Detecção Automática de Plágio em Ambiente Educacional Virtual

Elizabeth Rocha[1], Claudia Tinós Peviani[1], Thiago Rafael Pretto[1], Willian Martins Silva[1], Neylson Gularte[1]

[1]Universidade Federal da Grande Dourados (UFGD)

Caixa Postal 533 - 79804-970 - Dourados - MS - Brazil

{elizabethrocha,claudiapeviani}@ufgd.edu.br,
{thiagorafaelpretto,00wms00,neylsond}@gmail.com

Abstract. *This article uses distance education to portray the context of the needs that underlie the field of intellectual production and how it strongly expresses genuine challenge attributed to the educational setting. Considering that the activities and interactions expressed in the academic activities of those who study through distance learning takes place in virtual environments, it appears that it would be a useful tool that could capture work that presented characteristics of plagiarism. In these terms, there is the computational challenge, which is to develop a software able to detect plagiarism in real time, enabling the student to have another chance for sending a file, with the aim of rehabilitation and recovery of authorial ethics.*

Resumo. *Este artigo utiliza a Educação a Distância para retratar o contexto das necessidades que permeiam o campo da produção intelectual genuína e como isso expressa forte desafio imputado ao cenário educacional. Considerando-se que as atividades e interações expressas nas atividades acadêmicas dos que estudam a distância acontecem em ambientes virtuais, verifica-se que uma ferramenta útil seria a que pudesse capturar trabalhos que apresentassem características do plágio. Nesses termos, surge o desafio computacional, que consiste em desenvolver um software capaz de detectar o plágio em tempo real, possibilitando ao aluno outra chance de envio do arquivo, com o intuito da reeducação e resgate da ética autoral.*

1. Introdução

A forte expansão da Educação a Distância (EaD) na última década, em termos de Brasil, pode ser atribuída, essencialmente, a duas realidades. A primeira consiste dos programas educacionais criados para oferta de cursos de grau superior, por universidades públicas, destinados à população essencialmente adulta, interiorana, como o Sistema Universidade Aberta do Brasil (UAB) [Mota 2010].

A segunda remete ao ritmo acelerado com que se constatam as mudanças das Tecnologias da Informação e Comunicação (TICs) e seus impactos nas várias dimensões do cotidiano contemporâneo. Um desses impactos remete ao caráter da *re-mixagem* que muda e amplia a "paisagem comunicacional e social contemporânea", conforme defende Lemos (2005).

Tais mudanças são perceptíveis e incorporadas na legislação que trata da EaD, que aos poucos a faz avançar da condição emergencial e complementar ao ensino presencial, para modalidade educacional com mediação didático-pedagógica, que possui "metodologia, gestão e avaliação peculiares", suportadas pelas TICs, conforme o

Decreto 5.622/05 que regulamenta o Art. 80 da Lei de Diretrizes e Bases (LDB) 9.394/96.

A partir do reconhecimento legal e da diversidade que envolve o cotidiano dos espaços que desenvolvem essa modalidade educacional, como aqueles que formam e informam, na perspectiva cultural da "sociedade da informação", diversas necessidades educacionais, comunicacionais e informacionais precisam ser sanadas a fim de garantir a adequada articulação entre o ensino e a aprendizagem.

Uma dessas necessidades remete ao caráter genuíno da produção intelectual por parte dos acadêmicos, sobretudo, dos que estudam a distância, em ambientes virtuais. Conforme dito por Silva (2011, p.01) "A Educação a Distância contemporânea se desenvolve essencialmente em ambiente Web. Isso se deve, de forma inconteste, ao avanço da tecnologia digital e telemática. No Brasil, por exemplo, o quantitativo de pessoas que possuem computadores e serviço de Internet tem aumentado significativamente".

É certo que uma das características mais marcantes dos que vivem em sociedades que utilizam a Internet, em 2012, é a possibilidade de acesso a informações variadas, atualizadas e em tempo real. A Internet é uma tecnologia dinâmica o suficiente para determinar o ritmo da velocidade nas comunicações, a variedade de espaços, a tendência de mercado e de comportamento no mundo digital.

Embora nem todas as sociedades disponham de Internet é possível afirmar que a utilizamos em escala global, pois dos 7 bilhões de seres humanos, 2 bilhões a acessam pelo menos uma vez por semana, conforme divulgado pela agência de telecomunicações da ONU, no fim de 2010.

Como toda tecnologia, a Internet pode ser usada para o bem e para o mal, pois depende do contexto de interesses dos que a utilizam. No que compete à Educação, a Internet tem favorecido a expansão da Educação a Distância (EaD) no mundo [Nunes, 2010]. Tal expansão se insere no rol das mudanças intensificadas nas últimas décadas, a partir da revolução da sociedade informatizada, que faz estremecer os alicerces da educação tradicional. Isso é bom ou ruim? Que ideologias e mudanças de paradigmas educacionais nos possibilitam o mundo digital?

São questionamentos que, diante das mudanças que ocorrem na Educação, precisam de respostas que ajudem a identificar as tendências educacionais, como por exemplo, o ensino híbrido, que mescla momentos letivos presenciais e a distância, em ambientes virtuais, modelo possível graças a experimentações educacionais diferenciadas, por parte de muitos docentes, bem como pelo avanço da Internet.

A expressividade dos números elencados anteriormente leva a refletir que as pessoas conectadas tem acesso a ferramentas que favorecem a edição e compartilhamento da informação. Se em 2005, Lemos apontava o aspecto da "Ciber-cultura-remix" e a defesa da recriação e da livre expressão, por grupos como a "Foundation for a Free Information Infrastructure", em 2012, temos sites sendo fechados pelo FBI, como no caso do *Megaupload*, sob o argumento da pirataria.

Então, as coisas não são tão livres quanto se pensa. Se a Internet é terra de ninguém é fundamental saber o que fazer e como usar as informações nela disponíveis para não incorrer na ação da pirataria ou do plágio.

2. Desafio Educacional

Estimular a produção intelectual genuína do acadêmico é um dos maiores desafios imputados ao cenário educacional. Os Ambientes Virtuais de Aprendizagem (AVAs) aumentam muito as possibilidades interativas entre professores e alunos. Além disso, favorece o acesso ao conteúdo, recebimento e envio de atividades. Há, ainda, o acompanhamento e mediação em fóruns, chats e wikis. Isso representa aumento considerável de dados e informações, na forma avaliativa, a ser corrigido pelo professor. [Tarja 2001].

Considerando o volume de avaliação gerada nos AVAs, uma preocupação forte, por parte do professor, geralmente se vincula ao caráter genuíno dos trabalhos e comentários apresentados pelos alunos. Considerando, ainda, que tudo o que se incorpora aos AVAs, como o MOODLE, "roda" na Internet, verifica-se que uma ferramenta útil é aquela que captura o plágio.

Ora, mas é certo que muitas ferramentas que capturam o plágio, sobretudo o *ipisis litteris*, já existem. Podem ser exemplificadas, o Turnitin e URKUND, que são pagas. Então, qual a novidade da proposta? Que esforços justificam o desenvolvimento de outro aplicativo com essa característica?

Antes de responder, vale a pena ressaltar que um aplicativo dessa natureza, além de facilitar o trabalho do professor, favorece ao aluno a oportunidade de refletir sobre sua própria produção intelectual, de modo conseguir um trabalho ético e original [Bastos 2009]. A facilidade encontrada pelos alunos, considerando as conexões, por meio dos hipertextos, das relações semânticas são situações sedutoras para os menos empenhados nas suas produções acadêmicas. O advento do computador ligado à Internet alterou nossas formas de criação e de comunicação, de tal modo que parece normal a muitos, que a cópia fiel do texto de outrem merece tanto crédito acadêmico, quanto aquele que levou meses para ser construído [Johnson 2001].

A oferta de cursos na modalidade a distância, por sua vez, necessita de um conjunto de tecnologias para interação entre alunos e professores. Realizar cursos a distância implica desenvolver boa parte das interações e realização das atividades, em Ambientes Virtuais de Aprendizagem, como o MOODLE, por exemplo.

Ao corrigir as atividades, contudo, o professor percebe tal qual no modelo presencial, que muitas produções dos alunos são cópias fiéis tiradas de materiais disponibilizados na Internet. Considerando que o professor já está com suas atividades vinculadas à Web, seria bastante proveitoso, em termos de confiabilidade do trabalho intelectual e aquisição do tempo, se houvesse um software, integrado ao AVA, que desse um relatório, em tempo real, relativo ao plágio, por parte do aluno.

Voltando às duas últimas perguntas, há duas características que evidenciam um contexto de utilidade que justifique o investimento nessa proposta do aplicativo anti-plágio. A primeira é que o plugin a que se refere esta proposta é open source e está integrado no AVA MOODLE. Isso representa custo zero para qualquer instituição de ensino, inclusive particulares.

A segunda, é que a proposta lançada aqui se refere a um aplicativo capaz de detectar, em tempo real, o plágio acadêmico, em sua forma textual, nas extensões doc, docx, rtf, odt, pdf, de forma específica ao ambiente MOODLE. Sabe-se que o plágio de qualquer natureza, especialmente o acadêmico é uma atitude anti-ética e que precisa ser

encarada e combatida com seriedade por todos. Embora a proposta da captura automática do plágio em si não seja inovadora, a proposta da integração com o MOODLE é, na medida em que se trata de uma ferramenta livre, além de ser em tempo real, proposta ainda não plenamente alcançada pelos aplicativos existentes.

Tem-se, deste modo, na detecção do plágio excelente oportunidade de reeducar a comunidade acadêmica e com isso resgatar e propagar a ética autoral. O diferencial consiste em utilizar o aplicativo para a conscientização de forma dialogada, a partir de uma integração entre as vertentes tecnológica e educacional, pois o intuito não é punir, mas reeducar o acadêmico que tiver seu trabalho identificado como plágio, conscientizando-o sobre a gravidade dessa ação e os impactos sobre o prisma jurídico.

3 – O processamento textual como desafio computacional na detecção de plágio

Falar em processamento textual na detecção automática de plágio *ipsis litteris*, ou seja, a cópia idêntica em partes ou integral envolve a abrangência de formatos de documentos, reconhecimento de padrões de componentes textuais e utilização de APIs de motores de busca. Esse conjunto de etapas se mostra forte desafio computacional na detecção de plágio.

Embora muitas instituições e docentes estabeleçam normas de escrita para a submissão de arquivos em atendimento às atividades avaliativas das suas disciplinas, sejam elas via AVA, correio eletrônico ou outros meios, boa parte dos professores não se preocupa, efetivamente, com o formato digital do documento, como *pdf*, *doc*, *docx*, *rtf* e *odt*, atentando-se, especialmente, ao cumprimento das regras estabelecidas, como padrão da ABNT ou outro específico. Com base nisso, o primeiro desafio computacional para a detecção automática de plágio, consiste de a ferramenta identificar qualquer tipo de texto submetido para avaliação, sem distorção da integridade do documento.

Com os diversos formatos para documentos de texto cada um com sua arquitetura e forma de armazenar dados, se faz necessária vasta gama de ferramentas para extração ou conversão do documento submetido para texto puro (*txt*). Embora as ferramentas existentes para detecção de plágio já implementem essa funcionalidade de conversão, ainda existe uma grande lacuna representando os formatos textuais menos utilizados.

Apesar de existir ferramentas ou bibliotecas específicas para conversão de arquivos para texto puro, existe outra escassez, no sentido de preservar a formatação original do texto, quando isso se faz necessário. Quaisquer formatações escritas pelo usuário são perdidas no processo de conversão, sejam elas tamanho, fonte ou cor.

Uma vez feito o tratamento do texto, com ou sem formatação, surge um novo desafio: Identificar quais componentes textuais representam o trabalho do aluno, plagiado ou original, e quais componentes são derivados de estruturas textuais que não devem implicar plágio, como por exemplo, um parágrafo onde existe uma citação indireta, uma referência a outro trabalho no fim do documento, ou ainda títulos e cabeçalhos do documento.

No caso de citações diretas e referências, fica evidente que apesar da intenção do autor em utilizar e referenciar corretamente o trabalho de terceiros, a ferramenta irá

apontar a prática do plágio, uma vez que a ferramenta não é capaz de distinguir esse referenciamento de uma cópia *ipsis litteris*.

No caso de títulos, subtítulos e cabeçalhos, o problema não está no fato de o acadêmico copiar um título de outro lugar, mas sim no contexto similar de produção intelectual desses componentes, uma vez que, por exemplo, o docente solicita um trabalho acadêmico para sua classe sobre determinado tema, e todos os alunos utilizam o mesmo título proposto pelo professor para escrever seus trabalhos. A ambiguidade gerada é a mesma no caso dos cabeçalhos, onde estes podem ser o padrão adotado por determinada instituição de ensino. Ou ainda conter informações que seriam facilmente encontradas globalmente, por exemplo, nome do curso, nome do acadêmico, data, dentre outras.

O uso das técnicas de reconhecimento de padrões seria uma possível solução para esse impasse, já que vem sendo aplicadas para resoluções de problemas em diversas áreas como a biometria, automação industrial, mineração de dados e o processamento textual, propriamente dito.

Existem padrões para esses componentes pré-estabelecidos por associações regulamentadoras, como a ABNT, que em divergência com outros padrões de norma culta, podem gerar a necessidade de um mesmo componente textual passar por várias validações até que ele seja identificado ou não. Sendo assim, se faz necessária a personalização de padrões que atendam às normas acadêmicas de cada instituição educacional.

Esses padrões poderiam então ser mapeados junto às entidades regulamentadoras e transcritos computacionalmente para expressões regulares, visto que estas estão presentes em todas as linguagens de programação de alto nível utilizadas para o desenvolvimento de ambientes virtuais de aprendizagem[1].

Quando identificados, os componentes textuais que não caracterizam plágio seriam extraídos do texto, mantendo apenas a parte textual com necessidade de ter sua originalidade avaliada. A produção resultante seria então dividida em sentenças que seriam analisadas sequencialmente através de requisições ao motor de busca.

Esse processamento textual descrito anteriormente compõe a primeira etapa do problema. Pelo fato de ser um serviço Web, o problema é agravado, uma vez que vários alunos utilizam a ferramenta de detecção para submissão de arquivos enquanto outros fazem requisições simultâneas ao servidor do AVA, afetando assim a escalabilidade [Sommerville 2011]. Como solução para a primeira etapa do problema a distribuição de carga entre vários servidores, resolveria a demanda de vários alunos.

Surge então a segunda etapa do problema, as requisições feitas ao motor de busca. A verificação *ipsis litteris*, exige que todas as sentenças sejam analisadas, mas o motor de busca impõe um limite de requisições por segundo. Como se tratam de muitas sentenças por texto, isso acaba gerando um gargalo entre os servidores do AVA e o motor de busca. Como solução para a segunda etapa do problema, poderia utilizar a técnica de *caching*, nas sentenças previamente consultadas. E com as sentenças novas criar uma fila para que possam ser realizadas as requisições ao motor de busca.

---

[1] PHP PCRE (Perl Compatible Regular Expressions) - http://php.net/manual/pt_BR/book.pcre.php

Uma fila de prioridade poderia ser a técnica usada para implementar a fila de sentenças, pois já é aplicada em sistemas operacionais, como por exemplo a fila de processos aguardando o processador para execução. Os processos mais prioritários são executados antes dos outros. Outro exemplo é a fila de pacientes esperando por um transplante de fígado em geral é uma fila de prioridade [Song 2008].

O objetivo desta discussão é propor uma ferramenta para detecção automática de plágio *ipsis litteris* capaz de identificar qualquer tipo de texto submetido para avaliação, com a garantia da integridade do documento.

4 – A proposta de uma ferramenta para detecção automática de plágio

Há pesquisas de ferramentas de detecção de plágio voltadas para o AVA MOODLE desde 2010. Existe uma iniciativa de um *plugin* denominado Crot, do pesquisador russo Sergey Butakov que baseia seu processamento textual na técnica de cálculo de *fingerprints* (*hashes*) do texto submetido pelo acadêmico [Butakov e Schebinin 2008]. A partir do texto original são gerados *hashes* que são comparados com os já existentes na base de dados do MOODLE, oriundos de outras submissões. Isso resulta um grande avanço em termos de tempo de resposta, visto que para este procedimento a ferramenta independe de um motor de busca e tudo ocorre no servidor do AVA.

Posteriormente, o texto é submetido a uma pesquisa na Web através da API de desenvolvimento do Bing. Nesta etapa é onde se encontra o gargalo da ferramenta, uma vez que ele precise realizar o *download* de todos os arquivos encontrados para uma análise de similaridade textual entre estes e o documento submetido. A ferramenta Crot não se adequaria a nossas necessidades pelo fato não executar a análise de forma instantânea, já que o texto é armazenado e submetido para análise perante solicitação do administrador do MOODLE.

Nossa proposta visa coibir a prática do plágio através da reeducação do acadêmico, informando este em primeira mão que foi encontrado plágio em seu trabalho. A ferramenta ideal deve apresentar o resultado da avaliação em questão de instantes, permitindo assim que o acadêmico fique ciente da existência do plágio, mas não de sua localidade. O aluno pode assim, optar em correr o risco de submeter um documento contendo produção intelectual de terceiros, ou reavaliar sua escrita e submeter novamente, uma vez que o MOODLE permite um envio prévio e outro definitivo.

No atual estado do projeto, após o processamento textual de remoção e divisão, o texto puro é submetido de forma *ipsis litteris* à pesquisa na Web através do motor de busca da Bing API, conforme exibido na figura 1. Os resultados são analisados e uma porcentagem de plágio é calculada baseada na razão entre a quantidade de sentenças submetidas e sentenças contendo plágio *ipsis litteris*.

Uma das próximas etapas decorrentes é a expansão do atual pacote de expressões regulares que realizam o reconhecimento de componentes textuais, como por exemplo, referência a um artigo publicado em evento ou a um capítulo de livro. Outro avanço que está em via de ser alcançado é a criação de uma fila de prioridade.

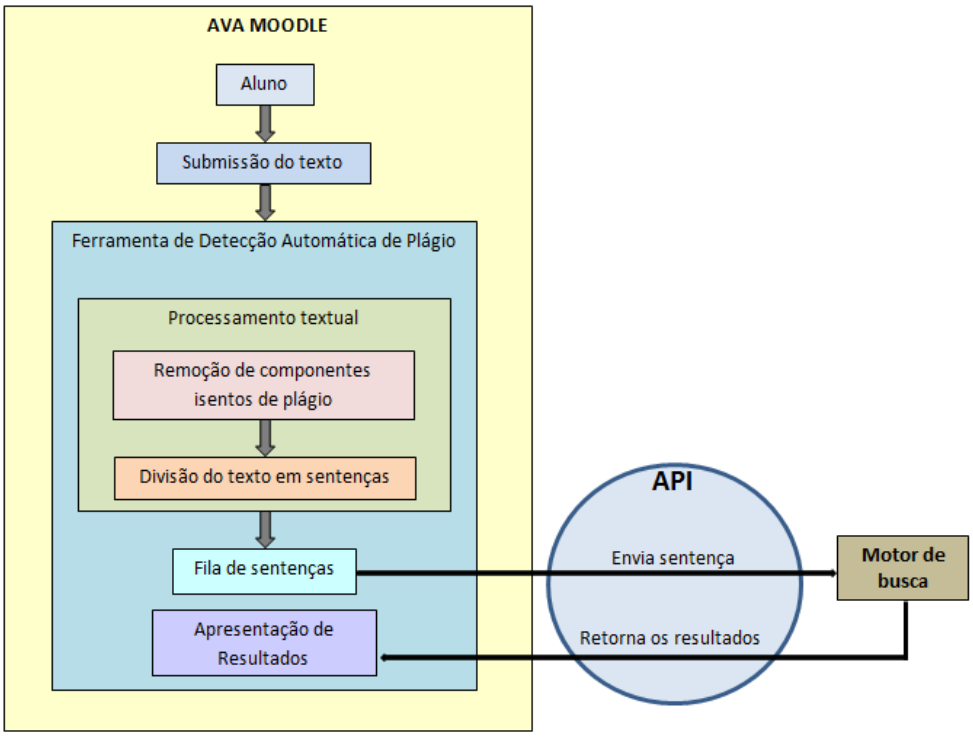

Figura 1. Etapas do processo de detecção de plágio na Web.

Essa poderia ser uma fila de prioridades, em que a prioridade pode ser baseada no algoritmo de escalonamento de processos SJF (*Shortest Job First*) que consiste no fato de arquivos menores tenha uma prioridade maior, a fim de diminuir o tempo de permanência na fila de espera ou ainda através de uma definição de prioridade dos arquivos submetidos, estabelecida pelo professor [Silberschatz 2008].

A ferramenta de detecção de plágio está sendo implementada e testada num ambiente experimental. A Tabela 1 apresenta alguns resultados obtidos através de testes feitos usando a ferramenta.

Tabela 1. Resultados obtidos através da ferramenta de detecção de plágio

Texto	Formato	Total de Sentenças	Total de plágio detectado	Falsos Positivos	Falsos Negativos	Tempo de Resposta (seg)
Texto 1	*doc*	182	4 (2.1%)	4 (2.2%)	-	17
Texto 2	*pdf*	90	3 (3.33%)	3 (3.33%)	-	13
Texto 3	*docx*	89	5 (5.62%)	5 (5.62%)	-	12
Texto 4	*doc*	24	22 (91.68%)	-	2 (8.33%)	5
Texto 5	*doc*	227	28 (12.33%)	4 (1.76%)	-	21

5. Considerações finais

A Educação a Distância juntamente com a Tecnologia da Informação e Comunicação vem mudando o cenário da educação no Brasil, como também o comportamento de professores e alunos. O acesso facilitado à informação via Web tem gerado muitos registros de cópia de produção intelectual entre acadêmicos.

Esse fato gera desconforto ao professor e as ações punitivas em coibição à prática do plágio estimulam o desenvolvimento de ferramentas livres para detecção automática de plágio. As ferramentas existentes, contudo, resolverem o trabalho do professor apenas em parte, já que precisa baixar todos os documentos para análise pelo software. Essa tarefa certamente consome um tempo e desgaste considerável por parte do professor.

Com a implementação desta ferramenta, pretende-se atender as expectativas dos professores quanto à confiabilidade do caráter genuíno para a correção das atividades avaliativas postadas no AVA MOODLE. Visando o aluno, a intenção é a reeducação quanto ao uso da ética autoral no momento da produção de conteúdo.

Dentro das etapas apresentadas no desenvolvimento da ferramenta é importante salientar no processamento textual, o refinamento do que foi escrito, a partir da remoção dos componentes textuais que não representam a produção do aluno em si, mas sim ao referenciamento generalizado de outras obras. A principal dificuldade está em abranger todas as normas técnicas disponibilizadas em cada contexto.

As limitações de consultas impostas pelos motores de busca têm se mostrado um dos problemas mais preocupantes, pois para a eficiência da ferramenta é necessário não limitar as consultas diárias. A solução seria uma política pública para que as instituições educacionais pudessem ter acesso ilimitado e gratuito a este serviço, fortalecendo, assim, o desenvolvimento de ferramentas integradas aos seus AVAs para detecção de plágio.

6. Referências

Mota, R. (2010) "A Universidade Aberta do Brasil", In: Litto, F. M. & Formiga, M. Educação a Distância: O Estado da Arte, São Paulo.

Bastos, A. F. et al. (2009) "O software antiplágio e o conhecimento docente: um estudo de caso". Monografia a apresentada à Faculdade Sete de Setembro. Fortaleza.

Johnson, S. (2001) "Cultura da interface: Como o computador transforma nossa maneira de criar e comunicar", Jorge Zahar Editor Ltda., Rio de Janeiro.

Tarja, S. F. (2001), Informática na educação: novas ferramentas pedagógicas para o professor da atualidade, Editora Érica.

Silberschatz, A., Gagne, G., Galvin, P. B. (2008), Sistemas Operacionais com Java, Elsevier, 7ª edição.

Silva, W. M. (2011), "Desenvolvimento Do Plugin Pointer No MOODLE 2.1 Para Captura De Plágio Textual Na Web". TCC defendido em dezembro de 2011 pela Faculdade de Ciências Exatas e Tecnológicas da Universidade Federal da Grande Dourados.

Sommerville, Ian (2011), Engenharia de Software, Pearson Education, 9ª edição.

Song, S. W. (2008) "Fila de Prioridade", http://www.ime.usp.br/~song/mac5710/slides/03prior.pdf, Abril.

Butakov, S. and Schebinin, V. (2008). Plagiarism Detection: The tool and the Case Study. In *IADIS International Conference e-Learning 2008, Amsterdam, The Netherlands,* pages 304-310. IADIS.