CLASSIFICACÃO DE USUÁRIOS BASEADO NA TROCA DE MENSAGENS ENTRE PERFIS

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 usuários baseados na troca de mensagens entre perfis 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 o XtraK4Me (Aprendizado Não Supervisionado). Este algoritmo é o resultado do trabalho descrito na tese de mestrado de Alexander Schutz.

Com o objetivo de aperfeiçoar a classificação de perfis já usual na rede social Scientia.Net utilizamos no desenvolvimento deste trabalho conceitos de Processamento de Linguagem Natural (PLN), Esta abordagem pega documentos textuais e aplica sobre estes uma serie de algoritmos estatísticos a fim de extrair as palavras chaves. As palavras chaves extraídas e atribuídas a cada

perfil dos usuários da rede social podem ser, com certa facilidade, adicionadas a lista de atributos usados na classificação de usuários já implementado no Scientia.Net.

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 - 3.6.8 - demanda Java 1.6).

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.

Metodologia

Para alcançar os objetivos e metas traçados neste projeto, ele foi dividido em cinco fases: Revisão Bibliográfica; levantamento do estado da arte sobre o tema; Implementação (Fóruns), Implementar uma Aplicação a qual vai classificar as trocas de mensagens nos fóruns do Scientia.Net; Implementação (Chat) Implementar uma Aplicação a qual vai classificar as trocas de mensagens nos chats do Scientia.Net; Integração, fase para a integração das aplicações desenvolvidas ao algoritmo de Redes Neurais já implementado no Scientia.Net; Avaliação, fase com a finalidade de avaliar o desempenho das técnicas desenvolvidas.

Durante a fase de Revisão Bibliográfica fizemos um levantamento bibliográfico da área de Aprendizagem de Máquina bem como sobre as tecnologias que poderiam ser utilizadas para construir as ferramentas do projeto, também estudamos algoritmos de Processamento de Linguagem Natural a fim de extrair parâmetros adquiridos da troca de mensagens entro os usuários da rede social, ainda nessa fase foi realizado um estudo da ferramenta WEKA a fim de que dispuséssemos do domínio desta ferramenta e do seu conteúdo.

Na fase de Implementação, em contraste com o descrito no plano de trabalho, decidimos implementar uma única ferramenta capaz de extrair palavras-chave de um dado banco de dados, seja este alimentado por Fóruns ou Chat. Essa abordagem foi escolhida, posto que a rede social Scientia.Net ainda não possui ativo a funcionalidade que da suporte a Fóruns ou mesmo Chat, contudo acreditamos que o modo como a ferramenta foi desenvolvida permita uma fácil adaptação ao Scientia.Net quando da incorporação das funcionalidades acima citadas.

Em seguida, na fase de Avaliação, posto que avaliar algoritmos de extração de palavras chave pode ser muito subjetivo, quando feito a partir da leitura e análise do texto original, faz sentido examinar seu desempenho quando dispomos de palavras-chave previamente selecionadas por um especialista no contexto, o qual detém critérios de relevância altamente dependentes de significado e intenções. Sabendo que as funcionalidades do Fórum e Chat ainda não estão disponíveis no Scientia.Net não dispomos de tais palavras previamente selecionadas.

Para minimizar essas dificuldades optamos por uma avaliação conjunta com a integração, onde as palavras-chave extraídas podem ser propriamente avaliadas quando introduzirmos na descrição dos perfis de usuários do Scientia.Net, e estes perfis forem utilizados como parâmetros de

entrada para o algoritmo de classificação de perfis já usual na rede. Nessa ocasião poderemos avaliar o desempenho das palavras-chave obtidas em aperfeiçoar a classificação de usuários.

Por fim, na Fase de Integração que dar-se ia a junção da aplicação desenvolvida com o algoritmo já implementado no Scientia.Net. Nesse ponto do projeto esperávamos utilizar os dados dos usuários da rede social para servir de entrada no Algoritmo de Aprendizagem de Máquina, contudo o Scientia.Net ainda não possui, até a data deste relatório, uma base de dados referente a troca de mensagens entre usuários adequado a aplicação do Algoritmo, de forma que foi utilizado a base de dados da Rede Social Enron para validar a solução proposta.

Resultados e discussão

Conforme mencionado anteriormente, durante a fase de implementação foi desenvolvido um aplicativo que pretende aperfeiçoar a classificação de usuários do Scientia.Net. Para se classificar um texto, o primeiro passo é a extração das informações necessárias do artefato usado na classificação. No contexto deste trabalho, referente Classificação de Textos, isso equivale a extrair da troca de mensagens entre os perfis palavras chave, ou seja, os atributos relevantes para se caracterizar os textos alvo, desta maneira diferenciando perfis uns dos outros conforme for as palavra chave extraídas. De forma mais simplificada: aplicamos o algoritmo de extração de palavras chave sobre as mensagens dos usuários retirando as palavras chaves que facilmente podem ser atribuídas ao perfil do usuário a ser classificado com a Rede Neural.

O Algoritmo de Aprendizado de Máquina usado para a extração das palavras chaves é baseado em aprendizagem não supervisionada. Esta abordagem pega documentos textuais e aplica sobre estes uma serie de algoritmos estatísticos a fim de extrair as palavras chaves. As palavras chaves extraídas e atribuídas a cada perfil podem ser, com certa facilidade, adicionadas a lista de atributos usados na classificação de usuários pela Rede Neural.

Conclusão

O Scientia.Net enquanto rede social acadêmica tem como objetivo reunir pesquisadores nas mais diversas áreas do conhecimento a fim 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 extrair as palavra chave já foi construída, restando agora, abrirmos a rede social em fase de teste a um público específico e restrito mantendo o controle sobre o ambiente do algoritmo, adquirindo uma base de dados referente a troca de mensagens entre os perfis de usuários, para que possamos ter um retorno do comportamento do algoritmo em uma situação real.

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: agosto.2012.

Palavras-chave: Aprendizado de Máquina. WEKA. Perfis.