COBENGE 2014
ENGENHARIA: Múltiplos saberes e atuações
16 a 19 de setembro | Juiz de Fora - MG

IDENTIFICAÇÃO DE GRUPOS DE ALUNOS EM AMBIENTE VIRTUAL DE APRENDIZAGEM UTILIZANDO ANÁLISE DE LOG BASEADA EM CLUSTERIZAÇÃO

Márcia Fontes Pinheiro[1] - eng.marciafontes@gmail.com

Luiz Cortinhas Ferreira Neto[1] - luizcf14@gmail.com

Haroldo Nazaré de Sá Junior[1] - hnsj86@gmail.com

Eulália C. da Mata[1] - eucmata@gmail.com

Antonio F. L. Jacob Jr.[1] – jacobjr@ufpa.br

Ádamo de Lima Santana[1] - adamo@ufpa.br

[1]Universidade Federal do Pará – UFPA
Laboratório de Planejamento de Redes de Alto Desempenho – LPRAD
Rua Augusto Corrêa, 01, Guamá, Caixa postal 479
CEP: 66075-110 – Belém – PA

Resumo: *Este trabalho propõe uma metodologia de pesquisa para avaliação de desempenho acadêmico dos alunos de um programa educacional de larga escala, no curso de inclusão digital ministrados semi-presencialmente pela plataforma Moodle, utilizando clusterização para separação dos alunos em quatro níveis (Exelente, Bom, Regular, Insuficiente) . Assim obtendo o caminho comum dentro da plataforma para os cinco clusteres (um cluster é formado por alunos que não se enquadram nos níveis descritos). Esta pesquisa foi motivada pela incerteza do aproveitamento dos alunos na plataforma de acordo com o perfil do aluno. O estudo de caso consiste na formação de agentes de inclusão digital, com ênfase na utilização de redes sociais online. No curso ministrado, as redes sociais são planejadas para funcionar como espaço de articulação de projetos comunitários envolvendo agentes de inclusão digital e comunidades e compartilhar soluções para problemas técnicos do dia-a-dia, indicando apropriação e envolvimento da comunidade nesses espaços.*

Palavras-chave: *Inclusão Digital, K-means, Avaliação de Desempenho, Educação, Perfil do aluno.*

1. INTRODUÇÃO

Com o advento das Tecnologias da Informação e Comunicação (TICs), Internet, dos sistemas de informação para Web e dos recursos multimídia, surgiram diversas ferramentas e plataformas com a proposta de Web 2.0.

No campo da educação, a utilização desses sistemas de ensino eletrônico – e-learning – proporcionam novas possibilidades de aprendizagem na metodologia de ensino. Esses sistemas são denominados como Ambientes Virtuais de Aprendizagem (AVAs) por serem plataformas que centralizam ferramentas, mas que permitem interação entre os usuários de forma assíncrona e síncrona através de Wiki, chat, fórum, etc.

Os AVAs armazenam todas as interações dos usuários dentro da plataforma, informações do desempenho do aluno, podendo guardar informações como quais atividades um estudante participou, materiais que foram lidos e escritos, testes ao qual foi submetido, chats que o mesmo participou, páginas acessadas na plataforma, etc. (MOSTOW *et al*., 2005), assim como informações pessoais sobre usuário, tais como perfil, em sua base de dados.

Dentre os AVAs, temos o Moodle, que é um sistema *open source* desenvolvido pelo australiano Martin Dougiamas para uso acadêmico. Engloba diversas ferramentas web 2.0 para facilitarem a interação entre os usuários (MATA et al., 2010) e é a plataforma mais utilizada no mundo em educação a distância (CAPTERRA, 2014).

De maneira geral, os AVAs armazenam uma grande quantidade de informação o qual é muito valiosa para analisar o comportamento dos estudantes (MOSTOW & BECK, 2006) e que pode ser obtida através de Descoberta do Conhecimento em Base de Dados (DCBD), da expressão em inglês Knowledge Discovery in Databases (KDD), que é a extração automática de padrões implícitos e interessantes a partir de grandes volumes de dados (KLOSGEN & ZYTKOW, 2002). Para encontrar padrões nesses dados são envolvidos a mineração de dados (*Data Mining*), extração do conhecimento, descoberta da informação e padrão de processamento dos dados (FAYYAD, 1996).

KDD é uma área multidisciplinar o qual várias técnicas computacionais de aprendizado de máquina podem ser aplicadas, tais como árvores de decisão, redes neurais artificiais, redes bayesianas, etc. As tarefas de KDD mais utilizadas são métodos estatísticos, visualização, clusterização, classificação e regras de associação. Estes métodos podem descobrir novos conhecimentos interessantes e úteis com base em dados de uso dos alunos, podendo ser útil não somente para os educadores, mas também aos próprios alunos, uma vez que pode ser orientada para diferentes fins por diferentes participantes no processo.

Este trabalho apresenta a categorização do caminho na plataforma Moodle que melhor represente os alunos com maior aproveitamento do curso de inclusão digital.

Nos cursos oferecidos não existe métrica para avaliação de desempenho que seja satisfatória em tentativa de solucionar neste trabalho é proposto como metodologia à aprendizagem de log (JANSEN, 2006) adaptada para plataforma Moodle utilizada em grande escala, cursos de incentivos públicos e privados semi-presenciais brasileiros, basicamente definida em 3 etapas a metodologia utilizada classifica os alunos em três níveis de aproveitamento: “BOM”,”REGULAR” e “INSUFICIENTE”, pela abordagem de log, citada detalhadamente na sessão 3, a partir das classes será feita a analise de padrões procurando o caminho médio realizado pelos cluster’s na plataforma.

2. REDE TELECENTROS.BR

No Brasil, dentre as políticas públicas para favorecimento da inclusão digital, têm-se investido na criação e na utilização de centros tecnológicos comunitários, ou telecentros, onde o acesso público às TICs é disponibilizado para as comunidades menos privilegiadas a um custo mínimo ou isento de custos. Neste contexto, além da implantação dos telecentros para acesso à Internet, têm-se a formação de agentes para inclusão digital como aspecto crítico. A proposta envolveu um programa de formação em larga escala que requereu mecanismos de controle e monitoramento para gestores e beneficiadores desse programa, a Rede Nacional de Formação para Inclusão Digital – Rede Telecentros BR (SILVA, 2013).

A Rede Telecentros BR até o ano de 2011 contava com a participação de cinco polos regionais (um polo para cada região do País), dois polos estaduais (nos estados de São Paulo e Ceará) e um polo nacional. Sob a responsabilidade dos polos regionais estava a formação dos agentes de inclusão digital (que são os monitores dos telecentros), gestores de telecentros (responsáveis pela administração do telecentro), tutores (que atuavam na formação dos monitores) e supervisores de tutoria (responsáveis pela supervisão e acompanhamento do trabalho dos tutores) (SILVA, 2013).

No período de fev/2010 a dez/2012, os membros dos polos de formação se articularam para construir e aplicar o Curso de Formação de Monitores dos Telecentros e a ativação das redes sociais de agentes de inclusão social atuantes nas comunidades. O projeto de formação dos agentes de inclusão digital (monitores) contemplou a oferta de um curso de 480 horas, disponibilizado na plataforma Moodle (Rede Telecentros.BR, 2013) e dividido em dois módulos: 80 horas para um uma breve apresentação dos conteúdos da formação e 400 horas com foco específico no desenvolvimento de projetos comunitários e aprofundamento dos conteúdos. Para o desenvolvimento dos projetos comunitários, os agentes de inclusão digital percorreram de acordo com seus interesses e necessidade, sem percurso pré-definido, diferentes temas: comunicação comunitária, redes, cultura digital, comunidade, telecentros. No desenvolvimento do curso, os agentes de inclusão digital contaram com o apoio de tutores e supervisores de tutores dos diversos polos regionais (SILVA, 2013).

A formação contribuiu para que os agentes de inclusão digital usassem as TIC como ferramentas para alavancar transformações sociais na comunidade em que estavam inseridos. Neste contexto, os agentes de inclusão digital deveriam alcançar domínio técnico e instrumental das ferramentas relacionadas às tecnologias da informação e comunicação, atuando de forma solidária, cooperativa e interativa com os seus colegas de formação e de trabalho, com condições de se reconhecer e de atuar como agente de transformação social na comunidade os quais estavam inseridos (SILVA, 2013).

2.1 Plataforma Moodle

O Moodle é o Ambiente Virtual de Aprendizagem (AVA) selecionado para o desenvolvimento do Curso de Formação de Monitores do Telecentros.BR. A Rede Nacional de Formação para Inclusão Digital produziu o curso de formação de monitores do Telecentros.BR com objetivo de propiciar o desenvolvimento de habilidades no uso de tecnologias da informação e comunicação para dar condições de transformações sociais na comunidade. O curso foi estruturado em dois eixos pedagógicos: acessos a conteúdos e atividades formativas; elaboração e implementação de projetos comunitários. Ofertado em duas fases: a primeira englobava conhecer o ambiente virtual e abordagem prévia do conteúdo que será aprofundado na fase dois que aborda

principalmente o projeto comunitário, onde cada tema estudado poderá servir de apoio para realização de ações com a comunidade. Entre os temas abordados: telecentros, comunidade, comunicação comunitária, inclusão digital, redes e cultura digital.

No ambiente virtual de aprendizagem – moodle foi organizado o curso para formação dos monitores em: Fase 1: Ambientação e Voo Rasante, como mostram respectivamente as Figuras 1 e 2; Fase 2: Projeto Comunitário, como mostra a Figura 3. Mas também foi necessário realizar um curso para formação de tutores que atuaram auxiliando e orientando os monitores durante a formação do telecentros.br. No curso dos tutores foi abordado educação a distância, articulação social e inclusão digital. Os tutores foram selecionados com base no conhecimento sobre tecnologia, valorizando a experiência em inclusão digital.

Figura 1. Fase 1: Ambientação

Figura2. Fase 1: Voo rasante

Figura 3. Fase 2: Projeto Comunitário

O processo de formação foi realizado em grupo de "n" monitores para um tutor, que fizeram o acompanhamento para o primeiro acesso a plataforma, incentivaram a participação no ambiente virtual e no desenvolvimento do projeto comunitário, assim como realizaram a avaliação dos monitores mensalmente. Os tutores tiveram o acompanhamento e auxilio dos supervisores de tutoria no processo de formação. Os supervisores eram pessoas que fizeram parte dos polos regionais, facilitadores da formação, acompanhavam os tutores e monitores no processo de formação,

realizaram também avaliação das turmas com base nas informações levantadas na plataforma Moodle e com as informações repassadas pelos tutores. A coordenação pedagógica realizou acompanhamento dos supervisores e tutores com base em informações obtidas na plataforma e no sistema de avaliação.

3. K-MEANS

K-means é um método de clusterização (organização dos objetos similares, em algum aspecto, em um grupo) pertencente a classe dos algoritmos de aprendizados de máquina. Tem como finalidade dividir um determinado número de objetos em áreas chamadas de cluster. Estas áreas são constituídas por uma coleção de objetos que são similares entre si e diferentes dos objetos pertencentes aos outros clusters. A distância de cada objeto para cada cluster vai determinar em que cluster o objeto ficará alocado. Cada objeto pertence ao cluster no qual possui o elemento central (centróide) mais próximo deste objeto.

O primeiro uso do termo *k-means* foi empregado por James MacQueen no ano de 1967 em seu artigo "Some Methods for Classification and Analisys of Multivariate Observations". (MACQUEEN, 1967) explica em seu artigo como o K-means funciona e demonstra que este algoritmo é facilmente programável e computacionalmente econômico, sendo possível processar grandes amostras em um computador digital.

A figura a seguir ilustra os passos do funcionamento do algoritmo k-means:

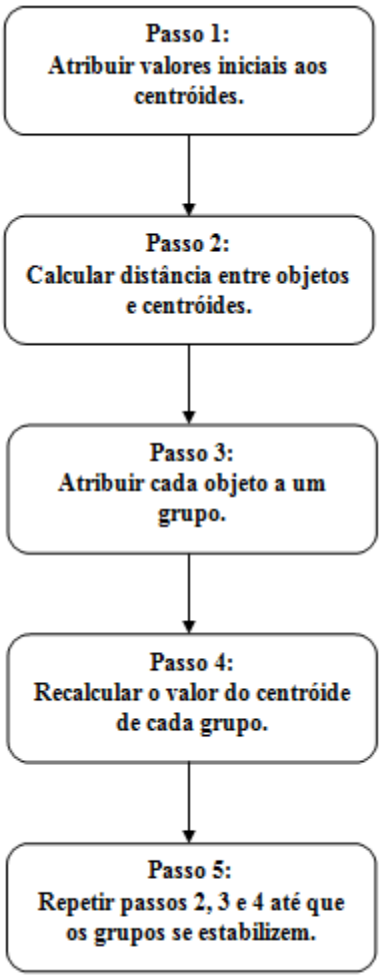

Figura 2. Passos do algoritmo K-means.

No primeiro passo, atribui-se valores aos k primeiros centróides seguindo algum critério. Geralmente é escolhido os k primeiros pontos da tabela para cada centróide. No entanto, estes valores também podem ser gerados aleatoriamente. Após atribuído os valores, há o cálculo da distância de cada objeto para cada centróide. O total de distâncias corresponde ao produto de N objetos por k centróiodes (N * k). Tendo a distância dos pontos para os centróides, os objetos são classificados de acordo com a classe a qual o objeto está mais perto do centróide, formando os clusters. Com os grupos formados, calcula-se novamente o valor do centróide, de acordo com a posição dos objetos pertencentes ao grupo. O novo centróide é calculado através da média de cada atributo de todos os pontos pertencentes a classe. Ao final deste processo, o algoritmo volta ao passo 2, repetindo os passos iterativamente até o ponto em que os centróides não mudem de posição.

Este algoritmo pode ser usado em diversas situações. (MACQUEEN, 1967) afirma que as possíveis aplicações do k-means estão incluídas em métodos para agrupamentos por similaridade, previsão não-linear, aproximação de distribuições multivariadas e testes não-paramétricos de independência entre várias variáveis.

4. METODOLOGIA

Para a realização deste trabalho foi utilizada a base de dados do Moodle utilizado na Rede Telecentros BR por aproximadamente mil e quatrocentos alunos (monitores) no curso de formação para inclusão digital promovido pela Rede Nacional de Formação para Inclusão Digital, a qual foi responsável pela construção, manutenção e aplicação do referido curso. A coleta de dados ocorreu de maneira implícita através do uso e interação dos monitores na plataforma Moodle e armazenamento na base de dados da plataforma. O Moodle armazena todas as interações dos usuários na plataforma em forma de *logs*: cada clique que o usuário realiza na plataforma para fins de navegação, bem como detalhes das atividades que os estudantes participaram.

O Moodle armazena os logs em uma base de dados relacional, com cerca de 145 tabelas inter-relacionadas, porém neste trabalho as informações utilizadas encontravam-se somente na tabela mdl_log responsável pelo armazenamento de eventos do sistema. O pré-processamento consistiu na limpeza das informações das demais tabelas do Moodle em conjunto com o filtro para eliminar alunos excluídos e administradores do sistema procurando mitigar a interferência destes grupos de usuários.

Apartir da etapa de pre-processamento a clusterização é permitida usando o algortimo K-means para duas dimensões onde são tratadas as colunas: “course” e “userid”, contando o número de acessos de cada usuário para cada curso, também são filtrados sómente eventos dentro da categoria “course”.Após a composição do *dataset* para o algoritimo de clusterização o mesmo séra executado para: k=5 pois são estabelecidos pelo programa quatro niveis de avaliações qualitativas: Excelente, Bom, Regular e Insuficiente, contudo foi acrescentada uma classe para que *outliers* (valores fora do esperado) possam ser minimizados nas outras classes (FAYYAD, U et al, 1996).

Os resultados do K-means são grupos definidos que passam pela analise de quantitativos de alunos,acessos e cursos.

Apartir do mesmo é possivel inferir caracteristicas comuns entre grande numero de usuarios no mesmo grupo e ate mesmo o grupo dos usuarios mais experientes na plataforma, chamados monitores.

5. RESULTADOS

Para o pré-processamento os numero de linhas da base foram resumidas para 13283 linhas, pois as linhas de log ficaram restritas as caracterisitcas citadas na metodologia.
Para o Kmeans foram encontrados os 5 clusteres com as seguintes caracteristicas:

- No cluster 1 foram contados 200848 acessos em 19 cursos distintos e 4046 alunos, cerca de 49,64 acessos por aluno e 2,61 acessos por aluno e por curso;
- No cluster 2 foram contados 69544 acessos em 2 cursos para 7 alunos, cerca de 9934,85 acessos por aluno e 4967,42 acessos por aluno e por curso;
- No cluster 3 foram contados 562578 acessos em 40 cursos e 1488 alunos, cerca de 378,07 acessos por aluno e 9,45 acessos por aluno e por curso;
- No cluster 4 foram contabilizados 349423 acessos em 30 cursos e 7603 alunos, cerca de 45,958 acessos por aluno e 1,53 acessos por aluno e por curso;

- No cluster 5 foram contados 179516 acessos em 24 cursos e 140 alunos, cerca de 1282,25 acessos por aluno e 53,42 acessos por aluno e por curso.

É importante ressaltar o tempo de execução em 15.69 segundos, para o pré-processamento e k-means. Todos os procedimentos foram executados no mesmo computador com os seguintes requisitos de hardware: Processador Core i7-3632QM @ 2.20Ghz, Memoria de 4GB @1333Mhz e sistema operacional Linux distribuição Ubuntu 14.04.

6. CONSIDERAÇÕES FINAIS

Como é objetivado neste trabalho os resultados demonstraram grupos distintos de usuários apartir da clusterização do log da plataforma Moodle, mais detalhadamente é possivel reconhecer com base nos resultados a distancia entre o maior e menor cluster isso ocorre pois no caso do cluster 2, pelo pequeno numero de usuários e grande numero de acessos o grupo demonstra ser formado provavelmente por usuários que administram o sistema, chamados Monitores, é facilmente identificado nos clusteres: 1,3 e 4 números próximos porem diferenciados pelo numero de acesso aos cursos no grupo 4 é possível inferir que estão localizados alunos com acesso regular ao sistema, já nos grupos: 3 e 1, respectivamente possuem crescimento de numero de acessos por aluno e por curso: maior para o primeiro e menor para o segundo.

Estes resultados são de extrema importancia para o gerenciamento dos cursos no moodle e analise quantitativa da analise qualitativa feita no programa Telecentros BR.

7. REFERÊNCIAS BIBLIOGRÁFICAS

CAPTERRA (2014). The top 20 most popular LMS Soft ware solutions. http://www.capterra.com/top-20-lms-software-solutions.UZJRP7WnCSp.

FAYYAD, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery: An Overview, in Advances in Knowledge Discovery and Data Mining, R. Uthurusamy, eds., MIT Press, Cambridge, Mass., 1996, pp. 1-36.

KATYAL, V.; AVIRAL; SRIVASTAVA, D. Elimination of Glass Artifacts and Object Segmentation. International Journal Of Computer Applications. [s.i.], abr. 2012.

KLÖSGEN, W., & Zytkow, J. Handbook of data mining and knowledge discovery. New York: Oxford University Press, 2002

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

MOSTOW, J., et al. An educational data mining tool to browse tutor–student interactions: Time will tell! In Proceedings of the workshop on educational data mining, 2005, pp. 15–22.

MOSTOW, J.; Beck, J. Some useful tactics to modify, map and mine data from intelligent tutors. Journal *Natural Language Engineering, vol.* 12, 2006, pp. 195-208

PICHILIANI, Mauro. **Data Mining na Prática:** Algoritmo K-Means. 2006. Disponível em: http://imasters.com.br/artigo/4709/sql-server/data-mining-na-pratica-algoritmo-k-means. Acesso em: 14 jun. 2014.

SILVA, A., et al. Análise de Redes Sociais para avaliação e monitoramento de programas de treinamento em larga escala baseados no uso de ambientes de aprendizagem e redes sociais online. II Brazilian Workshop on Social Network Analysis and Mining, 2013.

IDENTIFICATION OF STUDENT GROUPS IN A VIRTUAL LEARNING ENVIRONMENT USING LOG ANALYSIS BASED IN CLUSTERING

Abstract: *This paper proposes a research methodology for the new evaluation of student academic performance from an web based educational program on a large scale, in the course of digital inclusion offered in Moodle platform, using clustering for separation of students into three levels. So obtaining the common path on the platform for the three clusters. This research was motivated by the uncertain performance of the students on the platform in according to student profile. The case study consists in the formation of digital inclusion agents, with emphasis on the utilization online social networks. In the ministered course, social networks are designed to operate as a place of articulating community projects involving digital inclusion agents and communities and share solutions to technical problems of day-to-day stating appropriation and community involvement in these places.*

Key-words: *Digital Inclusion, K-means, Performance Evaluation, Education, Student Profile.*