Sistema de Mineração de Dados para Determinar Tendência Profissional

Souza, Gustavo F.
Centro Federal de Educação Tecnológica do Rio Grande do Norte, UNEDZN, CEFET-RN,
59112-450, Natal, RN
gustavo@cefetrn.br

Gorgônio, Flavius L., Santos, Sérgio P., Costa, José Alfredo F.
Laboratório de Sistemas Adaptativos, DEE, UFRN,
59072-970, Natal, RN
flavius@dca.ufrn.br, sergio_p_santos@hotmail.com, alfredo@dee.ufrn.br

Resumo: *A escolha da carreira profissional é um momento difícil na vida acadêmica de um estudante. Ao chegar ao último ano do ensino médio os alunos, muitas vezes, não tem condições de escolher corretamente o curso a ser realizado. O objetivo desde trabalho é utilizar informações sobre o candidato, como os cursos realizados, as notas nas disciplinas já cursadas no ensino médio e as influências aparentes, para predizer a área de atuação mais provável ou a distribuição de probabilidade para as principais áreas. Para tal, utilizaremos ferramentas de mineração de dados para determinar a área de conhecimento de maior probabilidade de afinidade para o aluno. Tais ferramentas permitem conhecer os grupos de afinidade dos alunos e, a partir da análise dos resultados obtidos, conjuntamente com outras ferramentas de análise estatística, sugerir opções relacionadas ao seu perfil. Espera-se que este trabalho possa auxiliar o estudante no processo de identificação do seu perfil profissional. O sistema não visa substituir o processo de orientação vocacional realizado pelos psicólogos, mas busca ajudar o estudante a se conhecer enquanto perfil profissional.*

Palavras-chave: mineração de dados, tendência profissional, análise de agrupamentos,

Introdução

O futuro é algo incerto para todas as pessoas. Ninguém pode prever o que acontecerá. Quem pode imaginar o que se passará no decorrer de alguns anos? Aos 16, 17 anos se imaginam milhões de possibilidades de futuro pessoal. É aí que está uma das grandes dúvidas da grande maioria dos jovens (e talvez a questão mais difícil do vestibular): "que curso fazer?" Esta é uma escolha decisiva, que vai determinar o que você fará nos próximos anos de sua vida, qual será a sua profissão. Momento crítico, em que pesa a responsabilidade da escolha que influenciará o restante da vida. É muita coisa para alguém tão jovem. Por isso, o número de abandonos e transferências de cursos em universidades no Brasil está crescendo. Cada vez mais jovens e imaturos, os estudantes ingressam em cursos sobre os quais não têm conhecimento suficiente. Desconhecem a profissão a que estão se candidatando. Assim, desistem e entram em um ciclo de desistência e mudança de curso, até descobrirem o que querem realmente. O processo acaba acontecendo na ordem inversa da natural, onde primeiro eles "praticam" o curso para depois conhecê-lo e saber se é isso ou não o que querem.

Escolher a profissão é ponto fundamental no projeto de vida de uma pessoa. É mais do que identificar uma vocação, baseada em sensações, informações, desejos, influências, etc. É, na verdade, definir com total segurança em que área do conhecimento e das relações humanas se deve atuar, visando a otimização da felicidade pessoal (satisfação, sucesso, autoconfiança, auto estima elevada, etc.), associada à máxima aplicabilidade da potencialidade intelectual/cultural [1].

A escolha de uma graduação acadêmica é um momento decisivo na vida, pois ocorre em ocasião na qual a personalidade ainda sofre os impactos finais da adolescência, não existindo um equilíbrio reflexivo ajustado a uma decisão de tal vulto, face aos comprometimentos envolvidos [2].

Além disso, o mundo do trabalho vem exigindo um investimento cada vez mais elevado em relação ao tempo de preparo para o ingresso na atividade profissional e com isso a responsabilidade na escolha da profissão se torna cada vez maior.

Nesse contexto, uma ferramenta computacional poderia ser de grande valia para ajudar os alunos a perceberem melhor informações sobre sua tendência profissional. Encontrar essas informações, analisar e visualizar dentro de um grande volume de dados é uma tarefa não-trivial para humanos [3]. Neste caso,

tanto em função do grande número de opções de cursos disponíveis atualmente, como pela grande quantidade de informações disponíveis sobre cada candidato.

Este trabalho visa apresentar uma metodologia, ainda em estudo, que possa auxiliar o aluno a identificar sua tendência profissional. Como a informação da área profissional não é conhecida a priori, o problema abordado pode ser considerado como um problema de clustering [3].

Metodologia

Inicialmente, as áreas de conhecimento foram organizadas em seis campos de aproximação, tomando como base a divisão interna feita por algumas instituições de ensino superior consultadas. Essas áreas abrigam os cursos que apresentam certa ligação entre si, sejam pela similaridade de conhecimentos, seja pela especificidade técnica e tecnológica.

As áreas de conhecimento são: Área 1: Ciências Biológicas, Área 2: Saúde, Área 3: Ciências Humanas, Área 4: Gestão e Negócios, Área 5: Ciências Exatas e da Terra e Área 6: Tecnológica. A Tabela 1 mostra alguns cursos abrigados por cada uma das seis áreas do conhecimento.

Tabela l: Cursos das Áreas de Conhecimento

Número	Área de conhecimento	Exemplos de Cursos
Área 1	Ciências Biológicas	Biologia, Aqüicultura, Biomedicina, Ecologia, Meio Ambiente.
Área 2	Ciências da Saúde	Medicina, Enfermagem, Fisioterapia, Farmácia, Odontologia, Nutrição, Educação Física.
Área 3	Ciências Humanas	Antropologia, Artes, Ciências Sociais, Filosofia, Geografia, História, Letras, Psicologia.
Área 4	Ciências Sociais Aplicadas	Administração, Biblioteconomia, Ciências Contábeis, Economia, Direito, Turismo, Pedagogia, Serviço Social.
Área 5	Ciências Exatas e da Terra	Estatística, Física, Geologia, Matemática, Química, Ciências da Computação.
Área 6	Tecnológica	Arquitetura, Cooperativismo, Engenharia Civil, Mecânica, Elétrica, de Materiais, da Produção, da Computação, Química, Têxtil e Agrônoma.

Essas áreas foram distribuídas em posições de um plano cartesiano de modo a traçar um hexágono, como mostra a Figura 1. Cada área está posicionada em um vértice da figura de tal forma que áreas com maior similaridade sejam colocadas mais próximas que outras com menor similaridade.

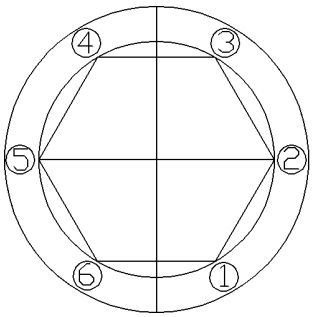

Figura l: Distribuição das áreas sobre o plano cartesiano.

O aluno, a priori, é representado na origem (0,0) do eixo de coordenadas. Neste ponto ele está eqüidistante de todas as áreas. Alguns atributos do indivíduo podem atraí-lo para determinada área e outros perdem repeli-lo.

Em seguida foram realizadas aplicações de questionários em três turmas de último ano do ensino médio, de aproximadamente 30 alunos de três escolas diferentes (uma escola estadual, uma técnica e uma particular de Natal), buscando a variabilidade dos resultados. Dentro do próprio questionário, pedimos para que os alunos indicassem o curso ou área profissional que pretendem seguir.

O sistema de determinação da área profissional

O sistema desenvolvido parte do princípio de que o aluno encontra-se na origem do sistema de coordenadas cartesiano e cada atributo deverá representar um vetor, que represente a influência desse atributo sobre o indivíduo na busca de uma área profissional.

Esse vetor contribui para a escolha da área profissional do individuo de acordo com a equação 01, que é a equação da soma de vetor ponderada por um fator α.

$$p(x,y) = p(x,y) + \alpha(p(x,y) - i(x,y)) \qquad \text{[Eq. 01]}$$

Caso haja repulsão é utilizada a mesma equação 01, mas o fator utilizado é β que atende as seguintes propriedades:

$$|\beta| < |\alpha| \quad \text{e} \quad \beta < 0 \qquad \text{[Eq. 02]}$$

As equações de atração e de repulsão são inspiradas nas equações de atualização de pesos da rede auto-organizáveis competitivas, que atualizam seus neurônios com um argumento não nulo apenas para o neurônio vencedor [4].

Estudos mais detalhados precisam ser realizados para a determinação do α e do β ótimos, bem como da relação entre eles. Pode ser considerado ainda, um par (α, β) diferente para cada atributo, o que pode melhorar ainda mais a eficiência do sistema. As equações da atração (eq. 01) e da repulsão são aplicadas diversas vezes a cada atributo, a fim de retirar as influências da ordem de apresentação dos atributos.

Outro atributo analisado são as notas obtidas pelos alunos nos anos anteriores, ou seja, no primeiro e no segundo ano do ensino médio, que foram normalizadas e transformadas em um único valor através do calculo da média aritmética.

Para incluir a influencia do rendimento acadêmico no sistema, programou-se a equação 03, mostrada que considera a própria média obtida como peso.

$$p(x,y) = p(x,y) + n^2(p(x,y) - i(x,y)) \qquad \text{[Eq. 03]}$$

onde n é a média normalizada do resultado obtido em cada disciplina.

O fator n^2 é o peso que pondera o passo do atributo. Foi utilizado a potencia para reduzir o passo, bem como acentuar diferenças entre notas como 6,0 e 8,0, no qual o 8,0 apresenta uma maior inclinação do que o 6,0.

Ao final das iterações de todos os atributos, obtemos a posição final para cada indivíduo no plano cartesiano (Pfinal), realizamos então o calculo da distância entre essa posição e a posição de cada área de conhecimento, de acordo com a equação 04. Observa-se que d é um vetor de 6 posições e representa as distâncias entre o individuo e as áreas de conhecimento.

$$d^i = \sqrt{(p^{final}{}_x - A_x^i)^2 + (p^{final}{}_y - A_y^i)^2} \qquad \text{[Eq. 04]}$$

Onde A^i_x é a coordenada x da área i;

Após o calculo do vetor **d** para cada individuo, realizamos a inversão dos valores a fim de obtermos uma escala linear e direta, uma vez que a escala de distância é inversa, isto é maior distância significa menor tendência. A equação 05 foi utilizada para realizar essa inversão. Por fim, a equação 06 é utilizada para atribuir aos valores um caráter percentual, facilitando assim o entendimento dos resultados.

$$t^i = \frac{1}{d^i} \qquad \text{[Eq. 05]}$$

$$t^i(\%) = \frac{100t^i}{\sum_{i=1}^{6} t^i} \qquad \text{[Eq. 06]}$$

Conclusões e considerações finais

O sistema de orientação vocacional apresentado faz uso de informações acadêmicas para estimar a tendência de área de conhecimento que deve ser seguida pelo individuo na sua vida acadêmica futura. É importante lembrar que diversos fatores podem influenciar nessa decisão e neste sistema não são consideradas. O sistema faz uso do hexágono das áreas apresentado e incrementa as informações na forma de um vetor de tendência. A somatória dessas tendências mostra a posição final que deverá ser comparada com as posições das áreas para determinar a tendência profissional.

Entre as vantagens desse sistema pode-se considerar o fato de se utilizar tipos de dados diferentes (numéricos e categóricos). Não houve como validar os resultados, uma vez que não tínhamos a priori a informação da área profissional dos alunos.

Ao compararmos os resultados obtidos no sistema com os dados informados pelos alunos, constatamos uma grande divergência, isso se deve ao fato dos alunos escolherem cursos por diversas razões, muitas vezes desprezando as aptidões e indicando cursos que estão na moda ou simplesmente por que pagam melhores salários.

Não é objetivo desse sistema a substituição do profissional competente para a orientação vocacional, mas o de auxilio, visto que em muitos casos informações objetivas podem contribuir para uma análise mais completa das escolhas profissionais. Como o custo e o tempo necessário para a realização do teste é muito baixo, sugere-se a realização do teste e caso haja divergência entre o interesse do aluno e o resultado ou conflito entre os percentuais, deve ser consultado um profissional da área e efetuada uma orientação mais detalhada.

A informação do curso a ser seguido não foi considerada, pois existe uma grande diversidade de cursos, alguns que só existem em algumas universidades e outros somente no exterior, além do fato de que diversos cursos apresentarem diferenças muito pequenas o que torna difícil o diagnóstico. Para o caso da escolha do curso, sugerimos que o aluno procure obter informações nas universidades ou com profissionais daquele curso para definir a escolha correta.

De uma forma geral, esperamos ter contribuído com a indicação de uma área e que o sistema possa ajudar alunos e profissionais na árdua tarefa de escolher o curso que irão realizar.

Uma sugestão de trabalhos futuros é o de ampliar a base de dados, no sentido de ter maior quantidade de alunos, bem como informações mais precisas e significativas. A utilização da internet para as respostas dos questionários é uma tática possível e que poderá facilitar a obtenção desses dados. A utilização do SOM na separação dos indivíduos deve ser mais investigada, a fim de determinar grupos de interesses baseado nos atributos.

Agradecimentos

Agradecemos aos professores José Ribamar (CEFET-RN), Elias Custódio (GEO Natal) e Lourdes (Atheneu Norte-Riograndense) e a Úrsula Reis e Odara Sá pelo apoio dado á pesquisa e pela cooperação na aplicação dos questionários.

Referências

[l] Pimenta, S.G. (1981) Orientação vocacional e decisão: estudo crítico da situação no Brasil. 2ed. São Paulo, Ed. Loyola.

[2] Ferretti, C.J. (1988) Uma nova proposta em orientação profissional. São Paulo: Cortez

[3] Larose, Daniel T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, New Jersey, 2004.

[4] Haykin, Simon (2001). Redes Neurais. Princípios e prática. Porto Alegre: Bookman.

.