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

MISS MARPLE – DESENVOLVIMENTO DE FERRAMENTA PARA AUXILIAR NA VERIFICAÇÃO E DETECÇÃO DE INDÍCIOS DE PLÁGIO COM BASE NO MÉTODO DIP – DETECTOR DE INDÍCIOS DE PLÁGIO

DISSERTAÇÃO DE MESTRADO

Catiane Priscila Barbosa Arenhardt

Santa Maria, RS, Brasil
2013

MISS MARPLE – DESENVOLVIMENTO DE FERRAMENTA PARA AUXILIAR NA VERIFICAÇÃO E DETECÇÃO DE INDÍCIOS DE PLÁGIO COM BASE NO MÉTODO DIP – DETECTOR DE INDÍCIOS DE PLÁGIO

Catiane Priscila Barbosa Arenhardt

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

Orientadora: Profª. Drª. Roseclea Duarte Medina

Santa Maria, RS, Brasil
2013

Ficha catalográfica elaborada através do Programa de Geração Automática
da Biblioteca Central da UFSM, com os dados fornecidos pelo(a) autor(a).

Barbosa Arenhardt, Catiane Priscila
MISS MARPLE - DESENVOLVIMENTO DE FERRAMENTA PARA AUXILIAR NA VERIFICAÇÃO E DETECÇÃO DE INDÍCIOS DE PLÁGIO COM BASE NO MÉTODO DIP - DETECTOR DE INDÍCIOS DE PLÁGIO / Catiane Priscila Barbosa Arenhardt.-2013.
71 f.; 30cm

Orientador: Roseclea Duarte Medina
Dissertação (mestrado) - Universidade Federal de Santa Maria, Centro de Tecnologia, Programa de Pós-Graduação em Informática, RS, 2013

1. Ferramenta detector indícios de plágio I. Duarte Medina, Roseclea II. Título.

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

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

***MISS MARPLE* – DESENVOLVIMENTO DE FERRAMENTA PARA AUXILIAR NA VERIFICAÇÃO E DETECÇÃO DE INDÍCIOS DE PLÁGIO COM BASE NO MÉTODO DIP – DETECTOR DE INDÍCIOS DE PLÁGIO**

elaborada por
Catiane Priscila Barbosa Arenhardt

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

COMISSÃO EXAMINADORA:

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

_____
Profª. Iara Augustin**, Drª. (UFSM) -** Examinadora

_____
Profª. Mára Lúcia Fernandes Carneiro**, Drª. (UFRGS) -** Examinadora

Santa Maria, 22 de abril de 2013.

DEDICATÓRIA

Famílias Barbosa, Arenhardt e Mazzutti... E amigos.

AGRADECIMENTOS

Primeiramente agradeço a Deus, por ser o doador da vida, e nosso ser superior que está sempre presente em todos os momentos de nossa vida.

Agradeço aos meus pais Elena e Pedro, em especial a minha mãe pelo incentivo, esforço e dedicação...

Agradeço à minha família de modo geral, pelo apoio e incentivo de cada um, em especial, aos meus avós Anita e Hirco e ao meu esposo Aldair...

A todos os professores do curso, especialmente a minha orientadora, Roseclea Duarte Medina, pelo tempo disponibilizado a mim, pela amizade e pelo conhecimento dividido.

Obrigada à secretaria do PPGI, na pessoa do Secretário Josmar, pela amizade e empenho em auxiliar na resolução de burocracias...

Obrigada aos colegas de pesquisa e desenvolvimento Ricardo Bianchin e Vinicius Leal Trindade pela fundamental colaboração no trabalho.

Obrigada aos colegas de laboratório e do GRECA, de modo especial à colega e amiga Solange de Lurdes Pertile, pelo apoio, conhecimentos e amizade divididos em diversos momentos...

Agradeço à Universidade Federal de Santa Maria, bem como todos os profissionais que se envolvem para proporcionar uma educação de qualidade e gratuita...

Agradeço, de forma geral, a todos que direta ou indiretamente contribuíram com um pedacinho deste sonho, todos ficarão guardados em meu coração de forma especial...

.

“Grandes coisas têm pequenos inícios.”
- PROMETHEUS

RESUMO

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

MISS MARPLE – DESENVOLVIMENTO DE FERRAMENTA PARA AUXILIAR NA VERIFICAÇÃO E DETECÇÃO DE INDÍCIOS DE PLÁGIO COM BASE NO MÉTODO DIP – DETECTOR DE INDÍCIOS DE PLÁGIO

AUTORA: CATIANE PRISCILA BARBOSA ARENHARDT
ORIENTADORA: ROSECLEA DUARTE MEDINA
Data e Local da Defesa: Santa Maria, 22 de abril de 2013.

O trabalho desenvolvido objetivou identificar as necessidades que ainda permeavam as ferramentas de análise e detecção de indícios de plágio para posterior desenvolvimento de uma nova ferramenta, denominada *Miss Marple*, a qual atendesse aos requisitos identificados no decorrer das pesquisas, além de dar continuidade ao trabalho desenvolvido denominado DIP - Detector de Indícios de Plágio. As pesquisas realizadas no decorrer deste estudo possibilitaram o levantamento das características de cada ferramenta estudada, possibilitando tracejar um comparativo entre as mesmas e o desenvolvimento da nova ferramenta. A validação da ferramenta desenvolvida foi realizada em duas modalidades de curso, presencial e a distância, além da avaliação de usabilidade da ferramenta.

Os resultados alcançados evidenciaram que a ferramenta *Miss Marple*, desenvolvida, apresentou bons resultados nas análises de percentual de indícios de plágio chegando a uma precisão aproximada de 100% e obteve o melhor tempo de processamento quando comparada com três das ferramentas que foram estudadas (Farejador de Plágio, Plagius detector e VIPER). Além do tempo de processamento e consistência da análise de indícios de plágio, outro fator que merece destaque é a construção do repositório de análise por arquivo submetido, o qual proporciona ao usuário o acesso aos textos com trechos similares que ficam armazenados localmente em seu hardware.

Palavras-chave: Ferramentas; Plágio; DIP; *Miss Marple*.

ABSTRACT

Master's Dissertation
Post-Graduate Program in Informatics
Federal University of Santa Maria

MISS MARPLE - DEVELOPMENT TOOL TO AID IN THE DETECTION OF VERIFICATION AND EVIDENCE OF PLAGIARISM METHOD BASED ON DIP - DETECTOR EVIDENCE OF PLAGIARISM

AUTHOR: CATIANE PRISCILA BARBOSA ARENHARDT
ADVISOR: ROSECLEA DUARTE MEDINA
Defense Place and Date: Santa Maria, April 22nd, 2013.

The work aimed to identify the needs that still permeated the tools of analysis and detect plagiarism for further development of a new tool, called *Miss Marple*, which met the requirements identified in the course of research, besides continuing the work called DIP - Evidence of Plagiarism Detector. The research conducted in this work allowed the survey of the characteristics of each tool enabling trace a comparative study between them and the development of the new tool. The validation of the developed tool was performed in two modes of travel, and distance, and by evaluating the usability of the tool.

The results showed that *Miss Marple* tool, developed, presented good results in the analysis of percentage signs of plagiarism coming to an accuracy of approximately 100% and got the best processing time when compared with three of the tools that were studied (Sniffer plagiarism, Plagius detector and VIPER). In addition to the processing time and consistency analysis of evidence of plagiarism, another factor that deserves mention is the construction of the repository file submitted for analysis, which provides the user access to the texts with similar passages that are stored locally on your hardware.

Key words: Tools; Plagiarism; DIP; *Miss Marple*.

LISTA DE FIGURAS

Figura 2.1.1 - Uso da Internet como fonte de pesquisa [FERRARESI *et. al*, 2008].....21
Figura 3.1.1.1-2 - Araponga [ARAPONGA, 2012] 25
Figura 3.1.1.3-1 - Interface do DOCCOP [DOCCOP, 2012].....27
Figura 3.1.1.4 -1 - EtBlast [EtBlast, 2012].....28
Figura 3.1.1.5 -1 - Farejador de Plágio [FAREJADOR, 2012].....29
Figura 3.1.1.6 -1 - Interface da ferramenta Plagiarisma [PLAGIARISMA, 2012].....30
Figura 3.1.1.7 -1 - Interface do Plagium [PLAGIUM, 2012].....31
Figura 3.1.1.8-1 - Interface do Plagius [PLAGIUS, 2012].....32
Figura 3.1.1.9-1 - Interface do VIPER [VIPER, 2012].....33
Figura 5.1-1- DIP – Versão Desktop. [PERTILE, 2011] 44
Figura 5.1-2 - DIP – Versão Moodle [PERTILE, 2011] 44
Figura 5.1-3 - DIP – Versão MLE-Moodle [PERTILE, 2011] 45
Figura 5.2.1-1 - Diagrama de caso de uso ferramenta.....48
Figura 5.2.2-1 - Diagrama caso de uso – usuário/ ferramenta.....49
Figura 5.3-1 - Cálculo de Similaridade [PERTILE, 2011].....50
Figura 5.3-2 - Cálculo do melhor indíce de similaridade [PERTILE, 2011] 51
Figura 5.3-3 - Síntese geral de funcionamento da ferramenta *Miss Marple* 52
Figura 5.4-1 - Interface do Miss Marple, iniciando a execução de uma análise.....523
Figura 5.4-2 - Execução de uma análise.....55
Figura 5.4-3- Feedback final da análise.....55
Figura 5.4-4 - Formação do repositório 55
Figura 5.4-5 – Interpretação do relatório de indício de plágio – Miss Marple.....55
Figura 3.1.2.5-1 - Verificação de indícios de plágio em arquivo sem plágio.....61

LISTA DE QUADROS

Quadro 3.1-1 - Ferramentas para detecção de indícios de plágio Adaptado de [SIBI, 2011] e [PERTILE, 2011] 36
Quadro 5.2-1 - Resumo das diferenças entre Método DIP e ferramenta Miss Marple 47
Quadro 6.2-1 - Questões de avaliação das ferramentas de softwares de detecção de plágio - íntegra. [NUNES et. al, 2012] 64
Quadro 6.2-2 - Comparativo das ferramentas utilizadas para testes 69

LISTA DE GRÁFICOS

Gráfico 5.1.1 - Resultado da análise de relevância dos arquivos [PERTILE, 2011].....45
Gráfico 6.1.1 - Precisão dos resultados – comparação entre ferramentas 59
Gráfico 6.1.2 - Tempo de processamento em Internet de 512Kb.....62
Gráfico 6.1.3 - Tempo de processamento em Internet de 3Mb 63
Gráfico 6.2.1- Ocorrência de falhas durante a execução do *Miss Marple* 65
Gráfico 6.2.2 - Consistência das referências encontradas 66
Gráfico 6.2.3 - Quesitos avaliados no checklist de usabilidade. 67
Gráfico 6.2.4 - Percentual de atendimento de requisitos de usabilidade.....68

LISTA DE ABREVIATURAS E SIGLAS

API – Application Programming Interface

AVA – Ambiente Virtual de Aprendizagem

CNPQ – Conselho Nacional de Desenvolvimento Científico e Tecnológico

CTRL C – Comando para copiar

CTRL V – Comando para colar

DIP – Detector de Indícios de Plágio

Doc – Extensão do MS WORD abreviação de documento

Docx – Extensão do MS WORD abreviação de documento

HTML – HyperText Markup Language

IDE – Integrated Development Environmen)

Pdf – Portable Document Format

ppt – Extensão do MS WORD abreviação de apresentação de slides

MEDLINE – banco de dados de referências da área da saúde

Rtf – Rich text Format

Txt – Abreviação de texto puro

Teleduc – Ambiente virtual de aprendizagem denominado Teleduc

URL – Uniform Resource Locator

UML – Linguagem de Modelagem Unificada - Unified Modeling Language

WEB – World Wide Web

SUMÁRIO

1 INTRODUÇÃO **15**
1.1 Motivação **16**
1.2 Objetivos **17**
1.3 Principais contribuições do trabalho **18**
1.4 Organização do texto **18**
2 REVISÃO BIBLIOGRÁFICA **19**
2.1 Caracterização e prática de Plágio **19**
2.2* Técnica de *Stemming **23**
3 FERRAMENTAS AUXILIARES NA DETECÇÃO DE INDÍCIOS DE PLÁGIO **24**
3.1 Ferramentas estudadas **25**
3.1.1 Ferramentas gratuitas. 25
3.1.2 Ferramentas pagas 33
3.1.3 Motores de busca 37
4 METODOLOGIA DO TRABALHO **39**
5 PROPOSTA DO DETECTOR DE INDÍCIOS DE PLÁGIO MISS MARPLE **43**
5.1 Método DIP – Detector de Indícios de Plágio **43**
5.2 Proposta: A ferramenta *Miss Marple* **46**
5.2.1 Modelagem da Ferramenta 47
5.3 Funcionamento da Ferramenta *Miss Marple* **49**
5.4 *Miss Marple*: Apresentação da ferramenta **53**
5.5 Desafios encontrados **57**
6 RESULTADOS **59**
6.1 Validação da ferramenta **59**
6.2 Requisitos de usabilidade **64**
7 Considerações finais **70**
REFERÊNCIAS **72**

1 INTRODUÇÃO

A difusão da Internet e junto a ela a quantidade de informações disponíveis formando um grande acervo virtual facilita acesso a uma infinidade de materiais, proporcionando ao usuário a possibilidade de usufruir de informações de maneira incorreta ou mal intencionada, sem dar os devidos créditos aos autores [MORAES, 2012].

No meio acadêmico, os alunos são colocados a frente da produção de materiais de pesquisa, tanto de artigos quanto de trabalhos de conclusão de curso, e estes estudantes, por muitas vezes, sentem-se inseguros com a escrita, são inexperientes, desconhecem sobre o que caracteriza cópia ilegal ou plágio e utilizam-se das informações de maneira incorreta, ou ainda, copiam ou compram trabalhos intencionalmente. O autor [BARNBAUM, 2002] escreve que a falta de conhecimento do que constitui o plágio leva muitos alunos a cometê-lo inconscientemente. Se o estudante não sabe exatamente o que é o plágio, não pode evitar fazê-lo.

Dentro destas perspectivas, é visível que aumente consideravelmente a atenção do professor e de revisores de texto em relação a autenticidade dos trabalhos, culminando no acréscimo de sua demanda de trabalho devido a maior atenção dedicada para a correção.

Comumente, os professores e revisores em geral (de congressos, periódicos, revistas, jornais, comissões editoriais), têm grandes quantidades de textos para analisar, e muitos acabam por não dispôr de tempo suficiente para um acompanhamento mais profundo de toda esta produção. Sendo assim, o uso de ferramentas para auxiliar na verificação de indícios de plágio se apresenta como uma boa alternativa para este fim, objetivando otimizar o tempo do professor ou revisor no controle da autenticidade das informações escritas.

Com a finalidade de desenvolver uma ferramenta para auxílio na verificação de indícios de plágio textual, primeiramente partiu-se da identificação do estado da arte nessa área, bem como, a elaboração de uma lista de ferramentas que foram pesquisadas bibliograficamente. Em seguida, foram realizados testes nessas ferramentas para identificação de suas funcionalidades. O levantamento bibliográfico das ferramentas foi realizado a partir de trabalhos já desenvolvidos em [SIBI, 2011], [LIMA e RESENDE, 2012], [PERTILE e MEDINA, 2011] e [SANTOS e FRANCO, 2010], e nos sites dos fabricantes [DOC COP, 2012], [EPHORUS, 2012], [ETBLAST, 2012], [FAREJADOR, 2012], [PLAGIARISMA,

2012], [PLAGIARISM.ORG, 2012], [PLAGIUM, 2012], [PLAGIUS, 2012], [PLAGIO.NET, 2012], [SCHOLARONE, 2012], [TURNITIN, 2012], [URKUND, 2012], [VIPER, 2012].

Na sequência, foram realizados testes nas ferramentas que apresentaram licença gratuita. Com base nos levantamentos bibliográficos e testes, identificou-se que as ferramentas de licença livre, necessitam de cadastro de usuário, fazem buscas por termos similares somente na Internet e não fazem análise cruzada de arquivos inteiros, além de oferecerem verificação de arquivos com extensões e tamanhos limitados.

A partir desse levantamento, foram identificadas e elencadas funções que precisavam ser aprimoradas, ou mesmo desenvolvidas, para que uma ferramenta utilizada na detecção de indícios de plágio trate um número maior de especificidades de textos. Sendo assim, se propôs o desenvolvimento de uma ferramenta de detecção de indícios de plágio textual, em arquivos com extensões .doc, .docx, .pdf e .HTML , utilizando técnicas de *stemming* (extração do radical das palavras e armazenamento em uma lista), o que possibilita a comparação de palavras com um mesmo radical. Além disso, inclui-se no desenvolvimento do trabalho a análise de referências cruzadas, a partir do *download* dos documentos suspeitos que serão armazenados em um diretório e este, por fim, formará um repositório de documentos suspeitos. Ao final, será realizada a comparação entre os arquivos culminando na geração do relatório apresentando os indícios de plágio.

1.1 Motivação

A tecnologia está inserida em todas as áreas, dentre elas, a área de pesquisa e desenvolvimento de recursos para auxílio na melhoria da qualidade da educação. Portanto, a fim de prezar pela boa qualidade no desenvolvimento dos trabalhos, devem ser bem escritos, trazer novas contribuições e usufruir de fontes de pesquisa de qualidade. No entanto, para escrita de trabalhos, se faz necessária à utilização de referências para endossar as pesquisas realizadas, sendo um dos fatores primordiais o cuidado na utilização de referências de maneira correta, prezando pelo crédito ao autor das ideias ou frases utilizadas no decorrer do trabalho, como forma de manter a autenticidade das informações.

O controle da autenticidade não é uma tarefa banal. Por exemplo, quando se trata de turmas de educação à distância, um docente tem a responsabilidade da disciplina que ministra em mais de um polo, somando uma quantidade considerável de alunos, o que dificulta a

verificação manual e individualizada de indícios de plágio nos trabalhos. Segundo [NEIL, 2004], [SANTANA e MARTINS, 2003] o problema do plágio na educação a distância é facilitado devido ao grande número de materiais que são disponibilizados online, com isso propiciando a prática do plágio e dificultando a inibição e verificação da autenticidade dos trabalhos devido ao número de alunos.

Existem diversas ferramentas para auxílio na verificação de indícios de plágio dos trabalhos, entretanto, a grande maioria possui funcionalidades restritas de extensão de documentos ou ainda de licença de uso. Este trabalho justifica-se pelo fato de haver contribuições para serem desenvolvidas dentro da área de detecção de indícios de plágio textual, através do acréscimo de demais funcionalidades e novas formas de análise de similaridade estudadas e desenvolvidas no método DIP – Detector de Indícios de Plágio [PERTILE, 2011]. Outro fator que contribuiu para o desenvolvimento desta ferramenta é o estudo e os testes realizados nas ferramentas descritas no decorrer do presente trabalho, que possibilitou a identificação de ausência de algumas funcionalidades que serão desenvolvidas e citadas no decorrer deste texto. Por fim, a pesquisa culminará no desenvolvimento de uma nova ferramenta que verifica o percentual de indícios de plágio em diversas extensões de documentos, tais como: .pdf, .doc, .docx, e HTML, além da adoção de técnicas de *stemming* e busca e análise em arquivos disponíveis na Internet e no repositório de documentos que é criado no decorrer de cada análise.

1.2 Objetivos

O objetivo principal deste trabalho é verificar as "contribuições do desenvolvimento" de uma ferramenta de análise de indícios de plágio no controle da autenticidade dos trabalhos acadêmicos, culminando no desenvolvimento desta nova ferramenta, proporcionando assim, a análise de diferentes extensões de documentos, bem como, o aprimoramento do método já desenvolvido, DIP – Detector de Indícios de Plágio.

Para alcance do objetivo principal, este é permeado pelos objetivos específicos que seguem:

- Analisar materiais que referenciem problemas na verificação de indícios de plágio;
- Realizar um estudo do estado da arte das ferramentas para detecção de indícios de plágio disponíveis para uso e traçar um comparativo;

- Realizar um estudo sobre a técnica de *stemming* de busca e comparação de palavras similares;
- Aprimorar e desenvolver novas contribuições para o DIP – Detector de Indícios de Plágio;
- Validar o módulo em disciplinas de graduação e pós-graduação, em cursos presenciais e à distância.

1.3 Principais contribuições do trabalho

O referido trabalho tem como principais contribuições:

- Desenvolver uma nova ferramenta de auxílio na verificação de indícios de plágio textual;
- Trazer novas contribuições para o método do qual se embasou para criação da nova ferramenta, denominado DIP – Detector de Indícios de Plágio;
- Aplicar o *checklist* sugerido em [NUNES et. al] baseado na *ErgoList*[1] e a norma *ISO 9126*[2] para avaliação de usabilidade.

1.4 Organização do texto

Este trabalho de pesquisa está dividido em sete capítulos, no capítulo 2 consta a revisão bibliográfica, que aborda temas como caracterização e prática do plágio, ferramentas de análise e detecção de indícios de plágio e ferramentas afins com este trabalho, métodos de detecção de indícios de plágio.

No capítulo 3 é apresentada a metodologia do trabalho. No capítulo 4 consta a proposta de desenvolvimento da ferramenta, suas características, funcionalidades e modelagem, além da implementação da ferramenta proposta, a interface em execução. Já no capítulo 5 encontram-se a validação e os resultados obtidos a partir dos testes, além de uma breve avaliação de usabilidade.

No capítulo 6, localizam-se as conclusões e trabalhos futuros, e concluindo, no capítulo 7, encontram-se as referências bibliográficas.

---

[1] Disponível em: <http://www.labiutil.inf.ufsc.br/ergolist>.
[2] Disponível em: <http://www.abntcatalogo.com.br/norma.aspx?ID=2815>.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo estarão contemplados os termos e conceitos que serão utilizados no decorrer do desenvolvimento do trabalho. As abordagens irão contemplar: caracterização e prática do plágio; ferramentas de auxílio na detecção de indícios de plágio; motores de busca. Todos os itens relacionados serão descritos nas seções a seguir.

2.1 Caracterização e prática de Plágio

O texto *Code of Practice on Plagiarism* define plágio como a utilização das palavras ou ideias de outra pessoa como se fosse seu, e cita como exemplos de plágio: copiar, traduzir um texto de um idioma para outro, parafrasear ou referenciar incorretamente [HANDBOOK, 2009]. Ainda em LACKES *et. al*, [2009, *apud* MEGEHEE e APAKE, 2008] encontra-se a afirmação que em torno de 70% das obras não são referenciadas corretamente, o que por sua vez, também, configura o ato ilegal de plágio. Segundo [SILVA, 2008], o problema do plágio dentro do meio escolar, vem desde o ensino fundamental, onde se copia textos de outrem parcialmente ou totalmente sem referenciar a fonte.

Alguns exemplos que podem ser considerados como plágios são citados a seguir [HANDBOOK, 2009; OLIVEIRA e OLIVEIRA, 2008]:

- Citação: trata-se da cópia idêntica das palavras e ideias do autor, sem fazer referência ao autor e a obra.
- Paráfrase: o escritor do trabalho transcreve com suas palavras as ideias do autor que deveria ser referenciado, com a finalidade de torná-las um pouco distintas do original, e, por sua vez, essas palavras não são referenciadas, caracterizando plágio.
- Resumo: é uma paráfrase mais curta, porém, não segue somente as ideias do autor, o escritor também expõe as suas, contudo, não referencia o autor do qual utilizou para fundamentar o texto.
- Referência: este tipo de plágio ocorre quando não se referencia a obra original e sim paráfrases presentes em uma obra secundária utilizada para a formulação ou fundamentação de um texto. Por exemplo, para construção do texto A utilizou as obras B e C, as quais eram resumos da obra D, a obra D deveria ter sido referenciada

acompanhada das obras B e C. Este tipo de plágio ocorre quando se referencia resumos, paráfrase ou citações ao invés de referenciar a obra original.

Os tipos de plágio são definidos por KIRKPATRICK [2007, *apud* OLIVEIRA, 2007], e são subdivididos em:

- Plágio Direto: cópia de uma fonte por completa sem usar citações ou referenciar o autor.
- Referência Vaga ou Incorreta: como o próprio nome traz, esse tipo de plágio acontece quando uma referência é feita de maneira incorreta, ou seja, o escritor não informa o início e o fim da referência retirada da bibliografia.
- Plágio Mosaico: este tipo de plágio é um misto de paráfrases com citações, ou seja, o escritor muda algumas palavras do autor e reformula os parágrafos, porém não faz referência à fonte, o que caracteriza o plágio.
- Plágio Extra Corpal: cópia de textos fontes externas, que não a sua, mas de um grupo em que este sujeito faça parte também.

O trabalho desenvolvido nesta pesquisa faz análise de dois tipos de plágio: Plágio Mosaico e Extra Corpal.

O plágio é uma prática bastante frequente, principalmente quando se trata de trabalhos científicos no meio acadêmico. As causas apontadas são as mais diversas, entre elas, destacam-se o acesso mais facilitado às informações devido a grande quantidade de dados proporcionados pela Internet, o desconhecimento por parte dos alunos e a falta de orientação sobre o que configura plágio, a inexperiência na escrita e também a desonestidade intelectual ao copiar informações sem dar os devidos créditos aos autores. [MORAES, 2004], [PLAGIO.NET, 2012].

Mais alguns fatores que podem influenciar na prática do plágio foram elencados por Barbastefano [2007, *apud* CALDEIRA e RODRIGUES, 2012]:

- Venda de trabalhos prontos pela Internet;
- Incapacidade para parafrasear autores;
- Depreciação do trabalho por parte do aluno;

- Consciência equivocada que a informação disponível na Internet é de livre acesso e utilização de todos; [GARSCHAGEN, 2006]
- Pesquisa no ensino fundamental, comumente caracterizadas por cópias e colagens de páginas da Internet; [GARSCHAGEN, 2006]
- Plágio bilíngue, devido o fácil acesso a tradutores;
- Desconhecimento sobre regras, legislação e regulamentações delimitando plágio.

Segundo [FERRARESI *et. al*, 2008], os universitários utilizam como sua principal fonte de informação a Internet (Figura 2.1.1). Porém, muitos destes acadêmicos, mesmo estando em nível de formação, não dominam a escrita de trabalhos científicos e precisam de orientações para a adoção de metodologia científica correta.

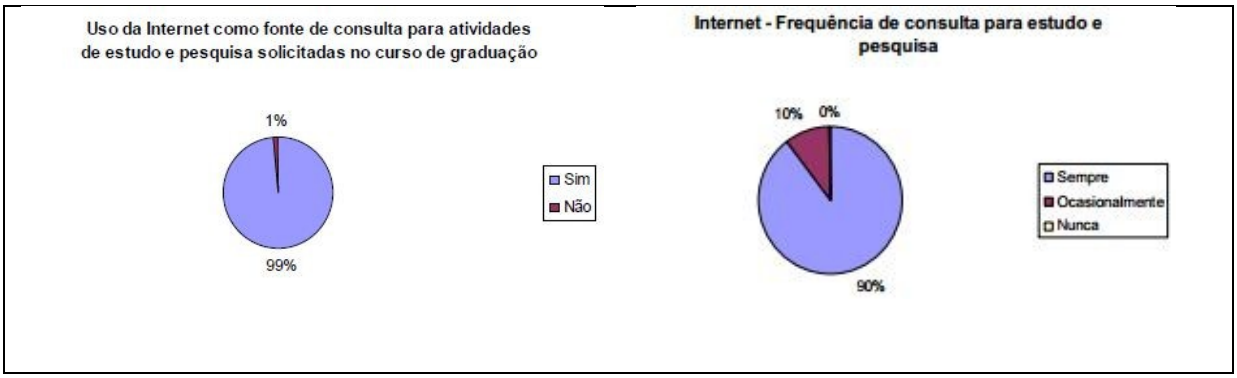

Figura 2.1.1 - Uso da Internet como fonte de pesquisa [FERRARESI *et. al*, 2008]

Em [FURTADO, 2012], encontra-se:

> Com o advento da *internet*, como já dissemos antes, e as extraordinárias facilidades que ela nos legou hodiernamente, essa situação se agravou, disseminando a ocorrência desses furtos virtuais. Deparamos-nos, então, com aquele plagiador que pratica a violação em proveito de si mesmo ou de outrem, sob encomenda, *comercializando* trabalhos acadêmicos prontos, maquiados pela leviandade de quem assim age. Mais do que um ilícito civil, uma vez que afronta direito de personalidade do autor, constitucionalmente garantido, atingindo a sua criação intelectual, nos deparamos também com um ilícito criminal gravíssimo, coberto ainda pela inteira reprovação moral a que se sujeita aquele que pratica o plágio. [FURTADO, p.01, 2012]

Conforme o que foi apresentando em **[**FERRARESI *et. al*, 2008], a Internet é a fonte de pesquisa mais utilizada, e a inexperiência para escrever seguida da falta de orientação, motiva o autor/aluno ir em busca de alternativas ou fontes de pesquisas que contenham as informações que esse sujeito busca. Este aspecto é endossado por [MORAES, 2004], que traz em seu texto a ideia de que a Internet potencializa a incidência do plágio, mas o responsável pelo ato do plágio é, sem dúvida, o ser humano, a Internet é apenas o instrumento de pesquisa, assim como outros instrumentos que estão disponíveis para uso (material impresso).

A dimensão do contexto de plágio também é enfatizada na mídia impressa e nos veículos de imprensa, conforme reportagem no [DIÁRIO DE CUIABÁ, 2012], que aborda que o ato de plagiar acaba enfraquecendo as pesquisas, pois muitas vezes, não são identificados no momento da finalização do trabalho, resultando em pesquisas e trabalhos duplicados. O texto ressalta ainda resalta que a facilidade de acesso a trabalhos na Internet de várias partes do país, dificulta o controle da autenticidade dos trabalhos por parte dos professores.

Segundo pesquisas realizadas por VASCONCELOS [2011] na Universidade Federal do Rio de Janeiro (UFRJ), dois fatores se destacam como sendo causadores do plágio: a facilidade de acesso às informações na Internet e o fator linguístico, ou seja, falta de desenvoltura e experiência para a escrita e insuficiência de conhecimento de uma língua estrangeira. A mesma pesquisa ainda aponta que a incidência de plágio triplicou entre a década de 1970 e 2007, tendo passado de menos de 0,25% para 1%.

Sendo assim, o fator autenticidade de pesquisas passou a receber uma atenção especial também do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), o qual criou uma comissão de controle de qualidade e autenticidade denominada de Comissão de Integridade na Atividade Científica do CNPq. Esta comissão desenvolveu diretrizes básicas para o controle de autenticidade da pesquisa científica. A comissão é formada por membros de diferentes áreas de pesquisa que avaliam manualmente os trabalhos e posteriormente submetem em softwares verificadores de indícios de plágio para apontamento das semelhanças textuais. Esse procedimento de análise manual tem como finalidade a identificação e comprovação do plágio, visto que o que pode ser apontado como plágio pelos softwares que analisam semelhanças entre trabalhos, e este pode não ser caracterizado como tal, devido às semelhanças descritivas da elaboração do trabalho. Após a dupla análise – manual e *software* – a comissão tipifica a conduta dos autores (falsificação, fabricação de

resultados, plágio e inclusão de autores sem colaboração intelectual no trabalho) e aplica as devidas penalidades. [CNPq, 2012]

Portanto, para auxílio na detecção de indícios de plágio é importante contar com ferramentas computacionais, como as que serão apresentadas no capítulo 3, as quais também foram estudadas em [LIMA e RESENDE, 2012].

2.2 Técnica de *Stemming*

A técnica de *Stemming trata-se* de uma metodologia de extração de radical das palavras, esta técnica possibilita a comparação mais detalhada entre palavras que contenham o mesmo radical, como por exemplo: Palha e Palhaço, Dúvida e Duvidamos. [ORENGO; HUYCK, 2001].

Segundo [XAPIAN, 2013], apenas palavras que compõem os idiomas dinamarquês, holandês, inglês, finlandês, francês, alemão, húngaro, italiano, norueguês, português, romeno, russo, espanhol, sueco e turco podem passar pelo processo de *stemming*. Estes idiomas possuem estrutura para estudo morfológico das palavras (regras), proporcionando a remoção correta de seus radicais.

No estudo morfológico das palavras são encontradas regras de remoção dos radicais. A remoção incorreta dos radicais, ou seja, a aplicação da técnica de *stemming* erroneamente, pode ocasionar a interpretação inexata das palavras, culminando na comparação de similaridade defeituosa.

Os motores de buscas utilizados na pesquisa por arquivos com suspeita de indícios de plágio tratam palavras com o mesmo radical como sendo sinônimos, não fazendo distinção entre elas.

Para a utilização desta técnica, contou-se com os recursos da biblioteca *Java Lucene,* a qual já possui funções pré-definidas que fazem o processo de *steminng.*

3 FERRAMENTAS AUXILIARES NA DETECÇÃO DE INDÍCIOS DE PLÁGIO

Para o desenvolvimento deste trabalho, foram estudadas 15 ferramentas de auxílio na verificação e detecção de indícios de plágio, com o intuito de levantar algumas características como:

a) Tipo de licença (livre ou paga);
b) Formato de apresentação e execução da ferramenta para o usuário (Web ou Desktop);
c) Tipos de extensões de arquivos analisados;
d) Tamanhos de arquivos suportados para análise;
e) Necessidade de cadastro para utilização;
f) Possibilidade de integração com ambientes virtuais de aprendizagem;
g) Geração e apresentação de relatório dos resultados da análise detalhando os indícios de plágio.

As ferramentas estudadas e/ou testadas foram: Araponga [SANTOS e FRANCO, 2010], DIP – Detector de Indícios de Plágio [PERTILE, 2011], DOCCOP [DOCCOP, 2012], Ephorus [EPHORUS, 2012], Etblast [ETBLAST, 2012], Farejador de Plágio [FAREJADOR, 2012], Plagiarism Detect [PLAGIARISM.ORG, 2012], Plagiarisma [PLAGIARISMA, 2012], Plagium – Online [PLAGIUM, 2012], Plagius Detector [PLAGIUS, 2012], ScholarOne [SCHOLARONE, 2012], Turnitin [TURNITIN, 2012], Urkund [URKUND, 2012], VIPER [VIPER, 2012]. Todas estas serão descritas nas próximas seções. O critério adotado para a seleção das ferramentas foi sua utilização na academia e a disponibilidade de informações em trabalhos já desenvolvidos em sites dos fabricantes, além da possibilidade de testes em algumas versões das ferramentas disponíveis.

3.1 Ferramentas estudadas

Nesta seção serão descritas as ferramentas estudadas para subsidiar o desenvolvimento deste trabalho. Para melhor entendimento, estas foram classificadas em dois grupos: ferramentas versão gratuita e versão paga.

3.1.1 Ferramentas gratuitas

As ferramentas gratuitas descritas neste trabalho em sua grande maioria foram testadas. Somente as ferramentas que não disponibilizam versões de testes é que foram analisadas a partir de informações bibliográficas coletadas de trabalhos já desenvolvidos ou através dos sites dos fabricantes.

3.1.1.1 Araponga

Ferramenta disponível para uso na Internet, requer cadastro para utilização. Desenvolvida na Universidade Federal de Itajubá – Minas Gerais, tem como sua principal característica a integração com o Ambiente Virtual de Aprendizagem (AVA) – *TelEduc.* Este AVA possibilita a criação e o acompanhamento de cursos de educação à distância ou como ferramenta auxiliar para cursos presenciais. A principal funcionalidade do Araponga é a verificação de indícios de plágio nos trabalhos submetidos pelos alunos no AVA. Araponga é ilustrado na Figura 3.1.1.1-1 [SANTOS e FRANCO, 2010]

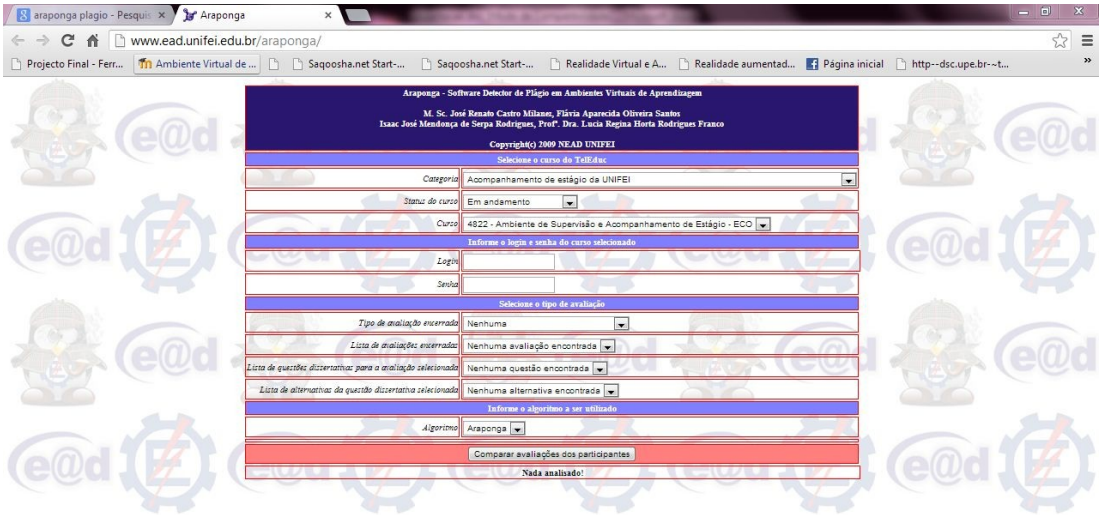

Figura 3.1.1.1-1 - Araponga [ARAPONGA, 2012]

A documentação da ferramenta não traz informações sobre tipos de plágio que detecta, e se é essencialmente interligada ao Teleduc. Estas características podem ser consideradas itens limitadores já que não são especificados na documentação da ferramenta. A instalação e testes da ferramenta foram impossibilitados devido à falta de disponibilidade da mesma para este fim.

3.1.1.2 DIP – Detector de Indícios de Plágio

Esta ferramenta pode ser utilizada tanto na Web, através da integração com o Moodle ou no desktop. Analisa arquivos com extensões .doc. Outra funcionalidade é a possibilidade de integração dessa ferramenta com o MLE – Moodle (*Mobile Learning Engine*), ambiente virtual de aprendizagem móvel [PERTILE, 2011]. Gera relatório ao final da análise contendo percentual de indícios de plágio por parágrafo bem como os endereços da Internet que contém o material suspeito em relação ao original (submetido para análise). Como esta ferramenta foi utilizada como base para o desenvolvimento deste trabalho, será melhor detalhada no decorrer do texto.

3.1.1.3 DOCCOP

Esta é uma ferramenta disponível para livre utilização, diretamente no navegador de Internet. Analisa documentos com extensões .doc e .pdf., em sua documentação não é esclarecido se há possibilidade de integração com ambientes virtuais de aprendizagem.

A quantidade de caracteres do documento que será analisado é limitada. O documento que será submetido à análise deve ser texto puro, e cabe ao usuário extrair este texto para ser submetido à ferramenta, através dos comandos de Copiar e Colar (*CTRL C e CTRL V*). No entanto, todas as imagens devem ser removidas, e qualquer tipo de arquivo digitalizado não é aceito na ferramenta. Em outras palavras, uma parte do pré-processamento do texto é feito de forma manual pelo usuário da ferramenta, o que exige que o usuário tenha conhecimento prévio sobre essa técnica, além de maior disponibilidade de tempo, acarretando na dificuldade de utilização. O relatório final é enviado via e-mail, mas esta etapa não é executada em tempo real, pois o envio de relatório obedece a uma fila de análise. A documentação da ferramenta não apresenta tempo limite de resposta, sendo que nos testes realizados, a demora foi de 120

minutos para um artigo de dez páginas. Na Figura 3.1.1.3-1, apresenta-se a interface do DOCCOP [DOCCOP, 2012].

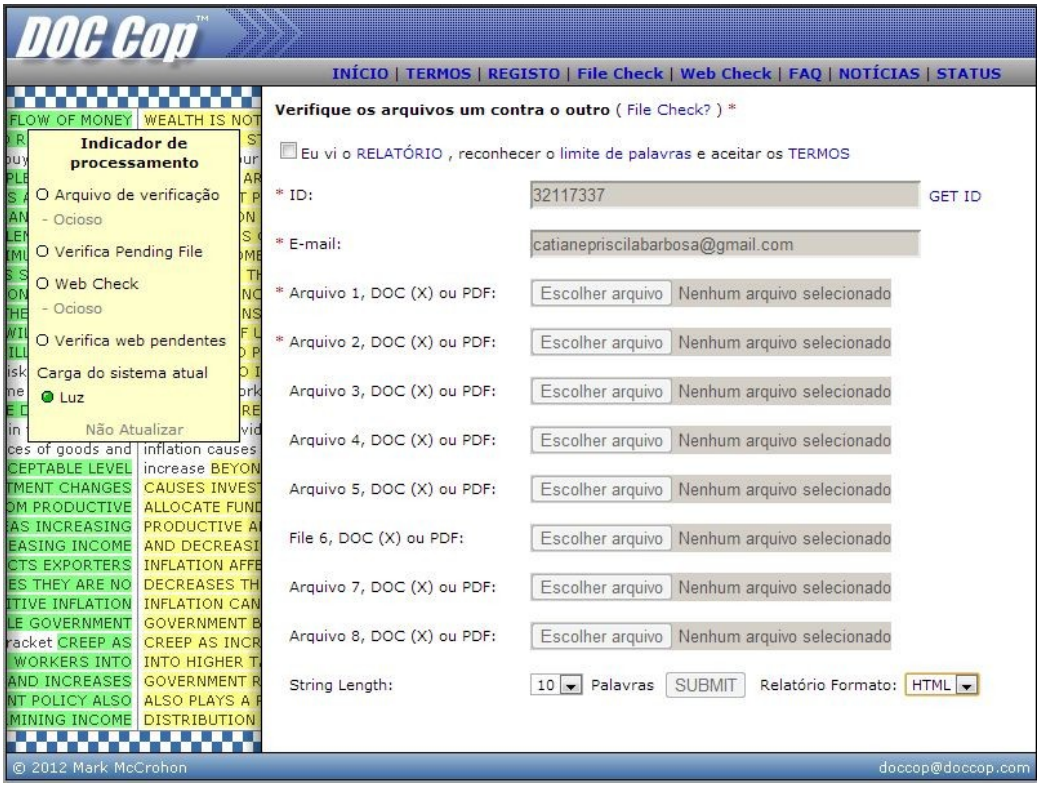

Figura 3.1.1.3-1 - Interface do DOCCOP [DOCCOP, 2012]

3.1.1.4 EtBlast

EtbBlast é uma ferramenta disponível na Web, que possibilita a análise de documentos produzidos especificamente na área de saúde. Faz a análise e comparação de similaridade dos textos em relação a maior base de documentos da medicina, denominada *MEDLINE*. Está em constante atualização devido o número de pesquisas desenvolvidas na área. Sua documentação não especifica tipos de extensão de arquivos que analisa e se há possibilidade de integração com ambientes virtuais de aprendizagem. Esta ferramenta está mantida em desenvolvimento dentro do Instituto de Bioinformática da Virgínia – Estados Unidos. [ETBLAST, 2012]. Apesar de realizar as pesquisas na maior base de dados da área da Saúde, a EtBlast apresenta algumas desvantagens, como por exemplo, a análise apenas de trechos de texto puro, ocasionando ao usuário a preocupação de trabalhar com o pré-processamento textual, removendo itens incompatíveis com a ferramenta (figuras, excesso de texto, excesso de espaços em branco). A EtBlast além de ser uma ferramenta de análise de indícios de plágio, também possibilita o desenvolvimento de novas aplicações a partir da utilização de

uma *API* de desenvolvimento relacionada a ela. Na Figura 3.1.1.1 - 4, é ilustrada a ferramenta descrita.

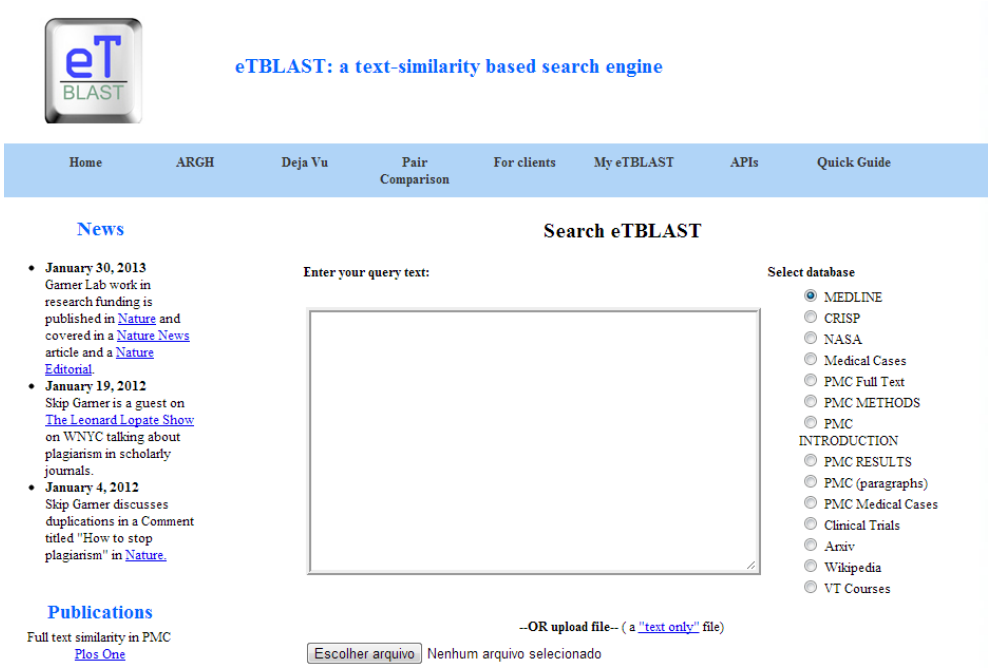

Figura 3.1.1.4-1 - EtBlast [EtBlast, 2012]

3.1.1.5 Farejador de Plágio

O Farejador de Plágio é um dos recursos mais populares atualmente. A ferramenta é disponibilizada em duas versões de licença, livre ou paga. Sob licença livre, faz análise de apenas 50% do arquivo submetido, e com tamanho máximo de 300Kb, suporta extensões .doc e .rtf. Segundo [PERTILE, 2011] é uma ferramenta que demanda um bom tempo de execução para análise dos materiais submetidos, ou seja, o custo de processamento é alto, sendo que analisa em torno de 30 a 40 páginas por hora. [FAREJADOR, 2012]. Na Figura 3.1.1.1- 5.1, é ilustrada a ferramenta descrita.

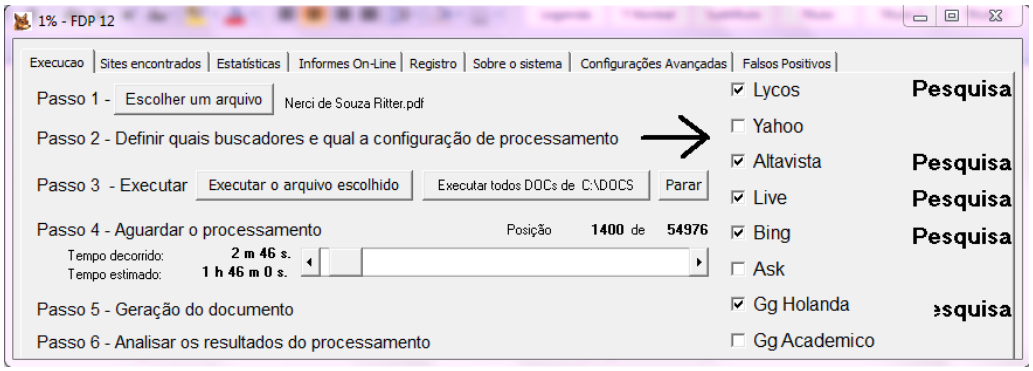

Figura 3.1.1.5-1 - Farejador de Plágio [FAREJADOR, 2012]

3.1.1.6 Plagiarisma

Ferramenta disponível na Web, gratuita, com número limite na quantidade de palavras para análise. Suporta arquivos com extensão .doc, sendo que, quando submetida extensão diferente, não apresenta erros nem resultados, ou seja, não apresenta *feedback* consistente para o usuário, o que poderá ocasionar desperdício de tempo na espera pelo resultado.

Plagiarisma não possibilita a integração com ambientes virtuais de aprendizagem, não requer cadastro para utilização e, ao final da análise, traz relatório com percentual de indícios de plágio [PLAGIARISMA, 2012]. A Figura 3.1.1.6-1 apresenta a interface da ferramenta.

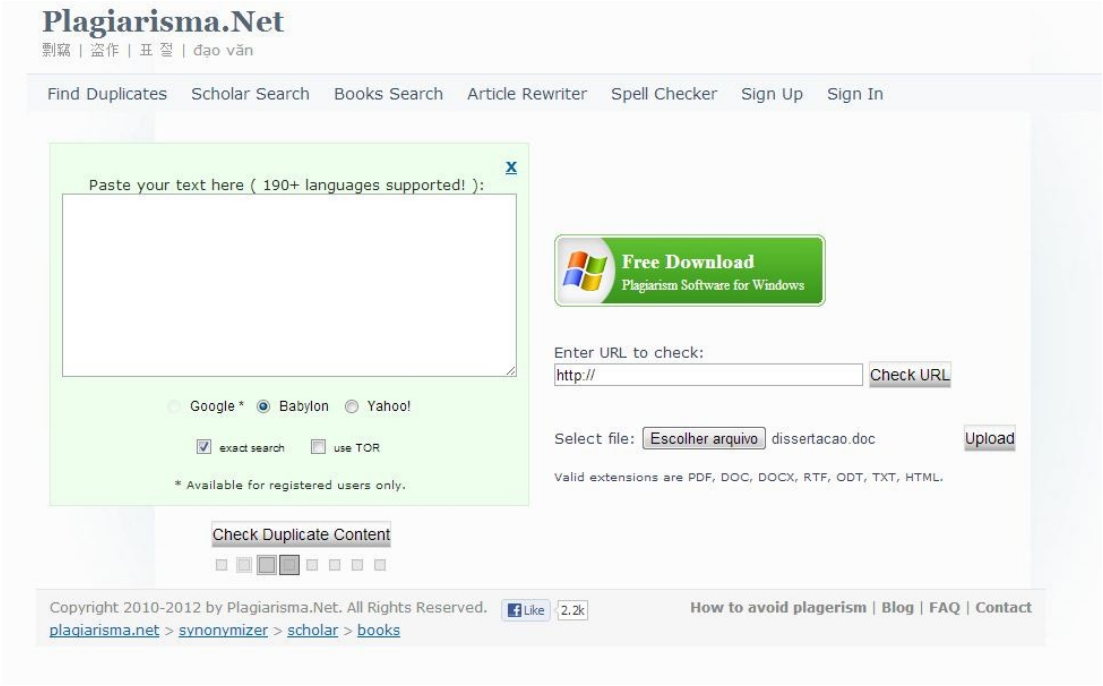

Figura 3.1.1.6-1 - Interface da ferramenta Plagiarisma [PLAGIARISMA, 2012]

3.1.1.7 Plagium – Online

A ferramenta Plagium – Online está disponível para utilização diretamente no navegador Web. Dispõe de duas opções de envio de textos para análise do percentual de indícios de plágio: a partir da submissão de, obrigatoriamente, pequenos trechos textuais, sendo que não aceita submissão de arquivos completos; ou envio de texto disponível na Internet através da *URL* para verificação de indícios de plágio. O fator limitante desta ferramenta é justamente o tamanho de arquivos suportados, limitando-se a pequenos trechos textuais ou textos já disponibilizados na Internet. Plagium – Online não possibilita a integração com ambientes virtuais de aprendizagem e não requer cadastro para utilização. A ferramenta apresenta relatório após a análise. [PLAGIUM, 2012]. A Figura 3.1.1.7-1 apresenta a interface do Plagium.

Figura 3.1.1.7-1 - Interface do Plagium [PLAGIUM, 2012]

3.1.1.8 Plagius Detector

Plagius Detector dispõe de dois tipos de licença, gratuita e paga. Sendo que, a gratuita, opera com limitação de verificação de 50% do arquivo submetido. Esta ferramenta está em processo de atualização, o que visa possibilitar a análise e comparação de arquivos que estejam armazenados dentro do computador pessoal do usuário, já que, até então, a análise se dava a partir de buscas na Internet. O Plagius Detector é uma ferramenta desktop e aceita diversos tipos de extensões de arquivos, tais como: doc, .pdf, .rtf, .HTML. Na versão paga, o Plagius Detector permite que o usuário configure suas preferências de pesquisa, limitando o número de palavras que serão pesquisadas, tamanho das frases, profundidade da verificação e número de varreduras no documento. Nos testes realizados e nos materiais pesquisados, não foram encontradas referências de possibilidade de integração com ambientes virtuais de aprendizagem. Esta ferramenta também apresenta relatório ao concluir a análise. [PLAGIUS, 2012]. A Figura 3.1.1.8-1 ilustra a interface do Plagius Detector.

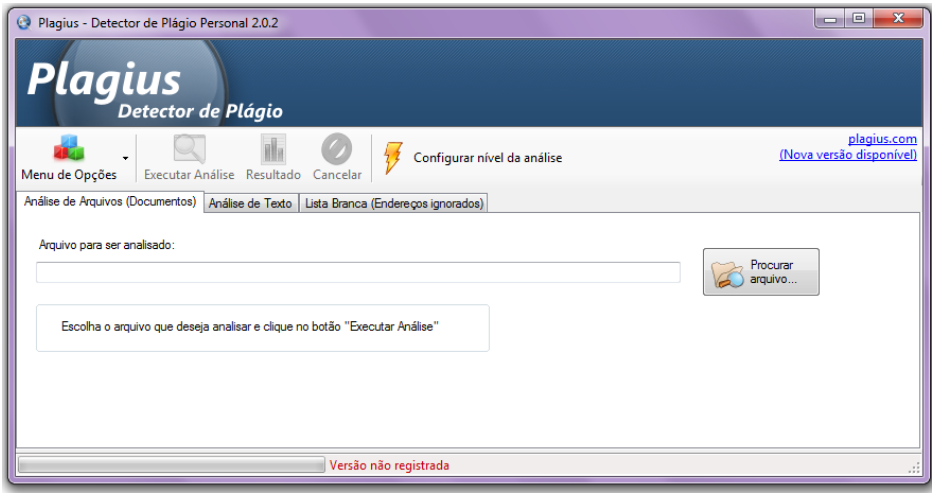

Figura 3.1.1.8-1 - Interface do Plagius [PLAGIUS, 2012]

3.1.1.9 VIPER

VIPER é uma ferramenta de auxílio na verificação e detecção de indícios de plágio, é desktop, e suporta extensões de arquivos .doc, .rtf, .html e .txt. A análise textual se dá através do processo de envio de sentenças para buscas na Internet, sendo que ao final do processo, é gerado um relatório em forma de tabela, contendo o percentual de indícios e a lista de URLs que apresenta material semelhante ao texto. Requer cadastro para sua utilização e não apresenta recursos para a integração com ambientes virtuais de aprendizagem. [VIPER, 2012]. Na Figura 3.1.1.9-1 é ilustrada a interface do VIPER.

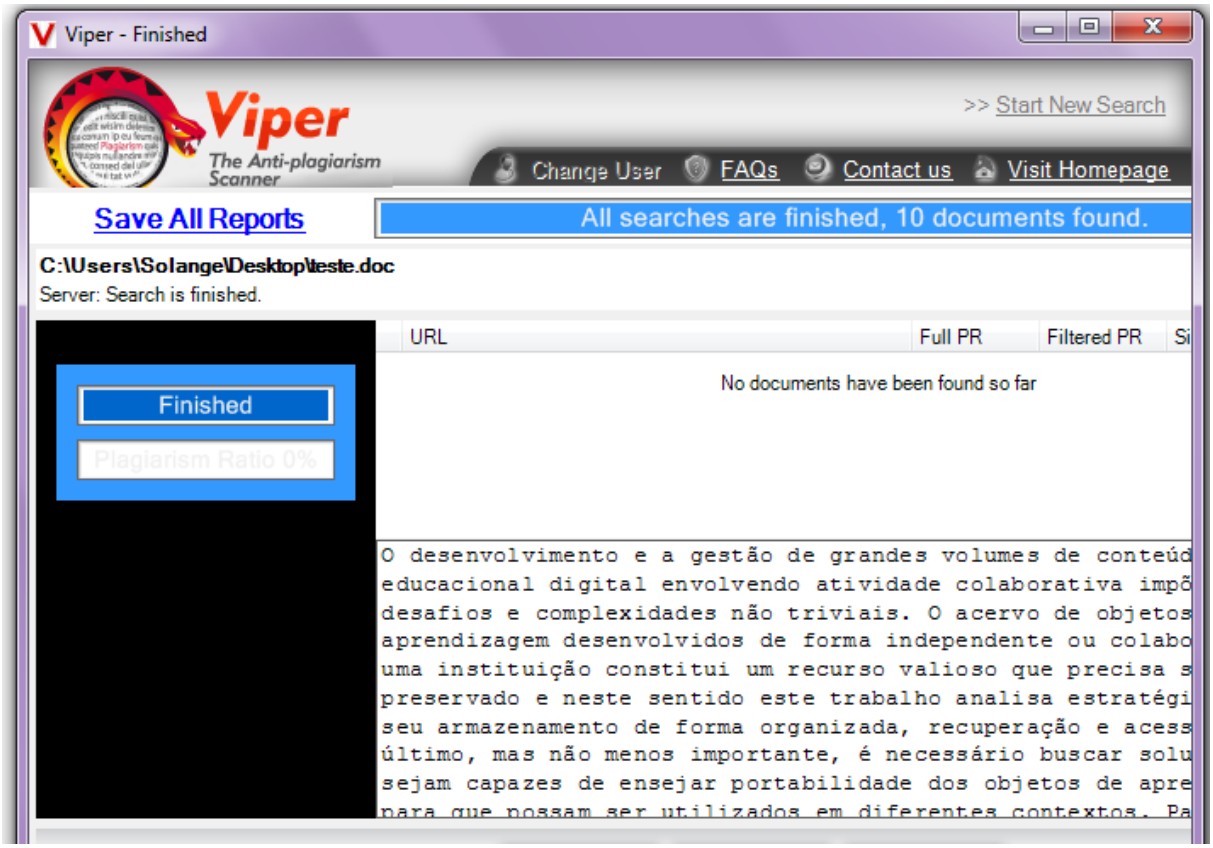

Figura 3.1.1.9-1 - Interface do VIPER [VIPER, 2012]

3.1.2 Ferramentas pagas

As ferramentas pagas descritas neste trabalho, em sua grande maioria, foram estudadas bibliograficamente, sendo que somente as ferramentas que também disponibilizavam de versões gratuitas é que foram testadas. A seguir, é apresentada a descrição das ferramentas pagas.

3.1.2.1 Ephorus

Ferramenta sob licença paga, funciona integrada a sites institucionais de escolas ou empresas que trabalham com o ramo de publicações. No momento que o autor envia sua produção textual pelo site institucional, esta já é submetida à análise de indícios de plágio. Este processo de envio para análise é invisível para o usuário. A análise é realizada a partir de buscas na Internet e também em seu grande acervo virtual, que é criado a cada nova submissão de arquivo (criação de repositório de documentos). Não foram encontradas evidências de integração com ambientes virtuais de aprendizagem. [EPHORUS, 2012].

Algumas de suas características podem ser apontadas como sendo desvantajosas em sua utilização, como por exemplo, ser licenciada e funcionar somente integrada ao site institucional, visto que, se esta página da Web estiver indisponível, a ferramenta também estará.

3.1.2.2 Plagiarism Detect

Ferramenta de análise de indícios de plágio desktop. Suporta extensões de arquivos .doc, .docx, .pdf, .rtf, .html, .ppt e .txt. A análise de indícios de plágio é feita a partir de buscas na Internet. O tempo de processamento de arquivos, segundo o desenvolvedor da ferramenta, é em torno de 120 segundos para três ou quatro páginas de texto. [PLAGIARISM.ORG, 2012]. A descrição desta ferramenta foi embasada na sua documentação, uma vez que não foi possível a realização dos testes para comprovação do tempo de processamento da análise, por se tratar de uma ferramenta paga. A ferramenta também apresenta relatório ao final da análise, e não pode ser integrada com ambientes virtuais de aprendizagem.

3.1.2.3 ScholarOne

É uma ferramenta direcionada para as mais diversas áreas de publicações, para análise de autenticidade de textos breves até livros completos. A aquisição da ferramenta é modulada, ou seja, o usuário adquire uma licença da ferramenta para análise de cada tipo de texto separadamente, livros, *abstracts*, jornais. Utilizada por instituições famosas, tais como IEEE - *Institute of Electrical and Electronics Engineers*, Universidade de Cambridge, Universidade de Oxford. [SCHOLARONE, 2012]. Em sua documentação da ferramenta não é descrito as extensões de arquivos analisadas, nem como é apresentado o resultado da análise de indícios de plágio. Não traz evidências de integração com ambientes virtuais de aprendizagem.

3.1.2.4 Turnitin

Ferramenta que pode ser usada tanto W*eb* quanto desktop. Possibilita também a integração com ambientes virtuais de aprendizagem. A análise de indícios de plágio é feita através de buscas na Internet e em um banco de dados contendo mais de 250 milhões de publicações, sendo que o processo de análise adota como parâmetro a identificação de

similaridade de palavras e/ou sentenças completas. Turnitin não especifica os tipos de extensões de arquivos e formato de apresentação dos resultados das análises. [TURNITIN, 2012]

3.1.2.5 Urkund

Esta ferramenta de verificação e detecção de indícios de plágio realiza a análise na Internet e em seu próprio banco de dados e pode ser integrada a ambientes virtuais de aprendizagem. O diferencial dessa ferramenta atribui-se ao fato de estar integrada ao e-mail profissional do docente, sendo que os trabalhos acadêmicos que forem enviados para este, antes passam pela verificação de autenticidade para, posteriormente, ser enviado ao professor. Caso o trabalho apresente indícios de plágio, é retornado ao discente para reformulação do texto. Por outro lado, se o texto não apresentar indícios de plágio, o docente recebe a atividade do aluno em sua caixa de entrada de e-mails. Os tipos de extensões de arquivos não são descritas em sua documentação. [URKUND, 2012]

Uma síntese das ferramentas estudadas, com funcionalidades e licenças, é apresentada no Quadro 3.1-1.

Ferramenta	Gratuita	Paga	WEB	Desktop	Extensões suportadas	Cadastro	Integrada AVAs	Apresenta relatório
Araponga [SANTOS e FRANCO, 2010]	Sim	Não	Sim	Não	Ø	Sim	Sim	Sim
DIP – Detector de Indícios de Plágio [PERTILE, 2011]	Sim	Não	Sim	Sim	.doc	Não	Sim	Sim
DOCCOP [DOCCOP, 2012]	Sim	Não	Sim	Não	.doc e .pdf	Não	Ø	Sim
Ephorus [EPHORUS, 2012]	Não	Sim	Sim	Ø	Ø	Sim	Sim	Sim
Etblast [ETBLAST, 2012]	Sim	Não	Sim	Não	Ø	Ø	Ø	Sim
Farejador de Plágio [FAREJADOR, 2012]	Sim	Sim	Não	Sim	.doc e .rtf	Sim	Ø	Sim
Plagiarism Detect [PLAGIARISM.ORG, 2012]	Sim	Sim	Sim	Sim	Ø	Sim	Ø	Sim
Plagiarisma [PLAGIARISMA, 2012]	Sim	Não	Sim	Não	Ø	Não	Ø	Sim
Plagium – Online [PLAGIUM, 2012]	Sim	Não	Sim	Não	.txt	Não	Não	Sim
Plagius Detector [PLAGIUS, 2012]	Sim	Sim	Não	Sim	.doc, .pdf, .rtf, .HTML, .txt	Sim	Ø	Sim
ScholarOne [SCHOLARONE, 2012]	Não	Sim	Ø	Ø	Ø	Ø	Ø	Sim
Turnitin [TURNITIN, 2012]	Não	Sim	Sim	Sim	Ø	Não	Sim	Sim
Urkund [URKUND, 2012]	Não	Sim	Sim	Ø	Ø	Ø	Sim	Sim
VIPER [VIPER, 2012]	Sim	Não	Sim	Sim	Ø	Sim	Ø	Sim

Legenda: Ø = não foram encontradas informações.

Quadro 3.1-1 - Ferramentas para detecção de indícios de plágio adaptado de [SIBI, 2011] e [PERTILE, 2011]

Após realizar o levantamento e o estudo das ferramentas que foram tratadas nesse capítulo, pôde-se observar que as características apontadas inicialmente como sendo critérios de avaliação ainda requerem melhorias. Identificaram-se como possibilidade de melhorias os seguintes fatores: diversificação nas extensões de arquivos analisados; otimização do tempo de análise; possibilidade de acesso aos arquivos que contenham alguma relação de plágio com o arquivo submetido através da criação de repositório de arquivos por análise; e melhoria na

qualidade da análise a partir da técnica de *stemming*. Também se identificou que as ferramentas, em sua grande maioria, requerem aquisição de licença por parte do usuário para que esta seja explorada em sua totalidade, como é o caso das ferramentas Farejador de Plágios, Plagius, e Plagiarism Detect. As três ferramentas citadas possuem versões de licença gratuitas, porém, limitam o tamanho máximo do arquivo submetido para análise de acordo com seu tipo de licença utilizada.

Por fim, as ferramentas estudadas não apresentam em sua documentação qualquer informação sobre a especificação da metodologia de análise de similaridade, sendo que a alternativa encontrada para este processo na bibliografia é a de Mineração de textos a partir do auxílio de motores de buscas na *Web,* os quais já trazem em seu desenvolvimento algoritmos que realizam o pré-processamento de similaridade textual, como por exemplo, o Google. [MORAIS, AMBRÓSIO, 2007]

3.1.3 Motores de busca

Motores de busca são softwares robôs que otimizam a consulta de milhares de fontes de informações que estão armazenadas em centenas de centros de dados espalhados pelo mundo [CENDÓN, 2001]. A busca pelos dados pode ser realizada a partir de palavras chave ou linguagem natural. As interfaces dos motores de busca caracterizam-se por serem páginas em HTML, como por exemplo, nos casos do Google e AltaVista. Segundo estudos apresentados por [CENDÓN, 2001], no ano de 2001, o Google era o motor de buscas mais requisitado pelos usuários, com cerca de 560 milhões de páginas e 56% de acessos.

Em nova pesquisa realizada em janeiro de 2012 pelo site espanhol [DESARROLLO, 2012], o *ranking* apresentou que o Google permanece em primeiro lugar com 82,68%, seguido do Yahoo com 5,82%, do motor de buscas chinês *Baidu* com 5,73% precedido pelo *Bing Microsoft* com 3,91% de usuários. Já em setembro de 2012 uma nova pesquisa foi realizada e publicada pela empresa desenvolvedora de páginas W*eb* [VenTICS, 2012] e mostra que ocorreram modificações nas posições do *ranking*, permanecendo o topo com o Google com 84,39% , em seguida o Yahoo com7,58%, precedido pelo *Bing* com 4,33% e, por fim, o *Baidu* com 1,74% de usuários.

Apesar de todos os motores de buscas apresentarem um número considerável de páginas armazenadas em seus bancos de dados, e essas estarem abarrotadas de informações, cada

motor de buscas retornará um resultado diferente de pesquisa quando utilizado, pois cada um possui seu algoritmo próprio, possibilitando as pesquisas de forma diferenciada em seu banco de dados.

As análises de indícios de plágio nas ferramentas estudadas utilizam-se de técnicas de mineração de texto a partir dos algoritmos inseridos dentro dos motores de buscas, sendo que, os arquivos empregados para comparação em relação ao arquivo suspeito, são originados do resultados da pesquisa realizadas por esses motores. Segundo [HISTÓRIA 2011 *apud* SENA, 2011], o Google permanece em primeiro lugar no ranking de motores de buscas, e atualmente possui em torno de um bilhão de páginas da Internet recheadas de conteúdos.

A partir dos estudos realizados no decorrer deste trabalho e a constatação de que ainda existem melhorias para serem desenvolvidas levando-se em consideração as ferramentas testadas e as estudadas, busca-se o desenvolvimento de uma nova ferramenta denominada *Miss Marple,* a qual possibilita a análise de indícios de plágio em arquivos com extensões. doc, .docx, .pdf e HTML. As análises são realizadas através de pesquisas na Internet a partir da *API* de pesquisa do *Google* - API *Google Search Ajax,* adoção de técnicas de *stemming* usando a biblioteca *Lucene* - JAVA para indexação dos termos e também análise de referência cruzada em um repositório de arquivos. A partir da nova ferramenta desenvolvida, este trabalho também contribuiu e aprimorou o desenvolvimento da ferramenta DIP – Detector de Indícios de Plágio [PERTILE, 2011], iniciando pelo levantamento das funcionalidades até então desenvolvidas (extensão de arquivos, metodologia de análise, tempo de processamento) e objetivando as novas contribuições que então foram acrescentadas (análise de uma diversidade de extensões de arquivos, criação do repositório de textos suspeitos e melhoria na qualidade de análise de similaridade de termos a partir da adoção de técnicas de *stemming*).

4 METODOLOGIA DO TRABALHO

Para o desenvolvimento da proposta do trabalho, o processo iniciou com o estudo do trabalho desenvolvido no Grupo de Redes e Computação Aplicada da Universidade Federal de Santa Maria, denominado de Método DIP [PERTILE, 2011]. Em seguida, foram realizados levantamentos das funcionalidades desenvolvidas e das novas contribuições que poderiam ser acrescentadas. Este processo foi feito a partir de estudos sobre o método e reuniões informais com a autora.

Na sequência, com o objetivo de atualização bibliográfica, partiu-se para o levantamento e testes das ferramentas atualmente utilizadas para auxiliar na detecção de indícios de plágio. Nessas ferramentas foram analisadas algumas características, como por exemplo: licença, extensões de arquivos submetidos para análise e tamanho máximo dos arquivos. As ferramentas utilizadas foram que estavam disponíveis na Internet para uso, ou que apresentavam versões gratuitas. Para as demais, foram utilizados de dados de trabalhos já desenvolvidos e também através de informações dos sites dos fabricantes. As ferramentas estudadas são descritas no Capítulo 3.

Após o estudo das ferramentas, partiu-se para a modelagem, utilizando a linguagem *UML* (Linguagem de Modelagem Unificada - *Unified Modeling Language*), com o intuito de definir as ações da nova ferramenta e as ações do usuário. Prosseguindo com o projeto de desenvolvimento da ferramenta, partiu-se para a definição do ambiente de desenvolvimento - *IDE* (*Integrated Development Environment*), na qual se optou pelo *NetBeans 7.1.2*, e a linguagem de programação adotada, JAVA. A escolha do ambiente e da linguagem se deu devido à gratuidade da *IDE* e ainda, pelo fato de a linguagem de programação possibilitar o desenvolvimento de projetos de software multiplataformas, ou seja, que podem ser executados com sucesso em diferentes sistemas operacionais, além de possibilitar a integração de bibliotecas prontas que otimizam o desempenho dos programas desenvolvidos.

Também com vistas ao desenvolvimento da ferramenta utilizou-se a biblioteca JAVA *Lucene* com a finalidade de potencializar fases de pré-processamento textual e valer-se de suas funcionalidades de *stemming* de palavras. Além das bibliotecas, recorreu-se ao uso da *API* de pesquisa *Google Search Ajax*.

As bibliotecas e API utilizadas estão descritas abaixo:

- *PDF Box*: biblioteca que trabalha com arquivos pdf, geração do relatório;
- *POI*: possibilita o trabalho com arquivos do MS Word;
- *HTTP CORE*: biblioteca que trabalha com arquivos HTML;
- *DOCX4J*: possibilita o trabalho com a nova extensão de arquivos do Word, .docx;
- *APACHE PDF BOX*: biblioteca que possibilita o trabalho com arquivos .pdf da disponíveis na Internet.
- *COMMONS MATH*: biblioteca utilizada para realizar os cálculos de percentual de similaridade entre os arquivos.
- *Google - API Google Search Ajax*: API de buscas do Google, que procura arquivos suspeitos com termos similares na Internet.

No decorrer deste trabalho desenvolveu-se uma ferramenta para verificação de indícios de plágio textual, o tipo de plágio abordado nessa pesquisa é o plágio extra corpal (cópia de fontes externas) e mosaico (cópia parcial de alguma fonte, com troca apenas de algumas palavras do texto).

Os hardwares utilizados no decorrer no desenvolvimento e testes foram um Notebook (Sistema Operacional *Microsoft Windows 7 - Service Pack 3* - Intel (R) *Core 2 Duo* (R), 4Gb de RAM), dois *Notebooks* (Sistema Operacional *Microsoft Windows 7 - Service Pack 3* - Intel i3, 4Gb de RAM), dois *Notebooks* (Sistema Operacional *Ubuntu* - Intel i3, 4Gb de RAM), um Computador Pessoal *Pentium 4* (Sistema Operacional *Microsoft XP* – 2GB de RAM).

O processo de validação do trabalho deu-se a partir da utilização da ferramenta desenvolvida em um curso de graduação presencial e um curso de pós-graduação à distância. Foram selecionados alguns artigos (oito de cada modalidade de curso) aleatoriamente submetidos para análise na ferramenta – *Miss Marple* – desenvolvida, além de dois artigos montados sem qualquer indício de plágio.

Com a finalidade de traçar um comparativo de tempo de processamento, e qualidade de análise, os mesmos artigos foram submetidos em outras três ferramentas (Farejador de Plágio, Plagius Detector e VIPER). A determinação das ferramentas, que compuseram o levantamento das comparações, foram selecionadas a partir de pesquisas realizadas em bibliografia tais como [PERTILE, 2011], [LIMA, 2011] e [USP, 2013] que apresentam algumas das características de cada uma dessas ferramentas suas disponibilidades de licença (versão gratuita), além de serem as mais conhecidas popularmente na academia [USP, 2013] e pelos usuários deste tipo de software.

Por fim, a interface da nova ferramenta desenvolvida – *Miss Marple* – foi avaliada por um público de 20 usuários, de nível superior, com conhecimentos básicos de informática, que utilizaram-se do *checklist* proposto por [NUNES *et. al*, 2012] que baseia-se nas fontes *ErgoList*[3] e a norma ISO 9126[4]. Essa proposta de avaliação de interface que se fundamenta nas fontes citadas, tem como objetivo investigar o atendimento de requisitos de usabilidade de software.

Os usuários que avaliaram a ferramenta foram distribuídos em dois grupos, sendo que no decorrer do processo de desenvolvimento da ferramenta, esta foi avaliada por um grupo de quatro discentes de um curso de pós-graduação, nível de mestrado, e que são da área da computação. Esses usuários fizeram apontamentos de melhorias na ferramenta, tais como qualidade de interface, manual do usuário e consistência dos resultados. Essa avaliação não foi descrita neste trabalho por não possuir caráter validativo.

Os usuários que validaram a ferramenta, para fins de levantamento dos dados para este trabalho, perfazem um público de 25 alunos, de diversos cursos de graduação, sendo que em sua grande maioria eram de cursos não relacionados com computação, possuem conhecimentos básicos de informática. Os alunos participaram da avaliação de espontânea vontade, a ferramenta lhes foi apresentada, bem como suas funcionalidades, e foram disponibilizados arquivos para testes. A escolha dos arquivos que foram submetidos para análise se deu pelos usuários que testaram a ferramenta.

A validação da ferramenta consistiu em submeter oito artigos desenvolvidos por alunos de um curso em nível de especialização de uma universidade "A" Federal, oito artigos de um curso de nível de graduação de uma universidade "B" privada, e dois textos escritos sem qualquer indício de plágio para fins de confirmação de autenticidade do *score* apresentado como percentual de indícios de plágio. Todos os textos submetidos para análise tinham entre 8 e 15 páginas e foram escolhidos aleatoriamente para os testes. A finalidade destes testes foi verificar a corretude de funcionamento da ferramenta. Para tal, após a análise de todos os arquivos, estes passaram por uma verificação manual, com o intuito de localizar e comparar o indício de plágio do arquivo original em relação ao resultado trazido pela ferramenta. A limitação de quantidade de arquivos para validação em 18 arquivos, no total, se deu devido às análises manuais comprobatórias da corretude da ferramenta, sendo que para cada arquivo

[3] Disponível em: <http://www.labiutil.inf.ufsc.br/ergolist>.
[4] Disponível em: <http://www.abntcatalogo.com.br/norma.aspx?ID=2815>.

analisado este foi conferido manualmente, confrontando o arquivo com os resultados/arquivos que compunham o repositório de cada análise.

5 PROPOSTA DO DETECTOR DE INDÍCIOS DE PLÁGIO MISS – MARPLE

Neste capítulo será descrita a proposta, desenvolvimento e limitações encontradas na construção e funcionamento da ferramenta *Miss Marple*. Primeiramente, na seção 5.1 é descrito o Método DIP de [PERTILE, 2011], que serviu de embasamento para o desenvolvimento das funcionalidades da ferramenta *Miss Marple*. Após, será exposto o desenvolvimento da ferramenta e suas funcionalidades.

5.1 Método DIP – Detector de Indícios de Plágio

O Método DIP trata-se de um método/ferramenta desenvolvida por [PERTILE, 2011], que analisa e calcula o percentual de indícios de plágio. São verificados arquivos com extensão .doc através de buscas realizadas na Internet, utilizando a *API de buscas do Google - API Google Search Ajax.* O cálculo da similaridade e o percentual de indícios de plágio são realizados levando em consideração o número de termos similares do arquivo submetido em relação ao *content* (breve descrição do que se trata o resultado associado a pesquisa) retornado de uma pesquisa. DIP é disponibilizado em três versões: desktop, acoplada ao AVA Moodle, e MLE Moodle.

O funcionamento do DIP acontece da seguinte forma: inicialmente, o usuário escolhe um diretório onde deseja salvar seus relatórios retornados pelo DIP. Em seguida, os arquivos que passarão pelo processo de análise são submetidos ao DIP, que envia parágrafos com um número "x" de palavras para a API de buscas do Google – *API Google Search Ajax,* a qual realiza a procura pelos termos similares no breve resumo (*content*) dos conteúdos da Internet.

Na sequência, os resultados encontrados são retornados para o DIP, e este, por sua vez, faz o processamento e o cálculo de similaridade e em seguida gera o relatório em .pdf. Neste relatório é apresentado o percentual de indícios de plágio de cada parágrafo bem como as URLs onde se encontram. Nas Figuras 5.1.1, 5.1.2 e 5.1.3, ilustra-se o DIP, versão desktop, integrado dentro do Moodle e também no MLE-Moodle, respectivamente.

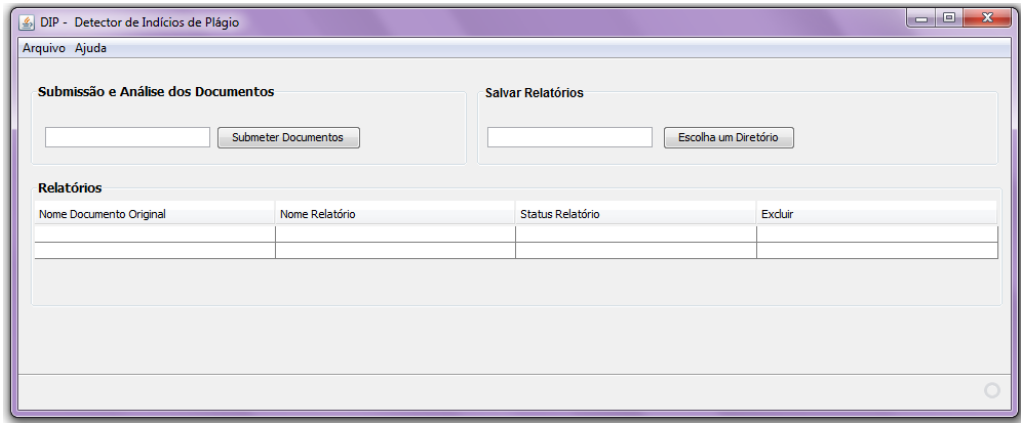

Figura 5.1-1- DIP – Versão Desktop. [PERTILE, 2011]

Figura 3.1.2.5.1-2 - DIP – Versão Moodle. [PERTILE, 2011]

Figura 3.1.2.5.1-3 - DIP – Versão MLE-Moodle. [PERTILE, 2011]

Após desenvolver o DIP, foram realizadas análises, que serviram como testes e comparativo do método desenvolvido em relação a algumas ferramentas. O resultado dessa análise pode ser observada no Gráfico 5.1.1 .

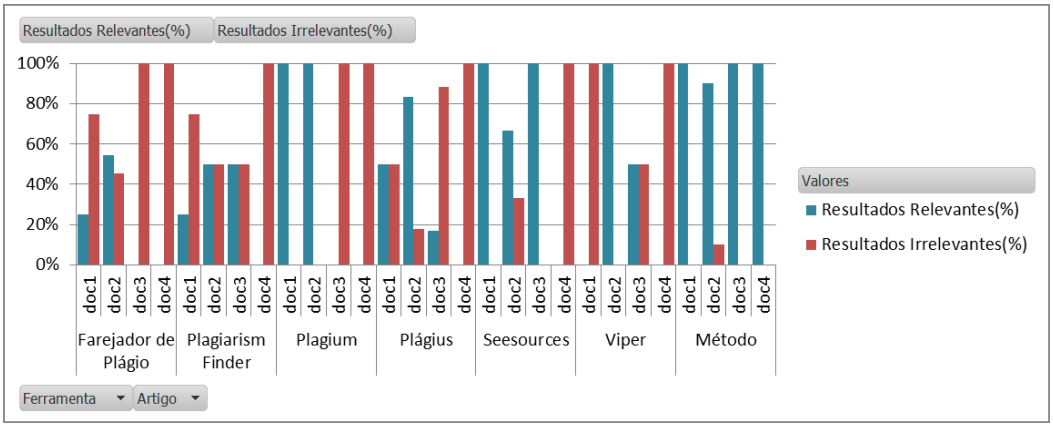

Gráfico 5.1.1 - Resultado da Análise de relevância dos arquivos. [PERTILE, 2011]

O Gráfico 5.1.1, ilustra a precisão dos resultados retornados na verificação de cada um dos documentos analisados por 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. Esta metodologia também foi adotada no desenvolvimento deste trabalho.

O método desenvolvido apresentou resultados satisfatórios em relação às demais ferramentas, obtendo resultados relevantes de 90% e 100%, sendo que somente no documento 2 o sistema apresentou um percentual de resultados irrelevantes de 10%. Esses documentos analisados compunham um acervo de documentos de um curso "x" de pós-graduação, nível de especialização. O conteúdo foi analisado por [PERTILE, 2011], e em sua pesquisa não é apontado os fatores pelos quais esta análise revelou dados irrelevantes neste arquivo (Documento 2) em específico.

Apesar de o Método DIP apresentar resultados satisfatórios, ainda permaneceu alguns quesitos faltantes em seu desenvolvimento que motivaram o desenvolvimento da ferramenta proposta – *Miss Marple* – como o acréscimo de mais extensões de arquivos para serem analisados, melhora na metodologia de análise e a criação do repositório dos arquivos analisados.

5.2 Proposta: A ferramenta *Miss Marple*

Conforme os estudos realizados das ferramentas existentes, além da exploração do Método DIP, identificou-se as possibilidades de desenvolvimento deste trabalho. Buscou-se conceber uma nova ferramenta de detecção de indícios de plágio, denominada *Miss Marple*[5], para auxiliar na verificação da autenticidade dos trabalhos acadêmicos e ou textos publicáveis. O Quadro 5.2-1, resume as alterações bem como as novas funcionalidades propostas neste trabalho.

[5] *Miss Marple* é uma detetive amadora, personagem de ficção presente nas obras de Agatha Christie.

Ferramenta	Extensão de arquivos	Análise	Utilização de Técnicas de stemming
Método DIP – Detector de Indícios de Plágio	apenas .doc	- Feitas a partir da API de buscas do Google - *API Google Search Ajax,* onde os arquivos são enviados para análise. A comparação e o cálculo de similaridade são feitos em relação ao *content* retornados da pesquisa	Não
Miss Marple – Detector de Indícios de Plágio	.doc, .docx, .pdf, ou HTML	- Feitas a partir da API de buscas do Google - *API Google Search Ajax* - Análise em relação aos arquivos que compõem o repositório	Sim

Quadro 5.2-1 - Resumo das diferenças entre Método DIP e ferramenta Miss Marple

5.2.1 Modelagem da Ferramenta

A modelagem possibilita que o desenvolvedor projete as funcionalidades e ações do software e do usuário, pois possibilita a visualização e a comunicação entre o desenvolvedor e o usuário a partir de diagramas [IBM, 2013].

Para definição das ações da ferramenta proposta e do usuário,optou-se pela modelagem utilizando a linguagem *UML - Unified Modeling Language*, conforme serão apresentados nas Figuras a seguir.

No decorrer, encontram-se dois diagramas de casos de uso (Figuras 5.2.1-1 e 5.2.1-2) da Ferramenta *Miss Marple*. Na, Figura 5.2.1-1, é exposto às ações da ferramenta durante sua execução.

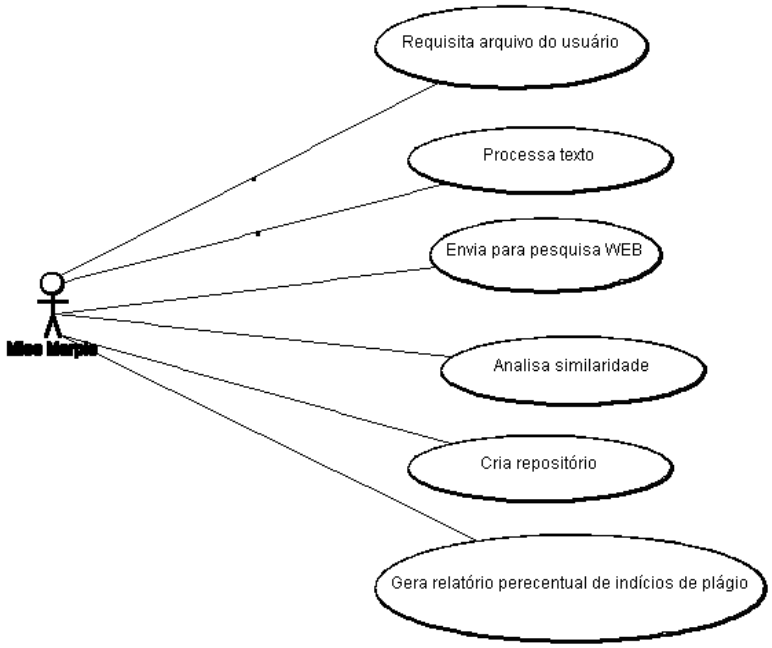

Figura 3.1.2.5.2.1-1 - Diagrama de caso de uso ferramenta

Neste diagrama (Figura 5.2.1-1) observam-se as ações da ferramenta *Miss Marple* ao executar uma análise. As ações da ferramenta ao ser requisitada são: requisitar o arquivo para o usuário, pré-processar esse texto, enviar para pesquisa na Web, analisar a similaridade entre o arquivo submetido e os encontrados na pesquisa e, em seguida, a criação do repositório com os arquivos que apresentavam índice de similaridade superior a 60%. Por fim, a geração do relatório para o usuário.

Já no Figura 5.2.2-1 são ilustradas as ações do usuário frente à ferramenta.

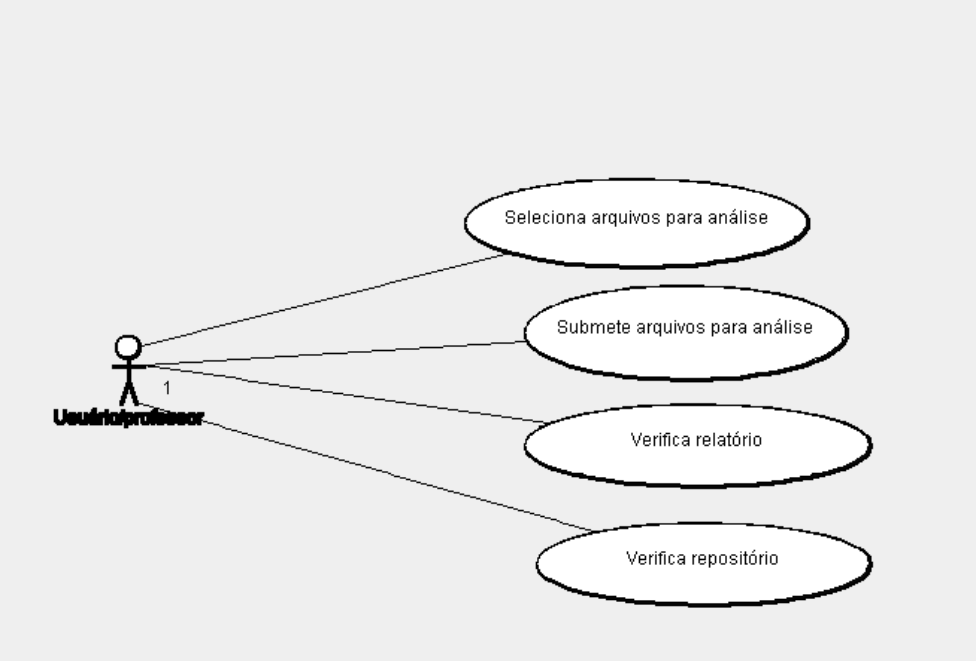

Figura 5.2.2-1 - Diagrama caso de uso – usuário/ ferramenta

O usuário seleciona os arquivos suspeitos de conter indícios de plágio, em seguida, submete-os para análise, aguarda as ações da ferramenta e o *feedback* da mesma, e, na sequência, analisa o relatório gerado, e para fins de comprovação do plágio, verifica os arquivos do repositório.

5.3 Funcionamento da Ferramenta *Miss Marple*

Inicialmente, todos os arquivos submetidos à ferramenta, independente do tipo de extensão utilizada, são convertidos em texto puro (.txt), durante o pré-processamento, e posteriormente, são enviados para a *API* de pesquisa do *Google - API Google Search Ajax* .

Na etapa de pré-processamento é onde ocorre a remoção das figuras, espaços, indexação de palavras e remoção das *stopwords* (palavras que são consideradas irrelevantes na análise de indícios de plágio, por exemplo: advérbios, artigos, conjunções, preposições e pronomes [DIAS, 2004]). Este é um dos diferencias da ferramenta, visto que em algumas outras, cabe ao usuário essa tarefa, como no caso da ferramenta DocCop.

Na etapa seguinte, com o envio dos textos para a *API Google Search Ajax* ocorre o processo de análise de similaridade com o *content* do Google *(breve descrição do que se trata a pesquisa)* e, em seguida, são coletadas as fontes/URLs, das quais são feitas o download dos arquivos que contenham pelo menos 60% de termos similares em seu *content.*

O percentual de similaridade fixado em 60% foi testado no Método DIP [PERTILE, 2011] e comprovado como sendo o melhor percentual.

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

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 5.3-1 - Cálculo de Similaridade [PERTILE, 2011]

A verificação da melhor porcentagem para a comprovação de indício de plágio se deu da seguinte forma: foram submetidos vários arquivos no Método DIP com diferentes índices de percentuais de similaridade, variando de 10% a 100%, com o retorno dessas submissões, os resultados foram analisados como relevantes e irrelevantes. A partir dessa classificação, foi possível identificar o índice de percentual que melhor apresentou resultados relevantes (S), que foi de 60%. O esquema do cálculo é apresentado na Figura 5.3-1: este mesmo cálculo de similaridade e percentual de cálculo de indícios de plágio foi adotado na ferramenta *Miss Marple*.

Esta verificação realizada por [PERTILE, 2011] é melhor ilustrada na Figura 5.3-2.

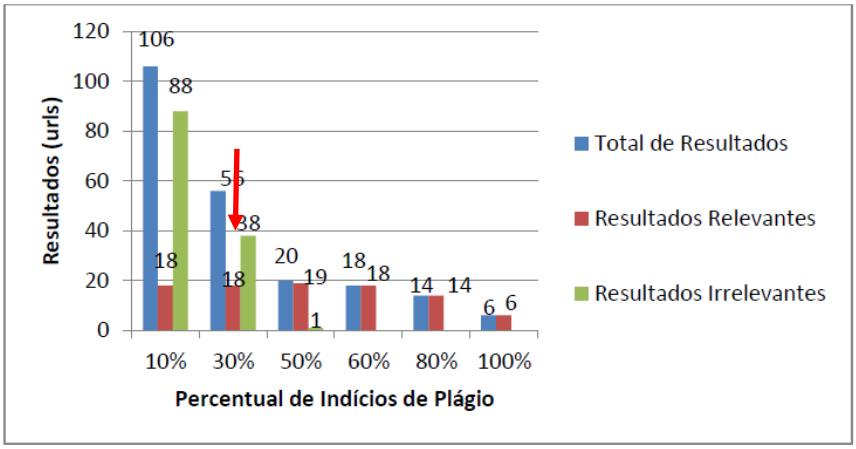

Figura 5.3-2 - Cálculo do melhor índice de similaridade [PERTILE, 2011]

A seta aponta na Figura 5.3-2 o melhor índice para cálculo de indícios de plágio, sendo que o percentual/índice considerável é de 60%. Segundo [PERTILE, 2011], este valor é o que melhor apresenta resultados concisos, ou seja, mais relevantes no processo de pesquisa realizado pela *API – Google Search Ajax*. Neste teste, foram submetidos 18 arquivos para análise, sendo que o percentual para comprovação de indício de plágio foi variado de 10% a 100%. Com percentual de 10% a 50% os resultados retornados pela ferramenta foram muito irrelevantes. Em outras palavras, a ferramenta trazia materiais sem qualquer relação com o documento submetido para análise. Já em percentuais acima de 60%, os resultados obtidos foram satisfatórios, sendo que dos 18 arquivos submetidos, retornaram materiais que efetivamente tinham relação com plágio. Portanto, o melhor índice para consideração de plágio, é acima de 60%.

Na ferramenta *Miss Marple*, utiliza-se deste mesmo percentual (acima de 60%) para buscas de arquivos similares, após as pesquisas é realizado o *download* dos arquivos suspeitos, estes são armazenados durante a análise em um diretório no espaço escolhido pelo usuário em sua máquina formando, ao final da análise, o repositório dos arquivos. Neste repositório é realizada a avaliação de referências cruzadas, ou seja, a contraposição entre o arquivo suspeito e o encontrado na busca, culminando na análise final de indícios de plágio.

Na fase de análise dos arquivos também é considerada a indexação dos termos, a partir da utilização da biblioteca *Lucene – Java*, que possibilita a aplicação de técnicas de *stemming* ao indexar termos.

A técnica de *stemming* possibilita a extração do radical das palavras, facilitando a comparação de termos que contenham o mesmo radical, por exemplo: Carro e Carroça, ambos têm o mesmo radical, mas são palavras distintas.

Ao final da dupla análise (na Internet e no repositório), é então gerado um relatório na tela do *Miss Marple com feedback em tempo de execução da análise* e em um arquivo .pdf que apresenta o percentual de indícios de plágio do texto analisado, localização das referências no computador e os endereços - URLs de onde foram baixados os arquivos do repositório.

Na Figura 5.3-3, é ilustrada uma síntese geral de funcionamento da ferramenta *Miss Marple*.

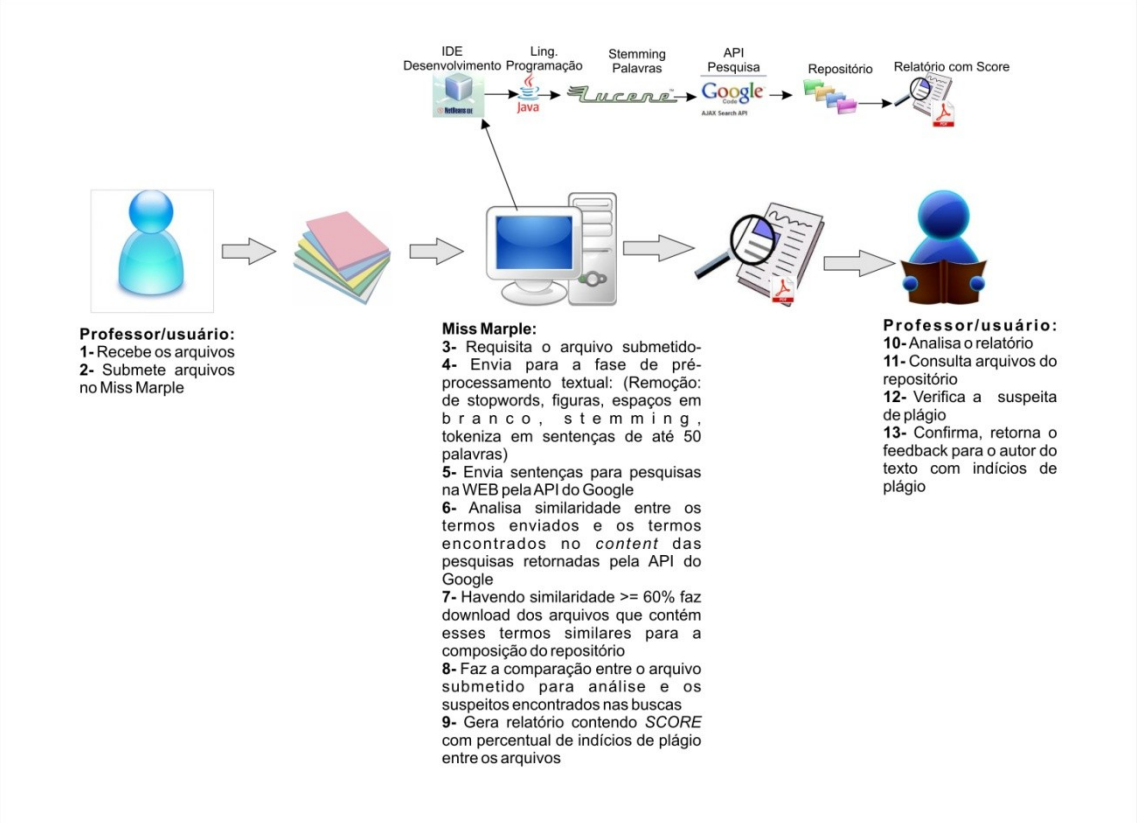

Figura 5.3-3 - Síntese geral de funcionamento da ferramenta *Miss Marple*

Os passos ordenados e escritos na Figura 5.3-3, são os seguintes:

Professor/usuário:

1- Recebe os arquivos.

2- Submete arquivos no *Miss Marple*.

***Miss Marple*:**

3- Requisita o arquivo submetido.

4- Envia para a fase de pré-processamento textual: (Remoção: de stopwords, figuras, espaços em branco, *stemming*, tokeniza em sentenças de até 50 palavras).

5- Envia sentenças para pesquisas na *Web pela API* do Google.

6- Analisa similaridade entre os termos enviados e os termos encontrados no content das pesquisas retornadas pela *API* do Google.

7- Havendo similaridade >= 60% faz download dos arquivos que contém esses termos similares para a composição do repositório.

8- Faz a comparação entre o arquivo submetido para análise e os suspeitos encontrados nas buscas.

9- Gera relatório contendo *SCORE* com percentual de indícios de plágio entre os arquivos.

Professor/usuário:

10- Analisa o relatório.

11- Consulta arquivos do repositório.

12- Verifica a suspeita de plágio.

13- Confirma, retorna o *feedback* para o autor do texto com indícios de plágio.

5.4 *Miss Marple*: Apresentação da ferramenta

Nesta seção será apresentada a interface da ferramenta *Miss Marple* em funcionamento.

A Figura 5.4-1 traz a tela inicial da ferramenta *Miss Marple*, com um arquivo submetido para análise.

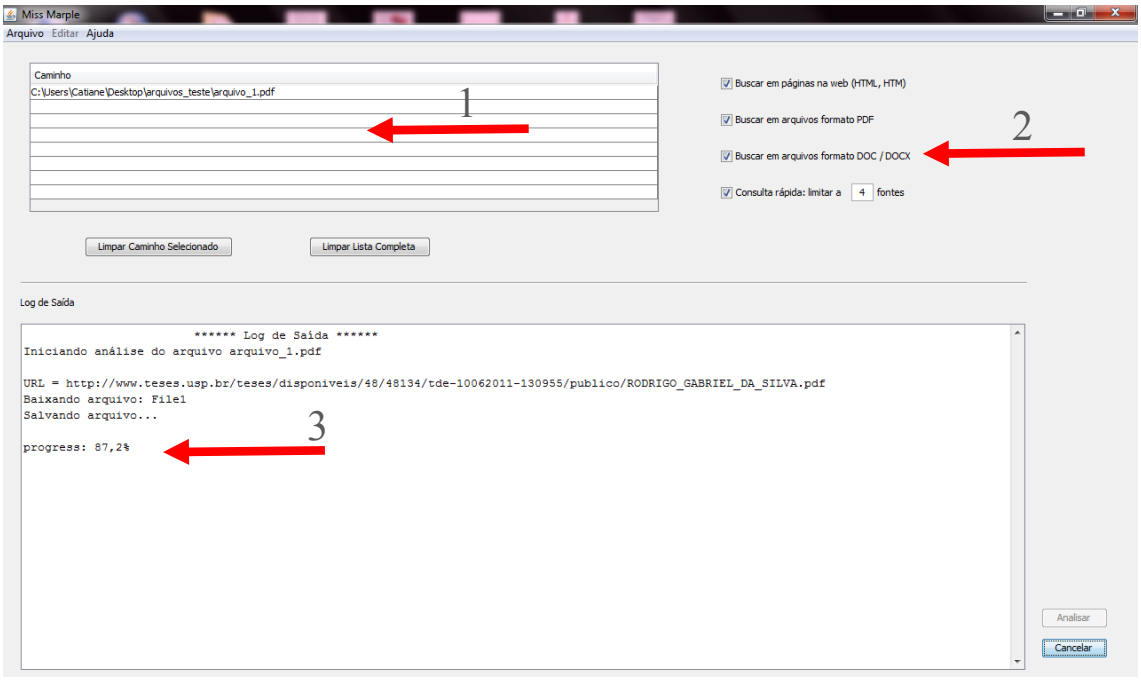

Figura 5.2.1. - Interface do *Miss Marple*, iniciando a execução de uma análise

Os campos destacados na Figura 5.4-1 são:

- Campo1: Nesta área fica localizado a lista de arquivos que estão sendo analisados em um determinado instante.
- Campo 2: Controles de tipos de buscas que o usuário deseja fazer (tipos de arquivos para compor o repositório), com a possibilidade de limitar o número de fontes pesquisadas.
- Campo 3: No decorrer da execução da análise é apresentado um feedback em tempo real das ações que estão ocorrendo durante o processo, neste espaço são descritos: Início da análise de cada arquivo, formação do repositório, gravação dos arquivos no repositório, final da análise, tempo decorrido durante o processo, endereços dos arquivos baixados, *escore* com percentual de indícios de plágio e similaridade de cada arquivo em relação ao submetido.

O *feedback* para o usuário em tempo real, informando-o das ações da ferramenta em um dado instante é outro diferencial desenvolvido nesse trabalho, visto que todas as ferramentas estudadas fornecem somente o tempo de processamento estimado para conclusão da análise e, ao final, o resultado de indício de plágio. Na ferramenta *Miss Marple* o usuário vai sendo informado da taxa de download dos arquivos (b), tamanho de

cada arquivo baixado (a), endereço de cada arquivo (c), decorrer da análise (d), composição do relatório (e). Conforme exposto na Figura 5.4-2, abaixo.

Na Figura 5.4-2, é ilustrado a execução de uma análise.

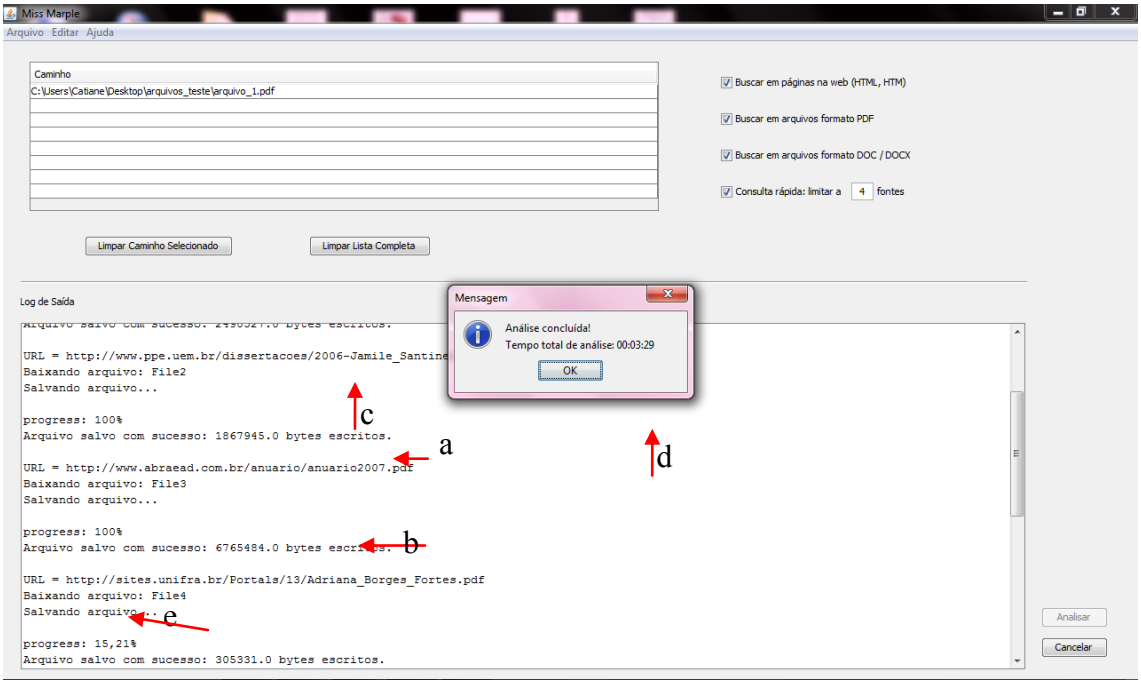

Figura 5.4-2 - Execução de uma análise

A seguir, na Figura 5.4-3 é ilustrado o *feedback* de saída do final da análise.

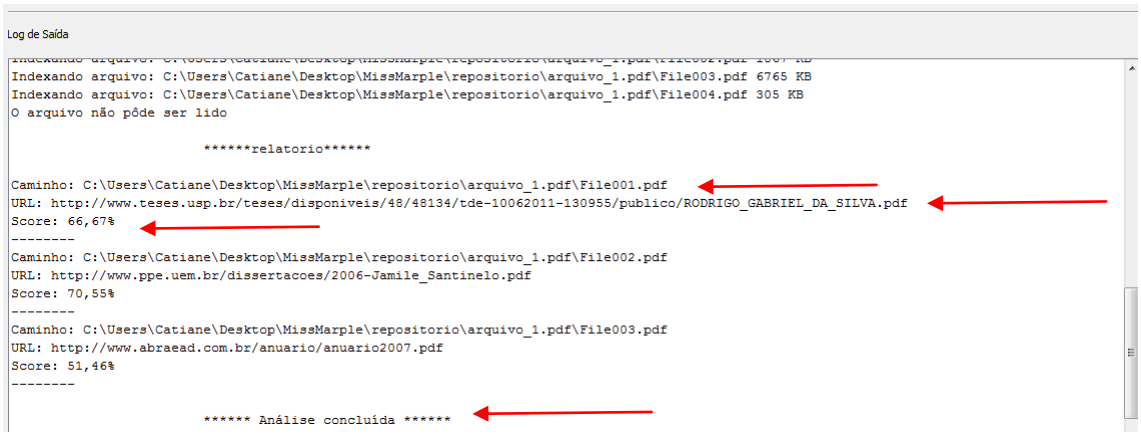

Figura 5.4-3- Feedback final da análise

Na figura 5.4.2, as setas apontam os endereços de cada um dos arquivos que foram impressos no relatório final, também é apresentado o *score* com o percentual de similaridade em relação ao arquivo submetido. Na última seta, localizada bem abaixo na Figura 5.4-3 o usuário é informado da conclusão da análise.

Em seguida, na Figura 5.4-4, ilustra-se a composição do repositório na máquina do usuário.

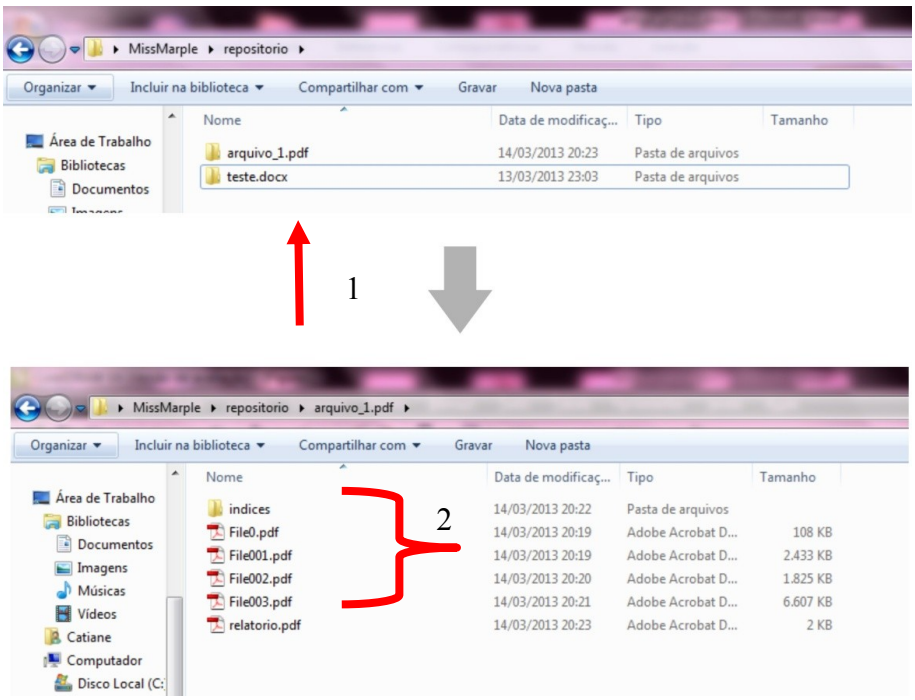

Figura 5.4-4 - Formação do repositório

A formação do repositório, conforme ilustrado na Figura 5.4-4, compreende primeiramente na alocação do diretório na máquina do usuário, conforme a seta número 1 aponta na Figura 5.4-4. No agrupamento de arquivos número 2, apresentado na mesma Figura apresenta-se os arquivos que compõem o repositório bem como o relatório final da análise, que contém as mesmas informações que são apresentadas no *feedback* em tempo real da análise na ferramenta.

O indício de plágio pode ser identificado pelo usuário do *Miss Marple* através da observação do relatório gerado ao final da análise, ou através do acompanhamento do *feedback*. A Figura 5.4.-5 ilustra a identificação do plágio.

Por exemplo, o usuário ao acompanhar a análise no *Miss Marple*, com o *feedback* em tempo real, irá verificar o seguinte:

O *arquivo_1.pdf* submetido para análise (**seta 1**) no *Miss Marple* tem 66,67% (**seta 2**) de indício de plágio em relação ao arquivo baixado do endereço (**seta 3**).

Figura 5.4-5 – Interpretação do relatório de indício de plágio – *Miss Marple*

Estes mesmos dados impressos no *feedback* (Figura 5.4-5), serão apresentados no relatório que ficará salvo dentro do repositório da análise de cada arquivo.

5.5 Desafios encontrados

No decorrer do desenvolvimento e testes do trabalho, foram encontrados alguns desafios que impactaram no resultado final.

Uma das dificuldades encontradas concentrou-se em torno da *Google - API Google Search Ajax*, pois trata-se de uma API em fase de reformulação, trazendo como consequência o limite de buscas diárias. Esse limite está fixado em 1000 buscas/dia. Uma possível solução seria o desenvolvimento de um algoritmo de buscas que realizasse o trabalho similar ao da API para solucionar o limite de buscas.

Outro fator desafiador é a dependência de velocidade de Internet e configuração de dispositivo tecnológico para um bom desempenho da ferramenta. Esses dois fatores são itens

que influenciaram no tempo de processamento de cada análise. As buscas por conteúdo similar em relação ao arquivo submetido para análise são feitas na Internet e a criação do repositório se dá dentro da máquina do usuário. É necessário espaço em disco para armazenamento dessas informações, ou então, agendamento de limpeza em disco. Este desafio foi detectado quando na decorrência dos testes em diferentes velocidades de Internet, conforme a análise realizada e descrita nos resultados deste trabalho, que apresenta a comparação de submissão de arquivos em velocidades de Internet de 512Kb e 3Mb.

Por fim, a execução do *Miss Marple* em servidores sem interface gráfica não ocorre de maneira satisfatória, sendo que a ferramenta foi desenvolvida na linguagem de programação JAVA – e esta linguagem traz consigo uma hierarquia de classes para carregamento da interface gráfica. No caso de servidores sem interface gráfica o ideal é se trabalhar com softwares que sejam executados diretamente e por completo em linha de comando ou através de *Web service*, sendo que o servidor, neste caso, atua apenas como provedor do serviço. Entretanto, o funcionamento da ferramenta ocorre satisfatoriamente em computadores pessoais, com sistemas operacionais gráficos, tais como: Windows nas distribuições XP, 7, Server 2008 e Linux distribuições Ubuntu e Debian.

6 RESULTADOS

Este capítulo contém a descrição da validação da Ferramenta *Miss Marple*, comparativo entre as quatro ferramentas, avaliação de interface, e sugestões de trabalhos futuros.

6.1 Validação da ferramenta

O processo da escolha dos arquivos e o perfil dos usuários que validaram a ferramenta são descritos na Metodologia deste trabalho.

Para a análise de autenticidade dos resultados e tempo de processamento, além da ferramenta desenvolvida, foram elencadas três ferramentas das quais foram estudadas no decorrer do trabalho (Farejador de Plágio, Plagius Detector e VIPER). O critério para seleção destas ferramentas deveu-se ao fato de todas disporem de uma versão gratuita e por serem comumente utilizadas dentro das instituições de ensino.

Para identificar a precisão dos resultados entre as ferramentas, especificamente para esta avaliação, selecionaram-se três artigos menores e um quarto texto sem qualquer indício de plágio, dentre os textos que já faziam parte dos arquivos selecionados para testes. Optou-se por essa alternativa devido à limitação de tamanho de arquivo nas ferramentas utilizadas na versão gratuita. As ferramentas escolhidas foram: VIPER, Farejador de Plágio, Plagius Detector e a ferramenta desenvolvida nesse trabalho, denominada *Miss Marple*.

Os resultados da análise são apresentados no Gráfico 6.1.1.

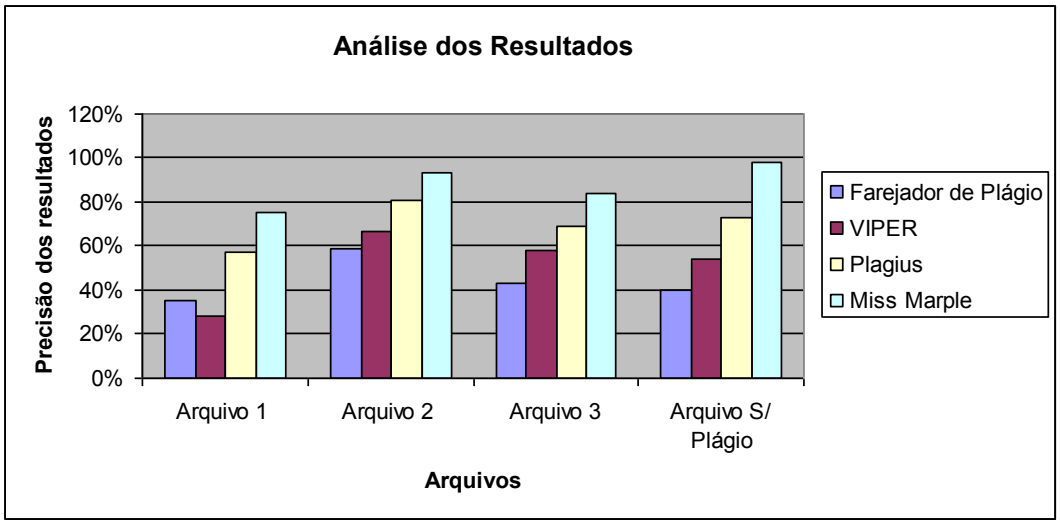

Gráfico 6.1.1 - Análise dos resultados – comparação entre ferramentas

Neste gráfico 6.1.1, identifica-se que a ferramenta *Miss Marple* foi a que melhor se destacou na precisão dos resultados, ou seja, que apresentou os resultados mais relevantes quando colocada em comparação com as outras três ferramentas, Farejador de Plágio, VIPER e Plagius detector. O fato da ferramenta *Miss Marple* apresentar os melhores resultados, deduz-se que se deve ao uso da técnica de *stemming,* visto que nas demais ferramentas não há registros de adoção da técnica em suas documentações. Ainda é possível identificar que a ferramenta que teve o segundo melhor desempenho foi a Plagius, seguida da ferramenta VIPER, precedida da Farejador de Plágios. Outro fator considerável é que todas as ferramentas apresentam limitações de tamanho de arquivo para análise (em torno de 300Kb), e a *Miss Marple* mesmo analisando arquivos por completo, sem tamanho específico, ainda obteve os melhores resultados.

Para fins de validação, foi submetido para análise dez arquivos sem qualquer indício de plágio. A ferramenta *Miss Marple* apresentou como resultado um *score* com 0% de indícios de plágio. Sendo assim, é possível inferir que a ferramenta comprovou sua precisão de análise. Na Figura 6.1.1 ilustra-se os resultados de dois dos dez (10) arquivos analisados e a conclusão da análise.

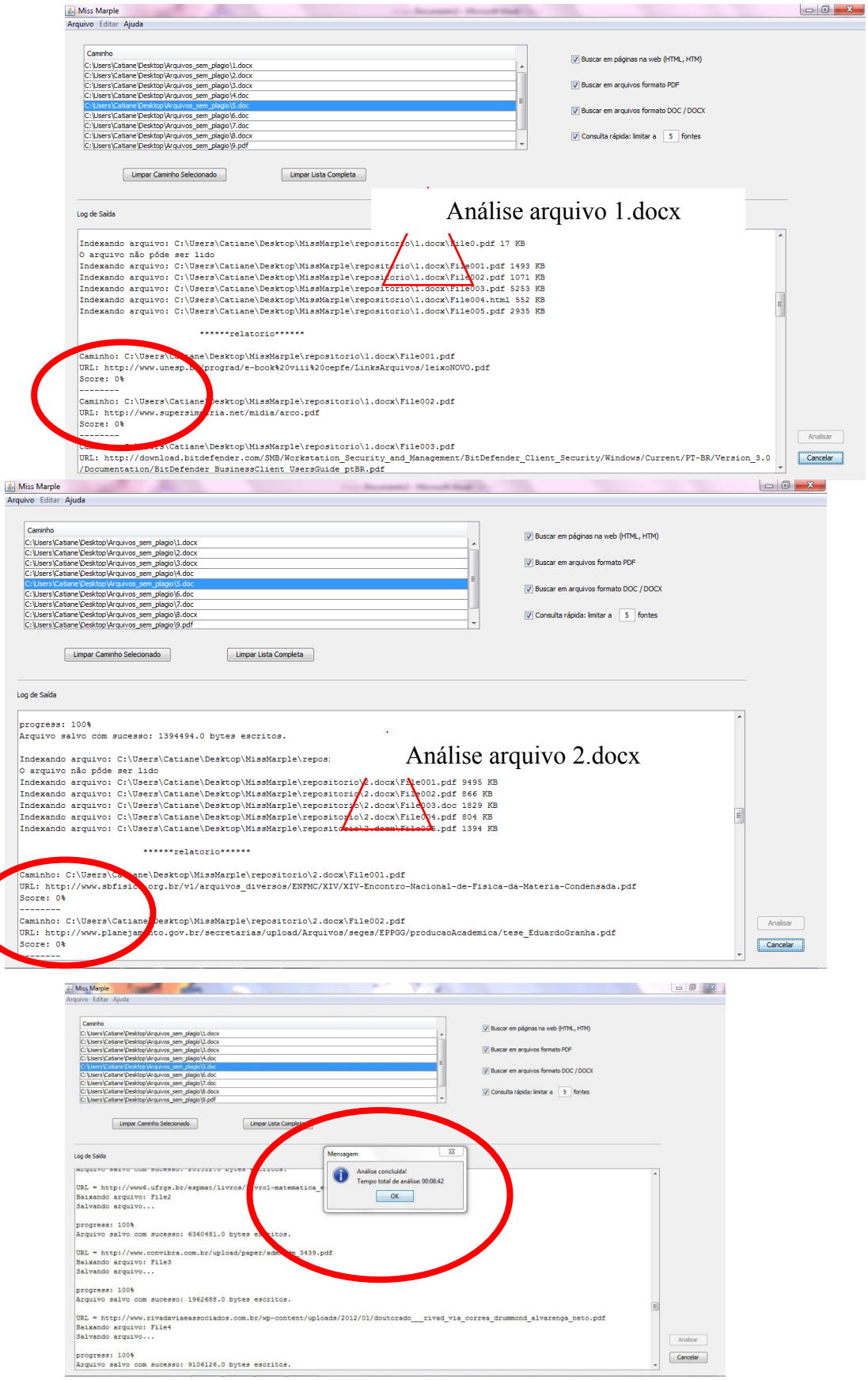

Figura 6.1-1 - Verificação de indícios de plágio em arquivo sem plágio

No decorrer das análises identificou-se que a ferramenta desenvolvida melhorou a precisão dos resultados em relação ao Método DIP – que antes variava de 71,42% e 98%. Agora, a precisão encontra-se na margem de 80,26% e 98%. Esta melhoria deve-se ao fato da inclusão do processo de *stemming* das palavras, o que, por sua vez, acaba por diferenciar as palavras com um mesmo radical.

Outro fator analisado com essas ferramentas foi o tempo de processamento em duas velocidades distintas de Internet, 512Kb e 3Mb, com o intuito de identificar se este fator seria impactante no tempo de análise e processamento. Todos os testes foram realizados no mesmo horário, porém em dias distintos. No Gráfico 6.1-2 e 6.1-3 encontram-se os resultados levados em consideração velocidade de Internet e tempo de processamento.

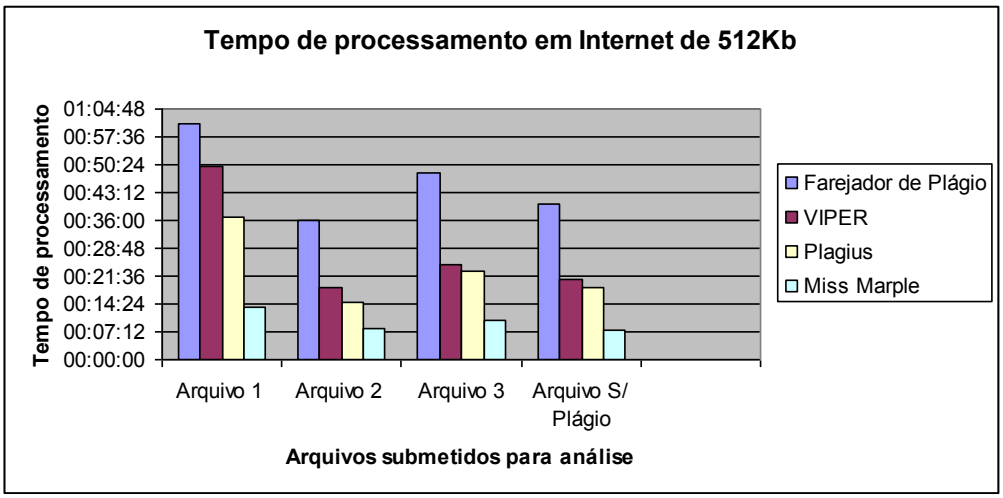

Gráfico 6.1-2 - Tempo de processamento em Internet de 512Kb

No caso da velocidade de 512Kb, apresentado no gráfico acima, a ferramenta que apresentou melhor tempo de processamento foi a *Miss Marple*, seguida pela Plagius e VIPER, e a ferramenta que apresentou em todos os testes maior tempo de processamento foi a ferramenta Farejador de Plágio. No Gráfico (6.1-3) são apresentados os resultados da comparação em Internet de 3Mb.

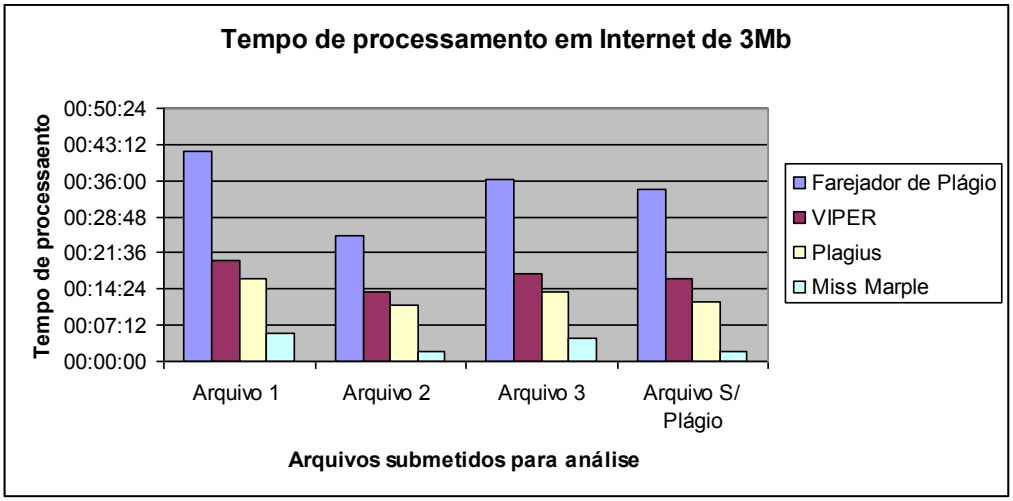

Gráfico 6.1-3 - Tempo de processamento em Internet de 3Mb

Os Gráficos 6.1-2 e 6.1-3 apontam a influência da velocidade de Internet contratada no tempo de análise, sendo que os mesmos arquivos analisados na Internet com 512Kb de velocidade na ferramenta *Miss Marple* tiveram o tempo fixado na média de sete minutos, em contraposição quando submetidos para análise em uma velocidade de 3Mb. O tempo de processamento decaiu para aproximadamente um minuto. Logo, a velocidade de Internet é um fator diferencial no tempo de processamento, já que todas as ferramentas disponíveis no mercado e utilizadas nas análises fazem a busca por textos suspeitos na *Web*.

Outra característica que influencia na velocidade de processamento da ferramenta *Miss Marple* é a possibilidade de o usuário limitar o tamanho de sua pesquisa por arquivos similares, sendo que as demais ferramentas usadas para a validação deste trabalho não apresentam essa opção. A vantagem de permitir ao usuário limitar o tamanho de sua pesquisa é que este consegue prever o tempo disponível para análise de cada arquivo além de levar em consideração sua velocidade de Internet contratada.

Nas análises apresentadas nos Gráficos 6.1-2 e 6.1-3 o limite de fontes de pesquisa foi fixado em 10 fontes por arquivo submetido para verificação de indício de plágio.

O *hardware* não é um fator determinante no tempo de processamento das ferramentas de modo geral, este requisito é importante apenas a partir do momento da criação do repositório de arquivos resultantes de cada análise, visto que isso requer alocação de espaço em disco para fins de armazenamento. Os hardwares utilizados para testes de validação da

ferramenta, foram todos computadores Desktop, Pentium IV, Sistema Operacional Windows XP.

6.2 Requisitos de usabilidade

A avaliação de usabilidade é uma tarefa importante, deste modo, esta metodologia foi aplicada na ferramenta desenvolvida e nas ferramentas utilizadas como base de testes (Farejador de Plágio, Plagius detector e VIPER).

Segundo a Norma ISO 9126[6], através da avaliação de interface, se podem identificar as falhas de comunicação entre a interface, o sistema e o usuário. Para tanto, tomou-se como base o *checklist,* ou lista de apontamentos, proposto por [NUNES *et. al*, 2012] que se baseia-se nas fontes ErgoList[7] e a norma ISO 9126[8]. As questões que formam o *checklist* são listadas no Quadro 6.2.1.

Perguntas
1- O *software* dispõe de todas as funções necessárias para a execução?
2- É permitida a exportação dos dados da análise?
3- Apresenta falhas com frequência?
4- O objeto é conciso nos resultados, passando confiança ao usuário?
5- Os trechos indicados de plágio estão corretos?
6- As referências apresentadas pelo sistema estão de acordo com os trechos copiados?
7- As indicações de plágio no texto são concisas?
8- O objeto é de fácil utilização?
9- É fácil de aprender a usar?
10- Os arquivos de instalação funcionam corretamente?
11- Todos os campos e mostradores de dados possuem rótulos identificativos?
12- Caso o arquivo a ser analisado possua um formato específico (PDF, DOCX, DOC), este formato encontra-se descrito?
13- O sistema fornece ao usuário informações sobre o tempo de processamento?
14- O usuário encontra disponível as informações necessárias para suas ações através do botão de comando “AJUDA”?
15- Na ocorrência de erros, o usuário pode acessar todas as informações necessárias ao diagnóstico e à solução do problema?
16- Quando, durante a análise do documento, o sistema torna-se indisponível ao usuário, devido a algum processamento longo, este é avisado desse estado do sistema e do tempo dessa indisponibilidade?

Quadro 6.2-1 - Questões de avaliação das ferramentas de softwares de detecção de plágio - íntegra. [NUNES et. al, 2012]

[6] Disponível em: <http://www.abntcatalogo.com.br/norma.aspx?ID=2815>.
[7] Disponível em: <http://www.labiutil.inf.ufsc.br/ergolist>.

Em todas as questões aplicadas, os usuários poderiam responder:

- Sim
- Parcialmente com restrições
- Parcialmente
- Parcialmente com muitas restrições
- Não

Além do questionário, foi disponibilizado para esses usuários um Manual de utilização da ferramenta *Miss Marple* com orientações básicas de utilização, visto que as demais possuíam em seus *sites* orientações de auxílio na utilização das mesmas.

A avaliação do *Miss Marple,* seguindo os critérios já descritos, é apresentado nos gráficos 6.2-1, 6.2-2 e 6.2-3.

O Gráfico 6.2-1 apresenta os resultados quanto ao questionamento sobre a frequência de ocorrência de falhas durante a utilização da Ferramenta *Miss Marple.*

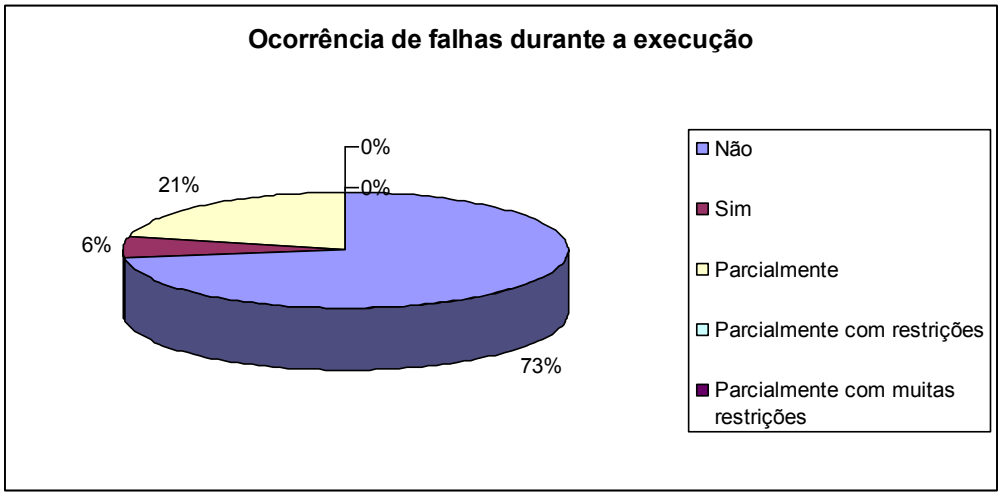

Gráfico 6.2-1- Ocorrência de falhas durante a execução do *Miss Marple*

Os resultados apresentados nesse gráfico são satisfatórios, uma vez que 73% dos usuários responderam que a ferramenta desenvolvida não apresenta falhas durante a execução, 6% responderam que a ferramenta apresenta falhas, devido à criação de diretórios duplicados no repositório quando realizada a análise de um arquivo com um mesmo nome, porém, a ferramenta alerta ao usuário questionando-o sobre a exclusão do diretório antigo ou

duplicado. Cerca de 21% dos usuários consideraram que as falhas encontradas podem ser classificadas como impacto parcial, pois a qualidade da análise não é comprometida.

O Gráfico 6.2-2, traz os dados referentes à consistência das referências/arquivos apontados como possíveis indícios de plágio, ou seja, a confirmação dos indícios de plágio do arquivo submetido na ferramenta em relação aos arquivos retornados da pesquisa na Web.

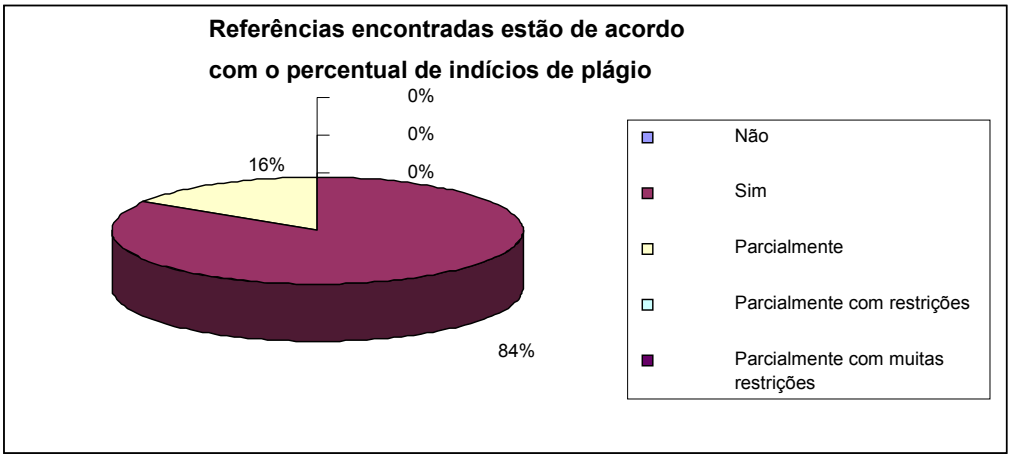

Gráfico 6.2-2 - Consistência das referências encontradas

Do total de entrevistados, 84% avaliou que as referências encontradas com semelhanças em relação ao arquivo submetido para análise, ou com indícios de plágio, são referências consistentes ou corretas, afirmando a qualidade da ferramenta desenvolvida. Já 16% dos usuários apontaram que a ferramenta traz resultados parcialmente consistentes.

O Gráfico 6.2-3 apresenta quesitos relacionados à facilidade de utilização da *ferramenta Miss Marple.*

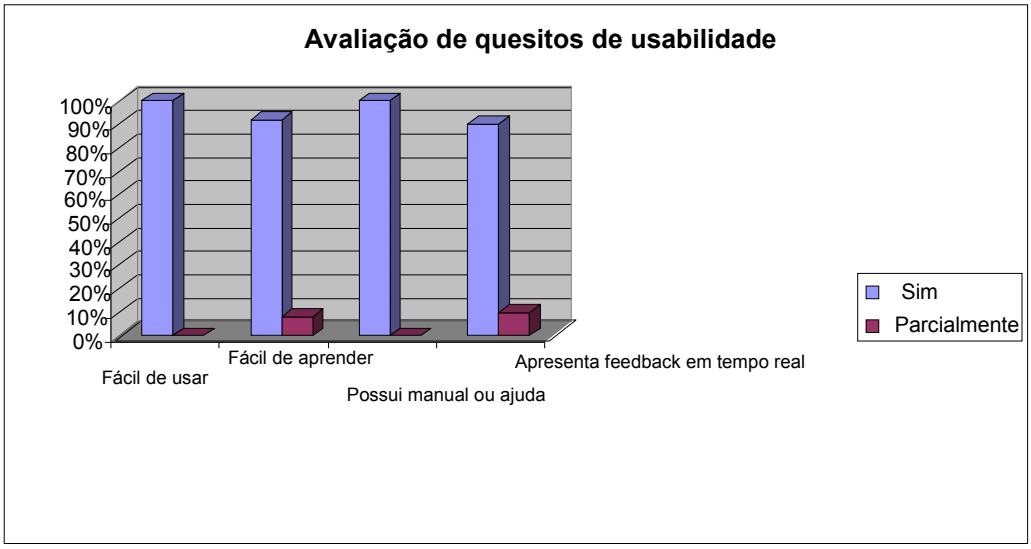

Gráfico 6.2.3 - Quesitos avaliados no *checklist* de usabilidade

Nestes questionamentos contemplados e apresentados no Gráfico 6.2.3, os usuários avaliaram a facilidade de uso da ferramenta, sendo que 100% dos participantes responderam que é de fácil utilização. O item facilidade de utilização está subdividido em outros requisitos como interface intuitiva, mensagens de avisos, manual de utilização e campo contendo ajuda dentro da ferramenta.

Em relação à facilidade de aprendizagem de utilização do software, 92% dos usuários responderam que a ferramenta é fácil de aprender e 8% respondeu que é parcialmente fácil. Este resultado deve-se ao fato de alguns usuários que testaram a ferramenta terem os conhecimentos muito básicos em informática, os quais apresentaram algumas dificuldades no trabalho de localização do repositório. A ferramenta dispõe de manual de ajuda e botão de ajuda, porém, os usuários ainda encontraram essa dificuldade durante a utilização.

Em relação à metodologia de apresentação de *feedback* retornado pela ferramenta, o qual possibilita o acompanhamento em tempo real de todas as ações que estão ocorrendo com o arquivo submetido através do modo textual impresso na tela da ferramenta a cada ação (Figuras 5.4.2 e 5.4.3), 90% dos usuários respondeu que o *feedback* é satisfatório, e 10% julgou que é parcialmente satisfatório, uma vez que o *feedback* não é representado por recursos gráficos, como por exemplo, barras de progresso.

Para concluir a avaliação de usabilidade, realizou-se um comparativo entre as ferramentas com a finalidade de identificar qual apresentava melhor índice de usabilidade. O questionamento realizado foi o seguinte: “Após avaliar essas ferramentas você considera que

se atende aos requisitos de usabilidade descritos no *checklist,* ou lista de apontamentos, proposto por [NUNES *et. al*, 2012] que baseia-se nas fontes *ErgoList*[9] e a norma ISO 9126[10],,

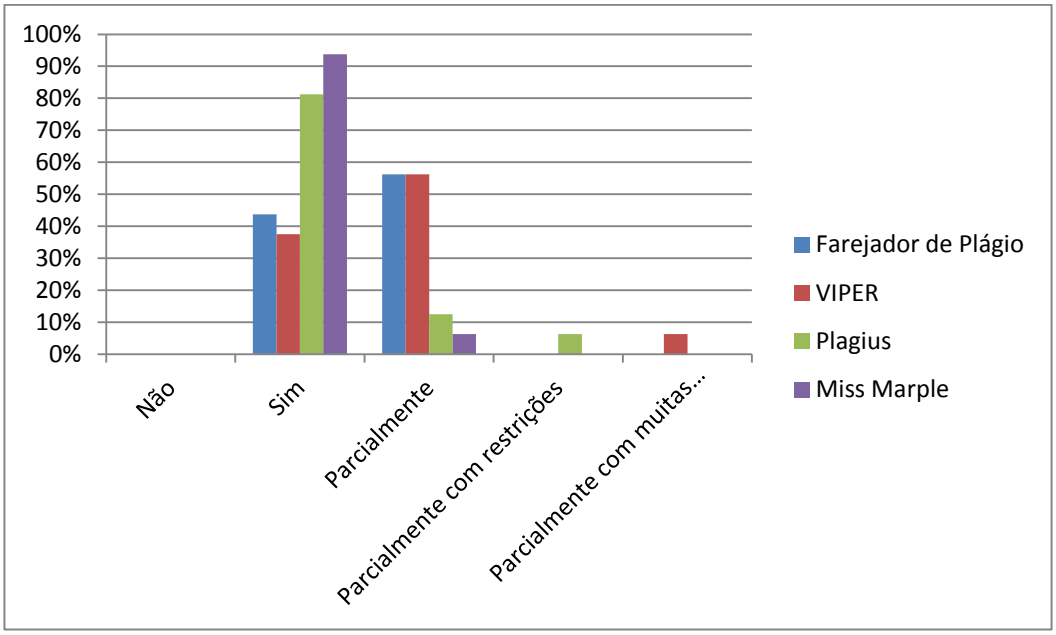

Gráfico 6.2.4 - Percentual de atendimento de requisitos de usabilidade

No Gráfico 6.2.4, ilustra-se o percentual de atendimento dos requisitos de usabilidade das ferramentas utilizadas para testes (Farejador de Plágios, Plágius Detector e Viper) bem como, a ferramenta desenvolvida *(Miss Marple)*. Os requisitos avaliados neste questionamento foram em relação ao *checklist* utilizado no decorrer deste trabalho. Os usuários responderam que a ferramenta Farejador de Plágio atende 43,75% dos requisitos de usabilidade propostos, em contraposição com a ferramentas em destaque *Miss Marple* que apresentou 93,75% de atendimento. Já a ferramenta Plagius apresentou 81,25%. Já a ferramenta VIPER ficou em última posição na classificação.

Segundo a avaliação, as ferramentas Farejador de Plágio e VIPER atendem parcialmente aos requisitos, com um percentual de 56,25%, seguidas pela ferramenta Plagius com 12,5% e, por fim, *Miss Marple,* que apresentou um percentual favorável em relação às demais, com 6,25%. Isto demonstra que no decorrer do seu desenvolvimento, prezou-se pelo atendimento aos requisitos de usabilidade. Por fim, somente a ferramenta Plagius Detector

[9] Disponível em: <http://www.labiutil.inf.ufsc.br/ergolist>.

apresentou um percentual de 6,25% classificado como parcialmente e com muitas restrições. Já a ferramenta VIPER, expôs 6,25% de não atendimento aos requisitos de usabilidade.

Finalmente, traçou-se um quadro comparativo (exposto a seguir) entre as ferramentas utilizadas no decorrer dos testes.

Ferramenta	Tipo de licença	Extensões de arquivos	Criação de repositório	Tamanho de arquivos – pré-processamento textual	Relatório	*Feedback* em tempo real
Miss Marple	Gratuita	.doc, .docx, .pdf, .HTML	SIM	Sem restrição / Sim	SIM	SIM
DIP	Gratuita	.doc	NÃO	Sem restrição / Sim	SIM	NÃO
Plagius Detector	Gratuita Paga	.doc, .docx, .pdf, .HTML	NÃO	Com restrição (500 palavras) / Sim	SIM	NÃO
Farejador de Plágios	Gratuita Paga	.doc e .rtf	NÃO	Com restrição / Sim	SIM	NÃO
VIPER	Gratuita	doc, .rtf, .html e .txt	NÃO	Sem restrição / Não	SIM	NÃO

Quadro 6.2-2 - Comparativo das ferramentas utilizadas para testes

Neste Quadro 6.2.2 destacam-se as contribuições da ferramenta *Miss Marple* desenvolvida, tendo como diferencial a não restrição do tamanho dos arquivos submetidos para análise, além da criação do repositório dos arquivos e *feedback* que foram encontrados no decorrer do processo em tempo real para o usuário.

CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo desenvolver uma alternativa de melhor controlar a autenticidade dos textos acadêmicos, culminando no desenvolvimento de uma nova ferramenta para auxiliar na verificação de indícios de plágio em produções textuais, tendo como base o método DIP – Detector de Indícios de Plágio [PERTILE, 2011].

Este método foi desenvolvido dentro do Grupo de Redes e Computação Aplicada da Universidade Federal de Santa Maria, com o intuito de auxiliar no processo de verificação de indícios de plágio nos trabalhos acadêmicos desenvolvidos dentro da Universidade. O Método desenvolvido atendia parcialmente as necessidades dos usuários, visto que analisava somente arquivos com extensões .doc e a análise era realizada de maneira simplificada. Sendo assim, era realizada apenas a comparação dos arquivos em relação a breves resumos do conteúdo das páginas Web, sem a comparação dos arquivos por inteiro. Desta forma, buscou-se o desenvolvimento de uma nova ferramenta (*Miss Marple*), com novos objetivos, e que culminasse no aprimoramento do método já desenvolvido, este, utilizado como base para o desenvolvimento da nova ferramenta.

Os objetivos propostos no trabalho foram alcançados, pois os resultados foram satisfatórios, principalmente quando comparados com outras ferramentas disponíveis para uso no mercado, como por exemplo, a Farejador de Plágios, VIPER e Plágius Detector, e o desenvolvimento de contribuições para o método DIP.

As contribuições deste trabalho foram: a) adição de um novo método de análise de similaridade a partir do método de *stemming* de palavras, que possibilita a diferenciação de palavras com o mesmo radical, melhorando a qualidade dos resultados de comparação de similaridade; b) criação de uma interface mais amigável com mais possibilidades e controles do usuário, possibilitando a este o controle de seu tempo disponível para submissões de arquivos através da determinação do número de fontes que deseja pesquisar e, com isso, determinar qual o melhor período do dia para análises e quantidade de espaço em disco disponível para a criação do repositório; c) inserção de análise de novas extensões de arquivos como .docx, .pdf e .HTM/HTML, o que amplia a quantidade arquivos que podem ser analisados, melhorando a qualidade pela busca de indícios de plágio; d) criação de um repositório local com os arquivos suspeitos baixados da Internet ao final de cada análise, agilizando/possibilitando ao usuário a consulta posterior desses arquivos que estarão

armazenados em sua máquina, caso necessite; e) possibilidade de o usuário limitar o número de fontes que esse deseja buscar em cada análise, proporcionando o controle de número de arquivos que compõe o repositório, e consequentemente, uma estimativa de tempo de análise de cada arquivo submetido. Esta contribuição permite ao usuário prever qual o tempo dedicará para as análises dos arquivos, destacando que, nas demais ferramentas estudadas, o fator tempo de processamento não era possível de ser previsto.

Este trabalho foi disponibilizado para a comunidade acadêmica no endereço eletrônico http://nte.ufsm.br/moodle2_UAB/ - Moodle Capacitação e enviado para disponibilização na página do Grupo de Pesquisa em Redes de Computadores e Computação Aplicada da Universidade Federal de Santa Maria – http://coral.ufsm.br/greca/?page_id=179&fb_ref=below-post&fb_source=message.

Ao concluir este estudo, sugere-se como trabalhos futuros, sugere-se o desenvolvimento de uma ferramenta com conceito de *Web Service* com a criação do repositório de arquivos em um Banco de Dados, localizado na nuvem *(cloud computing*), não necessitando de espaço em disco na máquina do usuário. Esta possibilidade aperfeiçoaria a pesquisa na análise de indícios de plágio, visto que inúmeras instituições de ensino ou publicações iriam compor o repositório das análises. Ainda há a necessidade de tratamento de plágio multilíngue, que já era tratado no Método DIP, mas que devido à descontinuidade da API de tradução, neste trabalho não foi abordado.

Por fim, outra sugestão de trabalho futuro seria a criação de filas de documentos para análise automática, sendo que o professor pudesse interagir apenas em dois momentos: na submissão dos arquivos e no envio do *feedback* para o aluno ao final da análise, quando este fosse comunicado pela ferramenta a ocorrência do plágio.

REFERÊNCIAS

BARNBAUM, C. PLAGIARISM: A Student's Guide to Recognizing It and Avoiding It. Valdosta State University, (2002). Disponível em: <http://www.valdosta.edu/~cbarnbau/personal/teaching_MISC/plagiarism.htm>. Acesso em: 15 de julho de 2012.

CENDÓN,. V. B. Ferramentas de buscas na WEB. (2001) Revista Ciências da Informação Brasília, v. 30, n. 1, p. 39-49, jan./abr.

CNPQ - Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em: http://www.cnpq.br/web/guest/noticias;jsessionid=04978B18195A99CE49903C1195C17637?p_p_id=engine_WAR_Engineportlet_INSTANCE_N14w&p_p_lifecycle=0&p_p_state=maximized&p_p_mode=view&p_p_col_id=column-3&p_p_col_pos=1&p_p_col_count=4&_engine_WAR_Engineportlet_INSTANCE_N14w_view=article&_engine_WAR_Engineportlet_INSTANCE_N14w_articleResourcePrimKey=263957&_engine_WAR_Engineportlet_INSTANCE_N14w_backURL=. Acesso em: 28 de março de 2013.

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.

DESARROLLO (2012). Disponível em : < http://www.desarrolloweb.com/de_interes/ranking-buscadores-enero-2012-6503.html>. Acesso em: 20 de março de 2013.

DOC COP. (2012). Disponível em: <http://www.doccop.com/>. Acesso em: 20 de junho de 2012.

EPHORUS. Ephorus: liderança na Europa, (2012). Disponível em: <http://www.ephorus.pt/home>. Acesso em: 21 de julho de 2012.

ETBLAST, 2012. Disponível em: < http://etest.vbi.vt.edu/etblast3/>. Acesso em: 20 de junho de 2012.

FAREJADOR. Farejador de Plágios, (2012). Disponível em: <http://www.farejadordeplagio.com.br/>. Acesso em: 21 de julho de 2012.

FURTADO. J. A. X. P. Trabalhos acadêmicos em Direito e a violação de direitos autorais através de plágio Disponível em: < http://www.egov.ufsc.br/portal/sites/default/files/anexos/5640-5632-1-PB.htm> Acesso em: 21 de julho de 2012.

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

IBM – (2013). International Business Machines Disponível em: < http://www-01.ibm.com/software/rational/uml/> Acesso em 10 de janeiro de 2013.

LIMA. C. E., RESENDE. P. M. A., 2012. Análise qualitativa e quantitativa entre as principais ferramentas de detecção de plágio. Disponível em: < http//: www.c3.furg.br/arquivos/download/04_lima_resende.pdf> Acesso em 15 de agosto de 2012.

LIMA, E. C. de. Análise de Técnica e Ferramentas de Detecção de Plágio, e Desenvolvimento de um Protótipo de Nova Ferramenta. Monografia de Conclusão de Curso - Universidade Federal de Lavras, Minas Gerais, 2011.

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. Disponível em: <http://www.faculdadesocial.edu.br/dialogospossiveis/artigos/4>. Acesso em: 02 de julho de 2012

MORAIS, M. E. A., AMBROSIO. L. P. A. (2007), Mineração de Textos. Instituto de Informática da Universidade Federal de Goiás. Disponível em: < http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em: 02 de julho de 2013

NEIL, R. (2004). Cheating in online student assessment: Beyond plagiarism. Online Journal of Distance Learning Administration, Volume VII, Number II, State University of West Georgia, Distance Education Center.

OAB – ORDEM DOS ADVOGADOS DO BRASIL: Disponível em: <http://www.diariodecuiaba.com.br/detalhe.php?cod=413526> Acesso em: 03 de janeiro de 2013

OLIVEIRA, M. et al.. Bibliotecas Digitais Aliadas na Detecção Automática de Plágio. Seminário Internacional de Bibliotecas Digitais Brasil. 2007. Disponível em: <http://libdigi.unicamp.br/document/?code=23482>. Acesso em: 05 de julho de 2012.

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.

ORENGO, V.; HUYCK, C. A stemming algorithm for the portuguese language. In: String Processing and Information Retrieval, 2001SPIRE 2001. PROCEEDINGS.EIGHTH INTERNATIONAL SYMPOSIUM ON. Anais. . . [S.l.: s.n.], 2001. p.186 – 193.

PLAGIARISMA. (2012). Disponível em: < http://plagiarisma.net/>. Acesso em: 20 de junho de 2012.

PLAGIARISM.ORG, 2012. What is plagiarism? Plagiarism.org. Disponível em: <http://www.plagiarism.org/plag_article_what_is_plagiarism.html>. Acesso em: 21 de julho de 2012.

PLAGIUM. (2012). Disponível em: <http://www.plagium.com/>. Acesso em: 12 julho de 2012.

PLAGIUS. Plagius - The ultimate in plagiarism detection, 2012. Disponível em: <http://www.plagius.com/s/en/default.aspx>. Acesso em: 21 de julho de 2012.

PLAGIO.NET, 2012. Disponível em: < http://www.plagio.net.br/pesquisa-e-publicacoes.html>. Acesso em: 02 de julho de 2012

PERTILE, S. L. ; MEDINA, R. D. . Desenvolvimento e Aplicação de um Método para Detecção de Indícios de Plágio. In: Simpósio Brasileiro de Informática na Educação, 2011, Aracajú. Anais do XXII SBIE - XVII WI, (2011). p. 1673-1682.

———. Desenvolvimento e Aplicação de um Método para Detecção de Indícios de Plágio. In: Conferência IADIS Ibero Americana WWW/INTERNET 2011, 2011, Rio de Janeiro. Conferência IADIS Ibero Americana WWW/INTERNET (2011).

PERTILE, S. L. . "*Desenvolvimento e Aplicação de um Método para Detecção de Indícios de Plágio"*. Dissertação apresentada ao Curso de Mestrado do Programa de Pós-Graduação em Informática, Universidade Federal de Santa Maria (UFSM, RS), 2011.

NBR ISO/IEC 9126-1: 2003. **Tecnologia de informação: Engenharia de software – Qualidade de produto**. Parte 1: Modelo de qualidade. Esta norma cancela e substitui a NBR 13596. Julho 2003.

NUNES, F. B. ; **VOSS, G. B.** ; MUHLBEIER, A. R. K. ; MEDINA, R. D. ; BERNARDI, G. ; BARBOSA, C. P. A. . Análise Comparativa Teórico-Prática entre Softwares de Deteção de Plágio. RENOTE. Revista Novas Tecnologias na Educação, v. 10, p. 1-10, 2012.

SANTANA, J. AND MARTINS, J. (2003)**.** Um sistema de detecção de plágio em ambiente de aprendizado virtual. pages 230–242. Em: Anais do Virtual Educa 2003, Miami.

SANTOS, A. O. F., FRANCO, R. H. R. L. (2010) Criação de Ferramenta de Detecção de Plágio em Ambiente Virtual de Aprendizagem. Dissertação apresentada ao Curso de Mestrado do Programa de Pós-graduação em Engenharia Elétrica, Universidade Federal de Itajubá-MG. Disponível em: < http://adm-net-a.unifei.edu.br/phl/pdf/0037064.pdf>. Acesso em: 02 de julho de 2012

SENA, A. (2011) Fontes de informação utilizadas pelos discentes do mestrado do Instituto de Educação Matemática e Científica da UFPA (IEMCIUFPA)

SIBI. Sistema Integrado de Bibliotecas - Universidade de São Paulo. (2011) Disponível em: < http://www.workshop.sibi.usp.br/relatorios/Lista_softwares_prevencao_plagio.pdf> Acesso em: 01 de agosto de 2012.

SCHOLARONE. (2012). Disponível em: < http://scholarone.com/>. Acesso em: 21 de julho de 2012.

TURNITIN. Prevent plagiarism, (2012). Disponível em: <http://turnitin.com/static/index.html>. Acesso em: 25 de julho de 2012.

URKUND. (2012). Disponível em: <http://www.urkund.com/int/en/>. Acesso em: 25 de julho de 2012.

USP. (2013). Disponível em: <http://www.escritacientifica.sc.usp.br/anti-plagio/>. Acesso em: 03 de julho de 2013.

VASCONCELOS, S. Questões éticas no ambiente científico. Disponível em: http://www.icb.ufrj.br/Revista-Bio-ICB/Acontece-no-ICB/Questoes-eticas-no-ambiente-cientifico-624.html. Acesso em: 28 de março de 2013.

VenTICS (2012). Disponível em: < http://www.ventics.com/ranking-buscadores-septiembre-2012/>. Acesso em: 28 de março de 2013.

VIPER. The Anti-plagiarism Scanner, (2012). Disponível em: <http://www.scanmyessay.com>. Acesso em: 25 de julho de 2012.

XAPIAN (2013). Disponível em: < http://xapian.org/docs/stemming.html>. Acesso em: 03 de julho de 2013.