Classificação Automática de Usuários de uma Rede Social utilizando Algoritmos Não-Supervisionados

Vinicius P. Machado[1], **Bruno V. A. de Lima**[1], **Sanches W. I. Araújo**[1]

[1]Departamento de Informática e Estatística– Universidade Federal do Piauí (UFPI)
Caixa Postal 15.064 – 91.501-970 – Teresina – PI – Brasil

vinicius@ufpi.br, brunovicente@ufpi.edu.br, sancheswendyl@hotmail.com

Abstract. *This article shows the profiles's classification in an academic social network, the Scientia.Net to allow users to interact with similar profiles. This classification is done automatically using machine learning algorithms, in this sense this paper shows results of the profiles selection using unsupervised algorithms in order to elect one who will be chosen to be applied in the social network. We approach the problem of profile's classification focusing on social networks citing implications about network's structure, and quality of links formed from this grouping.*

Resumo. *Este artigo mostra a classificação de perfis em uma rede social acadêmica, o Scientia.Net para permitir que usuários com perfis semelhantes possam interagir. Esta classificação é feita de forma automática usando algoritmos de Aprendizagem de Máquina, nesse sentido este trabalho mostra resultados da seleção de perfis usando algoritmos não supervisionados a fim de elegermos aquele que será escolhido para ser aplicado na rede social. Abordamos o problema da classificação de perfis focando nas redes sociais citando implicações quanto a estrutura da rede, e qualidade de links formados a partir desse agrupamento.*

1. Introdução

Na sociedade atual os sistemas de Redes Sociais se destacam como um meio de interação muito utilizado pelos usuários da internet. Neste cenário pode-se começar a refletir sobre uma rede social, não só capaz de unir pessoas com interesses diversos, mas também unir pesquisadores facilitando a comunicação e ajudando estes pesquisadores a ter acesso a informações acadêmicas relevantes, tais como artigos publicados em sua área de interesse e a ocorrência de eventos científicos.

O Scientia.Net é uma rede social que tem foco no ambiente acadêmico e visa agregar aos seus usuários itens de relevância acadêmica relacionados ao seu perfil. Com o Scientia.Net pretende-se ter um mecanismo que classifica os usuários de acordo com seu perfil acadêmico, permitindo-lhe que tenham contato com pesquisadores de sua área de interesse e de estudo. Em um trabalho anterior mostrou-se a criação de um mecanismo utilizando Redes Neurais Artificiais para a classificação dos usuários do Scientia.Net [Machado et al. 2011].

Neste trabalho apresenta-se um estudo comparativo de três métodos (algoritmos) de aprendizado de máquina não-supervisionados: Rede de Kohonen, Cobweb e K-means,

com o objetivo de apresentar uma classificação automática de usuários dentro do Scientia.Net baseada nos perfis acadêmicos. Os testes foram executados utilizando a base de dados do Scientia.Net, resultando assim em uma aplicação capaz de classificar e apresentar aos usuários outros pesquisadores interesses acadêmicos em comum.

2. Aprendizado de Máquina

O Aprendizado de Máquina pode ser descrito como o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática [Mitchell 1997].

A intuição humana não pode, nesse caso, ser totalmente abandonada, desde que o programador do sistema apresente os dados representados e o mecanismos usados para sua caracterização. Existem três tipos de aprendizado de máquina [Machado et al. 2011].

- O Aprendizado Supervisionado - onde existem dados de entradas com suas respectivas saídas, para serem apresentadas ao algoritmo de aprendizagem utilizado durante o processo de treinamento [de Pádua Braga et al. 2007].
- Aprendizado Não-Supervisionado - neste caso supõe que o conjunto de exemplos não está rotulado, com isto o sistema tenta classificar estes conjuntos agrupando os semelhantes em determinadas classes. [Russell and Norving 2004].
- Aprendizado por reforço - Consiste em mapear situações para ações de modo a maximizar um sinal de recompensa numérico. A ideia principal deste tipo de aprendizado é simplesmente captar os aspectos mais importante do problema colocando a um agente que vai interagir com o ambiente para alcançar uma meta [Sutton and Barto 1998].

Na próxima subseção são descritos os algoritmos utilizados neste trabalho.

2.1. Rede de Kohonen

A Rede de Kohonen ou Mapas Auto Organizáveis [Haykin 2001], tem como princípio a aprendizagem competitiva, simulando processos específicos do cérebro humano na aprendizagem por respostas sensoriais. Na rede de Kohonen os neurônios são organizados em uma grade, esta geralmente bidimensional. Essa grade tem a forma de uma superfície plana, onde os neurônios de saída estão organizados em linhas e colunas [Júnior and Montgomery 2007].

O treinamento de uma rede de Kohonen é competitivo e não-supervisionado. Este algoritmo organiza os neurônios em vizinhanças locais. Cada vez que um novo padrão é apresentado à rede, os neurônios competem entre si para ver quem gera a melhor saída. Escolhido o neurônio vencedor e seus vizinhos dentro de um raio ou área de vizinhança atualizam seus pesos. Uma observação relevante durante o treinamento, a taxa de aprendizagem e o raio de vizinhança são decrementados à medida que o algoritmo vai sendo executado [de Pádua Braga et al. 2007].

2.2. Algoritmo Cobweb

O Cobweb é um algoritmo de aprendizado de máquina, não-supervisionado que emprega um método de agrupamento [Fisher 1987]. O agrupamento é um método utilizado

para dividir um conjunto de dado sem grupos, sendo que a similaridade seria maximizada entre os objetos destes grupos, e miminizada entre membros de grupos diferentes [Goldschmidt and Passos 2005].

O Cobweb organiza incrementalmente os objetos em uma árvore que é uma estrutura onde cada um dos nodos da árvore representa um conceito que é resumido pelas distribuições dos valores dos atributos dos objetos pertencentes à subárvore do nodo. A raiz representa o conceito mais amplo que abrange todo o conjunto de objetos. [Ferreira et al. 2005]

2.3. Algoritmo de K-means

O K-means é uma técnica de aprendizado de máquina, não-supervisionado apresentada por MacQueen, 1967 e tem o objetivo de criar partições de uma população n-dimensional em k grupos em uma dada base de dados.

O algoritmo K-means utiliza um parâmetro de entrada k, que determina a quantidade de clusters, sendo que tais clusters possuem n elementos. Após a execução pretende-se obter uma alta similaridade dos elementos de um grupo e baixar similaridade entres os clusters criados pelo algoritmo [Sousa and Esmin 2011].

O algoritmo K-means pode ser resumido da seguinte maneira:

1. É escolhido aleatoriamente k objetos da base de dados como centros iniciais dos *clusters*;
2. Atribui-se cada objeto ao *cluster* ao qual o objeto é mais similar, de acordo com o calor médio dos objetos no cluster;
3. Atualiza-se as médias dos *clusters*, ou seja, calcula-se a média dos objetos para cada cluster;
4. Testa o critério de parada, e finaliza ou retorna ao item 2.

3. Redes Sociais

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 interesses comuns. Com esse objetivo surgem as redes sociais. Essas redes 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) [da Cunha Recuero 2004].

Dentro desse contexto podemos citar redes sociais formadas através da Comunicação Mediada por Computador (CMC) como Orkut, Facebook, Twitter e MySpace. Essas redes podem construir *clusters* (grupos de pessoas) focados em interesses comuns, onde é facilitado a troca de informações, levando naturalmente os participates a uma interação. Desses grupos podem surgir novas conexões entre os nós da rede social. É importante salientar que essas redes não estão paradas no tempo. Redes sociais são dinâmicas, estão em constante transformação [da Cunha Recuero 2004].

Quando observamos que estas conexões possibilitam a comunicação e a troca de informações entre as pessoas, 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, são poucas as ferramentas específicas para colaboração e disseminação de conhecimento para a área acadêmica, pode-se dizer como exemplo as redes sociais *Follow Science* [1], *Iamresearcher* [2] e a rede social acadêmica da *Microsoft* SOCL [3]. Além disso, existem diversas informações não estruturadas espalhadas por sites, blogs, bibliotecas virtuais e repositório de artigos que podem servir como literatura em muitas pesquisas científicas. Tais informações por muitas vezes passam despercebidas pelos pesquisadores, pois ferramentas de buscas atuais são dependentes do contexto das palavras-chaves utilizadas na busca e não no perfil de quem a procura.

3.1. ScientiaNet

Percebendo a força das redes sociais e sua capacidade de forjar novos laços sociais, este trabalho propõe uma estudo comparativo entre algoritmos de Aprendizagem de Maquina não-supervisionados aplicados a uma rede social como um ambiente de CMC, a rede social Scientia.Net.

O Scientia.Net é uma rede social acadêmica, baseada na internet que visa agregar de forma automática aos seus usuários itens de relevância relacionados ao seu perfil. Dessa forma o Scientia.Net é um agregador de informações contidas em diversos serviços da internet (fóruns, repositórios de artigos, sites, blogs e demais redes sociais). Além disso, a ferramenta provê a interação de seus usuários (estudantes, professores e pesquisadores) com base nos seus interesses em comum.

Esta rede social tem o foco voltado para o ambiente acadêmico e se diferencia das demais quando observamos o uso de algoritmos de Aprendizagem de Máquina na classificação de perfis. Com essa metodologia eliminamos parte da intuição humana na clusterização da rede, ou seja, automatizamos a escolha de ”amigos” dentro do sistema. Desta maneira estamos aptos a oferecer aos usuários perfis semelhantes ao seu para interação na ferramenta, fomentando a troca de informações no auxilio à pesquisa acadêmica. Porconseguinte podemos ver que a estrutura da rede fica em grande parte a cargo do agrupamento realizado pelo algoritmo bem como a qualidade dos links (conexões formadas entre os usuários) advindos da clusterização da rede.

Nesse sentido, podemos garantir a qualidade dos links a partir do momento que garantimos que a clusterização da rede fica a cargo do algoritmo não-supervisionado, isto é, o algoritmo agrupa perfis academicamente semelhantes propiciando ”amigos” relevantes como parceiros de pesquisa a todos os usuários. Nota-se que qualidade de link neste contexto se refere a quão relevante é um perfil para um determinado usuário, ou seja, se um perfil tem potencial para auxiliar a pesquisa de um determinado usuário.

Com isso, o Scientia.Net enquanto agregador de informações acadêmicas permite

[1] http://www.followscience.com
[2] http://www.iamresearcher.com/
[3] http://www.so.cl/

aos seus usuários uma melhoria na produtividade de suas pesquisas, além de fornecer mecanismos para interatividade e troca de conhecimento entre pesquisadores.

4. Ferramentas

4.1. Ferramenta Weka

Os algoritmos de aprendizado de máquina do Scientia.Net, foram implementados utilizando as bibliotecas do WEKA. O WEKA - *Waikato Environment for Knowledge* [of Waikato 2011] é um conjunto de bibliotecas Java de KDD[4] que possui uma série de algoritmos de aprendizado de máquina, preparação/mineração de dados e de validação de resultados. Foi desenvolvido na na Universidade de Waikato na Nova Zelândia, é um software livre e de código aberto disponível na Web[5].

O WEKA possui interface gráfica e seus algoritmos fornecem relatórios com informações analíticas e estatísticas dos dados em questão. Grande parte de seus recursos é acessível via sua interface, sendo que os demais podem ser utilizados através de API´s[6] em códigos criados por terceiros. A interface Gráfica (WEKA Explorer) pode ser visualizada na Fig. 1.

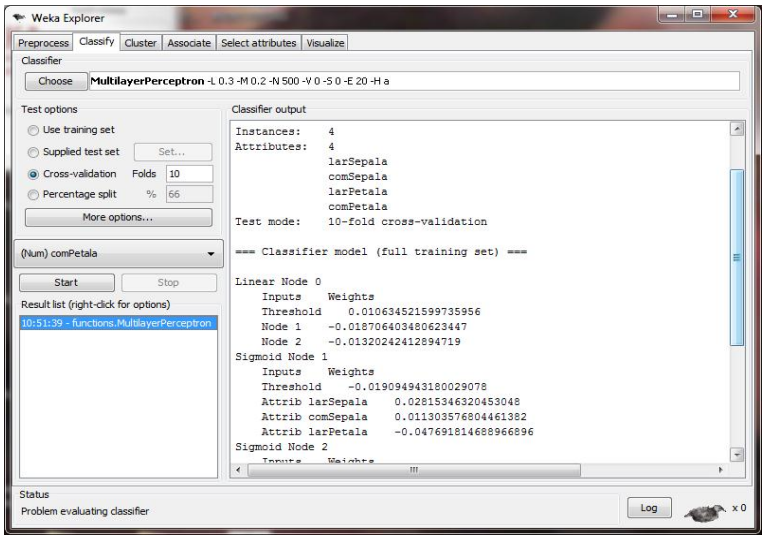

Figura 1. Janela do WEKA Explorer após a execução de um algoritmo de Aprendizagem de Máquina

4.2. Base de Dados Envolvida

A base de dados utilizadas neste trabalho foi criada baseada na estrutura do Scientia.Net que possui um total de 2000 usuários de 20 áreas distintas do conhecimento. A base de dados desta rede social foi estruturada utilizando o MySql.

[4]Knowledge-discovery in databases (Extração do Conhecimento) é um processo de extração de informações de base de dados, que cria relações de interesse que não são observadas pelo especialista no assunto, bem como auxilia a validação de conhecimento extraído.

[5]http://www.cs.waikato.ac.nz/ml/weka

[6]API - Application Programming Interface (ou Interface de Programação de Aplicativos) é um conjunto de rotinas e padrões estabelecidos por um software para a utilização das suas funcionalidades por programas aplicativos que não querem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços.

A geração desta base de dados se deu utilizando a ferramenta disponível no site Generatedata[7], onde foram convidados 20 pessoas de 20 áreas diferentes para auxiliar na geração dos dados. Desta forma cada uma das 20 pessoas geraram 200 usuários de suas respectivas áreas, simulando o cadastro de pessoas dentro do Scientia.Net. Uso dessa ferramenta se deu devido o ScientiaNet, ainda não possuir usuários reais suficientes para obter resultados relevantes.

A base de dados contém 8 atributos que formam a parte acadêmica do perfil dos usuários. Estes atributos são: graduação, área de interesse do usuário, mestrado e doutorado e seus respectivas subáreas.

5. Metodologia

Para a classificação dos usuários foram testados três algoritmos de aprendizado de máquina, descritos na seção 2. Os algoritmos citados foram implementados na linguagem Java, com o objetivo de utilizar as classes implementadas na aplicação WEKA. Com exceção da Rede de Kohonen que foi implementada separadamente devido não haver implementações deste métodos nas bibliotecas do WEKA.

A tabela da base de dados de usuários foi replicada para mais 9 tabelas com os mesmos registros alterando a ordem de inserção para executar os algoritmos. Os resultados das execuções foram reunidos para chegar a uma média dos resultados, chegando a um conjunto de resultados de classificação destes algoritmos.

Foram analisados a taxa de acerto (em porcentagem) dos algoritmos em questão e o tempo de execução destes algoritmos. Foi considerado o tempo de treinamento para os algoritmos supervisionados e tempo de geração de grupos para os algoritmos não-supervisionados. A taxa de acerto foi verificada observando a homogeneidade dos grupos gerados pelos algoritmos.

Após a análise dos algoritmos, foi utilizado o de classificação mais coerente e este foi implementado em uma aplicação que funciona em paralelo ao ScientiaNet. A aplicação permite o treinamento do algoritmo a cada 12 horas, realizando o aprendizado a medida que usuários forem cadastrando-se.

6. Resultados

Como foi dito na seção anterior, o desempenho dos algoritmos foi avaliado levando em consideração duas métricas: tempo de execução da classificação, desconsiderando o tempo que a aplicação busca os dados na base de dados, e porcentagem de acerto.

A tabela 1 mostra os tempos de execuções dos algoritmos para a classificação dos usuários considerando-se o tempo de geração de grupos.

O Cobweb apresentou a menor taxa de erro (Tabela 2), porém este apresentou resultados onde houve uma diferença na quantidade de grupos em relação aos outros algoritmos. Como o Cobweb cria os grupos de acordo com similaridade e não permite, como o k-means e a Rede de Kohonen, dizer quantos grupos se pretende ter, obteve-se uma quantidade média de 86 grupos gerados. Porém, destes grupos, após uma análise apenas 23 destes grupos apresentaram resultados satisfatórios. De acordo com estes resultados do Cobweb, os usuários que possuem área de interesse Engenharia de Software,

[7] http://www.generatedata.com

Tabela 1. Resultados dos tempos de execuções dos algoritmos

Algoritmo	Tempo de Execução (min)
Rede de Kohonen	5,77
Algoritmo de K-means	0,012
Algoritmo Cobweb	0,26

Banco de Dados e Pesquisa Operacional, foram divididos em sub grupos, diferenciando apenas o atributo Graduação.

Tabela 2. Resultados das taxas de erros e acertos dos Algoritmos

Algoritmo	Taxa de Acerto(%)	Taxa de Erro (%)
Rede de Kohonen	96,78	3,22
Algoritmo de K-means	86,74	13,26
Algoritmo Cobweb	98	2

Além dos 23 grupos, foram criados um grupo com 4 usuários de biologia e os grupos restantes foram criados com apenas um usuário ou nenhum usuários, totalizando uma média de 63 grupos sem nenhuma relevância na classificação. O Cobweb foi executado alterando o parâmetro *cuttof* [8] até atingir o valor 0.35 para obter o resultados atuais.

Outro algoritmo utilizado na classificação foi o Algoritmo K-means, utilizando um total de 20 centróides, que são a quantidades de grupos que se deseja ter ao final da execução. Assim o k-means também criou 18 grupos, sendo dois destes possuindo usuários de duas áreas de conhecimento. Tais grupos possuem, um com usuários de História e Direito e o outro com usuários de Medicina e Odontologia. O Algoritmo K-mens apresentou a pior taxa de erro entre os algoritmos utilizados nesta aplicação 2.

A rede de Kohonen utilizada neste trabalho, possui uma grade com 20 neurônios, pois haviam um total de 20 áreas de conhecimentos distintas, que representa a quantidade de grupos que se deseja obter,análogo ao Algoritmo K-means, sendo uma grade com dimensões 4 x 5, executando com 5000 iterações, com uma taxa de aprendizagem inicial de 0,5 e uma largura inicial de 0,8. Com esta configuração A Rede de Kohonen apresentou uma taxa de erro, que em relação ao algoritmo de K-means tornou-se aceitável e gerou um total de 18 grupos de usuários, e cada grupo é uma área de conhecimento distinta. Com exceção de dois grupos, onde a rede gerou um grupo com usuários de Direito e Economia e outro grupo com Medicina e Odontologia, assim como o Algoritmo K-means, devido as sub áreas em comuns desses desses usuários.

Com os resultados obtidos o algoritmo com melhor classificação foi implementado em uma aplicação que funciona em paralelo ao ScientiaNet. O Administrador do Sistema efetua o treinamento e o resultado deste é mostrado na tela de usuário do ScientiaNet como mostra a Figura 2.

[8]Este parâmetro define o grau de ramificação da árvore de classificação criada durante a execução do algoritmo, assumindo um valor no intervalo de 0 a 1. Quanto menor este valor mais ramos e sub-ramos serão criados pela árvore de classificação.

Figura 2. Tela de Perfil de Usuário do ScientiaNet.

7. Conclusão

Analisando os resultados do Algoritmo Cobweb, percebe-se que este gerou um quantidade de grupos exacerbada, pois uma grande quantidade dos grupos possui apenas um usuário, assim alguns usuários não iriam possuir contatos a ser recomendados pelo ScientiaNet. Os resultados do K-means, não apresentou grupos como o Cobweb, porém a taxa de erro desde algoritmo foi a maior entre os testados neste trabalho 2.

Apesar de apresentar o pior tempo de treinamento, a Rede de Kohonen mostrou-se eficiente, pois apresentou uma taxa de erro menor que o Algoritmo K-mens 2, e não apresentou grupos com apenas um usuário como o Algoritmo Cobweb, mesmo possuindo um tempo de treinamento muito alto. Com o objetivo de não permitir que o tempo de execução da Rede de Kohonen tenha grande relevância, o treinamento da Rede será executado em um servidor separadamente. Portanto acreditamos que a implementação da Rede de Kohonem auxiliará na classificação automática dos usuários do Scientia.Net.

Referências

da Cunha Recuero, R. (2004). Teoria das redes e redes sociais na internet: Considerações sobre o orkut, os weblogs e os fotologs. *XXVII Congresso Brasileiro de Ciências da Comunicação. XXVII INTERCOM*.

de Pádua Braga, A., de Leon Ferreira de Carvalho, A. P., and Ludermir, T. B. (2007). *Redes Neurais Artificiais: Teoria e Aplicações*. Rio de Janeiro, 2 edition.

Ferreira, G., Araújo, R., Orair, G., Gonçalves, L., Guedes, D., Ferreira, R., Furtado, V., and Junior, W. M. (2005). Paralelização eficiente deum algoritmo de agrupamento hierárquico.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. pages 139–172.

Goldschmidt, R. R. and Passos, E. (2005). *Data Mining: Um Guia Prático: Conceitos, Técnicas, Ferramentas, Orientações e Aplicações*.

Haykin (2001). *S.Redes Neurais: princípios e prática*. Porto Alegre.

Júnior, O. L. and Montgomery, E. (2007). *Redes Neurais: Fundamentos e Aplicações com Programs em C*.

Machado, V. P., de Lima, B. V. A., Arnaldo, H. A., and Araújo, S. W. I. (2011). Classificação automática dos usuários da rede social acadêmica scentia.net. *IV Congresso Tecnológico TI e Telecom .INFOBRASIL 2011*.

Mitchell, T. M. (1997). Machine learning.

of Waikato, U. (2011). *Weka 3 Machine Learning Software in Java*. University of Waikato.

Russell, S. and Norving, P. (2004). *Inteligência Artificial*.

Sousa, G. H. A. and Esmin, A. A. A. (2011). *Algoritmo de Enxame de Partículas Híbrido APlicado a Clusterização de Dados*.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning:An Introduction*. Cambridge.