, pois o número de produtos é muito superior ao de transações, sendo que ao utilizar valores maiores não seriam geradas regras, pois o número de vezes que cada produto (ou combinação de produtos) aparece nas vendas seria muito baixo em relação ao número total de vendas, não atingindo, assim, o índice de suporte mínimo.

Realizou-se o processo de Clusterização, no qual são gerados três *clusters* através do algoritmo K-means.

Para a geração das regras, incialmente não foi utilizada a Clusterização, o que significa que, independente do cliente para o qual estas se destinam, são avaliadas vendas a todos os clientes. Passando-se então, para a geração de regras utilizando-se da clusterizacão, sendo que, neste caso, são avaliadas apenas as vendas realizadas a clientes que pertencem ao mesmo *cluster* do cliente em questão.

Após a geração das regras de associação, é processada a geração das recomendações.

Para realização de experimentos foram utilizados os mesmos índices mínimos de suporte e confiança, bem como mantido o mesmo cliente ao qual são geradas as recomendações.

Antes da geração das regras, efetuou-se a Clusterização, criando-se três *clusters*. Abstraindo o resultado desta etapa para uma forma gráfica, tem-se a Figura 9, na qual o eixo horizontal representa a média de valores dos produtos adquiridos por cada cliente e o eixo vertical representa o valor total das compras de cada cliente.

Foram escolhidos os atributos médios de valores e total de compras já que clientes tendem a se interessar por produtos que seguem uma faixa de valores na qual usualmente compram. Mantendo clientes que têm os mesmos costumes em um mesmo *cluster*, recomendam-se produtos que seguem a mesma faixa de valores.

Os pontos vermelhos, azuis e verdes representam os diferentes *clusters* criados, já os pontos roxos representam os centroides de cada *cluster*.

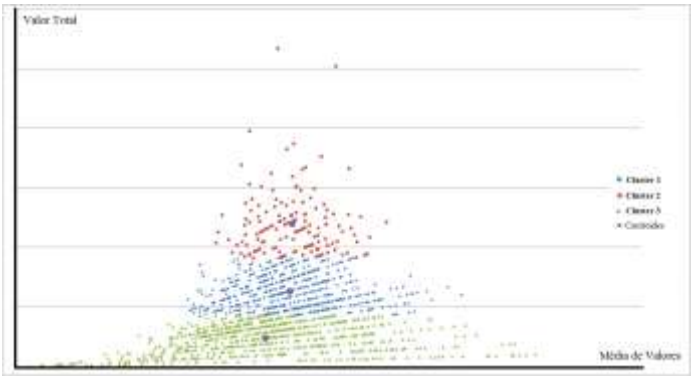

Figura 9 – Representação gráfica da Clusterização

As Regras de Associação são geradas de duas formas – com e sem o uso dos resultados da Clusterização. Um comparativo é apresentado Tabela 4.

Tabela 4: Comparativo da geração das regras de associação

	Sem Clusterização	**Com Clusterização**
Tempo de processamento[1]	54 minutos, 29 segundos e 999 milissegundos.	15 minutos, 3 segundo e 258 milissegundos.
Número de regras de associação geradas	30	3150
Número de recomendações	2	11

Observando-se os resultados destes experimentos, pode-se concluir que o tempo de processamento é menor ao utilizar a técnica de Clusterização para a geração das Regras de Associação. Isso se dá devido ao reduzido número de vendas para avaliação selecionadas na base de dados. A geração das regras de associação sem o uso de clusterização, por sua vez, avalia todas as vendas realizadas.

Referente ao número de regras de associação geradas nota-se que com o uso da Clusterização o número é muito maior em relação à geração de regras sem o uso da técnica supracitada. Através de análises realizadas durante a execução do algoritmo, nota-se que este número é mais elevado devido ao reduzido número de vendas que está sendo avaliado. Isto faz com que o índice de suporte de cada produto se eleve, já que este é calculado através do número de vezes que o produto aparece nas transações avaliadas.

[1] *Hardware utilizado: notebook com processador Intel Core I7-2670QM (2.20 GHz up to 3.10 GHz, 6MB L2 Cache), 6GB RAM, sistema operacional Windows 7 Professional SP1 x64*

Por exemplo, sem utilizar Clusterização, um produto que aparece em 350 vendas, de um total de 5000, terá seu índice de suporte calculado em 0,07 (350/5000). Utilizando a Clusterização, o mesmo pode estar presente em 300 das 1000 vendas selecionadas. Isto resulta em um índice de suporte de 0,3 (300/1000). Então, muitos produtos que não atingiam o suporte mínimo, com o uso da Clusterização passam a atingir este índice, participando da geração das regras.

Como consequência do número maior de regras de associação geradas, a quantidade de recomendações passa a ser maior.

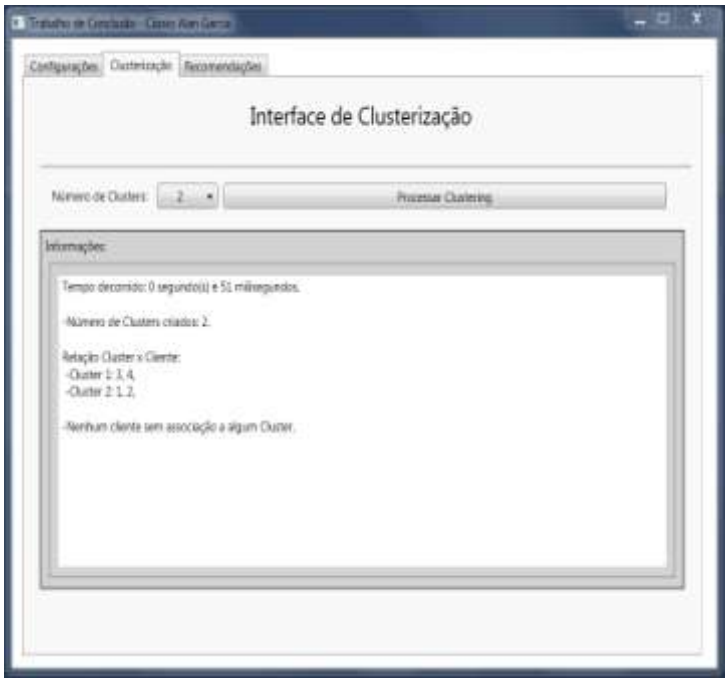

Figura 10 - Resultado da Clusterização

Este aumento do número de recomendações também pode ser mérito da qualidade das regras geradas, pois com o uso da Clusterização são avaliadas vendas realizadas a clientes que pertencem ao mesmo *cluster*, ou seja, que têm um perfil semelhante ao usuário que está recebendo as recomendações.

Como as regras geradas são mais direcionadas ao perfil de clientes de um *cluster* – e não de todos os clientes, um número maior de regras estará sendo útil para a geração das recomendações, isto ocorre avaliando o lado esquerdo da regra que atenda em sua totalidade os itens já adquiridos pelo cliente.

Dentre as recomendações encontram-se também produtos já adquiridos pelo cliente, pois mesmo que o produto já tenha sido adquirido, o cliente pode se interessar em comprar novamente. Vale ressaltar que produtos já adquiridos pelo cliente, são avaliados da mesma forma como os demais produtos, não sofrendo avaliação diferenciada dos demais.

Foi utilizada uma pequena base de dados que contém quatro clientes cadastrados, nove produtos e nove vendas, para fins de ilustração dos resultados obtidos. A Figura 10 mostra os resultados da Clusterização e a Figura 11 da geração de Regras de Associação e das Recomendações.

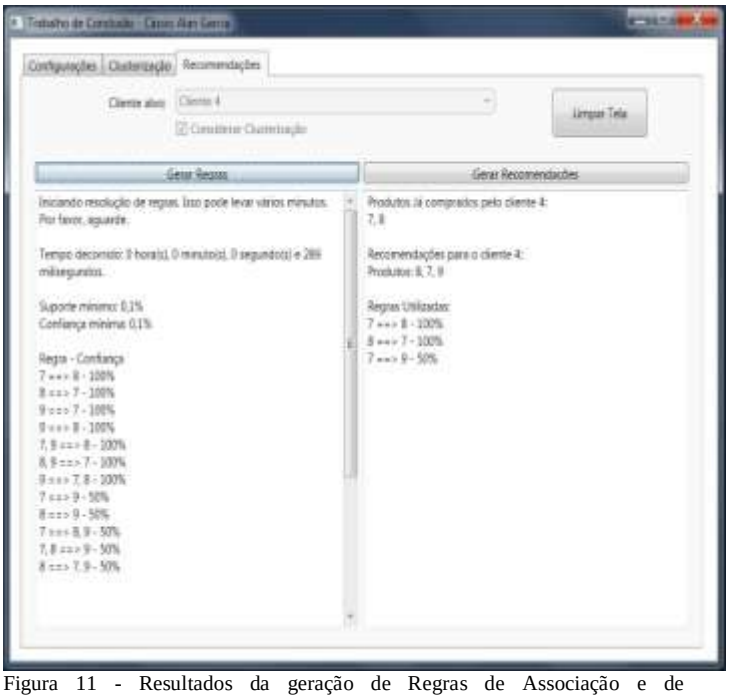

Figura 11 - Resultados da geração de Regras de Associação e de Recomendações

No caso desta base de dados nota-se que em 100% das vezes que o "produto 7" foi adquirido, o "produto 8" também foi (7 ➔ 8); 50% das vezes que foi comprado o "produto 7", foi também comprado o "produto 9". Com base nessas regras, geram-se as recomendações. Visto que consta no histórico de compras do "Cliente 4" os produtos 7 e 8, recomenda-se os produtos 7, 8 e 9.

este experimento, foram criados dois *clusters*, sendo que os clientes 3 e 4 pertencem ao *Cluster* 1 e os clientes 1 e 2 ao *Cluster* 2.

5 Conclusão

O campo de Descoberta de Conhecimento em Bases de Dados é focado no desenvolvimento de métodos e técnicas para tornar dados brutos em informação, mapeando dados de baixo nível (que podem ter grandes volumes dificultando o entendimento) em informações mais compactas, abstratas e mais úteis [17]. Sendo assim, a área de KDD torna-se adequada para a análise de dados, principalmente nos casos em que grandes quantidades estão envolvidas. A mineração de dados disponibiliza diversas tarefas que podem ser combinadas, a fim de unir vantagens e suprimir deficiências de

cada uma delas, desenvolvendo-se assim um sistema híbrido mais eficiente.

Nota-se o benefício na utilização de Clusterização junto à geração de Regras de Associação, no caso estudado, pois o número de vendas a serem avaliados na geração das regras é menor, diminuindo consideravelmente o tempo de processamento. Além das regras serem geradas com base em compras realizadas por clientes com características semelhantes, o sistema também recomenda itens que, possivelmente, terão uma maior aceitação do cliente que recebe as recomendações, pois ao se analisar o histórico de compras dos clientes nota-se que o índice de Confiança atende o mínimo especificado na etapa de configurações.

Este trabalho traz contribuições científicas para a área de Inteligência Artificial, no sentido de mostrar a possibilidade da integração das tarefas de Clusterização e Regras de Associação na área de Mineração de Dados, bem como a possibilidade da criação de um sistema híbrido de Inteligência Artificial – integrando Data Mining e um Sistema de Recomendação.

Relacionando com trabalhos já existentes na área de recomendação de produtos, o principal diferencial deste trabalho está na utilização da Recomendação Colaborativa, que provém da forma de utilização das Regras de Associação e, principalmente, por permitir avaliar produtos comprados por clientes semelhantes ao que estará recebendo as recomendações (devido à utilização da Clusterização), o que pode trazer um maior índice de acerto ao tentar oferecer produtos que realmente interessem ao cliente.

PRODUCTS RECOMENDER SYSTEM USING DATA MINING

ABSTRACT: This paper presents the study of data mining technique applied to recommender systems and the development of a computational system to indicate products to users according to their preferences and characteristics. With the use of techniques of Data Mining, Clustering and Association Rules, arising from the area of Artificial Intelligence, the developed system aims to integrate these two techniques, creating a hybrid system that provides information to generate recommendations.

Keywords: Data Mining, Recommender Systems, e-commerce.

Referências

[1] CAZELLA, Sílvio C.; NUNES, Maria A. S. N.; REATEGUI, Eliseo B. A Ciência da Opinião: Estado da arte em Sistemas de Recomendação. JAI: Jornada de Atualização em Informática da SBC. Rio de Janeiro: Editora da PUC Rio, 2010. Disponível em: < http://www.dcomp.ufs.br/~gutanunes/hp/publications/JAI4.pdf>, acessado em 15 de abril de 2012.

[2] JESUS, R. P.; ESCOBAR, M. Desenvolvimento de um Sistema de Recomendação de eventos com uso de Geolocalização. Ulbra: Canoas, 2011. (Projeto)

[3] FAYYAD, Usana M. (Coord.). Advances in knowledge discovery and data mining. Cambridge: MIT, 1996.

[4] CASTANHEIRA, L. G. Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões. UFMG, Belo Horizonte, 2008. (Dissertação). Disponível em: <http://cpdee.ufmg.br/defesas/349M.PDF>, acessado em 15 de abril de 2012.

[5] REZENDE, S. O. Sistemas Inteligentes – Fundamentos e Aplicações. Barueri: Editora Manole, 2003.

[6] MORESI, E. A. Ciência da Informação, v. 29, n. 1, p. 14-27, jan/abr. 2000.

[7] SANTOS, M. F.; AZEVEDO, C. Data Mining: Descoberta de Conhecimento em Bases de Dados. Lisboa: Editora FCA, 2005.

[8] SFERRA, H. H.; CORRÊA, A. M. C. J. Revista de Ciência e Tecnologia, v.11, n. 22, p. 19-34, 2003.

[9] VERMA. M. et al. A Comparative Study of Various Clustering Algorithms in Data Mining. GLNA Institute of Technology, Mathura. Disponível em <http://www.ijera.com/papers/Vol2_issue3/ID2313791384.pdf>, acessado em 15 de março de 2013.

[10] CORTES, S. C.; PORCARO, R. M.; LIFSCHITZ S. Mineração de Dados – Funcionalidades, Técnicas e Abordagens. PUC-RioInfMCC10/02, 2002. Disponível em: <ftp://139.82.16.194/pub/docs/techreports/02_10_cortes.pdf>, acessado em 15 de abril de 2012.

[11] LOPES, G. R. Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica. Porto Alegre: Programa de Pós-Graduação em Computação, 2007. Disponível em: <http://www.lume.ufrgs.br/handle/10183/10747>, acessado em 24 de abril de 2012.

[12] LICHTNOW D. et al. O Uso de Técnicas de Recomendação em um Sistema para Apoio à Aprendizagem Colaborativa. Revista Brasileira de informática na educação (RBIE), 14(3):49–59, 2006. Disponível em: <http://ceie-sbc.educacao.ws/pub/index.php/rbie/article/view/46>, acessado em 15 de abril de 2012.

[13] SERRANO, Maurício. Um Sistema de Recomendação para Mídias Baseado em Conteúdo Nebuloso. UFSCar, São Paulo, 2003. Dissertação (Mestrado). Disponível em: <http://www.bdtd.ufscar.br/htdocs/tedeSimplificado/tde_arquivos/3/TDE-2006-11-24T14:22:04Z-1269/Publico/DissMS.pdf>, acessado em 15 de abril de 2012.

[14] CAZELLA, S. C.; REATEGUI, E. B. Sistemas de Recomendação. XXV Congresso da Sociedade Brasileira de Computação. Unisinos: São Leopoldo, 2005. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.2811&rep=rep1&type=pdf >, acessado em 15 de abril de 2012.

[15] GARCIA, C. A., FROZZA, R. Estudo Comparativo de Aplicações em Mineração de Dados Aplicada a Sistemas de Recomendação. XIX Simpósio de Engenharia de Produção - UESP, 2012.

[16] FIGUEIRA FILHO, F.M., GEUS, P.L, ALBUQUERQUE, J.P. Sistemas de recomendação e interação na Web Social. In: I Workshop de Aspectos da Interacao Humano-Computador na Web Social, Porto Alegre, 2008. Disponível

em: <http://www.ic.unicamp.br/~fmarques/papers/websocial_ihc08.pdf>, acesso em: 21 de abril de 2012.

[17] FAYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. From Data Mining to Knowledge Discovery in Databases. Al Magazine Vol 17 Number 3, 1996. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131>, acesso em 14 de março de 2013.