CLASSIFICACÃO AUTOMATICA DE ARTIGOS CIENTÍFICOS UTILIZANDO REDES NEURAIS

Sanches Wendyl Ibiapina Araujo (voluntário da ICV/UFPI), Dr. Vinicius Ponte Machado (Orientador, Depto de Informática e Estatística – UFPI)

Introdução

A Internet já se consolidou como um dos maiores meios de disseminação de conhecimento permitindo o acesso das pessoas às informações nela disponíveis, contudo essas informações precisam estar acessíveis para as pessoas que possuem algum interesse nelas, nesse sentido pode se ver a internet como uma ferramenta capaz de reunir pessoas sob um denominador comum, com esse objetivo surgem as redes sociais.

Quando observamos que a conexão de pessoas possibilita a comunicação e a troca de informações, começamos a entender que é possível aplicar esses conceitos na área científica, onde a interação das pessoas é um fator importante para o avanço das pesquisas. Contudo, mesmo com todas as facilidades de comunicação proporcionadas pela Internet, quase não há ferramentas específicas para colaboração e disseminação de conhecimento para a área acadêmica.

O Scientia.Net é uma rede social baseada na Internet que visa agregar aos seus usuários itens de relevância relacionados ao seu perfil. O objetivo do nosso trabalho é o desenvolvimento de uma aplicação que possibilite a classificação automática de artigos científicos baseados nos perfis dos usuários do Scientia.Net. Com isso, o Scientia.Net enquanto agregador de informações acadêmicas permiti aos seus usuários uma melhoria na produtividade de suas pesquisas.

A fim de realizar esta classificação foi utilizado de Algoritmos de Aprendizagem de Máquina os quais tem como objetivo o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Existem três principais tipos de técnicas de aprendizagem de máquina:

O Aprendizado Supervisionado, que implica, necessariamente, a existência de dados de entradas e a indicação de uma saída a ser aprendida para ocorrer o processo de aprendizagem. [1]

Aprendizado Não-Supervisionado envolve a aprendizagem de padrões na entrada, quando não são fornecidos valores de saídas específicos. [2]

Aprendizado por Reforço que consiste em mapear situações (estados do ambiente) para ações (o que fazer) de modo a maximizar um sinal de recompensa numérico.

O algoritmo de Aprendizagem de Máquina utilizado nesse trabalho foi Redes Neurais Artificiais (Aprendizado Supervisionado). Redes Neurais Artificiais são sistemas que tentam simular o funcionamento do cérebro humano. São compostas por unidades de processamento simples chamados de Neurônios Artificiais.

Com o objetivo de solucionar o problema da classificação de usuários do Scientia.Net utilizamos no desenvolvimento deste trabalho as Redes Perceptrons de Múltiplas Camadas, estas redes possuem uma camada de entrada, contendo as entradas da rede, com uma ou mais camadas ocultas, que possui neurônios artificiais e uma camada de saída, que tem como objetivo resolver

problemas não lineares. Utilizamos também o Algoritmo Back-Propagation o qual treina as Redes Perceptrons de Múltiplas Camadas.

Na construção da aplicação de classificação foi utilizado a API do WEKA - Waikato Environment for Knowledge Analysis [3]. O WEKA é uma ferramenta de KDD - knowledge-discovery in databases, que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. WEKA foi desenvolvido na Universidade de Waikato na Nova Zelândia, sendo escrito em Java e possuindo código aberto disponível na Web (a atual versão - 3.4.3 - demanda Java 1.4).

O Fato de o WEKA ser escrito em Java e ter suas bibliotecas disponíveis, teve um peso relevante na decisão de utilizá-lo no Scientia.Net com isso os algoritmos podem ser utilizados em várias plataformas, deixando assim, o trabalho com uma boa portabilidade e melhor utilização no Joomla.

Metodologia

Para alcançar os objetivos e metas traçados neste projeto, ele foi dividido em quatro fases: Fase Teórica (FT); para levantamento do estado da arte sobre o tema; Fase de análise (FA), para o estudo sobre os principais Algoritmos acerca de Aprendizagem de Máquina bem como o estudo da ferramenta WEKA; Fase de Desenvolvimento (FD), para o desenvolvimento do método e da aplicação proposta; Fase de Integração (FI), para a integração da aplicação desenvolvida ao Scientia.Net.

Durante a Fase Teórica fizemos um levantamento bibliográfico da área de Aprendizagem de Máquina bem como sobre as tecnologias que poderiam ser utilizadas para construir a ferramenta do projeto. O resultado dessa fase foi a aquisição do conhecimento necessário para as fases seguintes.

Na Fase de Análise foi feito um levantamento dos principais algoritmos de Aprendizagem de Máquina observando tanto técnicas supervisionadas como não-supervisionados onde cada um deles foi minuciosamente analisados, também na Fase de Análise foi realizado um estudo da ferramenta WEKA a fim de que dispuséssemos do domínio desta ferramenta e do seu conteúdo. Essa análise dos algoritmos e da ferramenta é necessária para que seja possível definir como deve ser utilizada a API do WEKA para a construção da aplicação proposta no projeto.

Na Fase de Desenvolvimento, foi implementado uma aplicação que permitiu a classificação automática de artigos científicos, tomando por base os perfis de usuários do Scientia.Net, de maneira que pudemos apresentar artigos condizentes com a área de atuação do pesquisador cadastrado na rede social Scientia.Net.

Por fim, a Fase de Integração em que foi realizada a junção da aplicação desenvolvida com o Scientia.Net. Nesse ponto do projeto abrimos a rede social, em fase de teste, a um público específico e restrito mantendo o controle sobre o ambiente do algoritmo para que possamos ter um retorno do comportamento do algoritmo em uma situação prática.

Resultados e discussão

Conforme mencionado anteriormente, durante a fase de desenvolvimento foi construído um aplicativo que classifica artigos científicos a partir dos perfis de usuários do Scientia.Net.

Dentre todos os Algoritmos de Aprendizagem de Máquina suportados pela API WEKA encontramos as Redes Neurais Artificiais, algoritmo usado na implementação do aplicativo deste projeto. A fim de que pudéssemos testar o quão satisfatório é a classificação realizada pela Rede Neural, e tendo em vista que a ferramenta pensada com a finalidade de filtrar os dados uteis dos artigos científicos ainda não foi implementada, optamos por utilizar uma base de dados constituída de artigos científicos escolhidos aleatoriamente dentre áreas de interesse acadêmico selecionadas pelo Aluno Iniciando.

Esta base de dados possui 33 artigos cadastrados, foi utilizada dentro do processo de aprendizagem uma ferramenta conhecida com validação cruzada (cross validation), na qual os dados são particionados aleatoriamente em dois conjuntos, um para treinamento e outro para teste, sendo que o conjunto de treinamento é dividido em mais dois conjuntos disjuntos para estimação e validação. Tomando os 33 artigos 24 foram classificados corretamente (72.7273%) e 9 foram classificados incorretamente (27.2727%).

Conclusão

O Scientia.Net enquanto 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.

A aplicação responsável por classificar os artigos já foi construída, bem como a integração da aplicação ao Scientia.Net apresentado resultados satisfatório. Com o objetivo de melhorar o sistema de classificação de usuários em áreas de interesse comuns está outra linha de pesquisa dentro de redes sociais e aprendizado de máquina, a predição dos interesses de usuários baseado na troca de mensagem entre eles, isto se da através do processamento de linguagem natural aplicado ao conjunto de mensagens trocadas. Trabalhando nisso pode-se aperfeiçoar a classificação já realizada pelo Scientia.Net.

Referências bibliográficas

[1] A. P. Braga, A. P. L. F. Carvalho e T. B. Ludermir, Redes Neurais Artificiais; Teoria e Aplicações. 2ed, Rio de Janeiro, Brasil, 2007

[2] S. Russel e P. Norving, Inteligência Artificial. Rio de Janeiro: Elsevier, 2004.

[3] University of Waikato. Weka 3 Machine Learning Software in Java. Disponível em: http://www.cs.waikato.ac.nz/ml/weka Acesso: jan.2011.

Palavras-chave: Aprendizado de Máquina. WEKA. Artigos Científicos.