Área: CV ( ) CHSA ( ) ECET (X)

APLICAÇÃO DE ALGORITMO DE APRENDIZAGEM DE MÁQUINA NÃO-SUPERVISIONADO PARA CLASSIFICAÇÃO DE USUÁRIOS NA REDE SOCIAL ACADÊMICA *SCIENTIA.NET*

Heloína Alves Arnaldo (bolsista do PIBIC/UFPI), Vinicius Ponte Machado (Orientador, Depto. de Informática e Estatística – UFPI)

Introdução

Sistemas de Redes Sociais tornaram-se especialmente relevantes na internet devido a grande adesão de usuários aos vários sites *Web* que utilizam o conceito, como *Orkut*, *MySpace*, *Facebook* e *Flickr*. Esses sistemas funcionam com o princípio da interação social, ou seja, buscando conectar pessoas e proporcionar sua comunicação forjando laços sociais. Uma rede social é definida como um conjunto de dois elementos: atores (pessoas, instituições ou grupos; os nós da rede) e suas conexões - interações ou laços sociais (RECUERO, 2004, pag.12). A rede social, derivando deste conceito, passa a representar um conjunto de participantes autônomos, unindo idéias e recursos em torno de interesses compartilhados a partir das interações estabelecidas entre eles.

Os usuários das redes sociais formam bases de dados que proveem um importante meio de compartilhar, organizar e encontrar conteúdo, além de estabelecer contatos através de interesses comuns. Neste contexto foi criado o Scientia.Net – site de rede social que visa integrar informações contidas em diversos serviços da Internet (fóruns, repositórios de artigos, sites, blogs e demais redes sociais). Além disso, esta ferramenta visa promover a interação de seus usuários (estudantes, professores e pesquisadores) para fins acadêmicos, com base nos seus interesses em comum (MACHADO et al., 2011, pag.3).

Este trabalho apresenta uma aplicação desenvolvida para agrupar de forma automática os usuários do Scientia.Net através do algoritmo de aprendizagem de máquina não-supervisionado COBWEB (MITCHELL, 1997, pag.18). A aplicação foi criada a partir de pesquisas sobre algoritmos de aprendizagem de máquina e visa oferecer ao Scientia.Net um mecanismo de classificação que apresente a cada usuário do site, uma relação de outros pesquisadores com base nas suas pesquisas em comum. Com isso, pretende-se contribuir para a interação entre usuários de perfis semelhantes e assim melhorar na produtividade de suas pesquisas permitindo assim a troca de conhecimento.

Algoritmos que utilizam técnicas de Aprendizagem de Máquina melhoram automaticamente à medida que aprendem com experiências passadas (GOLDSCHMIDT; PASSOS, 2005, pag.237). Estes algoritmos têm como objetivo encontrar e descrever padrões a partir dos dados obtidos do ambiente. A tarefa principal é aprender um modelo a partir do ambiente e manter esse modelo consistente de modo a atingir as finalidades de sua aplicação. A tarefa de aprender consiste em escolher ou adaptar os parâmetros de representação do modelo.

Tal mecanismo de classificação automático de usuários contribui para que o usuário da rede social não desperdice tempo buscando os perfis dos pesquisadores com os quais deseja obter ou compartilhar informações, pois o algoritmo COBWEB se encarrega de identificar grupos de usuários por meio da classificação de padrões na base de dados do Scientia.Net.

Área: CV ( ) CHSA ( ) ECET (X)

O algoritmo de Aprendizagem de Máquina COBWEB, utilizado na implementação deste trabalho, foi adaptado do pacote *Waikato Environment for Knowledge Analysis* – WEKA, que consiste num conjunto de implementações de algoritmos de Aprendizagem de Máquina, desenvolvido na Universidade de Waikato na Nova Zelândia (WEKA, 2009).

Metodologia

Este trabalho foi dividido em três etapas:

- Etapa 1 - para levantamento bibliográfico sobre os algoritmos de *agrupamento* do *WEKA*, levantamento das tecnologias que poderiam ser utilizadas para construir a aplicação Web do projeto e realização de testes com os algoritmos do WEKA.
- Etapa 2 – para o desenvolvimento da aplicação para classificação de usuários proposta.

A ferramenta WEKA utiliza um formato próprio de arquivo de dados, o ARFF e o Scientia.Net utiliza banco de dados MySql para armazenar dados dos usuários . Dessa forma, surgiu a necessidade de desenvolver um conversor de dados MySql para o formato ARFF. Este conversor foi desenvolvido em Java e posteriormente foi integrado à aplicação responsável pela classificação automática de usuários, também desenvolvida em Java. Uma vez incorporada ao Scientia.Net esta aplicação irá de forma automática, através de uma conexão JDBC[1], se conectar ao banco de dados MySql do site, selecionar a tabela de usuários, e converter esta tabela para um arquivo no formato ARFF.

Após a implementação do conversor ARFF, o algoritmo COBWEB do WEKA foi integrado à aplicação. Para a realização dos testes com o COBWEB criou-se uma base de dados fictícia para o Scientia.Net composta por 60 usuários, dado que este projeto ainda está em fase de desenvolvimento. A classificação leva em consideração os valores dos atributos que formam o perfil acadêmico dos usuários. Estes atributos incluem: Graduação, Mestrado, Doutorado, Pós-Doutorado e suas respectivas subáreas (Figura 1).

nome	graduacao	mestrado	sub_mestrado	doutorado	sub_doutorado
Tania Araujo Revoredo	Ciencia da Computacao	Redes de Computadores	Redes de Senssores	Redes de Computadores	Senssoriamento Remoto
Marcio Alves de Macedo	Ciencia da Computacao	Engenharia de Software	Testes de Software	Engenharia de Software	Testes de Software
Gabriela Maria Ribeiro Cruz	Ciencia da Computacao	Redes de Computadores	Redes de Senssores	Redes de Computadores	Senssoriamento Remoto
Ewerton Vale Leal	Ciencia da Computacao	Engenharia de Software	Testes de Software	Engenharia de Software	Testes de Software
Bruno Vicente Alves de Lima	Ciencia da Computacao	Inteligencia Artificial	Redes Neurais	Inteligencia Artificial	Robotica
Eloisa Alves de Lima	Fisica	Fisica Quantica	Fisica Quantica	Fisica	Fisica Quantica
Lucas Rocha Araujo	Fisica	Fisica Quantica	Fisica Quantica	Fisica	Fisica Quantica
Luana Martins Alves	Engenharia Civil	Engenharia Civil	Construcao Civil	Engenharia Civil	Construcao Civil
Eduardo Barbosa de Lima	Engenharia Civil	Engenharia Civil	Construcao Civil	Engenharia Civil	Construcao Civil
Aurileide Maria Bispo Frazão	Quimica	Quimica Orgânica	Fotoquimica Orgânica	Quimica Orgânica	Fotoquimica Orgânica
Renato Duarte Fortes	Ciencia da Computacao	Engenharia de Software	Testes de Software	Engenharia de Software	Testes de Software
Ana Carolina	Quimica	Quimica Orgânica	Polimeros e Colóides	Quimica Orgânica	Sintese Orgânica
Eliana Silva Arnaldo	Ciencia da Computacao	Engenharia de Software	Testes de Software	Engenharia de Software	Testes de Software
Maria de Fatima Pereira Lima	Ciencia da Computacao	Engenharia de Software	Testes de Software	Engenharia de Software	Testes de Software
Lilian Rosalina	Quimica	Quimica Orgânica	Quimica dos Produtos Naturais	Quimica Orgânica	Quimica dos Produtos Naturais
Laires Andrade	Quimica	Quimica Orgânica	Fotoquimica Orgânica	Quimica Orgânica	Quimica dos Produtos Naturais
Lucas Brito de Andrade	Ciencia da Computacao	Inteligencia Artificial	Processamento de Imagens	Inteligencia Artificial	Processamento de Imagens
Cristina Santos	Quimica	Quimica Orgânica	Polimeros e Colóides	Quimica Orgânica	Sintese Orgânica
Pedro Vasconcelos Almeida	Ciencia da Computacao	Inteligencia Artificial	Processamento de Imagens	Inteligencia Artificial	Processamento de Imagens
Genivaldo Pereira Silva	Ciencia da Computacao	Redes de Computadores	Arquiterura de Sistemas de computação	nao	nao
Mateus Barbosa Araujo	Ciencia da Computacao	Inteligencia Artificial	Processamento de Imagens	Inteligencia Artificial	Processamento de Imagens
Marcos Silva	Quimica	Quimica Orgânica	Sintese Quimica	Quimica Orgânica	Sintese Quimica
Fernanda Alves de Lima	Ciencia da Computacao	Engenharia de Computação	Redes de Sensores sem Fio	nao	nao
Felipe Reis de Almeida	Ciencia da Computacao	Inteligencia Artificial	Processamneto de Imagens	Inteligencia Artificial	Processamneto de Imagens
Frustorso Euzébio	Quimica	Quimica Orgânica	Fisico-Quimico Orgânica	Quimica Orgânica	Fisico-Quimico Orgânica
Adriana Lima Duarte	Ciencia da Computacao	Redes de Computadores	Segurança em Redes de Computadores	nao	nao
Luiza Fernandes Couto	Ciencia da Computacao	Inteligencia Artificial	Processamento de Imagens	Inteligencia Artificial	Processamento de Imagens
Francizildo Anderson	Quimica	Quimica Orgânica	Sintese Orgânica	nao	nao
Francisco Evaldo	Engenharia Civil	Construcao Civil	Materiais e Componentes de Contrução	Construcao Civil	Instalações Prediais

Figura 1 – Tabelas de Usuários Fictícios do Scientia.Net

[1] *Java Database Connectivity* ou JDBC é um conjunto de classes e interfaces (API) escritas em Java que fazem o envio de instruções SQL para qualquer banco de dados relacional.

O primeiro passo para início dos testes consistiu em conectar a aplicação ao banco de dados MySql do Scientia.Net, selecionar a tabela de usuários e então gerar o arquivo ARFF correspondente . Pelo fato do COBWEB organizar a árvore de classificação de forma incremental, portanto sensível a ordem de entrada dos objetos, o arquivo ARFF com os dados dos 60 usuários foi aplicado ao algoritmo com 6 diferentes ordens aleatórias de entrada. Além do COBWEB, os outros algoritmos de agrupamento do WEKA foram submetidos a esses mesmo testes. O algoritmo COBWEB obteve 98,35% de acerto na classificação destes usuários. Este acerto refere-se à classificação dos usuários em suas classes correspondentes. Assim se um dado usuário possui valores de atributos que o torna próximo da classe A (Ciência da Computação), este não deve estar classificado na classe B (Medicina). O WEKA disponibiliza ferramentas que geram informações de seus algoritmos e que permitem verificar em qual classe foi instanciado cada objeto classificado.

O COBWEB foi o algoritmo que nos testes apresentou a maior taxa de acerto, e por isso foi selecionado para integrar a aplicação para classificação automática dos usuários do site Scientia.Net. Após escolher os atributos, criar o banco de dados, e testar o COBWEB, foi desenvolvido o site do Scientia.Net *Joomla* utilizando o componente *JomSocial*, como mostra na Figura 5, e na sua base de dados padrão foram adicionados os campos escolhidos como atributos para os usuários.

- Etapa 3- etapa de finalização do projeto, onde a aplicação Web desenvolvida foi incorporada ao *Scientia.Net*.

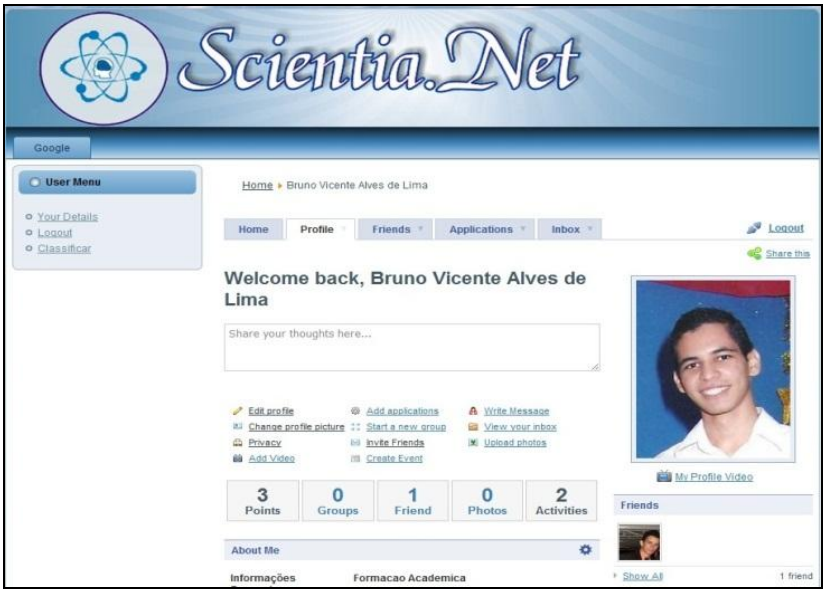

Figura 5 - Exemplo de tela do *Scienta.Net.*

Resultados e Discussão

Com a aplicação para classificação de usuários concluída, foi finalmente incorporada ao site Scientia.Net. Quando um novo usuário se cadastra no site a aplicação é executada e o Scientia.Net recebe em sua base de dados o resultado da classificação. Assim quando um usuário do site faz seu login, na página inicial pode ser visualizado o grupo de usuários cujos perfis são semelhantes ao seu.

Por exemplo: Aurileide Frazão com graduação em Química, com a sub-área de mestrado em Síntese Orgânica; Lilian Rosalina com formação em Química, com a sub-área em Fotoquímica Orgânica; Ana Carolina graduada em Química, sub-área Físico-Química Orgânica; Amanda Lorena também graduada em Química, com sub-área Polímeros e Coloides, todas classificadas pelo COBWEB no mesmo grupo. Assim o Scientia.Net apresenta, por exemplo, Lilian Rosalina, Lorena Amanda e Ana Carolina como sendo de interesse a Aurileide Frazão (Figura 2).

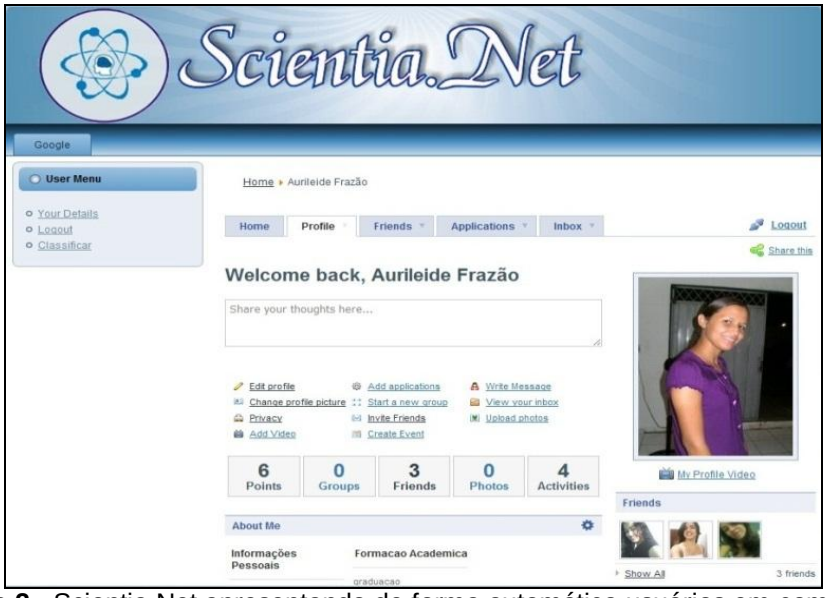

Figura 2 - Scientia.Net apresentando de forma automática usuários em comum ao perfil de outro usuário.

Conclusão

O Scientia.Net, rede social acadêmica, tem como objetivo reunir pesquisadores nas mais diversas áreas do conhecimento afim de possibilitar a troca de informações entre eles, e isto feito de forma automática através de algoritmos de Aprendizagem de Máquina. Neste trabalho foi utilizado o COBWEB.

O algoritmo COBWEB apresentou um bom desempenho na classificação dos usuários do Scientia.Net. Porém pelo fato de ter sido submetido a testes com uma base de dados considerada pequena, seu desempenho continuará sendo avaliado regularmente utilizando a base de dados real do Scientia.Net.

Referências Bibliográficas

[1] GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel**.** **Data Mining**: Um Guia Prático: Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. Rio de Janeiro, RJ: Elsevier, 2005.

[2] MACHADO, V.P.; LIMA, B.V.A.; ARNALDO, H.A.; ARAUJO, S.W.I. **Classificação Automática dos Usuários da Rede Social Acadêmica Scentia.Net**. In: IV Congresso Tecnológico TI e Telecom. INFOBRASIL 2011, Ceará, 2011.

[3] MITCHELL, T. **Machine Learning**. McGraw Hill, New York, 1997.

[4] RECUERO, R. C. **Teoria das Redes e Redes Sociais na Internet**: Considerações sobre o Orkut, os Weblogs e os Fotologs. In: XXVII Congresso Brasileiro de Ciências da Comunicação. XXVII INTERCOM, Rio Grande do Sul, 2004. Disponível em: <http://reposcom.portcom.intercom.org.br/bitstream/1904/17792/1/R0625-1.pdf>. Acesso: 18 jun. 2011.

Área: CV ( ) CHSA ( ) ECET (X)

[5] UNIVERSITY OF WAIKATO. ***Weka* 3:** Machine Learning Software in Java. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso: 03 jun. 2011.

Palavras-chave: Redes Sociais, Aprendizagem de Máquina, COBWEB, WEKA.