UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DESENVOLVIMENTO E APLICAÇÃO DE UM MÉTODO PARA DETECÇÃO DE INDÍCIOS DE PLÁGIO

DISSERTAÇÃO DE MESTRADO

Solange de Lurdes Pertile

Santa Maria, RS, Brasil

2011

DESENVOLVIMENTO E APLICAÇÃO DE UM MÉTODO PARA DETECÇÃO DE INDÍCIOS DE PLÁGIO

Solange de Lurdes Pertile

Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Informática, Área de Concentração em Computação Aplicada, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de
Mestre em Ciência da Computação.

Orientador: Profª. Drª. Roseclea Duarte Medina

Santa Maria, RS, Brasil

2011

P469d Pertile, Solange de Lurdes
Desenvolvimento e aplicação de um método para detecção de indícios de plágio / por Solange de Lurdes Pertile. – 2011.
72 f. : il. ; 30 cm

Orientador: Roseclea Duarte Medina
Dissertação (mestrado) – Universidade Federal de Santa Maria, Centro de Tecnologia, Programa de Pós-Graduação em Informática, RS, 2011

1. Informática 2. Software 3. Programação 4. Ambientes virtuais 5. Detecção automática 6. Indícios de plágio 7. Educação à distância 8. Moodle 9. Mle-Moodle I. Medina, Roseclea Duarte II. Título.

CDU 004.4

Ficha catalográfica elaborada por Cláudia Terezinha Branco Gallotti – CRB 10/1109
Biblioteca Central UFSM

Universidade Federal de Santa Maria
Centro de Tecnologia
Programa de Pós-Graduação em Informática

A Comissão Examinadora, abaixo assinada, aprova a dissertação de Mestrado

DESENVOLVIMENTO E APLICAÇÃO DE UM MÉTODO PARA DETECÇÃO DE INDÍCIOS DE PLÁGIO

elaborado por
Solange de L. Pertile

Como requisito parcial para a obtenção do grau de
Mestre em Ciência da Computação

BANCA EXAMINADORA:

_____
Roseclea Duarte Medina, Drª. (UFSM)
(Presidente / Orientadora)

_____
Lisandra Manzoni Fontoura, Drª. (UFSM)
(Examinadora)

_____
Patricia Alejandra Behar Drª. (UFRGS)
(Examinadora)

Santa Maria, 11 de Março de 2011.

DEDICATÓRIA

À minha família e ao meu namorado,
pelo amor, pela paciência e
por todos os momentos em
que estiveram ao meu lado
nessa jornada.

AGRADECIMENTOS

Primeiramente quero agradecer a minha família que sempre me apoiou, em especial á minha mãe pelo apoio efetivo nas horas boas e principalmente nas horas mais difíceis, quando nem tudo estava saindo como deveria.

Aos meus amigos pelo apoio efetivo e moral, e principalmente a meu amigo e colega Jaziel pela troca de experiência no laboratório, e claro pela sua ajuda, pois estava sempre à disposição para me ajudar nas horas de dúvidas na fase de desenvolvimento do trabalho.

Á minha amiga e orientadora Professora Roseclea pela orientação e apoio neste trabalho.

Ao meu namorado Eduardo pela dedicação, apoio e amor em todos os momentos.

Enfim, a todos que direta ou indiretamente contribuíram para a realização deste trabalho e em especial a Deus.

RESUMO

Dissertação de Mestrado
Programa de Pós-Graduação em Informática
Universidade Federal de Santa Maria

DESENVOLVIMENTO E APLICAÇÃO DE UM MÉTODO PARA DETECÇÃO DE INDÍCIOS DE PLÁGIO

AUTORA: Solange de L. Pertile
ORIENTADORA: Dr[a]. Roseclea Duarte Medina
Data e local da defesa: Santa Maria, 11 de março de 2011.

A distribuição e o acesso a informações por um número muito maior de pessoas na internet tem crescido de forma exponencial, o que vem dificultando o controle da originalidade de tais informações e facilitando o trabalho dos usuários plagiadores que fazem uso de tais informações de forma inadequada. Neste contexto que se destaca a importância de avaliar os textos produzidos nos cursos de pós-graduação e graduação, nas modalidades à distância e presenciais, que esta dissertação propõe um novo método para detecção de índicios de plágio em trabalhos acadêmicos, o qual realiza buscas por fragmentos similares com documentos da web. O método desenvolvido analisa o plágio mosaico, onde o autor copia partes de uma obra trocando somente algumas palavras sem dar crédito ao autor da obra original; e o plágio bilíngue, onde o conteúdo de um documento no idioma inglês é traduzido para o idioma português sem fazer referência à obra original. Além disso, o método foi implementado em uma ferramenta e integrada ao módulo de envio de tarefas da plataforma Moodle para acesso via desktop e pelo dispositivo móvel, visando potencializar aos professores os benefícios de sua utilização, já na postagem dos trabalhos no AVA; e implementada como um sistema computacional desktop para permitir aos usuários seu acesso também fora do AVA Moodle. Os resultados obtidos mostram que o método desenvolvido alcançou resultados satisfatórios em relação a outras técnicas encontrados na literatura, obtendo sobre uma coleção de 14 documentos índices de similaridades variando de 30,07% a 40% e com uma precisão nos resultados retornados entre 71,42% e 96,15%. Os resultados do experimento de um documento traduzido do inglês para o português teve uma precisão de 100% nos resultados retornados.

PALAVRAS-CHAVE: Detecção automática de indícios de plágio, Ambientes Virtuais de Aprendizagem, M-Learning, Educação à Distância, Moodle, Mle-Moodle.

ABSTRACT

Dissertação de Mestrado
Programa de Pós Graduação em Informática
Universidade Federal de Santa Maria

DESENVOLVIMENTO E APLICAÇÃO DE UM MÉTODO PARA DETECÇÃO DE INDÍCIOS DE PLÁGIO

AUTORA: Solange de L. Pertile
ORIENTADORA: Roseclea Duarte Medina
Data e local da defesa: Santa Maria, 11 de março de 2011

The distribution and access to information by a much larger number of people on the Internet has grown in a exponential, way has been hard the which control of the originality of the information and facilitating the work of plagiarists users who make use of such information inappropriately. It is in this context that stands the importance of evalute the texts produced in the post-graduate and graduate courses in the classroom and distance modalities, that this paper proposes a new method to detect signs of plagiarism in academic work, which performs to searche similar fragments with web documents. The method developed analyzes the mosaic plagiarism, where the author shares copies of a work by changing only a few words without giving credit to the original work, and the bilingual plagiarism, where the contents of a document in English is translated for Portuguese without reference the original work. In addition, the method was implemented in an integrated tool and to the sending task module of the Moodle platform for access by desktop and by the mobile device aiming to empower the teachers the benefits of its utilization in the posting of work in a AVA, and implemented as a desktop computer system to allow users the access also outside theAVA Moodle. The results showed that the developed method reached satisfactory results in relation to other techniques found in literature, getting over a collection of 14 documents indexs of similarities ranging from 30.07% to 40% and with a precision in the returned results between 71.42 % and 96.15%. The experimental results of a translated document from English to Portuguese had a 100% accuracy in the returned results.

KEYWORDS: Automatic detection of signs of plagiarism, Virtual Learning Environments, M-Learning, Distance Education, Moodle, Mle-Moodle.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AVAs Ambientes Virtuais de Aprendizagem

MLE *Mobile Learning Engine*

LMS *Learning Management Systems*

EAD Ensino à Distância

PDAs *Personal Digital Assistants*

ISAM Infra-estrutura de Suporte às Aplicações Móveis

TICs Tecnologias de Informação Aplicada a Educação

LES Laboratório de Engenharia de Software

TIDIA Tecnologia da Informação para o Desenvolvimento da Internet Avançada

FAPESP Fundação de Amparo à Pesquisa do Estado de São Paulo

SUMÁRIO

1 INTRODUÇÃO

A possibilidade de comunicação de qualquer lugar em tempo real através da internet faz com que a disponibilidade de informações digitais na web cresça consideravelmente, tornando acessível a qualquer pessoa conectada a internet um grande acervo de documentos em diferentes formatos disponibilizados em bibliotecas digitais.

Entretanto, a quantidade de informações distribuídas na mídia digital vem favorecendo a prática do plágio, uma vez que reduz o esforço de pesquisa e facilita a cópia dos plagiadores pelas simples operações do teclado ou do mouse para realização de uma pesquisa em motores de busca, selecionando, copiando e colando informações (OLIVEIRA *et al.*, 2007).

Conforme o autor Barbastefano (2007), a internet não é a única causa que leva a comunidade a praticar o ato de plagiar. Entretanto, além da facilidade de acesso à informação na internet, Wood (2004 *apud* Barbastefano, 2007) considera também que os alunos não consideram seu trabalho como válido ou merecedor de proteção intelectual e muitos acreditam que se o documento esta publicado na *web*, então a informação é disponível, verdadeira e livre.

Além disso, Stebelman (1998 *apud* Barbastefano, 2007) ressalta que não apenas a cópia de textos é um problema, mas a tradução também se configura em uso indevido, pela facilidade de acesso a programas de tradução. Este é um problema mais grave pela impossibilidade de rastreamento por ferramentas automáticas de busca.

Já os autores Brown e Howell (2001) apontam o desconhecimento de regras que delimitam o uso de citações e paráfrases sendo um dos principais aspectos que leva muitas vezes os autores a praticarem o ato de plágio sem intensão.

Existem algumas ferramentas, como a Turnitin (2010) e Urkund (2010), com objetivo de automatizar o processo de verificação de indícios de plágio. Geralmente estes sistemas indicam o grau de similaridade entre dois documentos, mas sempre se faz necessário à verificação manual de um humano para verificar se o documento suspeito pode ser considerado como um ato de plágio. Esta análise manual é essencial, já que não existe um sistema de detecção que verifique se o conteúdo considerado como indício de plágio esta ou não corretamente referenciado, ou seja, se o autor deu crédito à obra original. Portanto, o plágio continua a ser um crescente desafio, que afetam muitas áreas, geralmente publicação, educação e até mesmo negócios.

1.1 Contexto e Motivação

A prática de plágio em trabalhos acadêmicos tem se apresentado como um problema aos professores a cada dia, pois os alunos tem praticado este ato sem demonstrar o conhecimento das consequências que isto pode vir a causar. Além disso, a tarefa de verificar o plágio em uma grande quantidade de trabalhos se torna uma tarefa de difícil realização, principalmente quando se trata de trabalhos de conclusão de curso, onde a originalidade é imprescindível e no caso de utilização e citação de outras referências, os mesmos devem ser corretamente referenciados.

Segundo em entrevista realizada por Rabelo (2006), professores de graduação e pós-graduação afirmam que nem sempre identificar a cópia no trabalho entregue pelos alunos é uma tarefa fácil. Muitos professores optam por usar a internet como ferramenta para verificar a autoria dos textos. Além disso, consideram que a medida para verificação de trabalhos escritos pode ser muito trabalhosa, como ocorreu com um dos professores, que passava aos alunos trabalhos de 20 a 30 páginas como avaliação, e quando percebeu que haviam textos copiados resolveu adotar somente provas para compor a nota da disciplina, e afirma que já aconteceu de um aluno plagiar seu próprio texto.

De acordo com Rabelo (2006):

> O decano de Pesquisa e Pós-graduação da instituição, M. L., afirma que os casos de plágio em trabalhos finais são raríssimos. “Na graduação, é mais comum devido a uma porção maior de revisão bibliográfica”, analisa Pimentel [...]

Para Oliveira e Oliveira (2008) o plágio sempre foi um problema no ensino presencial e um grande desafio para os professores identificá-lo e coibi-lo. Mas esse problema também vem assumindo dimensões maiores no ensino a distância, pois a mídia digital de certa forma facilitou a prática do plágio, e a grande quantidade de alunos em cursos EAD acaba por inviabilizar o professor no processo de verificação de plágio de forma manual. Apresenta-se como exemplo, uma disciplina do curso de especialização em TICs, a qual possuí um único professor e cinco pólos com uma média de 40 alunos por pólo, ou seja, totalizando em torno

de 200 alunos. Sendo que o mesmo professor atende mais disciplinas a distância e a carga normal de disciplinas presenciais na gradução e pós-graduação.

Embora já existam sistemas de detecção automática de indícios de plágio, bem como, a Turnitin (2010) e a Urkund (2010) que permitem sua integração a ambientes virtuais de aprendizagem, ou até mesmo ferramentas para desktop, como a Viper (2010), Approbo (2010) entre outras, tais sistemas são, na grande maioria, de natureza privada e ainda requerem melhorias em termos de desempenho na precisão dos resultados e o custo de tempo na fase de análise da similaridade entre os documentos.

No entanto, o método proposto visa melhorar técnicas já existentes na literatura em termos de tempo de processamento na fase de verificação de indícios e na precisão dos resultados.

1.2 Objetivos e contribuições

O objetivo deste trabalho é desenvolver um método que agilize e facilite a análise da qualidade e originalidade de trabalhos acadêmicos. Para atingir este objetivo o método será capaz de receber como entrada um documento textual e identificar se ele contém trechos similares a outros documentos disponíveis na internet.

Esta pesquisa percorre, ainda, os objetivos específicos de:

- Análise das principais ferramentas de detecção de indícios de plágio e robôs de busca existentes na literatura;
- Um estudo sobre ambientes virtuais de aprendizagem, ambientes virtuais de aprendizagem móvel, assim como o ambiente Moodle e Mle- Moodle, os quais foram utilizados neste trabalho.
- Um levantamento dos tipos de plágio e sua prática em trabalhos acadêmicos.
- Pesquisa sobre o desenvolvimento de módulos para integração a plataforma Moodle e Mle- Moodle.
- Integração do método ao ambiente virtual de aprendizagem Moodle e Mle-Moodle.

Desta forma, a principal contribuição deste trabalho é melhorar consideravelmente a precisão dos resultados das abordagens existentes na detecção de indícios de plágio em um custo de tempo aceitável. Além disso, pretende-se também contribuir significativamente tanto com o *e-Learning* como com o *m-Learning* através da inclusão de uma ferramenta que realize

a verificação automática de possíveis indícios de plágio dentro do AVA Moodle e Mle-Moodle, e espera-se, com esta integração, preservar a originalidade dos trabalhos e proporcionar aos docentes que fazem uso do ambiente um melhor controle de tais tarefas, o que agilizará seu trabalho e lhe proporcionará um melhor uso do ambiente.

1.3 Organização do Texto

Para melhor compreensão desta pesquisa, o presente trabalho esta dividido em seis capítulos, sendo o primeiro capítulo a contextualização da dissertação em termos do tema, contexto e motivação, objetivos, contribuições e estrutura do trabalho.

Os capítulos 2, 3 e 4 apresentam a fundamentação teórica da dissertação, baseada em uma revisão bibliográfica, que abordou os seguintes conteúdos:

Capítulo 2: a definição de plágio e sua prática em trabalhos acadêmicos, além disso, apresenta-se uma classificação quanto aos tipos de plágio e uma análise de tecnologias existentes para uso na detecção, inclusive integrados a ambientes virtuais de aprendizagem.

Capítulo 3: apresentam-se alguns conceitos de ambientes virtuais de aprendizagem, assim como alguns exemplos existentes no estado da arte.

Capítulo 4: apresenta a metodologia e a estrutura do método desenvolvido.

Capítulo 5: discute-se a aplicação do método e apresentação dos resultados.

Por fim, são apresentadas as conclusões finais e trabalhos futuros, seguidas pela lista de referências bibliográficas utilizadas nesta pesquisa.

2 PLÁGIO NO MEIO ACADÊMICO

Conforme previsto pelo Artigo 184 do Código Penal a infração dos direitos autorais é crime e a punição prevista para aquele que infringi-la varia do pagamento de multa até reclusão de quatro anos, dependendo da forma como o direito do autor for ferido. Ou seja, o ato de plagiar, mesmo de forma aparentemente inocente, pode ser considerado como infração penal e ser passível de punição (LEGISLAÇÃO, 2010).

Com o advento da Internet e do acesso quase irrestrito a bancos de dados dos mais variados assuntos, o plágio está se tornando um grande problema, principalmente no meio acadêmico (SILVA e DOMINGUES, 2008). Esta grande quantidade de informações digitais facilita cada vez mais aos plagiadores realizarem a cópia de obras alheias, às vezes na íntegra, outras apenas de partes.

Conforme pesquisa realizada por Silva e Domingues (2008) onde alunos de Pós-Graduação *Lato sensu* responderam diversos questionários para análise de seu conhecimento sobre o a prática de plágio, continha uma questão para os alunos assinalarem quantas alternativas fossem necessárias para que fossem identificadas as fontes mais utilizadas por eles em pesquisas acadêmicas. O resultado mostrou que 85,1% dos estudantes utilizam a Internet, por meio de sites de busca, para a realização de pesquisas (Figura 1).

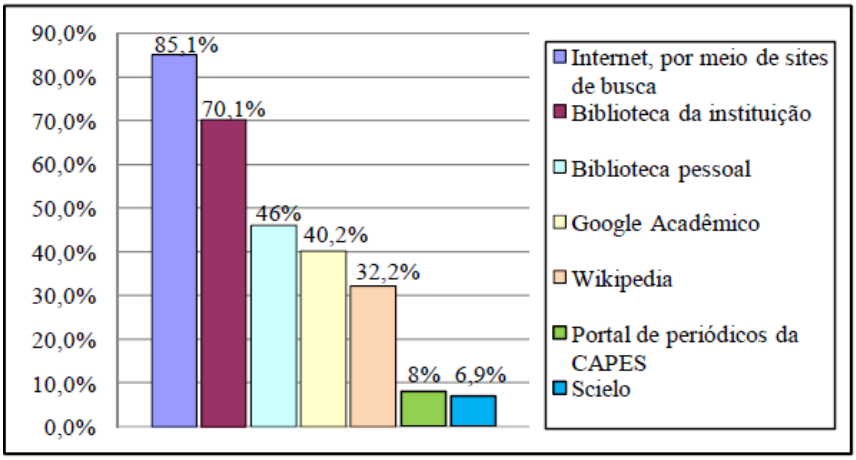

Figura 1: Fontes de pesquisas mais utilizadas pelos alunos
Fonte: (SILVA e DOMINGUES, 2008)**.**

Porém, a Internet é apenas uma ferramenta que facilita esta tarefa, mas o plágio não acontece em virtude do acesso à rede, mas segundo Moraes (2004), por falta de ética das pessoas.

> A Internet, sem dúvida, potencializa a incidência do plágio. Contudo, é preciso advertir: a proliferação da desonestidade intelectual nas universidades brasileiras não é culpa da Internet, poderosíssima máquina facilitadora da cópia. Culpá-la é interpretar estreitamente o problema. O responsável por essa grave crise ética é, obviamente, o próprio ser humano. Não pode a rede mundial de computadores ser tachada como vilã, até porque ela configura importante instrumento de pesquisa acadêmica e tende a ser cada vez mais valorizada na Sociedade da Informação em que vivemos (MORAES, 2004, p. 98).

O ato de plagiar excede à simples cópia de informações, por isso, buscaram-se na literatura vários conceitos citados por diferentes autores do que se pode considerar como sendo um ato de plágio.

Segundo Moraes (2004), o plágio pode ser considerado como a imitação de uma obra, considerada pela lei autoral como um verdadeiro atentado aos direitos morais do autor. O plagiador (ou plagiário) costuma não confessar o ato. Seja movido por inveja, seja por mera preguiça, o plagiário escamoteia e mente, desmoralizando o verdadeiro criador intelectual. Essa conduta é típica de nossa sociedade de aparência, na qual o importante não é ser, mas simplesmente parecer e aparecer. Moraes (2004) também defini o plágio como sendo quase sempre de parte(s) de obra alheia, e não de sua íntegra, visto que a prova judicial de obra completamente igual à outra consiste em tarefa que, muitas vezes, não exige maiores esforços.

Já Stein (2006), considera que o plágio é o ato de apresentar uma obra de qualquer espécie sem dar crédito aos seus verdadeiros autores, além de ser uma das formas mais graves de má conduta acadêmica.

Conforme Fonseca (*apud* Silva F., 2008, p.3):

> O plágio se caracteriza com a apropriação ou expropriação de direitos intelectuais. O termo “plágio” vem do latim “plagiarius”, um abdutor de “plagiare”, ou seja, “roubar” [...]. A expropriação do texto de um outro autor e a apresentação desse texto como sendo de cunho próprio caracterizam um plágio e, segundo a Lei de Direitos Autorais, 9.610, de 19 de fevereiro de 1998, é considerada violação grave à propriedade intelectual e aos direitos autorais, além de agredir frontalmente a ética e ofender a moral acadêmica.

Conforme descrito pela Handbook (2009), o plágio pode ser definido como a utilização das palavras ou idéias de uma outra pessoa como se fosse seu próprio trabalho, tem-se como exemplos, copiar, traduzir um texto de um idioma para outro, ou parafraser. Alguns exemplos que podem ser considerados como plágios são citados a seguir (HANDBOOK, 2009; OLIVEIRA e OLIVEIRA, 2008):

- **Citação:** É uma cópia palavra por palavra do que alguém disse ou escreveu. O uso de qualquer citação(s) de obras publicada de outras pessoas quer sejam publicadas em livros, artigos, na *Web* ou em qualquer outro formato, que não estejam claramente identificadas.
- **Paráfrase:** É utilização de palavras ou idéias de outra pessoa com poucas alterações ou parafraseada para torná-la diferente do original.
- **Resumo:** é feito, assim como a paráfrase, com as próprias palavras, porém, um resumo é consideravelmente mais curto e não segue a fonte ao pé da letra como a paráfrase, mas sem referênciar o verdadeiro autor no texto e nem na bibliografia.
- **Referência:** é o plágio em que há referência incompleta à obra original ou referência a um plágio (citação, paráfrase ou resumo). Um exemplo de plágio por referência é indicar em um texto como referências bibliográficas citações, paráfrases ou resumos em vez de referenciar os textos originais e seus verdadeiros autores.

Considerando as definições de plágio apresentadas, Kirkpatrick (2007 *apud* Oliveira; Oliveira, 2008) identifica a existência dos seguintes tipos de plágio:

- **Plágio Direto**: Consiste em copiar uma fonte palavra por palavra sem indicar que é uma citação e sem fazer referência ao verdadeiro autor.
- **Referência Vaga ou Incorreta**: Ocorre quando o escritor não indica onde começa e termina uma citação, ou seja, o escritor deve referenciar de maneira correta todos os trechos retirados da bibliografia de outro autor. Algumas vezes, um escritor faz referência a uma fonte uma vez, e o leitor presume que as sentenças anteriores ou parágrafos tenham sido parafraseados quando na verdade a maior parte do texto é uma paráfrase desta única fonte.
- **Plágio Mosaico:** neste tipo de plágio o escritor não faz uma cópia direta da fonte de outro autor, mas muda algumas palavras em cada sentença ou reformula um parágrafo, sem dar crédito ao autor original. Esses parágrafos ou sentenças não são citações, mas estão tão próximas de ser citações que eles deveriam ter sido citados ou, se eles foram modificados o bastante para serem classificados como paráfrases, deveria ter sido feito referência à fonte.
- **Plágio Bilíngue:** Carmo e Kennedy (2009) considera um ato de fazer passar o trabalho dos outros como o próprio a partir da tradução de uma obra para outro idioma sem dar a devida atribuição ao autor original.

2.1 Prática do Plágio

Segundo Liu *et al.* (2007), a prática de plágio pode ser aplicada de duas formas: intra-corpal, no qual um sujeito copia a tarefa de outro quando ambos estão realizando uma mesma tarefa, e extra-corpal, onde o sujeito copia de fontes externas, como, por exemplo, livro, artigo de revista, monografias ou internet.

No entanto, neste capítulo serão apresentados alguns relatos da prática de plágio vivenciados por professores de Universidades, e até mesmo por avaliadores de artigos em periódicos que se depararam com este tipo de problema.

Pode-se perceber que este não é um problema recente e nem só ocorre no Brasil, pois um estudo realizado por McCabe (2005) com mais de 80.000 alunos dos EUA e Canadá constatou que 36% dos estudantes de graduação e 24% dos estudantes de pós-graduação admitem ter copiado ou parafraseado frases da internet sem referênciá-las. Entre os diversos métodos de plágio que ocorrem com maior frequência na prática, Maurer, Kappe e Zaka (2006) menciona a tradução de conteúdo multilingue. Neste tipo de plágio o conteúdo de um documento em um idioma fonte é traduzido para outro idioma.

Garschagen (2006) relata em uma reportagem que "Plagiar nunca foi tão fácil e frequente nas universidades brasileiras, principalmente depois da popularização da internet. Os professores universitários são obrigados a duvidar de todos os trabalhos entregues pelos alunos". Segundo Silva F. (2008);

> Desse modo, na busca por caminhos mais fáceis e mais velozes, e tendo como aliada a natureza aparentemente pública do conteúdo *on-line*, além da disponibilidade/acessibilidade dos hipertextos digitais, na universidade essa prática tem-se dado de forma mais abrangente e acentuada, haja vista a velocidade na transmissão das informações – cruas ou refinadas – e a grande quantidade de textos/obras à disposição o leitor na internet: "Fica difícil não plagiar com tantas oportunidades" (GB), declara um graduando envolvido na pesquisa (SILVA F., 2008, p. 2).

Em uma pesquisa de campo na universidade pública do Estado da Bahia onde se realizou uma discussão sobre o plágio através de entrevista com 19 dos 20 alunos graduandos de Letras, apresentou que 36,84% assumem claramente já terem cometido plágio de textos; 21% plagiam, mas não assumem claramente e 41,1% dizem não ser a favor do plágio (SILVA F., 2008). Abaixo, segue como exemplo, a resposta de um dos alunos entrevistados em relação ao plágio.

> [...] Eu sou sincero. Plagiei semestre passado [...] eu sei que não é o caminho correto, mas desde que não seja prejudicial na minha construção do conhecimento. Aconteceu em uma disciplina que não considerava importante para mim, já que o curso de letras é muito abrangente e então sei o que é de meu interesse, o que acredito que seja de importância para mim e devo tentar aperfeiçoar-me; o que não era a disciplina na qual plagiei da *net* (JL) (SILVA F., 2008, p. 3).

Rabelo (2006) também relata em uma reportagem alguns exemplos da prática de plágio por alunos de graduação de uma universidade.

> Calouro de um curso da área de exatas, o estudante Marcos[1], 19 anos, mal ingressou na Universidade e já utilizou recursos inadequados para conseguir nota em uma disciplina. Ele confessa que, diante da falta de tempo, entregou ao professor um texto copiado da internet. O plágio em trabalhos acadêmicos não é novidade em instituições de ensino superior, públicas ou particulares. Na graduação – quando as exigências quanto a referências e citações são menores que em cursos de pós-graduação – não é difícil encontrar professores que tenham recebido trabalhos como o de Marcos (RABELO, 2006).

Apesar das estratégias aplicadas pelos professores para monitorar essa prática, os alunos não se intimidam e continuam praticando o ato de plagiar. Assim como Fabio[1], 20 anos, aluno do 3° semestre da área da Saúde, mostra que não falta habilidade para driblar as medidas adotadas pelos docentes.

> [...] Tive um professor que pedia trabalhos escritos à mão para evitar plágio. Mas não adiantou muito. Imprimi os textos da internet e copiei, conta Fabio[1]. [...] Você pega um texto de um amigo e muda algumas palavras[...] conta Marcos[1] (RABELO, 2006).

Recentemente, a IEEE (2010) relatou um ato de plágio praticado de forma descarada por um professor doutorando do *Institut Teknologi Bandung* (ITB), na Indonésia, onde o mesmo escreveu um artigo do seu trabalho de doutorado, o mesmo foi publicado na Conferência da IEEE e na Biblioteca Digital IEEE Xplore. Depois de uma acusação, a IEEE investigou e determinou que o artigo era uma duplicação quase completa do trabalho de um pesquisador austríaco que já havia sido publicado nos anais do 11 º Workshop Internacional. Com este ato, o doutorado do professor foi revogado e o mesmo inabilitado de publicar artigos na IEEE por três anos (IEEE, 2010).

---

[1] Nomes fictícios a pedido dos entrevistados.

2.2 Tecnologias existentes para Detecção de Indícios de Plágio

Na última década, sistemas de detecção de plágio surgiram para verificar o plágio em diferentes circunstâncias dependendo do tipo de documento eletrônico a ser avaliado. De acordo com Santana e Joberto (2003) os sistemas de detecção de plagio são divididos em duas categorias:

• **Sistemas de detecção de plágio por palavras:** Este tipo de detecção é realizado por cruzamento de palavras geralmente pré-estabelecido um tamanho mínimo de caracteres na palavra.

• **Sistema de detecção de plágio por sentenças**: Método no qual o cruzamento é feito por um conjunto de palavras, podendo ser ou não delimitada por sinal de pontuação. São mais sofisticados que o sistema de detecção de plágio por palavras.

Atualmente, existem diversas ferramentas de natureza privada ou de uso gratuito que se tornaram opções para muitos professores na verificação de plágio. No entanto, fez-se uma pesquisa das principais dessas ferramentas que tratam do plágio extra-corpal. Nesta pesquisa, buscou-se analisar três aspectos que no decorrer deste trabalho serão apresentados como principais características em sistemas de detecção de indícios de plágio, tais como, número de palavras que compõem um trecho/sentença analisada, cálculo de similaridade entre duas sentenças e o percentual de indícios de plágio aceitável.

2.2.1 Turnitin/Plagiarism.org/ iThenticate.com

Todos os três pertencem à mesma empresa (*iParadigms*), porém, possuem público-alvo diferente. É um sistema privado e totalmente *online* que rastreia a Internet identificando *sites* e bases de documentos que contenham trechos idênticos a um trabalho submetido para verificação de plágio e, ao final da análise, emite um relatório de originalidade desse trabalho. Atualmente, ele pode ser integrado aos CMS / LMS / AVA: *Bb, WebCT,* Moodle, *Angel Learning, Desire2Learn e Pearson* (TURNITIN, 2010).

Para detecção de plágio, primeiramente se faz uma impressão digital do documento submetido para análise, esta impressão é comparada com uma base de documentos e ao mesmo tempo é feita uma busca na Internet. Não se apresentam maiores detalhes do que

compõem a impressão digital do documento. São pesquisadas duas similaridades: palavras substituídas e frases adicionadas.

Para substituição de palavras, uma escala de 0 (sem similaridades) e 1 (cópia total) é utilizada. A distribuição adotada segue a Figura 2 (TURNITIN, 2010 *apud* MUSSINI, 2008)**.**

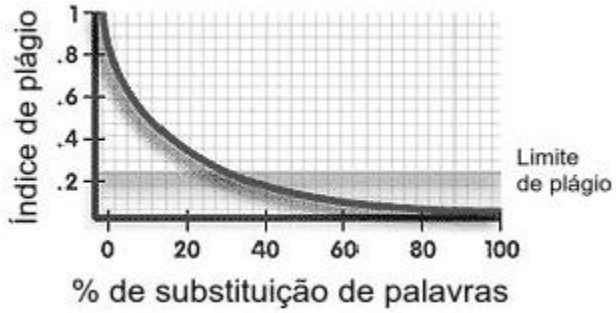

Figura 2: Gráfico de índice de plágio X porcentagem de palavras substituídas.
Fonte: (TURNITIN, 2010 *apud* MUSSINI, 2008)**.**

Na figura 2, a linha cinza é o limite para que um documento seja considerado cópia. Quanto maior a diferença entre os documentos comparados, menor o índice.

Conforme a Figura 3, apresenta-se casos onde se adicionam frases a um texto, onde o índice é 1, significa que nenhuma palavra foi adicionada, ou seja, a cópia é idêntica. Mesmo tendo todo outro documento adicionado, ainda é caracterizado um plágio, mesmo que em menor grau.

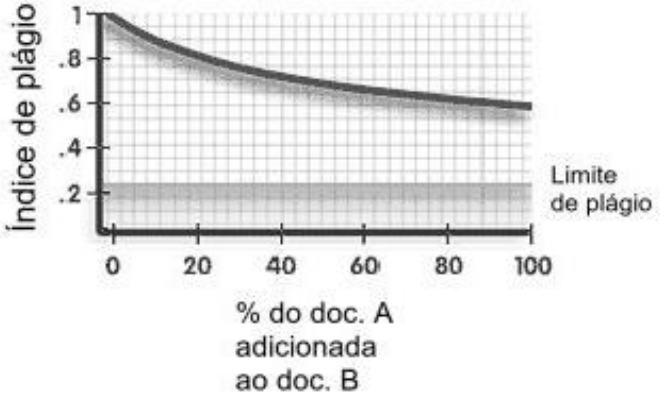

Figura 3: Gráfico do índice de plágio X porcentagem de cópia de duas fontes.
Fonte: (TURNITIN, 2010 *apud* MUSSINI, 2008)**.**

Não se encontrou referências sobre o tamanho da sentença/trecho que é enviada para análise e o percentual de indícios de plágio aceitável ou não aceitável. Além disso, esta ferramenta não disponibiliza de um tradutor para detecção de plágio bilíngue.

2.2.2 Ukund

Este sistema sueco é privado e faz a verificação de plágio com materiais disponíveis na internet e com uma base de dados do próprio Urkund. Esta base de dados é composta pelos próprios documentos submetidos pelos alunos. Este sistema pode ser integrado a diferentes LMS (*Learning Management System*), assim como, Fronter, Pingpong Blackboard och. Além disso, o trabalho deve ser enviado para ao e-mail do professor e antes de chegar à caixa de entrada de e-mail o documento passa pelos processos de verificação do sistema para então ser enviado ao e-mail do professor já com os resultados de indícios de plágio encontrados (URKUND, 2010). Não se obteve informações sobre técnicas utilizadas para o cálculo de similaridades. Além disso, não se encontrou referências sobre o tamanho da sentença/trecho do documento que é enviado para análise e o percentual de indícios de plágio aceitável ou não aceitável. Esta ferramenta também não disponibiliza de um tradutor para detecção de plágio bilíngue.

2.2.3 Ephorus

É uma ferramenta privada que oferece a integração de um campo para o envio de trabalhos dos alunos no site da própria instituição, ou seja, o aluno poderá submeter seus trabalhos digitalmente, e ao mesmo tempo ele estará enviando para o Ephorus (invisivelmente). Além disso, o Ephorus é facilmente integrado a diversos Ambientes Virtuais de Aprendizagem. Como exemplos pode-se citar: BlackBoard, Dokeos, Moodle, Fronter, IT's Learning, TeleTop e Workspaces (EPHORUS, 2010).

Não se obteve informações sobre técnicas utilizadas para o cálculo de similaridade. Além disso, não se encontrou referências sobre o tamanho da sentença/trecho do documento que é enviado para análise e o percentual de indícios de plágio aceitável ou não aceitável. Esta ferramenta também não disponibiliza de um tradutor para detecção de plágio bilíngue.

2.2.4 Seesources

É um sistema web gratuito e que não necessita de instalação, ao submeter o documento no *site* do sistema que o mesmo realiza a análise. Para verificação de indícios de plágio em um documento, o sistema extrai automaticamente assinaturas do texto original e realiza pesquisas na internet por referências similares, ordenando os resultados por relevância. Não se descreve com maiores detalhes do que compõem as assinaturas do texto. Esta ferramenta faz uso da *Google Search API* como mecanismo de busca para comparação do documento submetido (SEESOURCES, 2010).

Não descreve técnicas de similaridade utilizada e não permite a definição de nenhum parâmetro pelo usuário, ou seja, seus parâmetros são padrões do sistema. Ao final da análise do documento ele gera um relatório na própria interface web somente com as urls onde foram encontrados documentos similares (SEESOURCES, 2010). Esta ferramenta não esta mais em desenvolvimento, ou seja, parou na sua versão teste, e passou-se a desenvolver um sucessor, a ferramenta PlagScan mas que se transformou em uma ferramenta privada. A Figura 4 apresenta a interface do sistema Seesources.

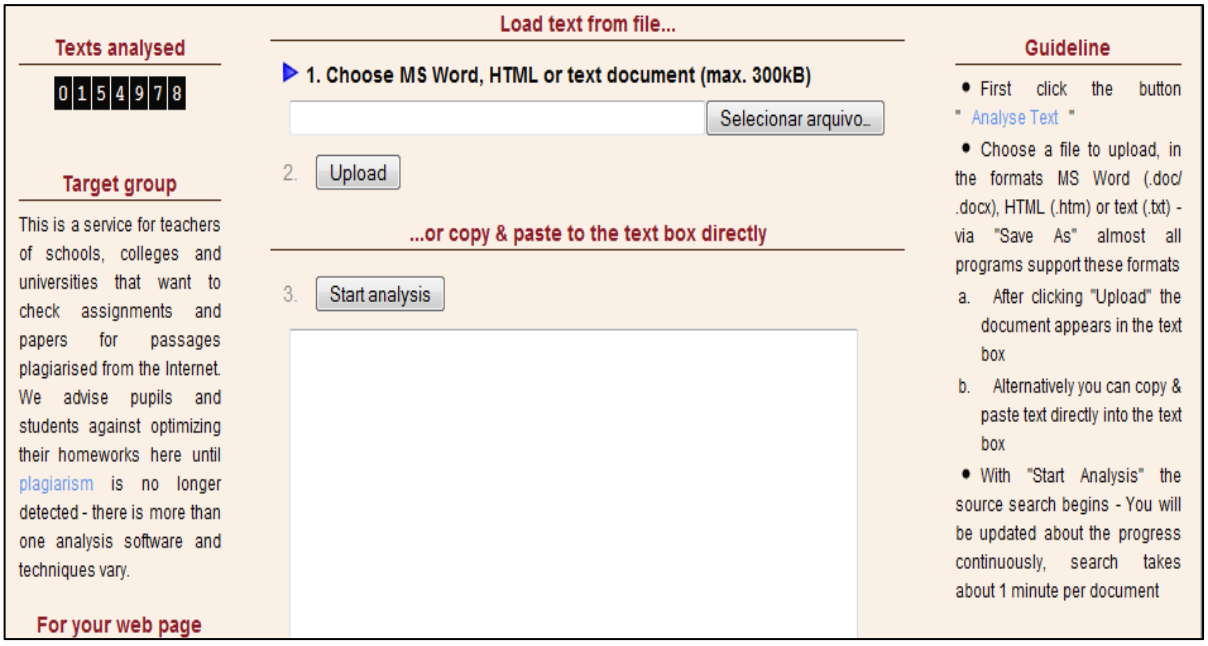

Figura 4: Tela do Sistema *Seesources*.

Além disso, não se encontrou referências sobre o tamanho da sentença/trecho do documento que é enviado para análise e o percentual de indícios de plágio aceitável ou não aceitável. Esta ferramenta também não disponibiliza de um tradutor para detecção de plágio bilíngue

2.2.5 Approbo

É uma aplicação para desktop gratuita que permite detectar plágios em trabalhos acadêmicos. O Approbo varre a *web* em busca de documentos similares, retornando um gráfico de semelhança. A ferramenta utiliza mecanismos de busca para encontrar os pontos de coincidência com o texto original. A partir desse ponto, o Approbo verifica palavra por palavra de todo o texto e mostra-o de forma gráfica, simplificando o trabalho ao utilizador (APPROBO, 2010). Não se obteve maiores informações sobre esta ferramenta pelo site da empresa desenvolvedora se apresentar fora do ar.

2.2.6 Viper

É uma ferramenta desktop gratuita que tem como função vasculhar a internet à procura de palavras-chave que sejam coincidentes com os textos apresentados como amostra. Após terminar esta busca, o Viper lista os resultados em uma tabela, com links de páginas que tenham sido detectadas como possíveis originários do texto (VIPER, 2010).

Este sistema faz uso de um percentual de plágio padrão do sistema, a qual não é descrita pelo autor. Realiza uma varredura no texto enviando sentenças para análise na *Web*. A cada sentença analisada são apresentadas as urls juntamente com o percentual de indícios encontrados (VIPER, 2010). A Figura 5 apresenta a interface da ferramenta Viper.

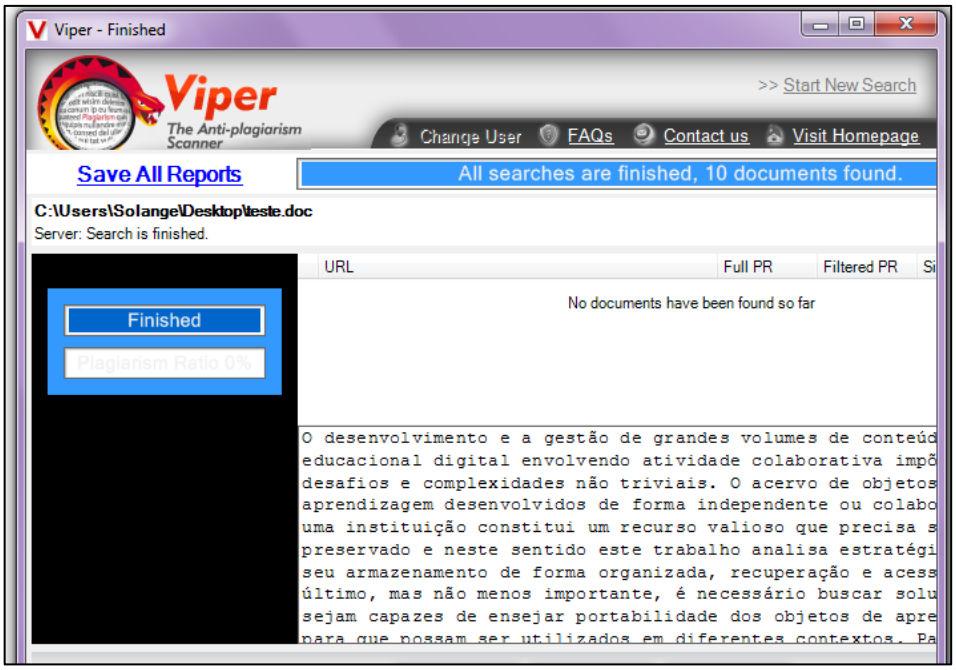

Figura 5: Interface da ferramenta Viper.

Não se obteve informações sobre técnicas utilizadas para o cálculo de similaridade e sobre o tamanho da sentença/trecho do documento que é enviado para análise. Além disso, esta ferramenta também não disponibiliza de um tradutor para detecção de plágio bilíngue.

2.2.7 Plagium

É um sistema *web* gratuito que esta em seu estágio beta para busca de textos por similaridade. Seu funcionamento é semelhante a um mecanismo de busca, pois permite somente que o usuário informe um trecho do texto ou um endereço (url) para análise, mas não permite a submissão de um documento em arquivo. Este sistema faz uso da *Yahoo API* para as buscas na *web* (PLAGIUM, 2010). A Figura 6 apresenta a interface do sistema Plagium.

Figura 6: Tela sistema Plagium.

Não se obteve informações sobre técnicas utilizadas para o cálculo de similaridade. Além disso, não se encontrou referências sobre o tamanho da sentença/trecho do documento que é enviado para análise e o percentual de indícios de plágio aceitável ou não aceitável. Esta ferramenta também não disponibiliza de um tradutor para detecção de plágio bilíngue.

2.2.8 Farejador de Plágios

É um sistema privado. Para verificação de indícios de plágio o sistema faz a varredura em todo o texto e a cada 10 palavras cria uma sentença, a qual é enviada para análise na *web*. Após analisar todo o texto os resultados são compilados e somente aqueles que tiveram ocorrências acima de quatro vezes são apresentados. O percentual de indícios de plágio aceitos não é relatado pelo autor (FAREJADOR, 2010). A Figura 7 apresenta a interface inicial da ferramenta farejador de plágio.

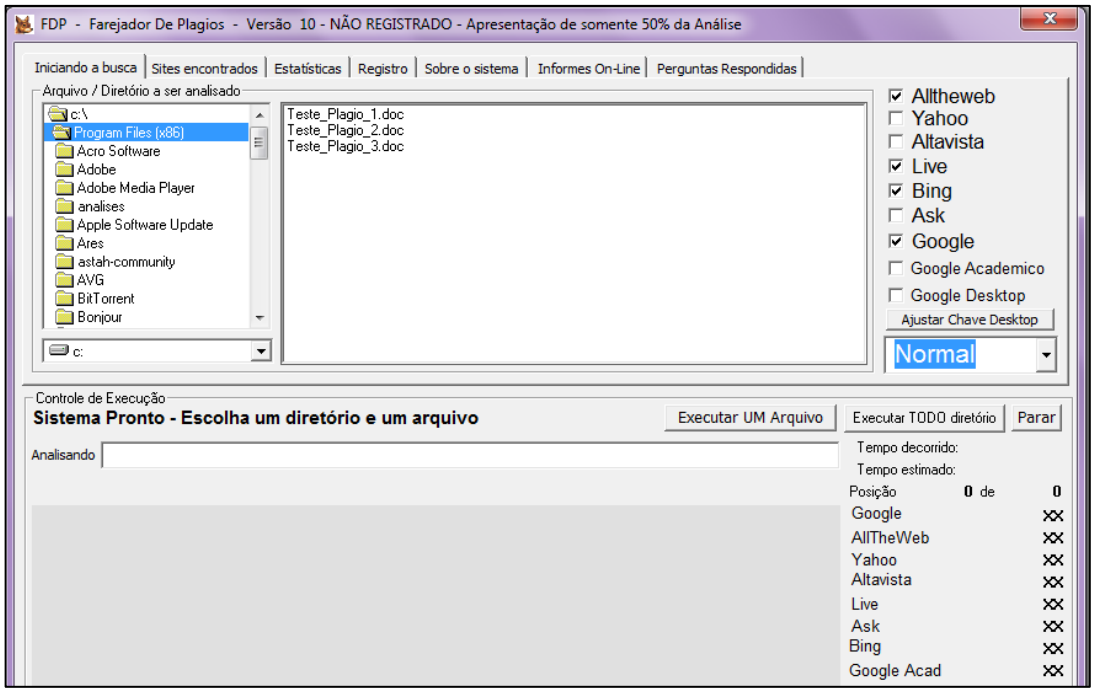

Figura 7: Interface da Ferramenta Farejador de Plágio.

O sistema segue as seguintes etapas para verificação de indícios de plágio em um documento:

- **Arquivo:** o usuário pode submeter um documento no formato .doc ao sistema.
- **Escolha dos buscadores:** permite a escolhe de um ou mais mecanismos de busca, tais como, yahoo, google, altavista.
- **Execução:** durante a execução todo o arquivo é pesquisado em trechos de 4 a 10 palavras e saltados. A quantidade de palavras de cada busca é definida pelo usuário, quando o mesmo escolhe nas configurações do programa entre rápida, normal, detalhada e rigorosa. Para cada configuração, uma forma de pesquisa é definida (PEZZIN, 2010):

Tabela 1: Quantidade de Palavras de cada busca.

	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 palavras
Rigorosa	4 a 5 palavras	4 a 6 palavras

Fonte: (PEZZIN, 2010).

O autor não descreve maiores detalhes sobre a escolhe de se analisar o texto seguindo estes parametros.

- **Organização dos resultados:** ao final da execução os sites encontrados são organizados e aqueles sites que extrapolarem o considerado “aceitável” serão apresentados

como plágio, os demais também são apresentados, mas nào são considerados, devido a pouca incidencia. Não é relatado qual o indice considerado aceitável pelo sistema.

- **Citação do arquivo analisado:** é gerado um relatório com o texto original e as referências encontradas a cada trecho que foi analisado.

A principal desvantagem deste sistema é de utilizar 10 palavras para compor uma sentença/trecho, ou seja, segundo os resultados apresentados na Figura 13, este tamanho de sentença retorna um número muito alto de resultados irrelevantes, dificultando na verificação final do documento por parte do usuário. Além disso, são apresentados como resultados somente as referências que aparecerem acima de quatro vezes, ou seja, se o autor plagiar uma única vez da mesma referência, esta então não será apresentada.

O sistema também apresenta um custo em tempo de processamento considerado alto, pois analisa a cerca de 30 á 40 paginas por hora, ou seja, ao analisar 10 trabalhos de 50 páginas cada, a ferramenta levaria em torno de 13 horas para analisar todos.

Não se obteve informações sobre técnicas utilizadas para o cálculo de similaridade e a mesma não disponibiliza de um tradutor para detecção de plágio bilíngue.

2.2.9 Plagius Detector

É uma ferramenta privada que disponibiliza uma versão grátis para teste, esta versão permite a análise somente de 50% do documento. Além disso, o sistema possibilita a alteração de parâmetros pelo professor, assim como: o número mínimo e máximo de palavras que uma sentença poderá ter para ser enviada para análise; o número de ocorrências que devem ocorrer no texto para que possa ser retornada ao resultado final.

Ao iniciar a análise, são extraídas várias frases do arquivo, as quais são pesquisadas na internet verificando sua existência. Ao final da análise, são apresentados como resultado o percentual total das expressões/sentenças que foram consideradas como indícios de plágio e as referências (urls/links) onde foram encontrados indícios de plágio por frase pesquisada (PLAGIUS, 2010).

A Figura 8 apresenta a interface do sistema plagius detector de plágio.

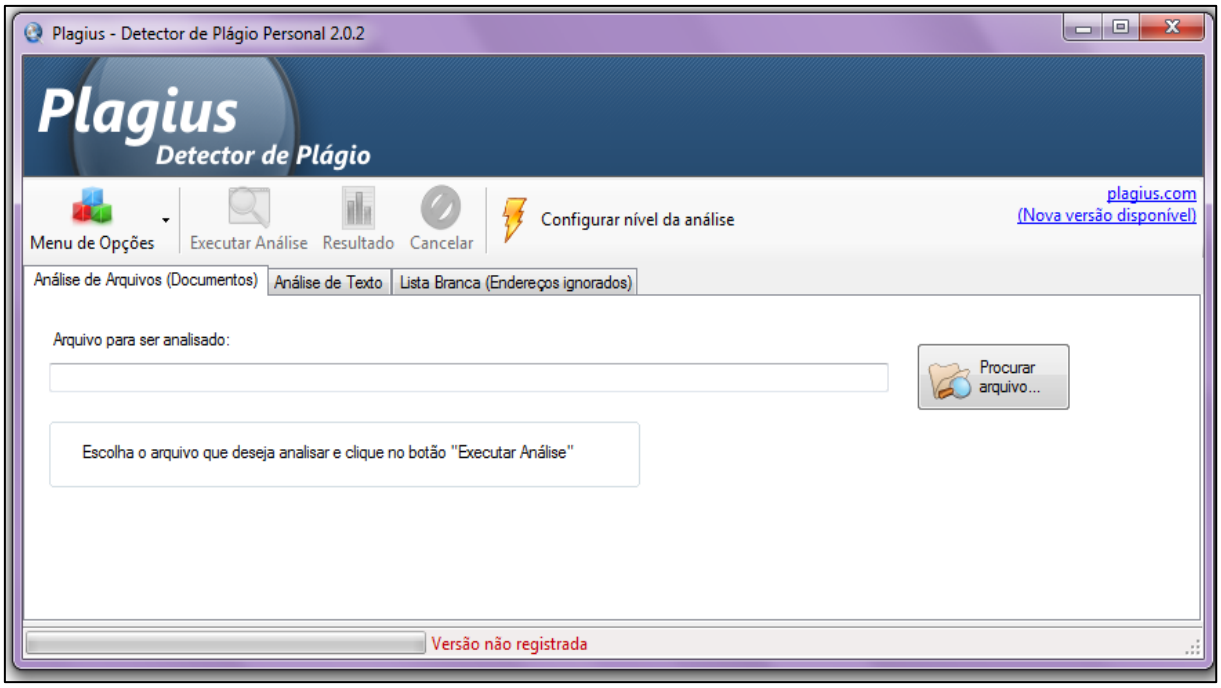

Figura 8: Interface do Plagius Detector.

A principal desvantagem deste sistema é que possibilita ao usuário a alteração do número de palavras que compõem uma sentença/trecho que será enviada para análise, um parâmentro que altera significativamente a precisão dos resultados retornados, ou seja, uma má escolha deste paramêtro poderá retornar um número indesejável de resultados irrelevantes podendo desta forma dificultar ainda mais o trabalho do professor ao analisar um documento.

Além disso, não se obteve informações sobre técnicas utilizadas para o cálculo de similaridade e ele não disponibiliza um tradutor para detecção de plágio bilíngue.

2.2.10 Plagiarism Finder

É uma ferramenta privada que faz uso de um percentual de plágio padrão do sistema, a qual não é descrita pelo autor. Permite ao usuário definir o valor máximo de palavras que uma sentença deve ter para ser analisada, mas limita este valor entre 1 e 8 termos. As referências encontradas são apresentadas a cada sentença analisada, e por fim um percentual de indícios de plágio encontrado no documento em geral (FINDER, 2010).

A principal desvantagem deste sistema é de permitir ao usuário a alteração do número de palavras que compõem uma sentença/trecho que será enviado para análise e limitar a alteração do número minimo de termos entre 1 e 8, ou seja, o usuário poderá optar em analisar

1 única palavra por vez, o que segundo os resultados apresentados na Figura 13, retornará um número muito alto de resultados irrelevantes, dificultando na verificação final do documento por parte do usuário. Além disso, o sistema apresenta as referências a cada sentença analisada, por exemplo, imagina-se que ao enviar 1 única palavra por vez para análise retorne uma url ou mais, pois cada trecho pode ter similaridades com mais de uma referência, o que retornará muitos resultados indesejáveis.

Não se obteve informações sobre técnicas utilizadas para cálculo de similaridades e não disponibilza de um tradutor para a verificação de plágio bilíngue.

2.2.11 Motores de Busca

Além das ferramentas descritas nas seções anteriores, muito dos professores também tem optado por motores de busca como auxílio nesta tarefa. Motores de busca são ferramentas da *web* que, utiliza-se de softwares ou programas de pesquisa para efetuarem os processos de busca, indexação e atualização de suas bases de dados, ou seja, são baseados no uso exclusivo de programas de computador para a indexação das páginas da web. A seguir são descritos alguns dos motores de busca mais utilizados atualmente.

Google: Atualmente, o maior motor de busca no mundo. Indexa 1,5 bilhões de páginas. Para fazer uma consulta no Google, basta digitar algumas palavras descritivas e pressionar a tecla “enter” (ou clicar no botão Pesquisa Google) para a sua lista de resultados relevantes.

O Google usa técnicas sofisticadas de identificação exata de textos para encontrar páginas que sejam tanto importantes como relevantes para a sua busca. Por exemplo, quando o Google analisa uma página, ele olha para o que as outras páginas que estão vinculadas à mesma têm a dizer sobre ela. O Google também prefere as páginas nas quais os termos de busca estejam próximos entre si.

Este mecanismo permite a busca por termos entre aspas, retornando somente páginas que incluam todos os seus termos de busca e ignora palavras e caracteres comuns (GOOGLE, 2010).

Yahoo: é um mecanismo que varre a *web* em busca de sites e os ordena segundo critérios de popularidade e relevância (YAHOO, 2010). O site de busca do Yahoo (http://br.docs.yahoo.com/searchtour.html) permite ao usuário aplicar filtros que devolvem

somente resultados mais similares ao que ele busca, assim como, ao fazer a pesquisa por uma palavra, o mecanismo de busca faz uma varredura e devolve como filtros em um menu os nomes de bibliotecas que possuem informações sobre o assunto, devolve complementos para a palavra, por exemplo, ao digitar testes ele retorna testes de gravidez, testes de conexão, entre outros, ajudando desta forma o usuário em sua pesquisa.

Altavista Brasil: Portal de buscas americano, oitavo colocado no *ranking* global da Média Metrix, que entrou no mercado brasileiro em dezembro de 2001. Em sua 8ª versão original, incluía 8 milhões de páginas da web brasileira, oferecia opções de pesquisa avançada, busca multimídia e tradução online. Na seção "Minha Pesquisa", os usuários podem armazenar até 25 buscas. Também permite personalizar os resultados das pesquisas (ALTAVISTA, 2010).

A Tabela 2 apresenta de forma resumida as principais características de cada uma das ferramentas descritas anteriormente.

Tabela 2: Características das Ferramentas de Detecção de Indícios de Plágio

Ferramentas	Source code	Web-based	Desktop	Public	Private	Tradutor	Min de Termos	Max de Termos	Índice Aceitável
Turnitin/Plagiarism.org/ iThenticate.com	x	٧			٧	Ø	x	x	x
Urkund	x	٧			٧	Ø	Ø	Ø	Ø
Ephorus	x	٧			٧	Ø	x	x	x
Seesources	x	٧		٧		Ø	x	x	x
Approbo	x		٧	٧		Ø	Ø	Ø	Ø
Viper	x		٧	٧		Ø	x	x	x
Plagium	x	٧		٧		Ø	x	x	x
Farejador de Plágios	x		٧		٧	Ø	x	x	x
Plagiarism Finder	x		٧		٧	Ø	٧	٧	x
Plagius	x		٧		٧	Ø	٧	٧	٧

Legenda: ٧Recurso disponível x Recurso não disponível Ø Sem informação

Pode-se perceber que na tabela apresentada acima, nenhuma das ferramentas analisadas apresenta um tradutor que verique indícios de plágio bilíngue. Além disso, as ferramentas Plagius e Plagiarism Finder apresentam uma desvantagem que é permitir ao usuário a alteração do número de palavras que compõem uma sentença/trecho que será enviado para análise, um parâmentro que afeta signitivamente na precisão dos resultados

retornados, ou seja, uma má escolha deste paramêtro poderá retornar um número indesejável de resultados irrelevantes podendo desta forma dificultar ainda mais o trabalho do usuário ao analisar um documento.

2.3 Trabalhos Correlatos

Este capítulo é dedicado à apresentação de algumas pesquisas acadêmicas tanto nacionais como internacionais no sentido de desenvolver técnicas para detecção de indícios de plágios, inclusive a integração dos mesmos em ambientes virtuais de aprendizagem.

Algumas dessas pesquisas podem ser encontradas em Franco e Milanez (2008), que propõem a integração do software Sherlock ao ambiente virtual de aprendizagem TelEduc para detecção de plágio em questões dissertativas realizadas dentro do ambiente. Os resultados encontrados foram entre 3% e 12% de similaridade, o que indica uma baixíssima incidência de plágio. Uma desvantagem deste trabalho, é que o método analisa somente o plágio intra-corpal, ou seja, não verifica indícios de plágio com documentos disponíveis na web, pois segundo os autores Silva e Domingues (2008) a maior fonte de pesquisa de alunos de pós-graduação é a internet. Além disso, o método não detecta plágio bilíngue.

No trabalho de Oliveira; Oliveira (2008) é proposta uma metodologia baseada em um modelo de representação vetorial, onde os pesos dos termos foram representados pelas frequências com que eles ocorrem nos documentos a que pertencem. Os resultados dos experimentos obtiveram índices de similaridades entre documentos variando de 60% a 98%. Esta metodologia não foi aplicada a nenhuma ferramenta para uso da comunidade e não realiza a análise de similaridade diretamente com documentos da *web*, ou seja, para realização dos experimentos foram selecionadas páginas retornadas pelo Google cujos conteúdos fossem relacionados aos temas dos trabalhos em análise. Além disso, o método não detecta plágio bilíngue.

Assim como Oliveira *et al.*, (2006), os autores Oliveira *et al.* (2007) adotaram um modelo vetorial como estratégia de representação dos documentos, onde os pesos dos termos foram representados pelas frequências com que eles ocorrem nos documentos a que pertencem e foram consideradas apenas as palavras que tiverem pesos maior que 50% do maior peso de um termo, ou seja, se o termo de maior peso for 3, somente serão considerados os termos que tiverem peso acima de 3/2 = 1,5. Para realização dos experimentos foram

criados duas coleções de documentos com textos retirados de livros, uma coleção de documentos originais e a outra de documentos considerados plagiados. Os resultados dos experimentos realizados sobre documentos publicamente considerados como sendo frutos de plágio indicaram índices de 69.60% a 95% de similaridade com os documentos originais. No entanto, o método não se aplicou em análise com documentos digitais e é um protótipo, ou seja, são resultados iniciais de um projeto que seria desenvolvido para prover a Biblioteca Digital de Monografias do Departamento de Ciências da Informação da Universidade Federal do Espirito Santo. Além disso, não verifica indícios de plágio bilíngue.

A proposta de Neill e Shanmuganathan (2004) foi desenvolver uma ferramenta em java que faz a busca de documentos similares na *web* através do *Google Web APIs – SOAP* (*Simple Object Access Protocol*). O autor não descreve a técnica e algoritmos utilizados por motivos de segurança, ou seja, para que os estudantes não tentem derrotar a ferramenta. Além disso, a proposta não descreve experimentos e resultados encontrados e não trabalha com plágio bilíngue.

Butakov e Scherbinin (2009) desenvolveu um algoritmo para análise de similaridade entre documentos locais (documentos do banco de dados) e globais (utiliza *Microsoft Live Search*). Para análise de dois documentos do banco de dados ele: Divide o documento em sentenças e remove caracteres sem sentido (caracteres de pontuação, espaços, etc.), para então calcular as impressões digitais dos documentos apresentados e armazenar os resultados em seu banco de dados. A comparação dos documentos é realizada a partir das impressões digitais. Os resultados dos experimentos em busca local variaram de 17% para 29% na detecção de plágio e a seção de plágio transversal variou de 38% a 100%.

Já para verificação com referências da web, faz-se os dois primeiro procedimentos e em seguida o sistema realiza buscas na *web* de pequenas frases do texto, calcula as impressões digitais de documentos a partir dos resultados da busca e, em seguida, faz-se a comparação do documento submetido com as impressões digitais dos documentos da *web*. O resultado dos experimentos de buscas globais de uma coleção de 155 documentos variou em 10% para 20% no nível de detecção de plágio, um baixo percentual de indícos. O autor não descreve de que é composta a impressão digital dos documentos e não apresenta informações sobre o percentual de indícios de plágio utilizado e a quantidade de termos que compõem cada sentença analisada. Além disso, não verifica indícios de plágio bilíngue.

A dissertação de mestrado de Mussini (2008) apresenta o *PeerDetect*, um sistema de detecção de plágio elaborado sobre uma infraestrutura P2P. A solução proposta apresenta duas abordagens. A primeira permite ao usuário detectar plágios em nós que pertencem a uma

rede P2P, procurando por documentos nesta rede, o que inclui documentos que podem não estar disponíveis na Internet. A segunda abordagem também faz uso de redes P2P para distribuir o trabalho da busca entre os nós participantes, mas é similar aos sistemas de detecção convencionais, os quais detectam plágios na Internet. O autor não apresenta informações sobre o percentual de indícios de plágio utilizado e a quantidade de termos que compõem cada sentença analisada. Além disso, não verifica indícios de plágio bilíngue.

A solução proposta por Stein e Eissen (2006) calcula um código *hash* para cada fragmento (chamado de *fingerprint*). Fragmentos idênticos terão a mesma *fingerprint*. A idéia por trás da proposta *fuzzy-fingerprints* está em gerar o mesmo *hash* para fragmentos semelhantes. O método desenvolvido neste trabalho verifica documentos no idioma inglês, mas não identifica indícios de plágio bilíngue.

No trabalho de Zou *et al.* (2010) é proposto um método que consiste em três etapas: pré-seleção, onde a tarefa é descobrir para cada documento suspeito uma pequena lista de documentos candidatos; a localização onde o autor faz uso do método de agrupamento para localizar fragmentos plagiados entre os documentos suspeitos e cada documento original, e a ultíma etapa é o pós-processamento, o qual descarta alguns fragmentos sem plágio do resultado final.

Recentemente, algumas pesquisas realizadas por Pereira (2010); Potthast *et al.* (2010), entre outros autores, vem sendo realizadas na área de detecção de plágio multilíngue, onde os documentos originais são traduzidos para diferentes idiomas, a fim de verificar se trechos de um determinado documento suspeito é uma tradução de um documento em outro idioma sem fazer referência à obra original. Neste caso, tais métodos não identificam plágios entre documentos que estejam no mesmo idioma, ou seja, ao submeter um texto em português, o mesmo é traduzido para outros idiomas e analisado a similaridade com documentos do idioma traduzido.

A Tabela 3 apresenta de forma resumida as principais características de cada uma das ferramentas descritas anteriormente.

Tabela 3: Características dos trabalhos correlatos

Trabalhos Correlatos	Tradutor	Min de Termos	Max de Termos	Índice Aceitável	Custo de Tempo
Franco e Milanez (2008)	x	Ø	Ø	Ø	x
Oliveira; Oliveira (2008)	x	Ø	Ø	Ø	x
Oliveira *et al.* (2007)	x	Ø	Ø	Ø	x

Neill e Shanmuganathan (2004)	x	Ø	Ø	Ø	x
Stein e Eissen (2006)	x	Ø	Ø	Ø	x
Zou *et al.* (2010)	x	Ø	Ø	Ø	x
Pereira (2010); Potthast *et al.* (2010)	√	Ø	Ø	Ø	x

Legenda: √Recurso disponível x Recurso não disponível Ø Sem informação

Conforme apresentado na Tabela 3, alguns dos principais trabalhos relacionados a esta pesquisa e as ferramentas descritas na seção 2.2, nota-se que tais métodos/ferramentas apresentam características semelhantes, mas ainda requerem melhorias em alguns termos, assim como: nenhuma das pesquisas apresentou resultados em termos de desempenho no custo de tempo na fase de análise de similaridade entre dois documentos; não utilizam mecanismos de tradução para detecção de plágio bilíngue; nenhuma das metodologias apresentadas relata resultados de testes que determinem a utilização do número de termos que representam uma sentença/trecho e o percentual de similaridade para obter melhores resultados. Além disso, as pesquisas aqui apresentadas são metodologias propostas, mas não aplicadas a ferramentas para uso da comunidade cientifica.

No entanto, busca-se neste trabalho desenvolver um método com base nos apresentados, mas complementando-os com as características que estes ainda não possuem e que seja capaz de retornar resultados precisos na fase de verificação de indícios de plágio de um documento suspeito em um tempo aceitável.

3 AMBIENTES VIRTUAIS DE APRENDIZAGEM

Ambientes Virtuais de Aprendizagem (AVAs) possuem interfaces que permitem a produção de conteúdos e diferentes tipos de comunicação, além do gerenciamento de banco de dados e controle das informações circuladas dentro do ambiente (SANTOS, 2003).

Segundo Almeida (2002);

> Ambientes digitais de aprendizagem são sistemas computacionais disponíveis na internet, destinados ao suporte de atividades mediadas pelas tecnologias de informação e comunicação. Permitem integrar múltiplas mídias e recursos, apresentar informações de maneira organizada, desenvolver interações entre pessoas e objetos de conhecimento, elaborar e socializar produções tendo em vista atingir determinados objetivos (ALMEIDA, 2002, p. 4).

As características dos ambientes virtuais vêm permitindo que um grande número de pessoas dispersas pelo mundo possa interagir em tempos e espaços variados. No entanto, alguns AVAs ainda assumem que tentam simular as clássicas práticas presenciais, utilizando-se de práticas utilizadas em experiências tradicionais de aprendizagem (ALMEIDA, 2002).

Os AVAs podem ser empregados como suporte para sistemas de educação a distância, bem como servir de apoio às atividades presenciais de sala de aula e ou diferentes ambientes por meio da internet ou intranet. Nesta seção serão apresentados os AVAs Moodle, Teleduc, AulaNet e Tidia – Ae.

TelEduc: O TelEduc é um ambiente de educação a distância que começou a ser desenvolvido em 1997, a partir de uma proposta de dissertação de mestrado do Instituto de Computação da Universidade Estadual de Campinas (UNICAMP). O desenvolvimento deste AVA foi realizado pelos pesquisadores do Instituto de Computação da Unicamp, junto com o Núcleo de Informática Aplicada à Educação (NIED) (RIBEIRO, REATEGUI e BOFF, 2007).

De acordo com Barbosa (2005, p.78 *apud* Ribeiro; Reategui e Boff, 2007) "Esse ambiente foi desenvolvido de forma participativa, ou seja, todas as suas ferramentas foram idealizadas, projetadas e depuradas segundo as necessidades relatadas por seus usuários".

AulaNet: O AulaNet é uma ferramenta de ensino a distância e um ambiente de software baseado na *web*, que foi desenvolvido inicialmente no Laboratório de Engenharia de Software do Departamento de Informática da PUC-Rio, e é constantemente atualizado pelo LES e pela EduWeb (www.eduweb.com.br) (AULANET, 2010).

O ambiente de criação e manutenção de cursos apoiados em tecnologias da internet pode ser utilizado tanto para ensino a distância como para complementação às atividades de educação presencial e formação de profissionais. Têm sido várias as entidades que recorrem ao Aulanet como ferramenta de apoio a ação de formação quer em entidades acadêmicas, quer em universos empresariais, sendo distribuído em países como Portugal, Brasil, Moçambique, Panamá, Canadá, Alemanha, África do Sul, entre outros (AULANET, 2010).

Tidia-Ae: O intuito do TIDIA-Ae é desenvolver um ambiente de colaboração e ferramentas de suporte e apoio ao ensino e aprendizagem com interações presenciais e à distância, síncronas e assíncronas. O sistema de aprendizado visa beneficiar instituições de ensino, empresas e fundações em suas atividades educacionais, possibilitando a expansão do alcance do aprendizado eletrônico.

As ferramentas desenvolvidas contemplam os três grandes grupos de ferramentas gerais de EaD, tais como, administração, coordenação e comunicação, além de ferramentas e conteúdos.

O projeto pertence ao programa geral do TIDIA (Tecnologia da Informação para o Desenvolvimento da Internet Avançada) financiado pela FAPESP. Sendo associado, ainda, ao IMS - *Global Learning Consortium e ao Sakai Foundation*, ambas são instituições internacionais que discutem de maneira colaborativa o uso da tecnologia e seus resultados nas atividades educacionais (TIDIA-AE, 2010).

Moodle: Este ambiente virtual de aprendizagem foi desenvolvido pelo australiano Martin Dougiamas, ele é aberto e gratuito permitindo que qualquer pessoa possa utilizá-lo e modificá-lo (RIBEIRO; MENDONÇA e MENDONÇA, 2007).

> O AVA Modular Object Oriented Distance Learning (Moodle) é uma plataforma, Open Source, ou seja, pode ser instalado, utilizado, modificado e mesmo distribuído. Seu desenvolvimento objetiva o gerenciamento de aprendizado e de trabalho colaborativo em ambiente virtual, permitindo a criação e administração de cursos on-line, grupos de trabalho e comunidades de aprendizagem Mendonça (RIBEIRO; MENDONÇA e MENDONÇA, 2007, p. 7).

Vem sendo utilizado por várias instituições no mundo, por ser um ambiente gratuito e de código fonte aberto. Com isso, possui uma grande quantidade de pessoas contribuindo para a correção dos erros e desenvolvimento de novas ferramentas, assim como a discussão sobre metodologias pedagógicas de usabilidade (RIBEIRO; MENDONÇA e MENDONÇA, 2007).

3.1 Ambientes Virtuais de Aprendizagem Móvel

Os ambientes virtuais de aprendizagem móveis apresentados nessa seção exploram elementos como consciência do contexto e da mobilidade do aprendiz (BARBOSA *et al.*, 2008). A seguir são apresentados os ambientes de aprendizagem móveis pesquisados neste trabalho e suas características:

CLUE *(Collaborative Learning support system with an Ubiquitous Environment):* é um sistema de compartilhamento de conhecimento e de colaboração em um contexto ubíquo controlado, voltado para o auxílio na aprendizagem da língua japonesa (OGATA e YANO, 2003).

CULE (*Context-Aware Ubiquitous Learning Environment for Peer-to-Peer Collaborative Learning*): um ambiente de aprendizagem ubíqua consciente do contexto. Ele provê serviços para acesso a conteúdo de forma adaptativa ao dispositivo, um sistema de anotações personalizadas a esse conteúdo e a formação de grupos virtuais, considerando o perfil, o contexto físico e virtual dos integrantes de um grupo (YANG, 2006).

LIP *(Learning in Process):* é um sistema cujo objetivo é prover consciência de contexto em um cenário de educação corporativa. O modelo de contexto usado em *LIP* tem como objetivo auxiliar na aprendizagem corporativa, mapeando as aplicações, tarefas e conteúdos em estudo pelo usuário. Com isso, baseado no perfil organizacional do usuário (como seu cargo, competências requeridas), o sistema tem como sugerir programas de aprendizagem mais eficientes, considerando seu contexto. A adaptação ao contexto se dá em função do dispositivo de acesso e do perfil do usuário, que integra o modelo de contexto (SCHMIDT, 2005).

GlobalEdu: uma infraestrutura para suporte a processos educacionais direcionado à educação ubíqua. O sistema é composto de módulos educacionais e de um agente pedagógico, que acompanha o aprendiz, assistindo o processo educacional, independente do dispositivo de acesso. Uma vez acessando a rede GlobalEdu, o aprendiz tem a sua disposição o agente pedagógico. Não existe a necessidade de um vínculo formal do aprendiz com um curso, por exemplo, para acessar as informações. As informações estão disponíveis no ambiente na forma de objetos de aprendizagem e elementos de contexto. O sistema sugere informações de contexto e conteúdos ao aprendiz, conforme a visibilidade determinada por ele (BARBOSA *et al.*, 2008).

Mle- Moodle (*Mobile Learning Engine Moodle*): é um ambiente de código-fonte aberto, totalmente gratuito e personalizável, vinculado ao AVA Moodle, às especificações podem ser adaptadas conforme necessário com WML, PHP e MySQL. O ambiente fornece simplesmente a interface para o dispositivo móvel (e não é em si um programa totalmente distinto), quaisquer alterações efetuadas ao Moodle são automaticamente convertidos para os dispositivos também (YINGLING, 2010).

É um sistema projetado para auxiliar o sistema *e-learning* através de dispositivos móveis, tais como, telefones, PDAs, smartphones para. O acesso ao Mle-Moodle é feito através do navegador de qualquer aparelho de celular, mas também pode utilizar o Mle-Cliente, que é um módulo especialmente desenvolvido para o processo de aprendizagem com dispositivos móveis.

No entanto, optou-se por utilizar o AVA Moodle por ser um ambiente gratuito e de código fonte aberto, possibilitando o desenvolvimento e integração de novas ferramentas. Além disso, disponibiliza o módulo Mle-Moodle, um módulo extensivo do Moodle utilizado para adaptação do ambiente para acesso via dispositivo móvel, aonde a ferramenta desenvolvida neste trabalho também será integrado, possibilitando os professores o acesso via dispositivo móvel.

4 DESENVOLVIMENTO E APLICAÇÃO DE UM MÉTODO PARA DETECÇÃO DE INDÍCIOS DE PLÁGIO

O desenvolvimento do método se deu principalmente pela necessidade de melhorar técnicas já propostas na literatura, tanto em termos de tempo de processamento na fase verificação/análise do documento como na precisão dos resultados retornados. No entanto, o método desenvolvido baseou-se em técnicas de processamento de linguagem natural e métodos de detecção de indícios de plágio já desenvolvidas na literatura, por autores como Zou, Long e Ling (2010), Stein e Eissen (2006), Mussini (2008), Butakov e Scherbinin (2009), Neill e Shanmuganathan (2004), Oliveira *et al*. (2007), Oliveira; Oliveira (2008), Franco e Milanez (2008), a qual buscou-se melhorá-las principalmente nestes dois aspectos.

O método proposto neste trabalho trata os seguintes tipos de plágio:

- **Plágio de forma extra-corpal:** onde o sujeito copia de fontes externas;
- **Plágio Mosaico:** onde o autor copia partes de uma obra trocando somente algumas palavras sem dar crédito ao autor da obra original e;
- **Plágio Bilíngue:** onde o conteúdo de um documento no idioma inglês é traduzido para o idioma português sem fazer referência à obra original, analisando a similaridade entre documentos suspeitos e documentos digitais disponíveis na internet.

Nesta seção será apresentada a metodologia do método desenvolvido, assim como, sua implementação sistema para integração ao ambiente virtual de aprendizagem Moodle, ao ambiente virtual de aprendizagem móvel Mle- Moodle e em um sistema desktop.

4.1 Metodologia

O trabalho proposto busca desenvolver um método capaz de detectar e recuperar documentos textuais da *web* que possuam uma relação de similaridade com um dado documento suspeito, visando contribuir para um reconhecimento mais eficaz de indícios de plágio e minimizar a sobrecarga do professor no acompanhamento de trabalhos acadêmicos.

A ferramenta foi desenvolvida na linguagem de programação java e php, utilizando-se do gerenciador de banco de dados PhpMyAdmin e das seguintes bibliotecas e APIs:

***iText*:** é uma biblioteca que permite criar e manipular documentos PDF, e que foi utilizada para a geração de relatórios.

***POI*:** utilizada para leitura e escrita de documentos da Microsoft, a qual foi utilizada para manipulação de documentos na extensão .doc.

***Translate API Google*:** com a utilização desta API os documentos são convertidos do idioma português para o inglês. Esta API permite a tradução em diferentes idiomas, mas neste trabalho fez uso somente do idioma inglês, por ser um dos idiomas mais utilizados no mundo.

***Google AJAX Search API*:** é uma API do google para buscas na *web*.

Com relação ao hardware, foram utilizados para esta pesquisa: Notebooks (Sistema Microsoft Windows 7 - Service Pack 3 - Intel (R) Core 2 Duo (R), 4Gb de Ram), iPhone 3G 16 GB e um servidor Windows Server Standard - Service Pack 2, Intel(R) Xeon, 8Gb.

4.2 Método de Verificação Desenvolvido

O método desenvolvido para a verificação de indícios de plágio trata de plágio mosaico, onde o escritor copia partes de uma referência alterando algumas palavras sem dar crédito à obra original; e o plágio bilíngue, onde o conteúdo de um documento é traduzido para outro idioma sem dar crédito à obra original. Para a verificação do plágio bilíngue, identifica-se se o conteúdo do documento original esta no idioma português, caso estiver, o mesmo é traduzido para o idioma inglês e enviado para análise, verificando se o conteúdo consta de uma tradução. Na detecção do plágio mosaico a técnica localiza somente as palavras idênticas entre o documento suspeito e o documento original, sem levar em consideração a ordenação das palavras. No entanto, não se implementou a detecção de similaridades entre uma determinada palavras com seu sinônimo e plural, alterações que são também utilizadas no ato de plagiar.

Para verificar a similaridade entre o documento suspeito com documentos disponíveis digitalmente na *web*, o conteúdo textual é dividido em sentenças e enviado para *API do Google* em buscas de referências similares. A comparação das sentenças entre o conteúdo dos dois documentos ignora a ordem dos termos.

Desta forma, a ferramenta aceita como entrada documentos textuais em diferentes idiomas e identifica se o mesmo possui ou não indícios de plágio.

A seguir descreve-se com mais detalhes as duas principais fases do método desenvolvido, onde há uma fase de pré-processamento e uma fase de verificação/análise de indícios de plágio de um determinado documento suspeito (Figura 9).

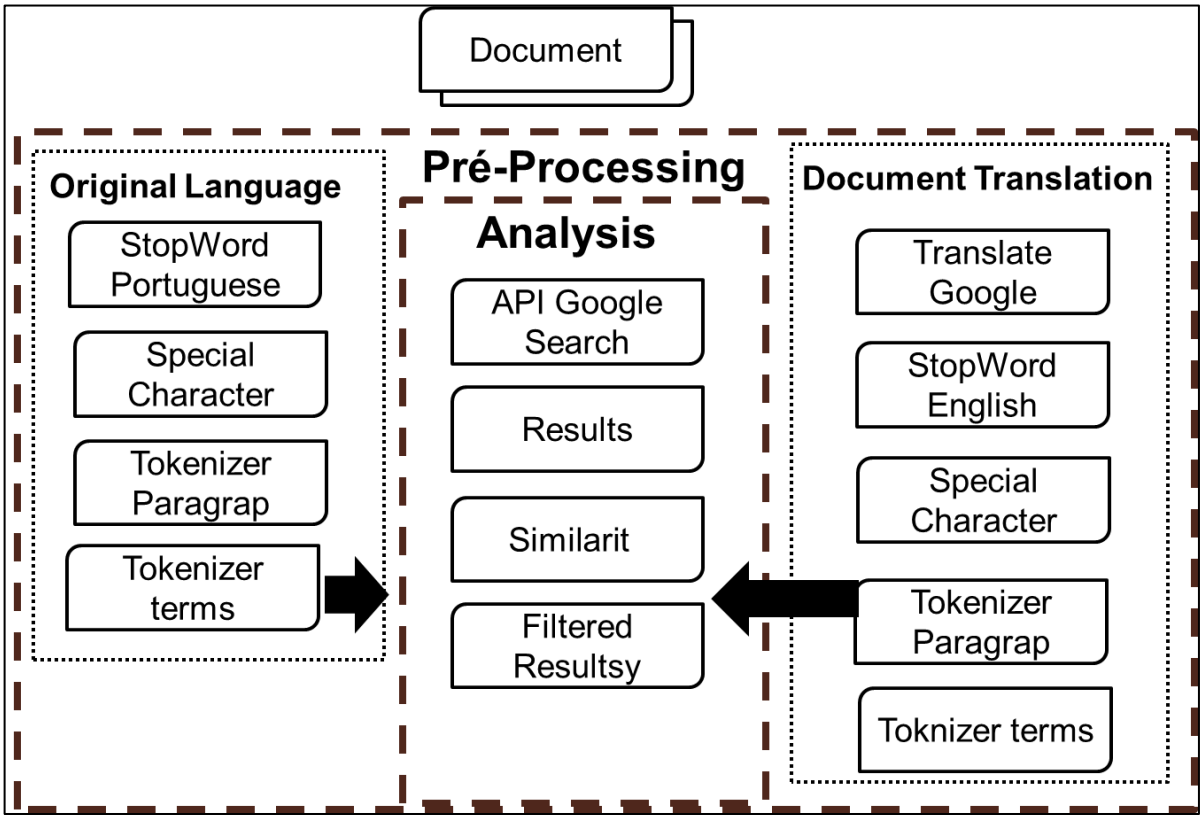

Figura 9: Arquitetura do Método Desenvolvido

4.3 Pré-Processamento

O principal objetivo desta fase é eliminar caracteres redundantes e palavras que não carregam nenhuma informação de maior relevância, assim como, *stopwords* e caracteres especiais, para reduzir desta forma o número de termos a serem analisados. As técnicas apresentadas a seguir se aplicaram no documento original (língua portuguesa) e no documento após a tradução do seu conteúdo para a língua inglesa. A seguir descreve-se com mais detalhes cada passo desta etapa.

4.3.1 Stopwords Portuguese/English

Nesta etapa adaptaram-se duas listas de *stopwords* de Dias (2004), uma lista com palavras em português e outra em inglês. A remoção das *stopwords* tem como objetivo eliminar palavras que não são representativas ao documento e consequentemente diminuir o número de palavras a serem analisadas. Pode-se considerar como *stopwords*: advérbios, artigos, conjunções, preposições e pronomes (DIAS, 2004). A Figura 10 mostra um trecho antes e depois da aplicação desta técnica.

Trecho Original
Com a internet, o aluno aumenta as conexões linguísticas, as geográficas e as interpessoais. As lingüísticas, porque interagem com inúmeros textos, imagens, narrativas, formas coloquiais e formas elaboradas, com textos populares.
Trecho sem Stopwords
internet, aluno aumenta conexões linguísticas, geográficas interpessoais. linguísticas, interagem inúmeros textos, imagens, narrativas, formas coloquiais formas elaboradas, textos populares.

Figura 10: Trecho sem *Stopwords*

A Figura 11 mostra um trecho traduzido do idioma original para o inglês e o trecho depois da aplicação da técnica com a lista de *stopwords* na lingua inglesa.

Trecho Original Traduzido
Because students often are confused about what is and is not plagiarism, I have prepared this handout to help you understand what is acceptable. There are some gray areas and if you have any questions, ask your instructor. Plagiarism is very serious and it can be grounds for failure in a
Trecho sem *Stopwords*
students often confused about plagiarism, prepared handout help you understand acceptable. There some gray areas you any questions, ask your instructor. Plagiarism serious grounds failure course.

Figura 11: Trecho sem *Stopwords English*

4.3.2 Special Character

Nesta etapa retiram-se os caracteres especiais, assim como, pontuação, espaços, etc. A Figura 12Figura 10 mostra um trecho antes e depois da aplicação desta técnica.

Trecho sem *Stopwords*
internet, aluno aumenta conexões linguísticas, geográficas interpessoais. lingüísticas, interagem inúmeros textos, imagens, narrativas, formas coloquiais formas elaboradas, textos populares.
Trecho sem Caracteres Especiais
internet aluno aumenta conexões lingüísticas geográficas interpessoais lingüísticas, interagem inúmeros textos imagens narrativas formas coloquiais formas elaboradas textos populares

Figura 12: Trecho sem Caracteres Especiais

4.3.3 *Tokenizer Paragrap*

Após a remoção das *stopwords* e caracteres especiais obtém-se um texto puro. Como as referências similares encontradas na análise são apresentadas por parágrafo, o documento textual então é desfragmentado em parágrafos para então ser aplicada a etapa de *Tokenizer Terms*.

4.3.4 Translate Google

Para a detecção de indícios de plágio bilíngue, os documentos suspeitos no idioma português são traduzidos para o idioma inglês e submetidos a pesquisas na *web* por documentos escritos na língua inglesa. Desta forma, se o usuário realizou um plágio a partir da simples tradução de um texto do inglês para o português o método será capaz de identificá-lo através desta etapa.

4.3.5 Tokenizer Terms

O sistema verifica indícios de plágio em documentos suspeitos por sentenças/trechos, ou seja, esta etapa é responsável pela divisão de cada parágrafo do documento em sentenças, para então serem enviadas para busca na *web* por referências similares. Para execução desta etapa, devem ser definidos dois parâmetros de entrada, que são: o número máximo e mínimo de termos/palavras que deverá compor uma sentença para que a mesma seja enviada para análise. Para verificar quais valores desses parâmetros produziram melhores resultados, foram efetuados testes com 15 documentos textuais, variando os valores dos parâmetros.

A Figura 13 apresenta a média dos testes considerando a variação dos parâmetros do número mínimo (MinT) e máximo (MaxT) de termos com os respectivos resultados relevantes e irrelevantes retornados por documento.

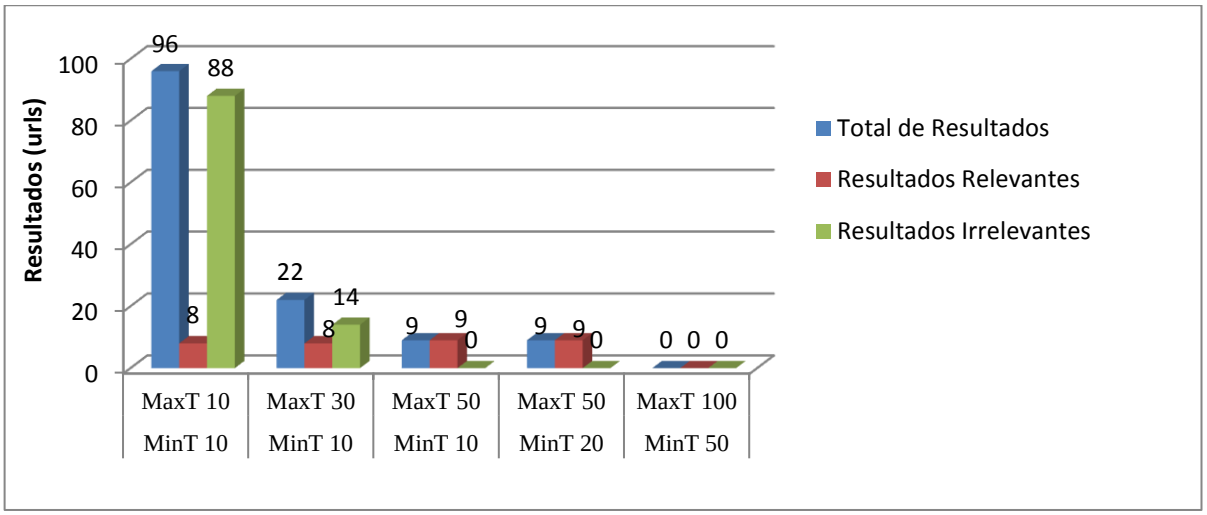

Figura 13: Análise das Sentenças

Conforme o gráfico apresentado, ao atribuir os valores MaxT= 10 e MinT=10, foram retornados 96 resultados, ou seja, o sistema localizou 96 referências com trechos similares ao documento suspeito, mas com estes valores como parâmetros o número de resultados irrelevantes atingiu 91,66%, um número inaceitável, pois dificultaria muito a análise final dos documentos por parte do usuário. Ao aumentar o valor do parâmetro MaxT= 30, pode-se perceber que o número de resultados irrelevantes diminui, apresentando um percentual de 63,63%, uma precisão ainda considerada muito baixa.

No entanto, observou-se que o parâmetro MaxT é que influência na precisão dos resultados, pois pode-se perceber que quando se utiliza um número muito baixo de termos, assim como MaxT=10 ou 30, são retornados um número elevado de resultados irrelevantes. Mas ao atribuir um valor mais alto, assim como MaxT= 50, atingiu-se uma precisão de 100%

de referências relevantes, uma precisão considerada perfeita. Já o parâmetro MinT, não tem muita influência nos resultados, mas quando se faz uso de um valor mais alto, assim como MinT= 50, onde somente serão analisadas sentenças que contém acima de 50 termos, as mesmas acabam não sendo analisadas, pois as vezes um parágrafo pode conter menos de 50 palavras.

A Figura 14 apresenta os testes considerando a variação dos parâmetros do número mínimo (MinT) e máximo (MaxT) de termos com seus respectivos tempos de processamento e na busca por documentos similares na *web*. Desta forma, pode-se observar que quanto maior a sentença, menor o tempo de processamento, e que quando se faz uso do parâmetro MinT=50 e MaxT=100, ou seja, quando a análise é feita em sentenças mais longas, o tempo de processamento na fase de análise do conteúdo é bem menor que os demais valores apresentados, mas segundo a Figura 13 a precisão dos resultados não é satisfatório.

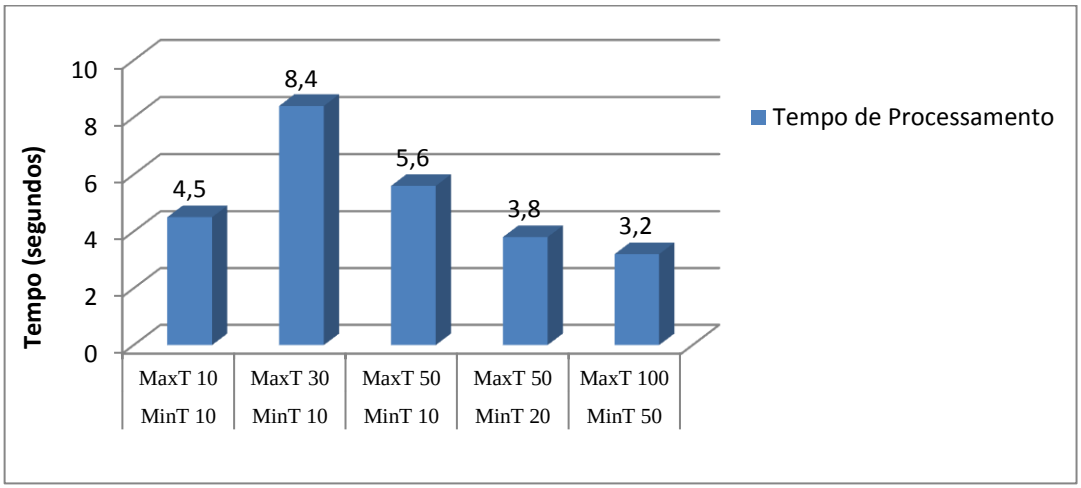

Figura 14: Tempo de Processamento

Conforme o gráfico apresentado, considera-se que ao atribuir os valores MaxT=100 e MinT=50 se obtém um melhor desempenho em termos de tempo de processamento, mas a precisão dos resultados é péssima, pois o gráfico da Figura 13 mostra que estes parâmetros não encontraram nenhum resultado, ou seja, não se identificou similaridades em um documento que continha plágio.

Além disso, pode-se observar no gráfico da Figura 13, que ao atribuir os valores MaxT=50 e MinT=10 ou MaxT=50 e MinT=20, os resultados retornados são os mesmos, mas ao analisar estes mesmos valores em termos de tempo de processamento, os parâmetros MaxT=50 e MinT=20 apresentam um custo de tempo melhor em relação aos demais parâmetros.

No entanto, a partir dos dois gráficos apresentados, concluiu-se que na maioria dos casos os melhores resultados tanto na precisão como em termos de tempo de processamento foram obtidos fixando os parâmetros com os seguintes valores: número mínimo de termos (MinT) = 20; número máximo de termos (MaxT) = 50.

4.4 Análise de Indícios de Plágio

Após a etapa de pré-processamento, onde se obtém somente os *tokens* com os termos relevantes para análise, os mesmos são submetidos para verificação de indícios de plágio com documentos da *web*, seguindo as seguintes sub-etapas:

4.4.1 API Google *Search*

Usou-se a *API Google Search Ajax* para realização desta etapa. Logo, cada *token* resultante da fase de pré-processamento é enviada para a *API Google Search Ajax,* a qual faz uma varredura na *web* em busca de documentos similares.

4.4.2 Resultados

Para cada *token* analisado pela *API Google*, são retornados no máximo 10 resultados similares (padrão da API). Os resultados retornados pelo google são compostos por:

- *Title*: Fornece o valor do título do resultado.

- *UnescapedUrl*: Fornece o url básico do resultado.

- *Content*: Fornece um pequeno *snippet*[2] de informações da página associada ao resultado da pesquisa.

A Figura 15 mostra um exemplo dos itens dos resultados retornados pela API Google.

[2] Um pequeno trecho.

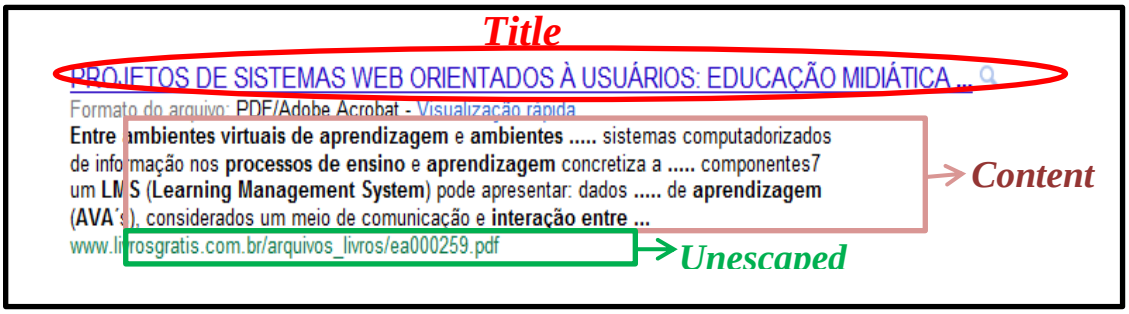

Figura 15: Resultado API Google

4.4.3 Cálculo de Similaridade

Para filtrar os resultados, e retornar somente os resultados mais relevantes, realiza-se nesta etapa um cálculo de similaridade, o qual analisa o número de palavras iguais entre a sentença do documento suspeito e os *contents* dos documentos recuperados da *web*, gerando um percentual de similaridade entre os mesmo.

Na Figura 16: Cálculo de Similaridade Figura 16 é mostrada a fórmula para o cálculo de similaridade utilizada neste trabalho.

$$S = \frac{Nº\ Palavras\ Iguais}{Nº\ Palavras\ Sentença} * 100$$

Onde: **S**= percentual de similaridade;
Nº Palavras Iguais = número de palavras da sentença que foram encontradas no conteúdo da url retornada pela API de busca do google.
Nº Palavras Sentença: quantidade de palavras que compõe a sentença que esta sendo analisada.

Figura 16: Cálculo de Similaridade

4.4.4 Resultados Filtrados

Após o cálculo de similaridade são consideradas e apresentadas somente as referências que ultrapassarem o percentual de indícios de plágio aceitável pelo sistema (Equação 1).

$$S \geq \text{indice de indícios de plágio aceitável}$$

Equação 1: Índice de plágio aceitável

Por padrão o sistema define o parâmetro de índice de similaridade aceitável por parágrafo do documento suspeito com o valor de 60%. Desta forma, o sistema analisará a similaridade entre sentenças de um documento suspeito com documentos da *web*, e a cada parágrafo apresentará somente as referências que ultrapassarem 60% de similaridade. Para verificar quais valores produziram melhores resultados, foram efetuados testes com 15 documentos suspeitos, variando os valores deste parâmetro.

A Figura 17 apresenta os testes considerando a variação deste parâmetro com seus respectivos resultados retornado.

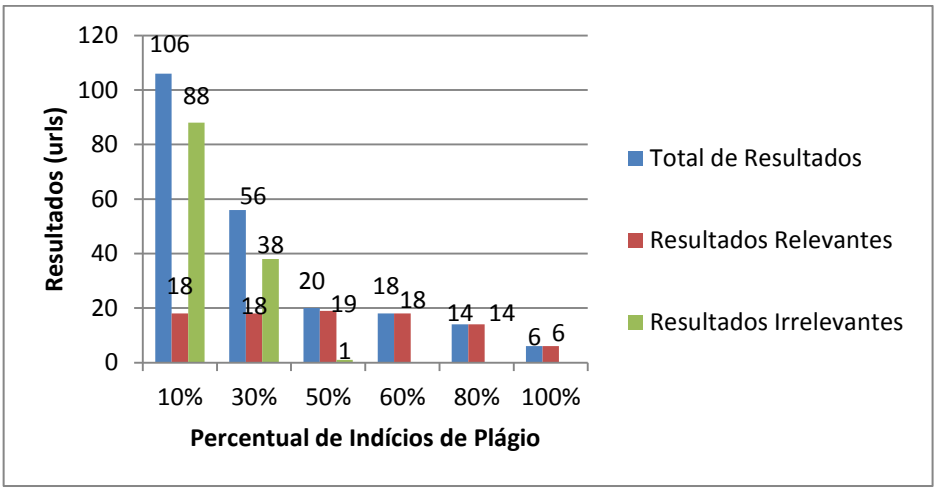

Figura 17: Testes de percentual índice de indícios de plágio

Conforme o gráfico, os valores atribuídos em 10% e 30% retornaram uma grande quantidade de resultados irrelevantes, e ao atribuir o valor de 50% apresentou-se somente 1 resultado irrelevante. Já os valores entre 60% e 100% apresentaram somente resultados relevantes, mas observou-se que, ao utilizar um percentual de 60%, obteve 18 resultados no total e todos positivos, mas ao analisar os mesmos documentos atribuindo os valores de 80% e 100%, o sistema retornou 14 e 6 resultados, diminuindo desta forma a quantidade de resultados em relação à verificação de indícios com o parâmetro em 60%, o que significa que o sistema não apresentou alguns resultados que haviam sido considerados como contendo similaridades na utilização do percentual de 60%.

Entretanto, chegou-se a conclusão de que atribuindo um valor muito baixo a este parâmetro, se obtinha um grande número de resultados irrelevantes e que ao atribuir um valor muito alto retornavam somente resultados relevantes, mas o método acabava deixando de apresentar alguns resultados que também eram considerados como indícios de plágios, mas com um percentual um pouco menor. Desta forma, com os testes realizados conseguiu-se

identificar que ao atribuir o parâmetro com 60% de similaridade, o sistema conseguia retornar uma ótima precisão.

4.4.5 Relatório

A cada parágrafo analisado são apresentados os resultados filtrados (urls/links) com o percentual de similaridade encontrado entre cada uma e o documento suspeito. Em seguida se calcula a média do índice de similaridade encontrado por parágrafo, a qual é realizada pela soma do percentual de similaridade encontrado em cada resultado e dividida pela quantidade de resultados filtrados, conforme ilustra a Equação 2. Por fim, um relatório é gerado em pdf com o texto original, e abaixo de cada parágrafo identificado como contendo indícios de plágio são impressos os resultados e o percentual de similaridade encontrado em cada um, bem como o percentual do parágrafo em geral. Além disso, o parágrafo é demarcado na cor vermelha.

$$\text{Indice Paragraph} = \frac{\text{Soma de todos os S}}{\text{Total de urls}}$$

Equação 2: Índice de similaridade

4.5 Método Integrado ao Moodle

Os Ambientes Virtuais de Aprendizagem (AVAs) tem sido empregados como facilitadores no processo de ensino e aprendizagem, tanto em termos de distribuição de informações quanto na disponibilidade de diferentes alternativas para diversificar as estratégias pedagógicas utilizadas (SANTOS, 2009).

A tarefa de verificar a originalidade de trabalhos submetidos aos AVAs de uma grande quantidade de alunos de forma manual acaba tornando o trabalho do professor extremamente cansativo e demorado, levando muitas vezes os professores a optar por aplicar tarefas nas quais os ambientes avaliam automaticamente os alunos, assim como, exercícios de múltipla escolha, associação de colunas e verdadeiro ou falso, deixando de lado as questões dissertativas e a produção de textos (FRANCO e MILANEZ, 2008). Mas há casos em que não se pode aplicar este tipo de tarefa, por exemplo, quando o aluno é submetido a desenvolver

seu trabalho de conclusão de curso, onde a originalidade é imprescindível, podendo conter citações alheias desde que sejam corretamente referenciadas.

Com objetivo de auxiliar os professores na verificação de tais tarefas, o método foi implementado em uma ferramenta e integrado ao recurso de envio de tarefas do Moodle, sendo necessária para esta integração o desenvolvimento e inclusão de um novo módulo para este AVA, o qual dá acesso ao professor aos relatórios de indícios de plágios encontrados nos trabalhos submetidos pelos alunos dentro do ambiente.

O módulo foi desenvolvido conforme as regras de inclusão de novos recursos ao Moodle, ou seja, utilizando a linguagem PHP e Mysql. Para integração da ferramenta ao ambiente virtual de aprendizagem, se faz necessário ter o Moodle na versão 1.9.9+ instalado no servidor, obter o pacote do sistema em java e o pacote do módulo detector de indícios de plágio. O pacote em java contém um executável, o qual precisa ser instanciado como um serviço do próprio servidor e inicializado, e para a instalação do módulo, deve-se descompactar o pacote com os arquivos do módulo e adicioná-lo ao arquivo *moodle/mod*. Para o funcionamento correto do sistema, além desta instalação, o professor também deverá dentro do curso, criar uma tarefa para envio de trabalhos onde o sistema detector se conectará automaticamente a este recurso, e também adicionar o módulo detector. Os pacotes para esta instalação estão disponíveis para download no endereço: ftp://200.18.45.197/.

O sistema realizará suas tarefas de forma transparente, ou seja, o mesmo não possuí uma interface gráfica por trabalhar como um serviço do Windows e estar diretamente ligado ao banco de dados do AVA Moodle. Os trabalhos submetidos no recurso de envio de tarefas do Moodle são armazenados em uma base de dados, e os documentos são analisados em um intervalo de tempo definido pelo usuário, mas como padrão do sistema este intervalo esta programado em 12 horas, ou seja, a cada 12 horas o sistema identifica quais documentos possuem a extensão .doc e que ainda não foram analisados, para então executar a análise de indícios de plágio. Após esta verificação, são gerados alguns dados no Módulo Detector de Indícios de Plágio dentro do Moodle, tais como, nome do aluno, o documento original e o relatório de indícios para download, além do status do relatório gerado, ou seja, se contém ou não indícios de plágio. O professor poderá estar alterando o parâmetro de intervalo de tempo de verificação dos documentos dentro do Módulo do Moodle.

A Figura 18 apresenta a integração do sistema ao recurso de envio de tarefas do Moodle.

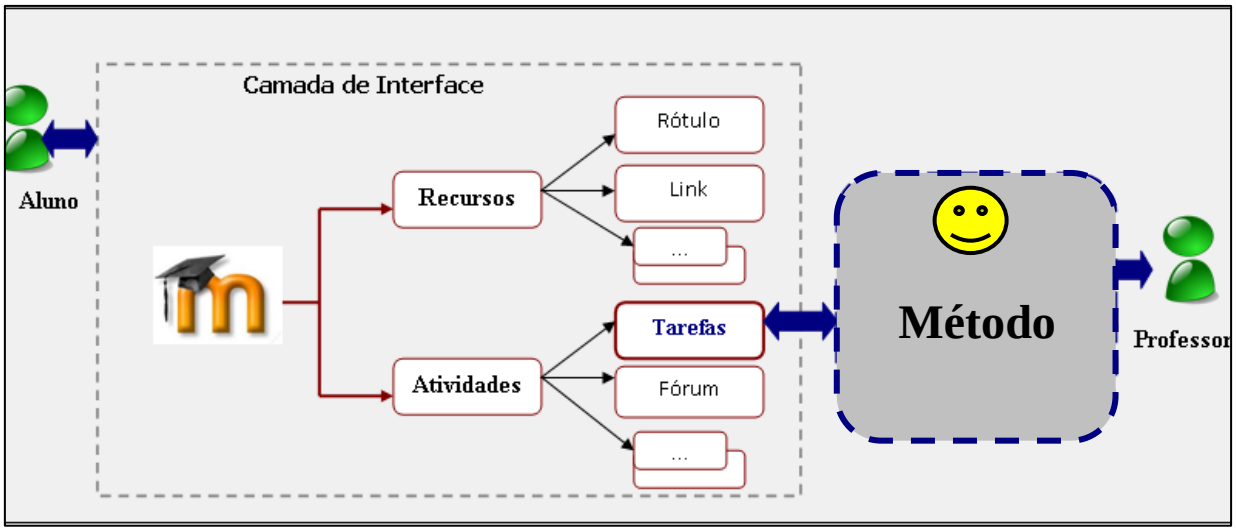

Figura 18: Integração do Sistema ao Moodle

O acesso ao módulo Detector de Indícios de Plágio é permitido somente ao professor e administrador do ambiente, para evitar que os alunos tenham acesso aos relatórios gerados. O caso de uso apresentado na Figura 19 mostra as interações permitidas pelo usuário dentro do módulo detector.

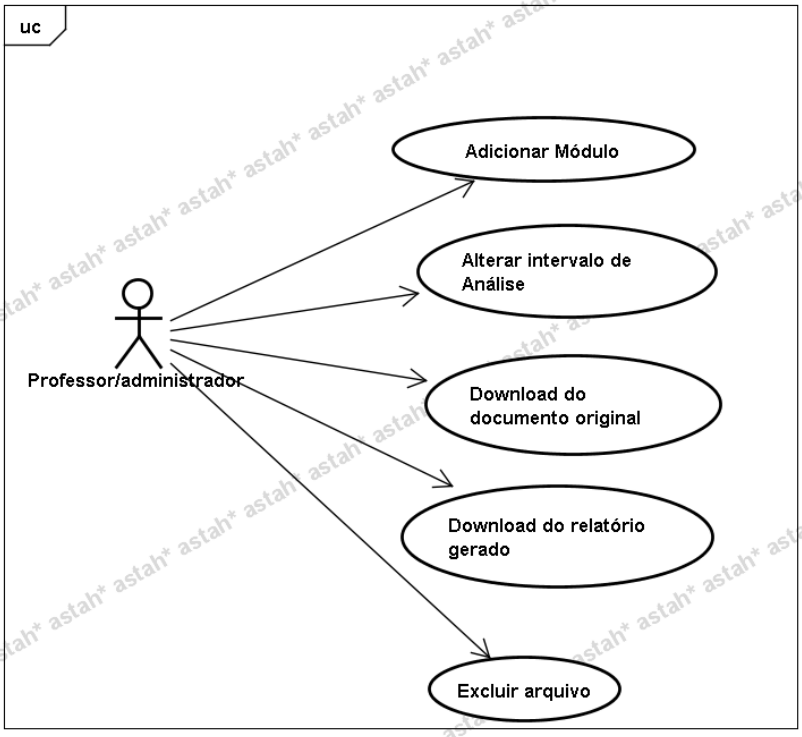

Figura 19: Caso de uso Módulo Professor/Administrador

A seguir são descritas as interações que o professor poderá fazer no Módulo.

Adicionar Módulo: O professor ou administrador do sistema poderá adicionar como uma atividade o módulo Detector de Indícios de Plágio aos cursos dentro do Moodle (Figura

20). Para realização desta interação, o sistema detector e o módulo detector já deverão estar instalados no servidor e no Moodle.

Figura 20: Inserção do Módulo Detector

Alterar intervalo de análise: Pode-se identificar um intervalo de tempo em horas para que o sistema entre em ação e analise os documentos submetidos. No entanto, não são analisados arquivos com o mesmo nome.

Visualizar Documento Original: Os trabalhos submetidos pelos alunos no recurso de envio de tarefas do Moodle também serão apresentados ao professor no módulo detector.

Visualizar Relatório: O relatório de indícios de plágio gerado é apresentado no módulo no formato pdf para acesso ao professor e download do mesmo.

Excluir: Permite a exclusão de um determinado documento que foi analisado.

A Figura 21 apresenta o Módulo Detector acessado dentro do Moodle via desktop.

Figura 21: Módulo do Sistema para o Moodle

4.6 Método Integrado ao Mle- Moodle

A integração da ferramenta desenvolvida para acesso ao *Mobile Moodle* teve por objetivo disponibilizar ao professor um alerta sobre os documentos já analisados pelo sistema, ou seja, o professor poderá ter acesso ao módulo detector e visualizar: o nome do aluno; se o trabalho já foi submetido ao ambiente; se o relatório foi gerado e se o mesmo possuí ou não indícios de plágio. No entanto, este módulo serve como auxílio ao professor, mas para abertura dos arquivos o usuário deverá estar acessando o Moodle via desktop.

Para integração da ferramenta ao Mle-Moodle, se fez necessário a adaptação do módulo desenvolvido para acesso através do dispositivo móvel, onde as tarefas submetidas no AVA acessada via desktop são refletidas no AVA móvel, ou seja, ao integrar o módulo detector ao Moodle o mesmo poderá ser acessado pelo dispositivo móvel para visualização dos recursos apresentados. A Figura 22 mostra o módulo acessado no Mle- Moodle pelo dispositivo móvel.

Figura 22: Módulo do Sistema para o Mle- Moodle

4.7 Método implementado em um Sistema Desktop

Além da integração da ferramenta desenvolvida ao ambiente virtual de aprendizagem Moodle para acesso via dispositivo móvel e desktop, o método também foi implementado em um sistema computacional com uma interface gráfica para uso fora do ambiente. Este está disponível como um executável sem a necessidade de instalação. A Figura 23 apresenta a interface do sistema.

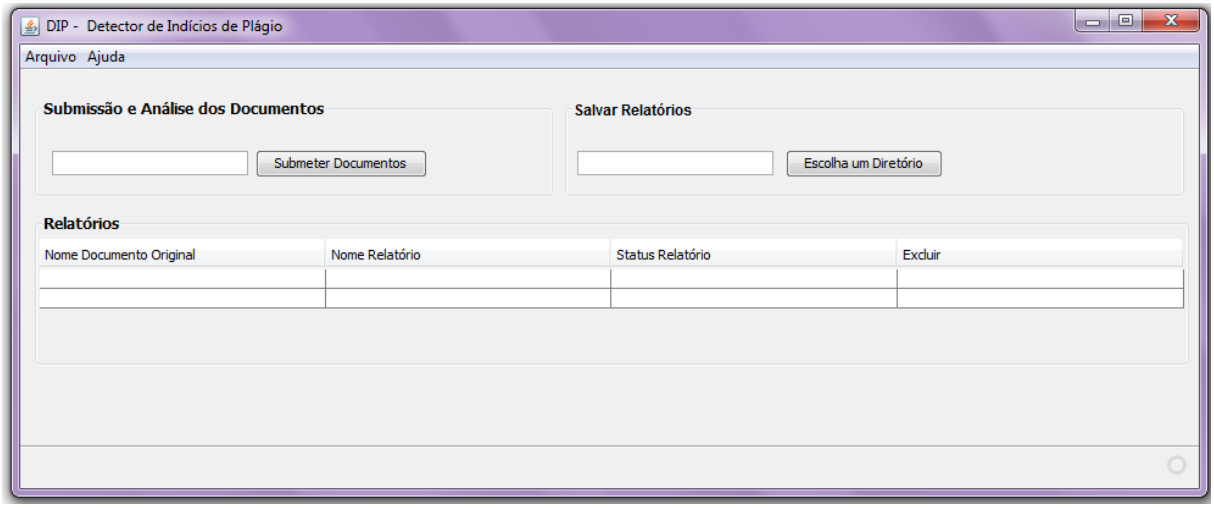

Figura 23: Sistema desktop

O sistema permite a inserção de vários documentos ao mesmo tempo para análise, o qual realiza a análise conforme a ordem de submissão. De acordo com esta ordem, a cada documento analisado o sistema retorna um relatório apontando o indício de plágio encontrado. O usuário tem a opção de escolher o diretório para salvar os relatórios.

5 RESULTADOS

Nesta seção serão descritos experimentos com o método proposto, além de uma comparação da ferramenta desenvolvida com algumas existentes na literatura.

5.1.1 Validação

Para a validação deste trabalho, aplicou-se o método em artigos reais apresentados como trabalho de final de curso de um grupo de alunos de um curso de pós-graduação de ensino a distância. A coleção de documentos é composta por 14 artigos reais do grupo de alunos, contendo entre 10 e 20 páginas cada; e o 15º artigo foi montado com textos traduzidos do idioma inglês para o português, com intuito de validar a fase de verificação do plágio bilíngue.

Aplicando-se o método proposto sobre esta coleção de documentos para analisar a similaridade entre trabalhos de alunos de um curso de pós-graduação e documentos disponíveis na internet, foram obtidos os seguintes resultados apresentados na Tabela 4.

Tabela 4: Resultados dos Experimentos

Resultados dos Experimentos			
Nome	**Resultados Relevantes**	**Indíce de Similaridade entre Documentos**	**Tempo de Processamento**
Artigo 1	71,42%	30,86%	8min 1seg
Artigo 2	88,23%	35,51%	22min 19seg
Artigo 3	87,09%	35,61%	20min 10seg
Artigo 4	80,00%	34,70%	6min 8seg
Artigo 5	79,06%	33,94%	10min 5seg
Artigo 6	98,00%	30,07%	3min 7seg
Artigo 7	92,30%	32,60%	5min 2seg
Artigo 8	90,90%	35,02%	2min 10seg
Artigo 9	89,13%	39,02%	9min 48seg
Artigo 10	80,00%	35,33%	10min 9seg
Artigo 11	83,33%	36,00%	6min 9seg
Artigo 12	96,15%	40,00%	7min 5seg
Artigo 13	73,91%	33,37%	2min 7seg
Artigo 14	75,00%	32,22%	6min 1seg
Artigo 15	100%	35,38%	11min 43seg

Os índices de similaridade podem variar entre 0% (dissimilaridade) e 100% (similaridade total) do documento em geral. Desta forma, para considerar o valor do índice de

similaridade como um indicador de plágio, segundo Shivakumar e Molina (1995) *apud* Oliveira e Oliveira (2008) , assume-se que:

Tabela 5: Avaliação de plágios pelos índices de similaridades.

Índices de similaridades entre 2 documentos	Conclusões
Menor que 33%	Não é plágio
Entre 33% e 67%	Há alguns indícios de plágio
Entre 67% e 90%	Há altos indícios de plágio
Acima de 90%	É plágio total

Fonte: Shivakumar e Molina (1995) *apud* Oliveira e Oliveira (2008).

No entanto, os resultados demonstram para os 14 artigos do grupo de alunos indícios de plágio entre 30,07% e 40%, que correspondem ao número de palavras iguais entre o artigo suspeito e o documento localizado na web. Segundo os autores Shivakumar e Molina (1995) *apud* Oliveira e Oliveira (2008) estes índices devem ser considerados como não havendo nenhum plágio ou há baixos indícios de plágio, mas o método desenvolvido calcula a similaridade entre trechos do documento suspeito com trechos de referências da web, apresentando a similaridade por parágrafo analisado, ou seja, se o sistema encontrar somente um parágrafo com um alto indício de plágio este será identificado e apresentado como plágio. Portanto, o índice de plágio encontrado entre parágrafos variou entre 62,82% e 84,27%, ou seja, encontraram-se documentos digitais que continham parágrafos com até 84,27% de similaridade ao documento suspeito.

Entretanto, onde o sistema apresentou um percentual de 30,07% de indícios no documento suspeito, foram identificados parágrafos com até 62,82% de indícios de plágio, ou seja, um valor considerado pelo Shivakumar e Molina (1995) *apud* Oliveira e Oliveira (2008) como baixo, mas conforme os testes devem ser considerados como indícios de plágio, pois o autor copiou 62,82% de um trecho de uma determinada referência, um percentual considerado um ato de plágio caso não tenha dado crédito ao autor da obra original. Já no documento apresentado com um percentual de 40% de indícios no documento em geral, foi identificado pelo sistema parágrafos com até 84,27%, um valor considerado com fortes indícios de plágio.

Para avaliar a verificação de plágio bilíngue, montou-se o artigo 15 da Tabela 4, o qual continha um texto traduzido do inglês para o português. Nesta validação, encontrou-se um percentual de indícios de plágio de 35,38% entre o conteúdo do documento suspeito com referências da *web*, um percentual considerado por Shivakumar e Molina (1995 *apud* Oliveira e Oliveira, 2008) como um baixo indício de plágio, mas que neste trabalho pode-se considerar como alto, pois o método chegou a identificar parágrafos com até 100% de indícios de plágio,

ou seja, parágrafos copiados de forma direta. Logo, notou-se que ao traduzir o texto do idioma português para o inglês mudam-se frequentemente as palavras para sinônimos, plural, entre outras palavras, e como o método verifica a similaridade entre palavras idênticas, obteve um desempenho baixo no cálculo de similaridade, mas não interferindo na precisão dos resultados encontrados, pois a precisão foi de 100%.

Como a coleção de documentos era real e não estavam identificados os trechos originais e plagiados, a avaliação da precisão dos resultados se deu a partir da verificação manual dos resultados retornados, ou seja, abriu-se cada referência e analisou-se se a mesma possuía ou não o trecho identificado pelo sistema como similar ao documento suspeito.

Nos experimentos dos 14 artigos avaliados obteve um percentual entre 71,42% e 98% na precisão dos resultados, ou seja, o sistema conseguiu detectar um grande número de indícios de plágio retornando um baixo índice de resultados irrelevantes, o que agilizará a verificação final por parte do usuário. Os resultados irrelevantes são referências da web retornadas pelo sistema, mas que não continham trechos similares ao documento suspeito.

O tempo de processamento foi outro aspecto validado na verificação dos artigos, onde se obteve uma média de 8 minutos e 13 segundos na verificação dos 15 documentos, um custo considerado baixo em relação as demais ferramentas estudadas neste trabalho (Figura 25), pois o sistema realiza a análise do documento duas vezes, ou seja, uma no seu idioma original e outra do conteúdo traduzido.

5.1.2 Comparação com outras ferramentas

Uma comparação do método desenvolvido foi realizada com algumas das ferramentas apresentadas na seção 2.2. Buscou-se comparar dois dos principais aspectos que um sistema de detecção de plágio deve possuir: precisão nos resultados e o desempenho no tempo na fase de análise de indícios de plágio.

Cada uma das ferramentas em análise apresentam diferentes características em termos de interface e parâmetros de entrada. Portanto, para poder realizar a análise de forma sucinta, considerou-se nas que permitiram a entrada de parâmetros, assim como, o número mínimo e máximo de palavras que devem conter uma sentença e o índice de indícios de plágios aceitáveis. Como algumas das ferramentas comparadas são de natureza privada, utilizou-se uma versão para testes na realização desta validação, as quais limitavam a verificação de um

documento com pequeno conteúdo textual. Desta forma, os documentos utilizados na verificação foram construídos da seguinte forma para possibilitar as comparações:

doc1: é totalmente plagiado da web, ele é composto de três parágrafos e 240 palavras, os dois primeiros parágrafos foram retirados do mesmo autor, e o último de um segundo autor.

doc2: contém 284 palavras, e foi criado a partir de trechos retirados do artigo de um aluno de um curso de especialização a distância. O primeiro parágrafo contém um alto índice de plágio, já os demais não possuem nenhuma similaridade com outros documentos.

doc3: contém 218 palavras, também foi montado a partir do artigo de um aluno do curso de especialização. O primeiro parágrafo é totalmente retirado da *web*, já o segundo parágrafo foi retirado da conclusão do artigo e mesmo assim contém um percentual alto de indícios de plágio.

doc4: é composto por um texto que foi traduzido do inglês para o português.

Para melhor compreensão desta comparação, a Figura 24 ilustra a precisão dos resultados retornados na verificação de cada um dos documentos analisados pelas 7 ferramentas. A análise do percentual de resultados relevantes foi realizada de forma manual, ou seja, foi aberta cada referência indicada como contendo similaridades ao documento suspeito e verificado se a mesma era ou não indício de plágio.

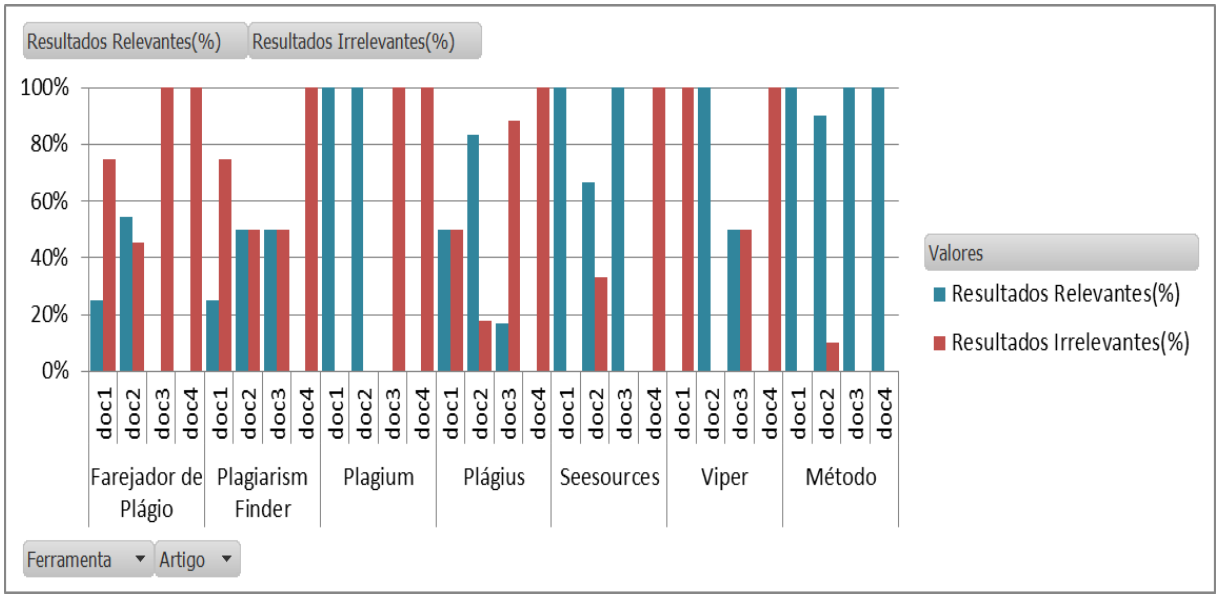

Figura 24: Análise dos Resultados das Ferramentas

O método desenvolvido apresentou resultados satisfatórios em relação às demais ferramentas, obtendo resultados relevantes de 90% e 100%, somente no documento 2 o sistema apresentou um percentual de resultados irrelevantes de 10%. Já os demais sistemas

analisados tiveram uma variação entre 0% e 100% na precisão dos resultados. No entanto, pode-se observar no gráfico da Figura 24 que os sistemas Farejador de Plágio, Plagiarism Finder e Plágius apresentaram um número elevado de resultados irrelevantes na verificação dos 4 artigos, entre 20% e 37% de relevância. A ferramenta Plagium apresentou uma média de 50% de resultados relevantes e apresentou 100% de irrelevância na verificação dos artigos 3 e 4; a Viper obteve a média de 38% em resultados relevantes, ou seja, apresentou um número maior de referências irrelevantes, e na análise dos documentos 1 e 4 foram encontrados somente referências irrelevantes, um total de 100% de irrelevância; e a ferramenta Seesources apresentou 67% de referências relevantes, não encontrando referências relevantes somente no artigo 4.

Desta forma, é possível inferir que nenhuma das ferramentas conseguiu identificar resultados relevantes no artigo 4, o qual era composto de um texto para verificação de indícios de plágio bilíngue. No entanto, percebe-se que tais sistemas não possuem um tradutor incluído para este tipo de análise, e conclui-se que o trabalho desenvolvido nesta dissertação realiza esta tarefa, tornando mais completo o processo de verificação de indícios de plágio, pois principalmente no Brasil utiliza-se muito de textos traduzidos do idioma inglês para o português.

Em termos de tempo de processamento na fase de análise de indícios de plágio com documentos disponibilizados na internet, o método desenvolvido também obteve um bom desempenho comparado com as demais ferramentas, apesar de realizar a verificação em dois idiomas, no idioma original dos documentos que atende ao idioma português e a tradução do conteúdo da língua portuguesa para a inglesa. A Figura 25 apresenta um gráfico com o desempenho de cada um dos sistemas comparados.

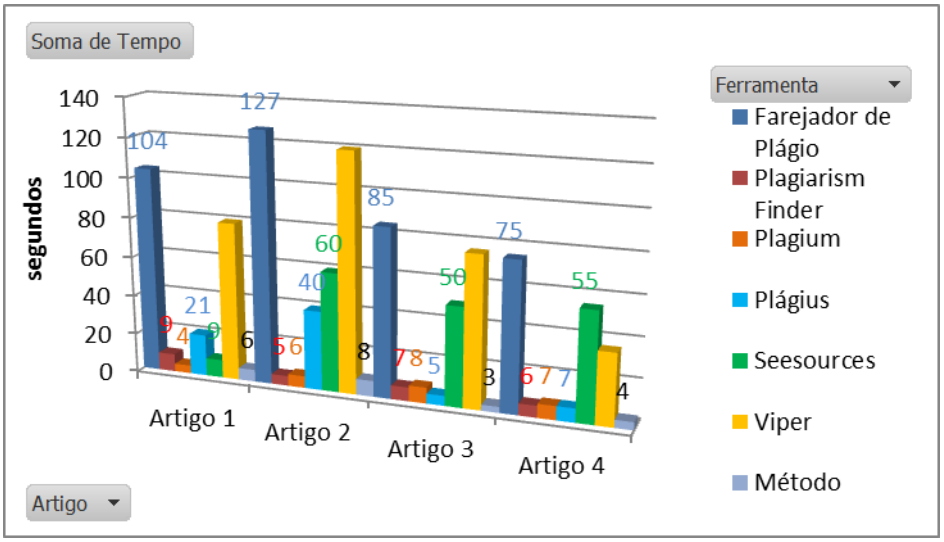

Figura 25: Tempo de Processamento entre as Ferramentas

Conforme o gráfico, a análise de indícios de plágio pelo método dos 4 artigos retornaram resultados entre 3 e 8 segundos, ou seja, uma média de 5,25 segundos por documento analisado, já as demais ferramentas variaram entre 4 e 127 segundos. A ferramenta Farejador de Plágio obteve um desempenho muito baixo em relação às demais, chegando a um custo de tempo na fase de verificação de indícios de 1 minuto e 7 segundos por artigo analisado, sendo que para a análise do artigo 2 que continha 287 palavras o sistema levou 2 minutos e 11 segundos. Outra ferramenta que também obteve um alto custo de desempenho em termos de tempo de processamento foi a Viper, que levou 2 minutos na análise do artigo 2.

Com base nos valores apresentados na Figura 25, considera-se que ao analisar 10 artigos contendo 5 mil palavras cada, ou seja, em torno de 15 páginas cada, o método desenvolvido levaria em média 17 minutos e 7 segundos, o que varia nas demais ferramentas entre 21 minutos e 1 segundo da ferramenta Plagium e 5 horas e 5 minutos na verificação pelo sistema Farejador de Plágio.

Para relatar com mais detalhes o desempenho das ferramentas estudadas neste trabalho, buscou-se alguns exemplos práticos para a avaliação do custo de tempo de cada uma delas. Desta forma, considerou-se uma turma de um curso de pós-graduação de ensino a distância, onde o professor precisaria analisar cerca de 50 trabalhos contendo 5 mil palavras cada um, ou seja, as ferramentas teriam que verificar possíveis indícios de plágio em 250 mil palavras. Neste exemplo, as ferramentas apresentaram um péssimo desempenho comparando-as com o método desenvolvido, ou seja, as ferramentas que obtiveram um melhor desempenho para processar todos os documentos foi o método desenvolvido que levou 1 hora e 4 minutos, a Plagium com 1 hora e 7 minutos e a Plagiarism Finder que levou 1 hora e 8 minutos. Já as demais ferramentas levaram um tempo muito mais elevado, variando entre 5 horas e 12 minutos da Plágius e 27 horas e 4 minutos do Farejador de Plágio.

Apesar do desempenho em tempo de processamento das ferramentas Plagium e Plagiarism Finder serem considerados um bom desempenho, eles obtiveram um péssimo desempenho em relação à precisão de resultados retornados, o que se pode observar na Figura 24, onde a ferramenta Plagium conseguiu identificar resultados somente em dois artigos, ou seja, dois dos artigos que continham plágio o sistema não conseguiu identificar nenhuma referência com similaridade, logo, estes dois artigos passariam como não contendo plágio. Já o Plagiarism Finder identificou um percentual de 68,75% de resultados irrelevantes, o que se considera como um número muito alto de irrelevância na precisão dos resultados. Além disso,

nenhuma das ferramentas conseguiu identificar referências similares ao artigo que continha o conteúdo traduzido do idioma inglês para o português, sendo que o mesmo possuía plágio.

No entanto, percebe-se que o custo de tempo na verificação de indícios de plágio em uma grande quantidade de documentos acaba se tornando insatisfatório na utilização das ferramentas pesquisadas.

6 CONCLUSÃO E TRABALHOS FUTUROS

O principal objetivo deste trabalho é oferecer um sistema de detecção de plágio eficaz, que verifique similaridades entre um determinado documento suspeito com referências disponíveis na web, retornando resultados precisos em um tempo aceitável.

Além da relevância do tema, considerando as ferramentas que foram analisadas neste trabalho, os resultados desejados foram alcançados tanto no custo de desempenho, que obteve em média 5,25 segundos por artigo analisado, bem superior aos encontrados nas demais ferramentas, assim como, a Plagium que levou 6,25 segundos, a Plagiarism Finder com 6 segundos, Plágius em 18,25 segundos, Seesources em 43,5 segundos, Viper em 1 minuto e 3 segundos e o Farejador de Plágio em 1 minuto e 6 segundos.

Outro aspecto analisado foi à precisão dos resultados retornados por cada uma das ferramentas, onde o método desenvolvido identificou uma média de 97,50% de resultados relevantes entre os quatro artigos analisados, um ótimo resultado em relação às demais ferramentas, assim como, a Plágius que identificou apenas 37,49%, a Seesources 66,66%, Viper 37,50%, Plagium 50%, Plagiarism Finder 31,25% e o Farejador de Plágio apresentou o pior desempenho na precisão dos resultados, identificando somente 19,88% de resultados relevantes.

Além disso, nenhuma das ferramentas e métodos estudados nesta pesquisa relata a identificação de plágio bilíngue, uma das principais vantagens deste trabalho, que utiliza uma biblioteca de tradução do Google para tratar este tipo de plágio, o qual é muito frequente nos dias de hoje, pois é muito fácil utilizar qualquer mecanismo de tradução para se apropriar de um conteúdo em um idioma e adaptá-lo a outro.

Um dos principais problemas encontrados durante esta pesquisa foi à dificuldade em se obter informações sobre técnicas utilizadas para a verificação de similaridades pelas ferramentas, pois a maioria é de natureza privada e não divulga esta informação, dificultando também na comparação entre as mesmas, pois nem todas disponibilizavam uma versão para teste e outras limitavam o tamanho do documento. Desta forma, realizou-se uma comparação com artigos curtos em sistemas gratuitos e em sistemas que disponibilizavam sua versão para testes.

Visando proporcionar o uso deste trabalho pela comunidade científica, o mesmo já está disponível de duas formas: para integração ao ambiente virtual de aprendizagem Moodle, com intuito de agilizar e facilitar a análise da qualidade e originalidade dos documentos já

postados pelos alunos no ambiente, e para integração ao Mle-Moodle, proporcionando ao professor o acompanhamento da análise dos trabalhos a partir do seu dispositivo móvel. Além disso, o sistema também esta disponível para utilização como um programa desktop.

Considerando os ótimos resultados alcançados pelo método proposto, conclui-se que o mesmo atingiu plenamente os objetivos propostos, diferenciando-se de sistemas similares pela precisão e rapidez dos resultados, bem como por implementar um método que identifica indícios de plágio bilíngue e por disponibilizar a opção de ser integrado ao recurso de envio de tarefas do AVA Moodle e Mle- Moodle, facilitando sua utilização por usuários do ambiente e também por estar disponível como um programa executável, sem a necessidade de instalação.

Alguns trabalhos futuros podem ser realizados para melhoria e complementos da arquitetura proposta, como:

- Passar a versão do sistema desktop para *web*.
- Incluir a análise em documentos em outros formatos, assim como, pdf, open office. Pois o mesmo somente analisa documentos na extensão .doc.
- Incluir a verificação de indícios de plágios para outros tipos de plágio, bem como, plágio direto, plágio por referência e plágio multilíngue, onde deve-se identificar conteúdos que são traduzidos em diferentes línguas, pois o método somente identifica quando se copia um conteúdo traduzido do inglês para o português.
- Aplicar técnicas de radicalização (*stems*) onde diversas palavras que designam variações indicando plural, flexões verbais ou variantes são sintaticamente similares entre si. Por exemplo, as palavras “real”, “realidade”, “realeza” e “realizado” têm sua semântica relacionada. Ou seja, esta etapa faz parte da detecção de plágio mosaico mas não foi implementada.
- Desenvolver um algoritmo para verificar se o parágrafo considerado como indício de plágio esta dando crédito à obra original. O método até o momento analisa a similaridade entre dois documentos sem identificar se o mesmo é um ato de plágio verdadeiro.
- Testar o método com documentos em outros idiomas, pois neste trabalho testaram-se somente trabalhos no idioma português.

7 7 BIBLIOGRAFIA

ALMEIDA, M. E. B. D. Tecnologia e educação a distância:abordagens e contribuições dos ambientes digitais e interativos de aprendizagem, 2002. 14 f. Disponivel em: <http://www.pr.senai.br/portaldelibras/uploadAddress/tecnologia_e_educacao%5B51791%5D.pdf>. Acesso em: 25 out. 2010.

ALTAVISTA. Visão Geral da Empresa Altavista, 2010. Disponivel em: <http://br.altavista.com/>. Acesso em: 31 out. 2010.

APPROBO , 2010. Disponivel em: <http://approbo.citilab.eu/approbo.jsp>. Acesso em: 22 jul. 2010.

AULANET. Solução tecnológica para acções de formação presenciais , 2010. Disponivel em: <http://www.aulanet.pt/>. Acesso em: 31 out. 2010.

BARBASTEFANO, R. G. PERCEPÇÃO DO CONCEITO DE PLÁGIO ACADÊMICO ENTRE ALUNOS DE ENGENHARIA DE PRODUÇÃO E AÇÕES PARA SUA REDUÇÃO, Florianópolis – SC, p. 1-18, dez. 2007. ISSN 1676 - 1901. Disponivel em: <www.producaoonline.ufsc.br>. Acesso em: 30 out. 2010.

BARBOSA, D. N. F. et al. Em direção a Educação Ubíqua: aprender sempre, em qualquer lugar,com qualquer dispositivo. **In: CINTED - Novas Tecnologias na Educação**, v. 6, n. 1, 2008.

BROWN, V. J.; HOWELL, M. E. The Efficacy of Policy Statements on Plagiarism: Do They Change Students' Views? **Research in Higher Education**, v. 42, n. 1, p. 103-118 , fev. 2001.

BUTAKOV, S.; SCHERBININ, V. The toolbox for local and global plagiarism detection. **In: Jornal Computers & Education**, v. 52, p. 781–788, 2009.

CARMO, M.; KENNEDY, D. M. Bilingual Plagiarism in the Academic World. **Ethical Practices and Implications in Distance Learning**, p. 320-327, 2009.

DIAS, M. A. L. Extração Automática de Palavras-Chave na Língua Portuguesa Aplicada a Dissertações e Teses da Área das Engenharias, 2004. 127 f. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica e de Computação, Campinas, SP.

EPHORUS. **Ephorus:** liderança na Europa, 2010. Disponivel em: <http://www.ephorus.pt/home>. Acesso em: 22 jul. 2010.

FAREJADOR. Farejador de Plágios 10, 2010. Disponivel em: <http://www.farejadordeplagio.com.br/>. Acesso em: 24 out. 2010.

FINDER, P. Plagiarism Finder, 2010. Disponivel em: <http://www.plagiarism-finder.com/en-index.htm>. Acesso em: 24 out. 2010.

FRANCO, L. R. H. R.; MILANEZ, J. R. C. Implantação de um software detector de plágio para análise das questões dissertativas do ambiente virtual de aprendizagem TelEduc, São Paulo, 2008. ISSN 1806 - 1362. Disponivel em: <http://www.abed.org.br/revistacientifica/_brazilian/edicoes/2008/2008_Edicao_pesquisa.htm>. Acesso em: 25 jul. 2010.

GARSCHAGEN, B. **Observatório da Imprensa - No Mínimo**, jan. 2006. Disponivel em: <http://www.observatoriodaimprensa.com.br/artigos.asp?cod=366ASP006>.

GOOGLE. Aprenda o básico sobre o Google, 2010. Disponivel em: <http://www.google.com.br/intl/pt-BR/help/basics.html>. Acesso em: 31 out. 2010.

HANDBOOK, A. (Brasil). **07.11 - Code of Practice on Plagiarism**, v. 1, 2009.

IEEE, O. Punishment for Plagiarism. **In: IEEE - The Institue**, v. 34, n. 3, 2010. Acesso em: 2010.

LEGISLAÇÃO. Código Penal Brasileiro. Decreto-Lei n.º 2.848, de 7 de Dezembro de 1940. DOU de 06 de Maio de 2010, 2010.

LIU, Y.-T. et al. Extending Web Search for Online Plagiarism Detection. **In: Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on**, p. 164-169, ago. 2007.

MAURER, H.; KAPPE, F.; ZAKA, B. Plagiarism - A Survey. **In: Journal of Universal**, p. 050-1084, 2006.

MCCABE, D. L. Cheating among college and university students: A North American. **In: International Journal for Educational Integrity**, 2005.

MORAES, R. O plágio na pesquisa acadêmica: a proliferação da desonestidade intelectual. **In: Revista Diálogos Possíveis - Faculdade Social da Bahia**, Bahia, n. 1, p. 92-109, jun. 2004. Disponivel em: <http://www.faculdadesocial.edu.br/dialogospossiveis/artigos/4>. Acesso em: 10 ago. 2010.

MUSSINI, J. A. NOVAS ARQUITETURAS PARA DETECÇÃO, Paraná, 2008. 107 f. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Paraná. 2008.

NEILL, C. J.; SHANMUGANATHAN, G. A Web-Enabled Plagiarism Detection Tool. **In: IEEE Computer Society**, Pennsylvania State Univ., University Park, PA, USA, v. 6, n. 5, p. 19-23, 2004. ISSN 1520-9202.

OGATA, H.; YANO, Y. How Ubiquitous Computing can Support Language Learning. **In: Proc. of KEST**, p. 1-6, 2003. Disponivel em: <http://www-yano.is.tokushima-u.ac.jp/ogata/clue/ogata-kest2003.pdf>. Acesso em: 31 out. 2010.

OLIVEIRA, E. et al. Classificando Automaticamente Documentos Digitais no Site de Notícias do UOL. **In: Seminário Nacional de Bibliotecas Universitárias**, Salvador, 2006.

OLIVEIRA, M. G. D. et al. Bibliotecas digitais aliadas na detecção automática de plágio. **V Simpósio Internacional de Bibliotecas Digitais**, São Paulo, 2007.

OLIVEIRA, M. G. D.; OLIVEIRA, E. Uma Metodologia para Detecção Automática de Plágios em Ambientes de Educação a Distância. **In: Congresso Brasileiro de Ensino Superior a Distância – ESUD 2008**, Gramado, RS, 2008. 1-20.

PEREIRA, R. C.; MOREIRA, V. P.; GALANTE, R. A New Approach for Cross-Language Plagiarism Analysis. **In: Proceedings of the CLEF 2010 Conference on Multilingual and Multimodal**, Padua, Italy, p. 15-26, 2010.

PEZZIN, M. Z. Metodologia estatística computacional de detecção automatizada do plágio autoral - Uma proposta de interpretação s resultados do programa Farejador de Plágio, 2010, 2010. Disponivel em: <http://www.farejadordeplagio.com.br/metodologia_do_farejador.pdf>. Acesso em: 12 jan. 2011.

PLAGIUM. Plagium, 2010. Disponivel em: <http://www.plagium.com/>. Acesso em: 24 out. 2010.

PLAGIUS. Plagius - The ultimate in plagiarism detection, 2010. Disponivel em: <http://www.plagius.com/s/en/default.aspx>. Acesso em: 15 out. 2010.

POTTHAST, M. et al. Cross-Language Plagiarism Detection. **In: Language Resources and Evaluation**, n. Online First, jan. 2010.

RABELO, C. Idéias roubadas. **UnB Agencia**, Brasília, 2006. Disponivel em: <http://www.secom.unb.br/unbagencia/ag0706-27.htm>. Acesso em: 7 out. 2010.

RIBEIRO, J. P. A.; REATEGUI, E.; BOFF, E. Integrando um Agente Pedagógico para Recomendação de Tutores a um Sistema de Gerência de Cursos. **In: Revista Novas Tecnologias na Educação**, Porto Alegre, v. 5, n. 1, jul. 2007. ISSN 1679-1916.

SANTANA, J. D. M.; JOBERTO, S. B. M. Um Sistema de Plágio em Ambiente Virtual de Aprendizagem. **Anais do Virtual Educa 2003**, Miami, 2003. p. 230–242.

SANTOS, E. O. Ambientes virtuais de aprendizagem: por autorias livre, plurais e. **In: Revista FAEBA**, v. 12, n. 18, 2003.

SANTOS, L. M. A. A Inserção de um Agente Conversacional Animado em um Ambiente Virtual de Aprendizagem a partir da Teoria Cognitiva, Porto Alegre, 2009. 115 f. Tese (Informática na Educação) - Universidade Federal do Rio Grande do Sul, UFRGS.

SCHMIDT, A. Potential and Challenges os Context-Awareness for Learning Solutions. **In: Proceedings of the 13th Annual Workshop of the SIG Adaptivity and User Modeling in Interactive**, Saarbrücken, out. 2005. p. 63-68.

SEESOURCES. SeeSources.com - instant, automatic & free text analysis, 2010. Disponivel em: <http://www.plagscan.com/seesources/analyse.php>. Acesso em: 22 jul. 2010.

SILVA, A. K. L. D.; DOMINGUES, M. J. C. D. S. Plágio no Meio Acadêmico: percepção de alunos de pós-graduação sobre o tema. **VI Simpósio de Gestão e Estratégia em Negócios**, Seropédica, RJ, 2008.

STEIN, B. A. S. M. Z. E. Near Similarity Search and Plagiarism Analysis. **in From Data and Information Analysis to Knowledge Engineering**, Springer: Berlin, p. 430-437, 2006.

STEIN, B.; EISSEN, S. M. Z. Near Similarity Search and Plagiarism Analysis. **in From Data and Information Analysis to Knowledge Engineering**, Springer: Berlin, p. 430-437, 2006.

TIDIA-AE. O Projeto, 2010. Disponivel em: <http://tidia-ae.iv.fapesp.br/node/3>. Acesso em: 31 out. 2010.

TURNITIN. Prevent plagiarism, 2010. Disponivel em: <http://turnitin.com/static/index.html>. Acesso em: 25 jul. 2010.

URKUND, 2010. Disponivel em: <http://www.urkund.com/int/en/>. Acesso em: 25 jul. 2010.

VIPER. The Anti-plagiarism Scanner, 2010. Disponivel em: <http://www.scanmyessay.com>. Acesso em: 7 29 2010.

YAHOO. Como funciona a busca do Yahoo!/Cadê?, 2010. Disponivel em: <http://help.yahoo.com/l/br/yahoo/bizex/bizex-44.html>. Acesso em: 31 out. 2010.

YANG, S. J. H. Context Aware Ubiquitous Learning Environments for Peer-to-Peer Collaborative Learning. **In: Educational Technology & Society**, v. 9 (1), p. p. 188-201, 2006.

YINGLING, M. Mobile Moodle. **In: Journal of Computing Sciences in Colleges**, v. 21, n. 6, p. 280 - 281, 2010.

ZOU, D.; LONG, W.-J.; LING, Z. A Cluster-Based Plagiarism Detection Method. **in CLEF 2010**, Padua Italy, 2010.