Metodologia estatística computacional de detecção automatizada do plágio autoral - Uma proposta de interpretação dos resultados do programa Farejador de Plágio

Maximiliano Zambonatto Pezzin, MSc
<maximilianopezzin@gmail.com>
Universidade do Contestado - UnC Concórdia - 10/2010

Resumo

Este documento aborda um tema polêmico e de difícil definição. A complexidade e as diversas variáveis envolvidas na abordagem do que é plágio literário, acabam por gerar um vácuo justamente na definição e na afirmativa. Afinal, um texto foi ou não plagiado? Este documento deve ser visto como uma proposta de definição de plágio, não exatamente em números, mas com base na metodologia do programa Farejador de Plágios. As variáveis e técnicas utilizadas pelo programa são a base da proposta do autor, e devem ser vistas como uma proposta, visto não haver (ressalte-se que não há) nenhum aspecto legal ou jurídico envolvido do processamento do farejador, ficando a cargo e responsabilidade do leitor a definição final da existencia do plágio.

1. Introdução

A lei 9610 que dispõem em seu artigo primeiro “Esta Lei regula os direitos autorais, entendendo-se sob esta denominação os direitos de autor e os que lhes são conexos”. No entanto, apesar de completa na questão de definir as obras literárias, as contextualizações e definições apresentadas são, por assim dizer, esquivas. Pode-se dizer que a lei é absolutamente vaga no tocante a definição formal de algo que é tão dúbio: afinal, o que é plágio? Como definir o que foi copiado? Como agilizar o processo de identificação de plágios?

2. O direito autoral, no Brasil e no Mundo

Algumas organizações de grande representatividade, tal como a WIPO (World Intellectual Property Organization) apresentam de forma bastante clara a necessidade de combate do plágio e realizar a defesa do direito autoral. O Dr Uchtenhagen, em seu

compêndio, com foco em obras musicais e culturais, comenta: “The document provides short explanations, on the different steps, conditions and on the various parameters indispensable for the creation of a collective management organization for musical works. It includes an interesting work plan indicating in a time frame the different stages necessary for achieving such goal”.

Estas palavras, voltadas inicialmente ao direito autoral musical, são perfeitas a este contexto, pois uma obra musical e uma obra literária, são, a principio, equivalentes no sentido da criação e propriedade intelecutal Assim, pode-se estender a ideia do direito autoral para as diversas áreas da criação humana. Estes estudos são as bases filosóficas trabalhadas e defendidas pela WIPO, uma organização mantida pela ONU que tem por meta garantir os direitos intelectuais em todos os âmbitos, modelos e instâncias.

A defesa do direito autoral é defendida também por inúmeros pesquisadores destas nossas terras tupiniquins, tal como MENDES, que relaciona as economias emergentes com a necessidade de pesquisa e desenvolvimento, ressaltando que mais que meros usuários tecnológicos, devemos cumprir nossa missão de gerar conhecimentos e inovar com idéias.. “ Over the last 20 years, emerging economies have generally benefitted from global trends in research and development (R&D) which have helped to boost national development and nurture broader international cooperation. An emphasis on innovation, its promotion and associated intellectual property (IP) aspects are key features of any policy that seeks to effectively promote economic growth and development. “. É nitida a visão do partner WIPO que devemos atuar de forma criativa, promovendo e defendendo o IP (Intelectual Propriety) de forma a gerar um cultura que gere um crescimento sustentável.

O direito autoral é tratado de formas diferentes, nas diversas academias de ciências, nos quatro cantos do mundo, e uma simples análise expõe que o respeito pela propriedade intelectual é tanto maior quanto maior for o grau de desenvolvimento (não necessariamente acadêmico) de uma população. Paises ditos desenvolvidos tendem a prover maiores garantias de IP.

Para ROMANCINI, o plágio caracteriza-se como uma falsa atribuição de autoria, uma apropriação indevida de trabalho de um autor por outro indivíduo (o plagiador). A cópia de idéias ou conteúdos de trabalhos de outra pessoa. É interessante notar que a origem etimológica da palavra (do grego “plágios” ao latim “plagiu”) carrega acepções que ilustram o conceito: “oblíquo”, “dissimulado”, “trapaceiro”.

Na França, o plágio ou plagiat é considerado crime, sendo aplicadas sanções penais e cíveis, tal como é claramente apresentado e exposto por Poirier, em seu site: “Pour ce type de procédure, il est préférable de s'adresser à un avocat, ce qui est obligatoire dans la procédure civile, devant le Tribunal de grande instance. La contrefaçon peut donner lieu à des sanctions pénales et à des sanctions civiles (paiement de dommages-intérêts).” Ou seja, a contravenção, como é definida pelo Ministère de la Culture, é um caso de polícia, e assim deve ser tratado.

De acordo com o Advogado autoralista brasileiro, Rodrigo Moraes, “ ... a primeira lei específica sobre Direito Autoral entrou em vigor em 1710, na Inglaterra, no período da Rainha Ana (Statute of Anne), e visava proteger obras literárias. Foi denominada Copyright Act. Daí ter surgido a expressão copyright, utilizada ainda hoje nos países de língua inglesa ... “

Surfando e lendo as palavras de defensores do dito ‘autor criativo’, em uma breve busca em textos Croatas, encontramos o prof Pažur, que, com muita propriedade nos diz, em croata: “S pojavom masovnog tržišta za tiskanu i kasnije snimljenu i emitiranu riječ, autori postaju značajan ekonomski čimbenik u trgovini1.Pravni balans prebacuje se s izdavača na autora čemu su posebno pridonijeli međunarodni zakonski akti počevši od Bernske konvencije iz 1886”, o que, com uma breve tradução “... com o advento do mercado de massa para impressos e gravados mais tarde e palavra transmitida, os autores devem ser considerar os fatores economicos significativos à sua produção, sendo atribuido um valor jurídico ao autor propriamente dito, que são atribuidos com base actos jurídicos internacionais que vão desde a Convenção de Berna de 1886 ...” mostra que não é de hoje a preocupação com a dita, propriedade intelectual.

Há de se enfatizar e enaltecer algumas ações, tal como a proposta pela ABNT, com a norma padronizadora, referente a citações, (NBR 10.520:2002), que são muito bem vindas, pois tendem facilitar e definir modelos de representação das fontes primárias, o que acaba por expor o plágio e evidenciar a falta da citação da fonte.

Finalmente, deixemos claro que o combate ao plágio e a defesa da propriedade intelectual ocorre em praticamente todas as sociedades, sendo mais visíveis e evidentes justamente onde a preocupação com a educação e o desenvolvimento intelectual e criativo são colocados como prioridade.

3. A metodologia do farejador

O programa Farejador de Plágio (http://www.farejadordeplagio.com.br), carinhosamente descrito por muitos como FDP, é uma ferramenta criada em 2006, que conta com milhares de usuarios, que já realizaram a analise de cerca de 520.000 documentos. O principio do farejador é bastante simples. Com base na leitura de um documento e em algumas configurações básicas do usuário, são executadas pesquisas sequenciais de trechos continuos do documento em sites de busca. Ao findar as buscas, são aplicadas diversas técnicas de processamento de dados a fim de realizar o apontamento do que é ou não copiado da internet.

3.1 Variaveis do programa

O programa tem por entradas de dados, o documento a ser analisado, a

definição dos buscadores utilizados, e o padrão de consulta, que pode ser: Rapida, Normal, Detalhada e Rigorosa.

Já as variáveis de saída são, o documento analisado, uma tabela de links encontrados (apresentados em curva ABC), e dados estatístico:percentual de participação dos mais usados, numero de áreas suspeitas, número de trechos suspeitos e quantidade total de ocorrências em cada buscador utilizado.

As variáveis de saida são geradas pelas rotinas de processamento, e só devem ser analisadas e interpretadas após a finalização de todo o processamento. As variáveis, propriamente ditas são:

Trechos Pesquisados	Número total de trechos pesquisados, cada trecho é enviado para cada buscador escolhido para uso
Sites semelhantes	Número total de resultados encontrados em todos os sites de busca, em todas as pesquisas realizadas.
Buscadores	O número de resultados de cada buscador é apresentado, separadamente.
Areas Suspeitas	Uma área suspeita é apontada sempre que pesquisas sequencias retornam a presença de semelhancas continuas. Geralmente uma incidência maior de plágio está associada a áreas de plágio contínuas. Outro fator que gera áreas suspeitas é plágio permutado, onde frases são alternadas. A forma de análise do programa facilita e expóe facilmente estas ocorrências.
Sites Suspeitos	Todos os sites os quais coincidirem mais de 4 vezes serão mostrados, como possiveis fontes de plagio. Se coincidir 4 vezes, e for em uma área suspeita, tambem será mostrado
1 a 2 mais usados	Este fator (não índice) indica o percentual dos 2 sites que mais aconteceram em relação ao total de registros coincidentes indexados na busca. **(Ver 4.3.3)**
1 a 5 mais usados	Este fator (não índice) indica o percentual dos 5 sites que mais aconteceram em relação ao total de registros coincidentes indexados na busca. **(Ver 4.3.3)**
1 a 10 mais usados	Este fator (não índice) indica o percentual dos 10 sites que mais aconteceram em relação ao total de registros coincidentes indexados na busca. **(Ver 4.3.3)**
Paginas / minuto	Taxa de leitura / pesquisas por minuto, que indica a velocidade das pesquisas
Confirmações /minutos	Taxa de confirmação, taxas elevadas indicam uma maior incidência em um trecho específico. A taxa NÃO é linear, ou seja, se ocorrerem muitas em uma área, a taxa será maior

A análise parcial dos dados, durante o processamento podem dar uma visão distorcida do documento, sugere-se, sempre, aguardar o término do processamento.

3.2 Funções de Processamento

A rotina de processamento do farejador é baseada em um modelo de processamento em força bruta. Todas as informações são processadas internamente no computador de análise, e, como é de imaginar, necessita intensivamente do acesso a WEB. Cada informação encontrada é gerenciada e armazenada em estruturas de dados na memória do computador, e ao fim do processamento, os dados são organizados e apresentados.

O gerenciamento, armazenamento e comparativos realizados em tempo real,

associados ao controle dos diversos objetos de programação, a interação com os diversos sites de pesquisa, a interpretação dos dados dos sites e, finalmente, a aplicação das heurísticas de análise continuada, como é de se esperar, consome uma parcela considerável do processamento do computador.

O processamento ocorre na forma clássica de um sistema de informações: variáveis de entrada, processamento e geração de informações de saída. O que difere o FDP é justamente o caráter dinâmico dos processamentos do programa, que, com alta interação com a internet, possibilita e permite que dados dinâmicos interfiram diretamente nos resultados do programa.

3.3 Gerenciamento das variáveis de processamento

O programa, ao longo do processamento, coleta dados e armazena todas as informações em chamadas variáveis de processamento. Considerando um documento de 100 páginas, com busca no modo NORMAL, gerará cerca de 2000 pesquisas, que, usando 8 buscadores, com 10 respostas por buscador, chegará a um número máximo de 160.000 resultados indexados.

Todas estas informações devem ser armazenadas em memória, sendo guardadas informações de localização, trecho, site, ordem e site de busca. Para se ter uma idéia do volume de informações, e quantidade de memória utilizada, cada resultado indexado poderia utilizar até 200 bytes, o que acarretaria o uso de mais de 3 Gbytes de dados. Isto mostra a grande quantidade de dados a ser gerenciada, e a dificuldade em manipular estes dados em um ambiente não computacional.

3.4 Heurísticas de busca e organização de dados

Dentre as diversas as abordagens de solução implantadas no programa, o processamento em força bruta é talvez a principal caracteristica do programa. O caso é que a análise e busca do plágio exige que todo o documento seja lido, e buscas esporádicas ou aleatórias poderiam apresentar falhas, ao ignorar um trecho que justamente seja copiado.

Desta forma, considerando o texto como uma sequência de palavras, o programa fará buscas sequenciais, tais como o professor o faria, copiando uma parte de uma frase e a inserindo em uma ferramenta de busca.

No caso do programa, a quantidade de palavras de cada busca é definida pelo usuário, quando o mesmo escolhe nas configurações do programa entre RAPIDA, NORMAL, DETALHADA e RIGOROSA. Para cada configuração, uma forma de pesquisa é definida:

	Quantas pesquisar por ciclo	Quantas pular em cada ciclo
Rápida	8 ou 9 palavras	8 a 10 palavras

Normal	6 ou 7 palavras	7 a 8 palavras
Detalhada	5 ou 6 palavras	5 a 7 palvras
Rigorosa	4 a 5 palavras	4 a 6 palavras

Em versões anteriores do programa, era permitido o usuário definir estes valores manualmente, no entanto configurações errôneas do programa acarretavam em falhas de processamento. Assim, foram definidos os modelos descritos acima, de forma a evitar falhas e equívocos de processamento. Por experiência, relatos e testes realizados, o modelo pré-configurado é o que apresenta os melhores resultados.

3.5 Máquina de inferência e organização de conjuntos

Considerando que a análise dos links indexados só faz sentido ao final da passagem por todo o documento, a rotina principal de processamento realizará o apontamento dos plagios no próprio documento analisado.

Entretanto, alguns resultados parciais podem ser apresentados e outros suprimidos, com base em algumas premissas básicas relacionadas a informações históricas, armazenadas de acordo com modelos pré-processados. Desta forma, diversos sites genéricos e repetitivos são suprimidos das análises, diminuindo a carga de processamento, agilizando o estudo e facilitando a análise dos resultados finais do programa.

Ao longo do processamento, diversas informações de controle e gerenciamento são armazenadas e geridas de forma a possibilitar o tratamento das variáveis de entrada e processamento. Estas informações combinadas com o tratamento dos links permite realizar apontamentos de eventuais coincidencias típicas de trechos copiados intencionalmente.

Algumas formas de cópias podem ser ditas não intencionais, visto ocorrerem de forma esporádica e aleatória. Estas ocorrências são ignoradas pelo FDP, por serem consideradas de difícil comprovação. São consideradas não intencionais por ocorrerem poucas vezes, o que dificulta realizar uma afirmação da quebra da direito autoral.

Computacionalmente falando, a geração de modelos mostra trata-se de um conjunto restrito a poucos elementos não conexos que não geram um modelo que se encaixe na definição do uso de ideias alheias como suas, ou seja, no entendimento do FDP, não seria plágio, por definição absoluta.

3.6 Lógica Fuzzy na definição de limites dos conjuntos de resposta

A lógica fuzzy, ou teria nebulosa, é uma forma de processamento de dados amplamente utilizado nas mais diversas áreas das ciências, desde a medicina até a mecanica, metalurgia, farmácia e contabilidade.

Dentre as vantagens da utilização desta técnica, destaca-se o tratamento

mais ‘humano’ das variáveis, permitindo que se considere e dê valor parcial as variáveis analisadas, não excluído algo próximo ao resultado limite, e agregando uma certa incerteza nas variáveis que resultaram positivas, ou seja, os limites dos resultados deixam de ser claros, passando a ser considerados inclusive dados que não estejam dentro do conjunto de resposta e, por vezes, eliminando dados que sejam considerados válidos, por serem considerados de menor valor, com base nos valores dos demais conjuntos resposta.

Neste momento há de se perguntar: Mas como o programa consegue validar dados que estejam fora da regra de pertinência e excluir outros resultados considerados relevântes ?

O caso é que com base nos demais conjuntos, é possivel definir um grau de relevância aos dados. Uma forma de explicar o modelo é de processamento fuzzy do FDP é com uma eleição com dois turnos. Dois candidatos detem grande maioria, mas os resultados dos minoritários podem influenciar no resultados, e mesmo com valores pouco expressivos, poderão alterar o resultado, ao impedir que um candidato obtenha os 50% necessarios. Ou seja, mesmo uma quantidade de 0,6% teria uma grande importância para um candidato que tenha feito 49,8%.

A idéia fuzzy permitiu que fossem evoluída a forma de análise do FDP, e facilitado consideravelmente o trabalho de análise dos resultados gerados pelo programa, ao mesmo tempo que permite uma evolução constante dos resultados gerados, aos fazer a análise acumulativa dos diversos resultados.

4. A proposta de interpretação

Como já foi falado, não há uma lei que defina plágio. Mesmo recomendações são vagas. O WIPO faz considerações e apontamentos de como tratar informações semelhantes, oriundas de fontes diversas, mas é prudente e receosa em definir plágio.

Com base na experiência do autor, e as diversas análises e estudos realizados, que culminaram na modelagem computacional descrita no capitulo 3, o que será apresentado a seguir é uma proposta de análise. Deixo claro, neste momento que considero um insensato apresentar minhas ponderações de análise como sendo algo definitivo, completo ou que venha a ser utilizado como modelo ideal de análise.

Como já comentado inicialmente, trata-se de uma mera proposta, que pode e deve ser evoluída com base na experiência mútua e cooperativa de outros pesquisadores que tenham o combate ao plágio como uma bandeira, e o estímulo a pesquisa e inovação como uma meta a todos seus pupilos.

4.1 Tipos de plagio detectado

As formas de plágio detectadas e propostas pelo FDP são classificadas em:

a) Trechos: Contínuos ou fragmentados - as áreas suspeitas

b) Frases esparsas em documentos extensos - coincidências esparsas
c) Sites mais utilizados
d) Permutas de fragmentos de parágrafos, frases ou orações
e) Frases modificadas com alto grau de similaridade
f) Texto disperso e alternâncias de ordem de frases
g) Uso de informações em sites de indexação elevada
h) Similaridades em erros fonético/sintaticos em orações similares

Novamente enfatizo que apresento uma proposta, formulada para permitir o processamento de um modelo computacional de compilação heurística genérica, trabalhando com as variaveis, técnicas, modelos e rotinas apresentadas no capitulo 3.

Tendo por base estas regras, torna-se necesário que o FDP gere os índices e fatores que permitam a análise das formas de plágios definidas. Estes valores servirão de base para que um avaliador consiga definir o grau de similaridade do documento analisado com os demais documentos encontrados na internet.

4.2 Índices e fatores do FDP

O cruzamento dos resultados obtidos pelo FDP, depois de todo o documento ser analisado, traz consigo a necessidade de se utilizar uma matriz esparsa de termos, trechos e links, os quais, combinados e compilados, permitirão a criação de uma visão ampla e geral do contexto do documento, ou seja, o que temos neste documento que também existe em outros documentos na internet ?

4.3 O que o FDP considera plágio

As propostas de definição de plágio a seguir apresentadas foram definidas a fim de permitir a análise computacional autonoma do documento, bem como possibilitar a geração de conjuntos e vetores básicos de análise.

Mais que sugerir cautela no uso desta proposta, pede-se que seja usado o BOM SENSO na análise dos resultados, principalmente quando o resultado acabe por afirmar que um documento tenha sido plagiado, e um eventual trabalho acadêmico seja incorretamente taxado como plagiado.

A seguir serão apresentados os oito casos computacionalmente descritos como plágio pelo FDP. Será utilizado um trecho de um artigo aleatório, que permite apresentar e expôr as oito situações. O trecho em questão será apresentado a seguir, sendo suas frases numeradas, para referencia nos tópicos de analise 4.3.1 a 4.3.8

De forma a analisar e demonstrar de forma clara, um texto de exemplo será utilizado, onde apontamentos e demonstrações serão utilizadas. No texto do exemplo, simulado, 10 pesquisas ocorrem, marcadas em azul. Tal como no programa, em

azul_sublinhado tem-se os trechos apontados como plagio e os **{links}** após os sublinhados são os locais de possível origem dos textos encontrados.

Texto de Exemplo:

[1]Sem hesitar, as obras musicais, fotografias e as audiovisuais, em face da subsistência (programas [2] de computador) e hardwares (máquinas) que ***{www.ldphi.com/artigoluizcarlos.pdf}*** *permitem com facilidade seu armazenamento, cópias, distribuição e alterações com fins econômicos no [3]e-commerce por mediação da rede mundial de computadores,* ***{www.ldphi.com/artigoluizcarlos.pdf}*** ***{www.tecnologiasa.com.br/tecnologias.pdf}*** *por qualquer indivíduo, trazem grandes perdas pecuniárias aos [5]autores de softwares*

No tablado científico, há probabilidade de negociação de banco de dados ou textos com resultados de experimentos e [6]pesquisas, em específico, pautados com princípios ativos ***{www.ldphi.com/artigoluizcarlos.pdf}*** *usados no preparativo de medicamentos, muito preciosos para a indústria [7]farmacêutica.*

Quanto aos programas de computador (softwares), por sua origem digital, isto é, compostos por binários [8] numéricos (0 e 1), desorelo comenta acerca da ***{www.scribd.com/camarotti_souza.pdf}*** *facilidade de clonagem, rompimento de códigos de segurança, distribuição, [9]transmaissão, armazenamento, etc. Habitualmente, nota-se* ***{www.scribd.com/artusi_2008.pdf}*** *na internet ofertas de programas de jogos, seja em CD-ROM's – entregues pelos correios – ou, [10]facilmente, posterior a confirmação de pagamento* ***{www.ldphi.com/artigoluizcarlos.pdf}*** ***{www.scribd.com/artusi_2008.pdf}*** *e fornecimento de uma senha, oferecida via eletrônica.*

Os diversos casos de plágio dtectados pelo programa são agora expostos e explanados, com base no texto de exemplo.

4.3.1 Trechos: Contínuos ou fragmentados - as áreas suspeitas

O plágio em trechos é o mais fácil de ser detectado e apontado, bem como o mais difícil de negar. Neste caso, é detectada a semelhança de mais de 4, 5 ou 6 trechos do documento analisado junto a documentos na internet. É a forma mais simples da detecção, onde, ao final da analise do FDP, são verificadas nas variáveis de processamento, quantas vezes cada determinado site foi encontrado, montando uma curva ABC e apontando, para cada site, quantas ocorrências do mesmo, foram apontadas.

Estes casos são os apontados na análise do texto exemplo, como as citações ***{www.ldphi.com/artigoluizcarlos.pdf}*** ***{www.ldphi.com/artigoluizcarlos.pdf}*** **,** nestes dois casos, o que se tem são sequências positivas de coincidências, que acabam por gerar áreas de suspeita, pois são trechos sequenciais com uma mesma fonte ou fontes.

Toda área suspeita é preocupante, principalmente se ocorrerem fora de áreas de citação e gerarem sequências superiores a 3 ou 4 links repetidos, pois indicaria a cópia de uma área estensa de texto.

4.3.2 Frases esparsas em documentos extensos - coincidências esparsas

Este é o caso complementar ao apresentado no 4.3.1, onde uma fonte, de forma absolutamente esparsa é apontada, que é o caso de ***{www.tecnologiasa.com.br/***

tecnologias.pdf} {www.scribd.com/camarotti_souza.pdf} onde o texto coincide com sites, de forma não contínua, mas em pelo menos outras 4 ou 5 vezes ao longo de todo o documento, ou seja, provavelmente trechos foram copiados de ajustados, gerando cópias descontinuadas.

A falha desta forma de apontamento de plágio está justamente na possibilidade de que, em um dado assunto, a continuidade textual ocorrerá e poderá gerar frases coincidentes com outros documentos, justamente por terem muitas similaridades relativas aos conteudos trabalhados.

Sugere-se cuidado e cautela com estes dados, principalmente quando o número de ocorrências, apontadas na curva ABC for inferior a 8 vezes. Este é o caso de plágio detectado quase impossível de se encontrar de forma manual, e uma grande contribuição no tocante a busca pelo FDP. Considero um caso de plágio indireto, que apoia e fortalece os plagios diretos.

4.3.3 Sites mais utilizados

Com base na curva ABC apresentada pelo programa, há de se considerar que os sites mais encontrados (topo da lista ABC) são os de maior probabilidade de constarem como plágio, justamente por aparecerem muitas vezes.

Como interpretar o fator dos mais usados:

	1 a 2 sites	1 a 5 sites	1 a 10 sites
até 10 pag	1% pouco 3% verificar + 6% preocupante	1% pouco 4% verificar + 8% preocupante	2% pouco 4% verificar + 9% preocupante
10 a 30 pag	1% pouco 3% verificar + 5% preocupante	1% pouco 4% verificar + 7% preocupante	2% pouco 4% verificar + 8% preocupante
30 a 60 pag	0% pouco 2% verificar + 4% preocupante	1% pouco 3% verificar + 5% preocupante	1% pouco 3% verificar + 6% preocupante
60 ou + pag	0% pouco 1% verificar + 3% preocupante	0% pouco 2% verificar + 4% preocupante	1% pouco 3% verificar + 5% preocupante

Enfatizo que estes são valores de referência, e devem ser excluidas as referencias devidamente citadas, leis e jargões, o que pode, muitas vezes, alterar os percentuais. Sugere-se cautela e atenção na análise destes dados.

Estes números devem ser analisados com muito cuidado, pois em documento técnicos, legislativos e com alta bagagem conceitual, as citações mesmo indiretas devem ser excluidas, para não se gerar falsos alertas de plágio, e acabar por gerar falsas acusações de plágio.

Este tipo de analise não é indicada para o modo de analise rápida com

documentos menores de 10 páginas, apesar de funcionar em casos de plágios evidentes, com percentuais elevados de coincidência.

4.4.4 Permutas de fragmentos de parágrafos, frases ou orações

Algo bastante comum e corriqueiro é copiar diversos parágrafos de várias fontes e realizar a permuta de frases ou fragmentos de frases. Apesar de complicar bastante a conexão e detecção por métodos tradicionais e manuais, é facilmente detectável pelo FDP, principalmente em seu modo RIGOROSO.

Obviamente não são geradas áreas suspeitas, mas aparecerão como coincidências esparsas, mas constarão normalmente na curva ABC do programa. Por experiência, pode-se afirmar a predisposição do escritor em realmente tentar burlar a autoria, ou seja, pretendia-se realmente e intencionalmente efetuar o uso de idéias alheias, pois alem de copiar e não referenciar, manipulou-se de forma proposital os parágrafos escritos.

Alguns programas e sites propõe-se inclusive a realizar a geração automática de textos com base em textos de referência, obviamente sem citar as referencias. Por inocência ou falta de inteligência, algumas pessoas não consideram a mistura de frases de documentos como plágio, o que certamente é um grande erro.

É interessante ainda, a forma clara que o FDP apresenta os resultados, visto que nestes casos de permuta de fragmentos, há uma grande tendência a repetição de autores gerando conjuntos, que, apesar de desconexos, evidenciam com uma clareza indiscutível a falta de referenciação dos autores das ideias.

4.4.5 Frases modificadas com alto grau de similaridade

Este tipo de plágio é muito mais complexo, tanto em se encontrar como na sua comprovação, pois não trata-se das mesmas palavras, e sim das mesmas ideias, o que, por definição de propriedade intelectual, é o que realmente importa.

O detalhe é que, a nível de processamento, a sutilesa da alternância de adjetivos e substantivos pode passar despercebida, até mesmo porque as ferramentas de busca já aplicam esta ideia em suas buscas, assim, o FDP acaba herdando essa capacidade de buscar pela ideia, e não pelo texto literal.

Este tipo plágio, é considerado por este autor, como um plagio direto, mas há de se considerar que o fato de alterar algumas letras ou palavras pode ser visto por alguns como plágio e por outros como uma obra inédita.

Uma forma de dirimir a dúvida e permitir uma afirmativa quanto a ser ou não plagio, a utilização de frase com alto grau de similaridade e utilizar-se do princípio da propriedade intelectual. O FDP busca e tem a capacidade de apontar origens onde a diferença entre os textos seja sutil ou pequena.

O autor sugere extrema cautela na afirmativa de plágio utilizando-se unicamente deste argumento, pois apesar de claro, pode gerar controversias relativas à compreensão de realmente tratar-se de plágio, pois é altamente subjetivo definir plágio em uma ideia que foi alterada.

O interessante é utilizar os plagios por similaridade como complemento a plagios diretos, de forma a reforçar a ideia que um texto teria sido copiado. Estaria agregando força a ideia do plagio direto encontrado, pois seria mais um indício de que existem problemas no desenvolvimento e criação do documento.

4.4.6 Texto disperso e alternâncias de ordem de frases

Muito parecido com o item 4.4.4 , é bastante comum que pegue um texto qualquer e o espalhe, colocando, por exemplo, uma ou duas frases por folha, ao longo de 20 ou 30 páginas. Desta forma, consegue-se uma falsa impressão de volume com certa rapidez, pois, mesmo o texto ficando desconexo, algo dificilmente perceptível, pois continua havendo concordância temática, o que para algumas pessoas é visto como normal.

Pessoalmente, considero quase impossível que manualmente detecte-se plagio de uma fonte ao longo de várias páginas, quando somente uma frase esteja presente por página, mas, novamente, para o FDP, isto náo é um problema, pois trabalha com o indexador de páginas, e consegue ‘pegar’ esta malandragem facilmente.

Como não há alterações nos documentos originais, considero este um plágio direto, e a repetição por diversas páginas implicaria em uma afirmativa de má fé e interesse de fato em copiar as ideias.

4.4.7 Uso de informações em sites de indexação elevada

Um artifício desconhecido por muitos e que passa facilmente despercebido é o uso de documentos que não estejam no inicio das respostas dos principais indexadores, tal como o google e o bing. Isto porque, como o resultado não é direto, e correspodneria á pagina 30 ou 40, acabaria por passar despercebido, visto não tratar-se de um documento conhecido, o que geraria uma falsa impressão de que o documento é autêntico.

O FDP consegue buscar estes documentos, e concede a estes documentos o mesmo peso de documentos no topo da lista de indexadores do sites de busca. Para isto basta que o usuário vá no Internet Explorer, e configure manualmente UMA vez os buscadores para que este retornem mais páginas de busca. O resultado será uma pesquisa mais lenta, porem muito mais eficiente, tanto em resultados como em prcisão do apontamento dos plagios.

Mas atenção, sugere-se o uso destas configurações mais rigorosas caso tenha-

se uma suspeita de que o trabalho possa ter sido burlado neste sentido, pois o tempo de resposta, dependendo da velocidade da internet, pode ser muito maior.

Vários relatos de usuários denotam o uso deste artifício e que os resultados foram muito interessantes, alterando bastante os resultados finais. Caso haja a supeita, sugere-se alterar o numero de respostas dos buscadores, bem como o valor padrão de 10, pode ser retornado a qualquer momento, inclusive durante o uso do FDP.

4.4.8 Similaridades em erros fonético/sintaticos em orações similares

Esta forma de identificar plágio é, no minimo, curiosa e inusitada.

Considerado um documento que contenha diversos erros ortográficos, sintáticos e semânticos. Os resultados do FDP podem, e facilmente conseguem buscar por este documento em um documento durante sua análise, justamente devido as discrepâncias existentes entre uma escrita correta e este, que contem erros.

O resultado que se obtém é justamente o apontamento do site com base nas diversas falhas de escrita. Este tipo de plágio, alem de fácil de detectar, acaba por delatar o autor, pois alem dos textos constarem em outros documentos, apresentam um elevado grau de discrepância a nivel da qualidade de escrita. Este, considero como sendo um plágio direto, de fácil detecção e difícil explicação.

5. Conclusões

A experiência do desenvolvimento, manutenção e uso do programa ao longo dos anos, bem como a valiosa e indescritível experiência trocada com alguns mestres da arte de ensinar, me trouxeram e ajudaram a criar uma visão própria do que é o plágio. Algo por muitos considerado de pouca relevância ou ‘preteritamente consertável‘, é visto por muitos, e por mim, como uma responsabilidade exclusiva do tutor.

Em outras palavras, o aluno, pupilo, discente ou acadêmico está ali para aprender. Caso este acadêmico cometa o ERRO de copiar, e o docente considere isto normal, o irresponsável não é quem está aprendendo, pois, enfatizo, está aprendendo. A tarefa de ensinar e explicar o erro é exclusivamente do DOCENTE. Cabe ao discente buscar pela forma mais fácil de resolver o problema. E ao docente explicar que nem sempre a forma mais fácil é a melhor forma.

Quanto a metodologia de definição de plágio apresentada, ressalto, uma última vez: “trata-se de uma metodologia baseada em técnicas computacionais, que, baseadas em variáveis e regras de processamento, busca e aponta coincidências, apresentando-as de forma a agilizar a análise dos resultados e definir com clareza se um documento é realmente desenvolvido por um determinado autor, sugere-se que cada resultado seja analisado separadamente”.

Referências Bibliograficas (afinal , devemos citar os autores , correto ?)

LEI Nº 9.610, DE 19 DE FEVEREIRO DE 1998. disponivel em https://www.planalto.gov.br/ccivil_03/Leis/L9610.htm#art115

WIPO World Intellectual Property Organization disponivel em www.wipo.int em 2/11/2010

Uchtenhagen, Ulrich The Setting-up of New Copyright Societies disponivel em http://www.wipo.int/freepublications/en/copyright/926/wipo_pub_926.pdf em 8/11/2010

Mendes,. Ricardo Camargo Innovation Promotion in Brazil Wipo Magazine September 2010

Romancini, Richard. *A praga do Plágio Acadêmico*. Revista Científica da FAMEC, ano 6, nº 06/2007, ISSN-1677-4612.

Poirier, François Université Paris 13. Disponivel em: http://www.univ-paris13.fr/ANGLICISTES/POIRIER/Plagiat.htm em 7/11/2010

Pažur, Ivana AUTORI ZNANSTVENIH RADOVA I AUTORSKO PRAVO Institut Ruđer Bošković, Croace