Monitoramento Automático de Mensagens de Fóruns de Discussão Usando Técnica de Classificação de Texto Semi-Supervisionado

Roberto L. de Oliveira Júnior[1,2], **Ahmed A. A. Esmin**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Lavras (UFLA)
Lavras, MG - Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG - Brasil

robertolojr@dcc.ufmg.br

ahmed@dcc.ufla.br

Abstract. *The foruns are resources very used in E-learning to information exchange and approach between teachers, tutors and students. Automatic monitoring environments to these forums are essentials tools for management and consequently in the support of quality improvement of E-learning courses. However, the development of these kind of environments with data mining techiniques and, in particular, supervised classification is a hard task for many reasons, among them, lack of labeled data to be used in the construction and training of these classifiers. This paper aims show an automatic monitoring environments of forums with the semi-supervised algorithm SVM-KNN. The results of the algorithm evaluation are satisfactory and its use in a system already developed creates new possibilities to management and the automatic monitoring of forums in E-learning courses.*

Resumo. *Os fóruns de discussão são recursos muito utilizados dentro de modalidade EaD para a troca de informações e na aproximação entre professores, tutores e alunos. Ambientes de monitoramento automáticos para estes fóruns de discussão são ferramentas indispensáveis para o gerenciamento e consequentemente no apoio da melhoria da qualidade de cursos de EaD. Entretanto, o desenvolvimento desse tipo de ambientes com a utilização de técnicas de mineração de dados e em especial a classificação supervisionada é uma tarefa difícil por vários razões, entre elas, por falta de dados já rotulados para serem usados na construção e no treinamento destes classificadores. Este trabalho tem como objetivo mostrar um ambiente automático de monitoramento de fóruns de discussão com a utilização de algoritmo semi-supervisionado SVM-KNN. Os resultados da avaliação do algoritmo são satisfatórios e o seu uso num sistema já desenvolvido cria novas possibilidades para o gerenciamento e o monitoramento automático dos fóruns em cursos de Educação a Distância.*

1. Introdução

O uso de meios computacionais para melhorar a qualidade dos cursos de Educação a Distância (EaD) e criar novas ferrramentas de gerenciamento vem despertando o interesse acadêmico, pois nessa modalidade de ensino, os professores não possuem contato físico com os alunos, necessitando de novos meios de interação e avaliação. Segundo [Moran et al. 2000] o pouco contato físico pode prejudicar o desempenho de um curso a distância, dificultando a avaliação do professor quanto a compreensão ao nível de apresentação de conhecimento que o aluno está tendo.

Assim, é comum o uso de Ambientes Virtuais de Aprendizagem (AVAs) nessa modalidade, que tratam-se de um ambiente em que materiais, tarefas, discussões, comentários, dúvidas, etc. podem ser publicados possibilitando a aproximação entre professores, tutores e alunos. Nesses ambientes, um dos recursos mais utilizados são os fóruns de discussão por serem um instrumento que permite articular debates e discussões entre os atores envolvidos no processo de ensino/aprendizagem [de Oliveira Júnior et al. 2011].

Devido a essa função dos fóruns de discussão é importante que haja um rigoroso gerenciamento desse recurso, que possui como característica o desenvolvimento das discussões de forma assíncrona, ou seja, para que o diálogo ocorra não é necessário que todos os participantes estejam *online* o tempo todo. Essa característica garante uma dinamicidade ao curso e ao mesmo tempo dificulta o controle e acompanhamento por parte de tutores e professores, que podem deixar passar mensagens importantes sem serem analisadas.

[Elia and Chamovitz 2009] e [Cavaroli and Coello 2005] desenvolveram metodologias para que os atores dos fóruns de discussão (professores, tutores e alunos) aliem as mensagens trocadas com o objetivo de produzir discussões mais ricas, interessantes e permitir a sua avaliação. Este método, apesar de efetivo, insere um fator de custo, a inspeção manual de cada mensagem enviada, tornando um trabalho massivo e demorado. Com o uso de técnicas de mineração de dados como, por exemplo, classificação, regras de associação e agrupamento pode-se diminuir a quantidade de mensagens a serem inspecionadas, minimizando a quantidade de supervisões. Além de permitir a análise em diversos contextos como, por exemplo: Identificação de dúvidas, reclamações, sugestões, impressões sobre a disciplina/curso, etc.

No trabalho de [Romero et al. 2008] são apresentadas algumas possibilidades do emprego dessas técnicas. [Lin et al. 2009] descreve uma proposta de sumarização dos fóruns de discussão em tópicos utilizando mineração de textos. [Gerosa et al. 2003] apresentou algumas possibilidades que podem ser utilizadas pelo docente de um curso na coordenação de fóruns educacionais, analisando aspectos relativos à estruturação do discurso e à categorização de mensagens, sem haver necessidade de inspecionar o conteúdo de cada uma das mensagens individualmente. Em [Vieira et al. 2005] foi criada uma ferramenta que monitora o diálogo dos usuários, analisando as características particulares das suas participações, e classificação do ponto de vista de aprendizagem. Utilizou-se nesse trabalho dois algoritmos de classificação, Redes Neurais Artificias e Árvores de Decisão, em conjunto, para que além de acurácia satisfatível obtivisse uma explicação do porquê dos resultados.

Porém, para construir uma aplicação que utilize o método de classificação é

necessário haver uma quantidade razoável de dados para treinar um algoritmo, sendo que o processo para adquirir esses dados muitas vezes é demorado e caro. Em [de Oliveira Júnior et al. 2011] é constatada a dificuldade em adquirir dados para treinamento nos fóruns de discussão de AVAs. Nesse contexto, o método de aprendizado semi-supervisionado se encaixa bem, pois segundo [Zhu and Goldberg 2009], um algoritmo semi-supervisionado aprende a partir de um pequeno número de dados rotulados juntamente com informações e estruturas internas contidas em um grande número de dados não rotulados.

O aprendizado semi-supervisionado possui aplicação na classificação de dados textuais como no trabalho desenvolvido por [Nigam et al. 2000], em que foi desenvolvido um algoritmo utilizando EM (Expetation-Maximization) para classificação textual. Em [Joachims 1999] foi introduzido o algoritmo TSVM (Transductive Support Vector Machine) para classificação textual, onde é feita uma explanação de porquê o TSVM é melhor adaptado para classificação textual e são mostrados resultados experimentais que melhoram substancialmente a precisão em relação a métodos indutivos, especialmente quando há poucos dados de exemplos

Portanto, nesse trabalho o ambiente de monitoramento de fóruns desenvolvido em [de Oliveira Júnior et al. 2011] foi reestruturado, substituindo o algoritmo de classificação, SVM, por um algoritmo de aprendizado semi-supervisionado, o SVM-KNN [Li et al. 2010]. Características do sistema como fácil integração com outros sistemas e a rápida identificação de questões, dificuldades e demais problemas que constantemente surgem por professores e tutores permaneceram.

O restante desse trabalho encontra-se organizado da seguinte forma: na Seção 2 é apresentado o algoritmo SVM-KNN; na Seção 3 é detalhada a metodologia utilizada nesse trabalho; na Seção 4 são apresentados os resultados e feitas algumas discussões; na Seção 5 é feita a conclusão do trabalho.

2. O Algoritmo SVM-KNN

O algoritmo SVM-KNN é um algoritmo de aprendizagem semi-supervisionado proposto por [Li et al. 2010] que utiliza o SVM (*Support Vector Machine*) [Cortes and Vapnik 1995] e KNN (*K Nearest Neighbor*) [Cover and Hart 1967]. O objetivo desse algoritmo é inicialmente ter um classificador SVM fraco, treinado com poucos dados previamente classificados, e através da inserção de documentos não classificados na base de treinamento, presentes em uma região de fronteira, melhorar iterativamente os resultados do SVM. Os documentos da região de fronteira são introduzidos pelo KNN e este processo representa muito mais do que somente o crescimento do conjunto de dados rotulados, pois os dados de fronteira fornecem informações adicionais sobre a distribuição nessa região [Li et al. 2010].

O SVM-KNN pode ser visto como uma implementação do algortimo EM, porém com a melhoria na seleção dos dados que irão, iterativamente, sendo inseridos ao treino. Como dito anteriormente, os dados escolhidos são aqueles que agregam informações sobre a distribuição na região de fronteira entre classes, sendo candidatos a vetores suporte para o SVM.

3. Metodologia

Em [de Oliveira Júnior et al. 2011] o Ambiente de Monitoramento é utilizado em um contexto de gerenciamento do fórum, ou seja, as mensagens são classificadas em Positivo ou Negativo, onde Negativo é atribuído às mensagens que contém dúvidas, conteúdo indevido, insatisfação e Positivo às mensagens que não contém dos padrões de Negativo. Porém nesse trabalho o contexto a ser monitorado é mais restrito, pois deseja-se apenas identificar as mensagens que contém Dúvidas ou que são Esclarecimentos à dúvidas.

A restrição utilizada nesse trabalho não causa perda de generalidade, uma vez que não há classificação entre Dúvida e Esclarecimento disponível nos fóruns do AVA. A definição utilizada para cada uma das classes é:

- Dúvida: mensagens que tratam de assuntos relativos as disciplinas do curso, comentários relativos a possíveis erros na elaboração de enunciados e o funcionamento de recursos do AVA;
- Esclarecimento: mensagens contendo informações sobre horário de atendimento dos tutores, sobre como proceder para executar uma determinada tarefa e esclarecimento de dúvidas sobre questões relativas aos exercícios.

3.1. Coleta

Com o contexto de monitoramento definido, foram coletadas 2.268 mensagens de um AVA Moodle, das quais 199 foram classificadas por um especialista. Ao final da classificação manual pelo especialista obteve-se a distribuição entre classes mostrada na Tabela 1.

Tabela 1. Organização dos dados

Classe	Quantidade
Dúvida	103
Esclarecimento	96
Não classificadas	2169
Total	2368

3.2. Preprocessamento e Indexação

Após a coleta das mensagens e classificação manual de algumas delas, passou a etapa de pré-processamento e em seguida a indexação. Inicialmente foi feita a conversão das mensagens para um vetor de características, em que cada palavra (*token*) compõe uma dimensão desse vetor [Hotho et al. 2005].

Com as mensagens devidamente representadas foi feita a remoção de *Stop Words*, ou seja, foram removidas palavras que não possuem contribuição estatística para a diferenciação entre documentos, como por exemplo, artigos, conjunções, preposições, etc [Hotho et al. 2005]. Posteriormente foi feita a representação dos documentos utilizando o Modelo de Espaço Vetorial [Salton et al. 1975], utilizando como conversão a métrica TF-IDF [Manning et al. 2008].

Após a conversão para o Modelos de Espaço Vetorial utilizando TF-IDF foi realizado o processo de *feature selection* reduzindo de 922 atributos para 200, mais o atributo de classe dos documentos.

3.3. Avaliação do Algoritmo

Nessa fase foi feita a organização dos dados para execução dos experimentos e a definição dos parâmetros para os algoritmos. As bases para execução dos experimentos foram separadas em três tipos: Base de Treinamento, Base de Testes e Base de Documentos não Rotulados.

Para a construção das bases de Treinamento e Testes foi utilizado um sistema de permutação e gerados 10 conjuntos diferentes de Treinamento e Testes. O processo de permutação gerado foi feito da seguinte forma: utilizando os 199 documentos classificados, foi feito o embaralhamento desses documentos. Com o conjunto embaralho foi feita a separação deste nos dois outros conjuntos necessários: Treinamento e Teste. O primeiro conta com 99 documentos, enquanto que o segundo possui 100 documentos. Todos os 10 conjuntos gerados possuem distribuições entre as classes diferentes e são diferentes entre si. A Base de Documentos não Rotulados foi construída com 500 mensagens, escolhidas aleatoriamente, dentre aquelas que não possuíam classificação.

Utilizando cada uma dessas bases foram executados dois experimentos, os quais foram feitos para analisar a melhor configuração de parâmetros do algoritmo SVM-KNN. Este algoritmo recebe 3 parâmetros: K - utilizado pelo KNN, N - quantidade de elementos de cada classe próximos a região de fronteira que serão escolhidos, M - tamanho limite da base de treinamento em relação ao total de dados disponíveis. Assim, nos experimentos os parâmetros K, N e M foram combinados utilizando os seguintes valores:

- M: 0,25;
- K: 1, 3, 5, 7;
- N: 1, 3, 5, 7, 9, 11;

A análise dos resultados foi feita utilizando as médias de Precisão, *Recall* e F_1-Measure. Um comparativo do tamanho das bases de treinamento finais também será mostrado.

3.4. Alteração do algoritmo

A aplicação desenvolvida em [de Oliveira Júnior et al. 2011] está organizada em camadas, como apresentado na Figura 1. Assim, foi necessária a alteração apenas na camada "Classificação/Re-indução", no que se refere à chamada do algoritmo e parâmetros passados.

O processo de treinamento do algoritmo SVM-KNN recebe, além de um conjunto de treinamento, um conjunto com dados não classificados. Após treinado, o classificador final é um SVM, porém diferente do SVM treinado apenas com o conjunto de treinamento. Esta adaptação foi feita criando apenas mais um parâmetro na tarefa de treinamento para permitir a passagem do conjunto de dados não classificados.

A interface de visualização dos resultados, correspondendo à camada "Visualização" foi mantida e pode ser vista na Figura 2. Na interface quando uma mensagem é classificada como "Dúvida" esta mensagem é definida como "Negativo" e quando classficada como "Esclarecimento" esta é definia como "Positivo".

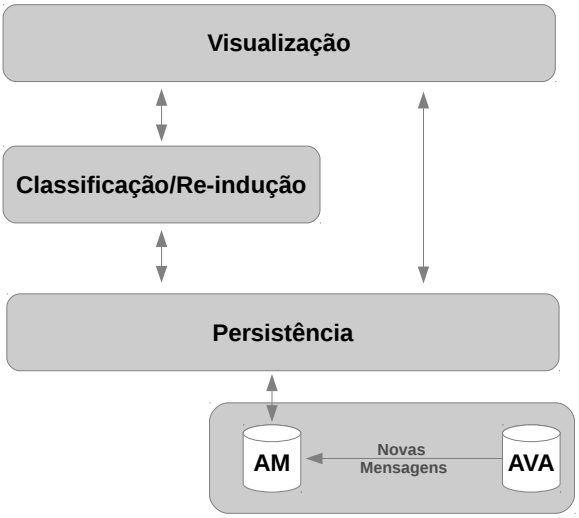

Figura 1. Arquitetura do Ambiente de Monitoramento.

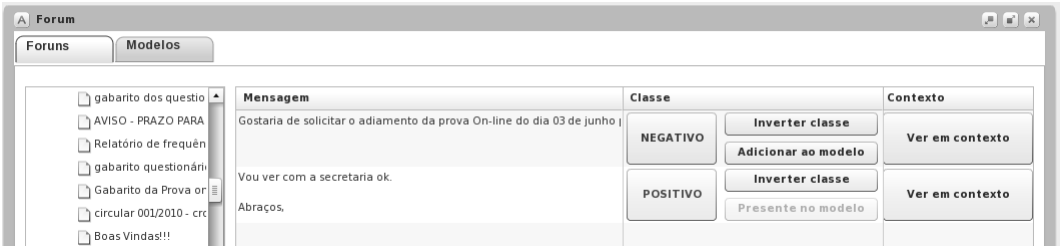

Figura 2. Tela da interface de apresentação das mensagens no Ambiente de Monitoramento de fóruns.

4. Resultados

Adiante serão detalhados os resultados do experimento de avaliação do SVM-KNN, em que na tabela apresentada os valores entre parênteses são os desvios padrão dos resultados para cada configuração. Realizou-se a análise estatística dos resultados, utilizando o Teste T, onde encontrou-se 5% de significância. A primeira linha mostra o resultado do algoritmo SVM utilizando apenas o conjunto inicial de 99 documentos rotulados, sendo este a base de comparação para o SVM-KNN. A quantidade de documentos utilizados está detalhada na Tabela 2.

Tabela 2. Detalhamento das quantidade de dados

Tipo	Quantidade	Porcentagem
Rotulados	99	16,52%
Não Rotulados	500	83,48%
Total	599	100%

No experimento realizado os valores de K e N foram combinados utilizando $M = 0,25$, ou seja, o tamanho final do conjunto de treino será por volta de 25% do total de documentos disponíveis, em torno de 150 documentos. Na Tabela 3 estão os resultados

alcançados, pelos quais é possível observar que, para todas as configurações, o SVM-KNN superou os resultados do SVM. Com destaque para a configuração k=1 e n=3 em que houve melhora de 37% na métrica F_1-Measure. Essa configuração foi a utilizada para na implementação da nova camada "Classificação/Re-indução".

Tabela 3. Resultados da avaliação do algoritmo SVM-KNN em comparação com o SVM.

k	n	#Rotulados	Precisão	*Recall*	*F-Measure*
		99	79,75% (0,05)	67,29% (0,13)	62,66% (0,18)
1	3	165	87,15% (0,03)	86,30% (0,03)	86,30% (0,03)
1	11	153	86,46% (0,03)	85,40% (0,03)	85,40% (0,03)
3	11	153	85,93% (0,03)	84,90% (0,03)	84,86% (0,03)
5	7	153	86,46% (0,03)	85,40% (0,03)	85,40% (0,03)
5	11	153	86,08% (0,04)	84,90% (0,04)	84,85% (0,04)

Além da melhora de Precisão, *Recall* e F_1-*Measure* a base de treinamento cresceu pelo menos 25% em relação a base original. Com isso, o modelo de classificação final do SVM-KNN obteve grau de generalização maior em relação ao SVM treinado apenas com 99 documentos.

Através do processo de re-indução do modelo, essas taxas de acerto podem aumentar por meio da inclusão de novos dados para o treinamento, com alto grau de confiabilidade de sua classificação. Outro ponto importante onde o processo de re-indução auxilia é na adaptação quanto a alteração da distribuição dos dados, causada pelo surgimento de novos ou desuso de termos.

5. Conclusões e Trabalhos Futuros

Ambientes de Monitoramento para fóruns de discussão são uma importante ferramenta para melhor a qualidade de cursos de EaD, pois são os fóruns de discussão um dos recursos mais utilizados para trocas de informação e aproximação entre professores, tutores e alunos. Porém, a criação desse tipo de ambiente com a utilização de técnicas de mineração de dados pode, muitas vezes, ser complicada.

O trabalho desenvolvido por [de Oliveira Júnior et al. 2011] trata-se de um sistema para o Monitoramento Automático das mensagens de fóruns, porém dada a falta de dados para o treinamento do algoritmo SVM esse sistema alcançou baixas taxas de acerto quando comparado à literatura. Percebendo isso, decidiu-se então utlizar um outro paradigma de aprendizado de máquina, o aprendizado semi-supervisionado com o algoritmo SVM-KNN e como visto na Seção 4, este algoritmo alcançou taxas superiores ao SVM analisando o mesmo contexto.

A substituição do SVM pelo SVM-KNN foi feita de forma pouco invasiva, necessitando de pequenas alterações e não impactando na forma como os dados são apresentados e armazenados.

Esses fatores abrem novas possibilidades da utilização deste ambiente no monitoramento de fóruns em diversos contextos e a facilidade da sua adaptação em novos cenários, uma vez que este ambiente necessita de poucos dados para o seu treinamento e para iniciar o efetivo monitoramento com taxas de acerto satisfatórias.

Como trabalhos futuros pretendemos realizar investigações sobre a classificação automática e em demanda/on-line de chats dentro de AVA.

Agradecimentos

Os autores agradecem à Fapemig e ao CNPq pela concessão de bolsas bem como pelo apoio financeiro.

Referências

Cavaroli, J. T. and Coello, J. M. A. (2005). Sea: Um sistema emissor de alertas para fóruns de discussão, baseado na categorização de mensagens e avaliação pelos pares. In *SBIE-WIE 2005*.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. 10.1023/A:1022627411411.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21 –27.

de Oliveira Júnior, R. L., Esmin, A. A. A., Coelho, T. A., Araujo, D. L., Alosno, L., and Giroto, R. (2011). Uma ferramenta de monitoramento automático de mensagens de fóruns em ambientes virtuais de aprendizagem. In *SBIE-WIE 2011/SBIE Trilha 1*.

Elia, M. and Chamovitz, I. (2009). Cqmsg - classificador e qualificador de mensagens: um instrumento para apoio à avaliação de fóruns temáticos. In *SBIE-WIE 2009*.

Gerosa, M. A., Pimentel, M. G., Fuks, H., and de Lucena, C. J. P. (2003). Coordenação de fóruns educacionais: Encadeamento e categorização de mensagens. In *SBIE 2003*, pages 45–54.

Hotho, A., Nürnberger, A., and Paass, G. (2005). A brief survey of text mining. *LDV Forum*, 20(1):19–62.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209.

Li, K., Luo, X., and Jin, M. (2010). Semi-supervised learning for svm-knn. *Journal of Computers*, 5(5).

Lin, F.-R., Hsieh, L.-S., and Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, 52:481–495.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Moran, J. M., Masetto, M. T., and Behrens, M. A. (2000). Mediação pedagógica e o uso da tecnologia. *Novas Tecnologias e Mediação Pedagógica*.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39:103–134.

Romero, C., Ventura, S., and García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18:613–620.

Vieira, A. C. H., Tedesco, P. C., Timóteo, A., and Lima, A. (2005). Analisando diálogos para classificação de padrões utilizando redes neurais artificiais e Árvores de decisão. In *SBIE-WIE 2005*.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.