Avaliação dos impactos de abordagens de pré-processamento em técnicas de filtragem colaborativa no repositório educacional Merlot

Henrique L. dos Santos[1]; Cristian Cechinel[2]

[1]*Universidade Federal de Pelotas – hldsantos@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – contato@cristiancechinel.pro.br*

1. INTRODUÇÃO

Objetos de aprendizagem (OAs) são unidades auto-contidas de aprendizado importantes na composição de programas de educação a distância (EaD).
A forma como estes recursos educacionais são disponibilizados pode permitir que os alunos possuam certa liberdade ao definir qual caminho de aprendizagem seguirão, de acordo com suas preferências, conhecimentos anteriores e habilidade.

Coleções de objetos de aprendizagem são normalmente armazenadas e organizadas em Repositórios de Objetos de Aprendizagem (ROAs). Os repositórios existentes podem apresentar diferenças em vários aspectos (localização do objeto, especificidade da área de cobertura do repositório, padrão de metadados adotado) e atender diferentes comunidades, países ou níveis de educação.

Estes repositórios podem conter dezenas de milhares de OAs de diferentes categorias e níveis de aprendizagem, dificultando a visualização geral e a busca por conteúdo por parte de seus usuários. Em busca de uma solução para tal questão, pode-se assumir que as avaliações e os comentários adicionados pelos usuários aos objetos podem contribuir na busca e recuperação de informação (CECHINEL et al., 2011).

No presente trabalho, busca-se identificar se ações de pré-processamento na base de dados de avaliações do repositório Merlot (www.merlot.org) melhoram a recomendação baseada em filtragem colaborativa entre usuários. Vale ressaltar que técnicas de filtragem colaborativa pura já foram aplicadas e avaliadas nesse repositório anteriormente por (CECHINEL et al., 2013), de forma que as propostas aqui testadas visam uma alternativa à abordagem clássica.

A hipótese aqui testada é de que a aplicação de métodos de pré-processamento, em especial a clusterização, não somente tornem a filtragem colaborativa escalável mas também melhorem a precisão das recomendações geradas. Nesse sentido, serão avaliadas taxas de erro *offline* e também a cobertura do espaço de usuários alcançada por cada método.

2. METODOLOGIA

Foram coletadas 9910 avaliações (que variam num intervalo fechado de 1 a 5) feitas por 3659 usuários sobre 4968 diferentes objetos de aprendizagem. Além disso, para possibilitar a aplicação das técnicas de pré-processamento, também foram coletadas algumas informações sobre esses usuários e OAs. Sobre os usuários foram extraídas todas as categorias às quais eles estavam associados no repositório e sobre os objetos foram extraídas suas descrições em forma de texto livre.

O primeiro método aplicado foi o de clusterização de usuários por meio de suas diversas categorias. Cada usuário foi representado por um arquivo de texto contendo, no modelo saco-de-palavras, todas as categorias em que ele se encontrava cadastrado no repositório. Um exemplo de representação de usuário pode ser mostrado como: *Arts*, *Music*, *Music Teaching, Music History*. Após essa categorização, cada usuário, ou cada texto simbolizando um usuário, foi transformado em um vetor de palavras, onde o valor de cada posição se referia ao valor TF-IDF da palavra no conjunto de arquivos gerados. Nesse caso, destacam-se palavras com frequência elevada no arquivo, porém não tão elevada no conjunto total de arquivos. Após isso, tendo cada usuário representado por um vetor, foi possível agrupá-los (clusterização) de forma que usuários teoricamente semelhantes em categorias educacionais permanecessem juntos. A quantidade de *clusters* geradas (*k)* foi uma variável do experimento, de forma que foram testados valores de 2 a 9, já que valores maiores acarretaram valores muito baixos na cobertura do espaço de usuários.

No segundo método, um processo análogo ao primeiro foi conduzido. Entretanto, nessa tentativa, foram clusterizados objetos de aprendizagem de acordo com suas descrições, também no modelo saco-de-palavras. O processo se deu de forma semelhante, com valores de *k* também variando de 2 a 9.

De posse dos *clusters* de usuários e de OAs, foi possível recortar a base de dados de avaliações, de forma criteriosa, para posteriormente gerar as recomendações para cada caso.

Para a geração de recomendações foi utilizada uma abordagem baseada em usuário, de filtragem colaborativa, onde os seus principais parâmetros, similaridade mínima entre usuários e tamanho de vizinhança, também foram variados, o primeiro de 0.2 a 0.9 e o segundo de 2 a 20. Para cálculo de similaridade, foi utilizada a relação *LogLikelihood* que mede a similaridade entre usuários de acordo com a contagem de eventos (avaliações) entre eles que ocorrem simultaneamente (usuários que avaliam o mesmo objeto) (DUNNING, 1993).

A fim de possibilitar uma medida avaliativa confiável, foram utilizadas métricas *offline* de avaliação. De forma que, para cada *cluster,* 90% de suas avaliações foram utilizadas para treinamento, e o restante para teste. Nesse sentido, recomendações geradas no treinamento são comparadas com os 10% de avaliações presentes no grupo de teste caso haja uma co-ocorrência, isto é, predição gerada para um usuário X sobre um OA Y também exista no grupo de teste, a diferença entre os valores de preferência (1 a 5) é, então, calculada. Mais especificamente, será apresentada a taxa de erro conhecida por RMS (raiz do erro quadrático médio), já utilizada anteriormente por (HERLOCKER et al., 2004) para medir a precisão de sistemas de recomendação. Além disso, para cada configuração possível, de tamanho de vizinhança e similaridade mínima, as recomendações foram geradas e tiveram seus erros calculados 50 vezes, assim, a média desses valores foi obtida.

Para todas estas etapas, foi utilizado o framework Apache Mahout versão 0.7 (http://mahout.apache.org/).

3. RESULTADOS E DISCUSSÃO

A Figura 1 mostra os resultados, em forma de box-plot, para ambos os experimentos. Como pode ser observado, a abordagem de clusterização de OAs supera a abordagem pura (k=1) em quase todos os valores de k, com exceção de k = 3. Para a proposta de clusterização de usuários, o mesmo padrão se repete.

Na comparação entre as duas abordagens, se verifica que os melhores casos ainda estão presentes na clusterização de OAs (especialmente para k=6 e k=8).

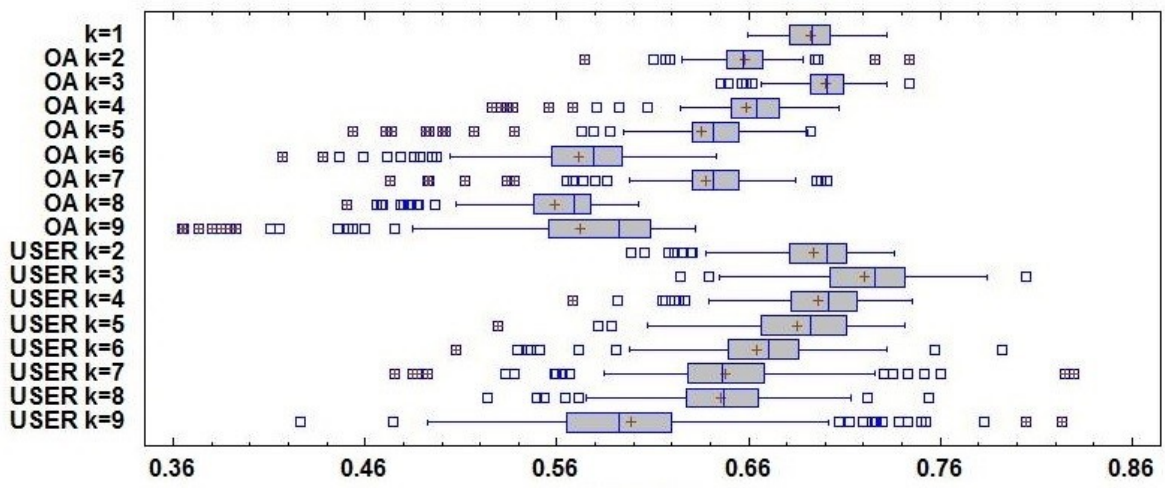

Figura 1. Box-plots para as duas abordagens e também para a abordagem pura (k=1) considerando o erro RMS médio.

Outra importante análise é a de que, a medida em que k aumenta, há um aumento também na dispersão das medidas do erro, indicando uma certa instabilidade na avaliação das recomendações. Isso pode ser justificado com ajuda da Figura 2, que mostra a porcentagem de cobertura do espaço de usuários de acordo com os valores de k. Nota se que, naturalmente, quando k aumenta (mais divisões são feitas no conjunto total de avaliações), menos usuários são atingidos pelas recomendações. Pode-se dizer que tanto essa baixa cobertura quanto a alta dispersão da taxa de erro (visto na figura anterior) é causada pela esparsidão em demasia na matriz de avaliações (usuários vs. OAs).

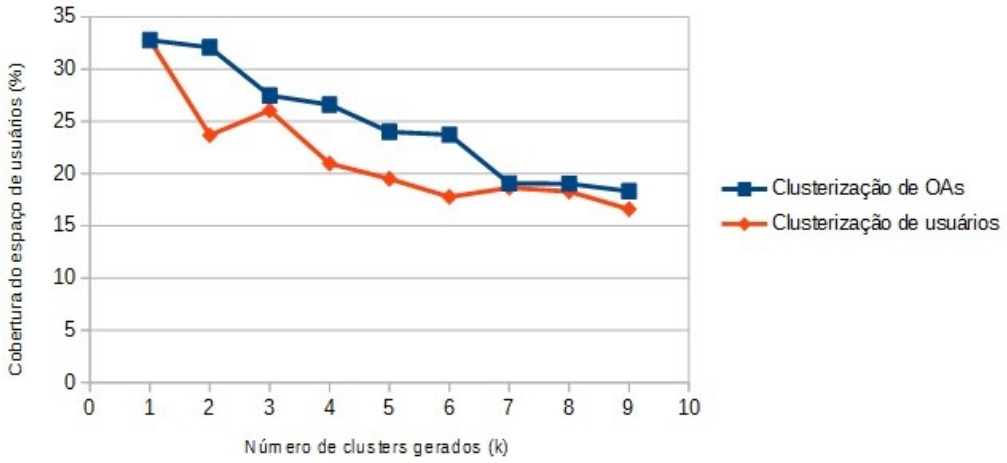

Figura 2. Cobertura do espaço de usuários para ambas as abordagens.

Balanceando as duas medidas (erro e cobertura) pode-se dizer que a melhor escolha encontra-se na clusterização de OAs para k = 6, que obtém uma cobertura de cerca de 25% e um erro RMS médio de 0.56.

4. CONCLUSÕES

O estudo de técnicas para melhorar o processo de recomendação em repositórios de aprendizagem tem se desenvolvido com certo destaque nos últimos anos. Em (TANG, 2005), por exemplo, se propõe um agrupamento de alunos através de seus interesses de aprendizagem a fim de atingir um alto grau de adaptação e personalização de suas atividades em um ambiente de *e-learning*. Outros trabalhos, como o de (VERBERT et al., 2012) argumentam que é necessário que o sistema recomendador considere informações referentes ao contexto educacional no qual o usuário se encontra.

No presente trabalho, mostra-se que uma estratégia de clusterização, tanto de usuários mas especialmente de objetos de aprendizagem, pode não somente melhorar a escalabilidade do sistema (consequência natural) como também tornar o recomendador mais preciso.

5. REFERÊNCIAS BIBLIOGRÁFICAS

CECHINEL, C.; SÁNCHEZ-ALONSO, S.; GARCÍA-BARRIOCANAL, E. Statistical profiles of highly-rated learning objects. **Computers & Education**, Amsterdam, v.57, n.1, p.1255-1269, 2011.

CECHINEL, C.; SICILIA, M.A.; SÁNCHEZ-ALONSO, S.; GARCÍA-BARRIOCANAL, E. Evaluating collaborative filtering recommendations inside large learning object repositories. **Information Processing & Management**, Amsterdam, v.49, n.1, p.34-50, 2013.

DUNNING, T. Accurate methods for the statistics of surprise and coincidence. **Computational Linguistics**, Cambridge, v.19, n.1, p-61-74, 1993.

HERLOCKER, J.; KONSTAN, J.; TERVEEN, L.; RIEDL, J. Evaluating Collaborative Filtering Recommender Systems. **ACM Transactions on Information Systems**, Nova Iórque, v.22, n.1, p5-53, 2004.

TANG, T.; MCCALLA, G. Smart Recommendation for an Evolving E-Learning System: Architecture and Experiment. **International Journal on E-Learning**, Norfolk, v.4, n.1, p-105-129, 2005.

VERBERT, K.; MANOUSELIS, N.; OCHOA, X.; WOLPERS, M.; DRACHSLER, H.; BOSNIC, I.; DUVAL, E. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. **Learning Technologies, IEEE Transactions on**. Los Alamitos, v.5, n.4, p-318-335, 2012.