Análise qualitativa e quantitativa entre as principais ferramentas de detecção de plágio

Eudes de Castro Lima, Antônio Maria Pereira de Resende

***Resumo* — O plágio é um problema crescente nas academias, pois desrespeita a propriedade intelectual e interfere no aprendizado do aluno. Como forma de inibir tal prática, surgiram diversas ferramentas de detecção de plágio, algumas focadas na detecção de plágio de código fonte e outras de documento de texto. Este artigo descreve análises qualitativa e quantitativa das principais ferramentas de detecção de plágio em documentos de texto. Na análise qualitativa, descreve-se e compara-se os principais recursos oferecidos pelas ferramentas. Na análise quantitativa aplicam-se testes de eficácia, sensibilidade e desempenho, sendo apresentado um quadro comparativo.**

Palavras-chave: Plágio. Ferramentas de detecção de plágio. Análise de ferramentas de detecção de plágio.

I. INTRODUÇÃO

Copiar fragmentos de textos ou textos completos sem citar os verdadeiros autores, caracteriza plágio. Apesar de ser uma prática crescente no meio acadêmico, ela é ilegal e não desejável por tratar-se de fraude.

A falta de dedicação juntamente com a falta de conhecimento sobre o tema são as principais causas que levam ao plágio. Muitas vezes, copiar da Internet é a maneira mais simples de alcançar os objetivos. Alguns alunos plagiam intencionalmente e outros o fazem inconscientemente. Segundo Barnbaum (2002) em [1], a falta de conhecimento do que constitui o plágio leva muitos alunos a cometê-lo inconscientemente. Se não sabe exatamente o que o plágio é, não pode evitar fazê-lo.

Identificar o plágio manualmente não é uma tarefa simples, demanda grande esforço e tempo dos professores que se veem envoltos de muitos trabalhos. Além disso, o crescimento das ocorrências de plágio criaram um nicho de mercado, propiciando o surgimento de várias ferramentas automatizadas para auxiliar na análise de plágio. Entretanto, diante da quantidade de ferramentas, selecionar qual utilizar tornou-se uma tarefa árdua.

Eudes de Castro Lima é graduado em Sistemas de Informação pela Universidade Federal de Lavras (UFLA). eudeslima@bsi.ufla.br

Antônio Maria Pereira de Resende é graduado em Matemática Aplicada a Informática pela Fundação de Ensino e Pesquisa de Itajubá (1995), possui mestrado em Engenharia Eletrônica e Computação pelo Instituto Tecnológico de Aeronáutica (1999) e doutorado em Engenharia Eletrônica e Computação pelo Instituto Tecnológico de Aeronáutica (2007). Atualmente é professor adjunto da Universidade Federal de Lavras (UFLA). tonio@dcc.ufla.br

Neste artigo, os autores apresentam os resultados de análises das ferramentas de detecção de plágio em documentos de texto. Aplicaram-se duas análises denominadas qualitativas e quantitativas. Por meio da análise qualitativa, compararam-se os principais recursos oferecidos pelas ferramentas. Por meio da análise quantitativa, aplicaram-se testes de eficácia, sensibilidade e desempenho, sendo apresentado um quadro comparativo.

O principal objetivo deste trabalho é fornecer uma visão geral das principais ferramentas de detecção de plágio em documentos de texto, fornecer características de cada ferramenta e um quadro comparativo que permita aos professores e interessados tomar a decisão de qual ferramenta utilizar.

O presente trabalho encontra-se estruturado em 4 seções, sendo a primeira uma breve introdução contendo os objetivos, definição do problema em estudo e a solução proposta. Na seção II, encontra-se a Revisão Bibliográfica onde são fundamentados os principais conceitos deste trabalho. A seção III detalha os passos para a seleção das principais ferramentas de detecção de plágio em documentos de texto. Na seção IV, é apresentada a análise comparativa realizada nas ferramentas de detecção de plágio selecionadas e os resultados obtidos. Por fim, na seção V, são apresentadas as conclusões obtidas neste trabalho e os trabalhos futuros.

II. REVISÃO BIBLIOGRÁFICA

A literatura sobre detecção de plágio em documentos de texto é reduzida quando comparada com a detecção de plágio em código fonte. São poucas as informações a respeito das ferramentas existentes. Conforme Lancaster (2003) em [7], as pesquisas relatadas geralmente são limitadas e os métodos e técnicas utilizados pelas ferramentas são omitidos ou pouco detalhados, dificultando descrever seu funcionamento e compará-los.

Segundo Tedford (2003) em [13], o uso das ferramentas de detecção de plágio no ensino superior foi destacado em 2001 quando Lou Bloomfield, professor de física na Universidade de Virginia, desenvolveu um programa para comparar *strings* de tarefas apresentadas com documentos presentes em um banco de dados. Em três anos foram analisados 1500 documentos e 158 estudantes suspeitos foram investigados.

Diferente dos trabalhos de Hage, Rademaker e Vugt (2010) em [3] e Kleiman (2007) em [6] que propõem uma comparação entre ferramentas de detecção de plágio em

código fonte, este trabalho propõe uma análise comparativa entre as ferramentas de detecção de plágio em documentos de texto.

A. O plágio e suas variações

O Plágio pode ter várias definições. Segundo Plagiarism.org (2011) em [10], os pontos seguintes são considerados plágio:

- transformar o trabalho de alguém em seu próprio;
- copiar palavras ou ideias de alguém sem dar crédito;
- não colocar a devida citação;
- dar informações incorretas sobre a origem de uma citação;
- copiar palavras ou ideias de uma fonte que compõe a maioria de seu trabalho.

Conforme Smith e Wren (2010) em [12], o plágio consiste na utilização de ideias ou trabalho de outrem, sem aviso ou autorização. Para Hartmann (2006) em [4], plágio consiste na reprodução parcial ou integral de uma propriedade intelectual e/ou artística.

O plágio pode se manifestar de diversas formas. Segundo Barnbaum (2002) em [1], os tipos mais frequentes são:

- **Copiar e Colar:** copiar e colar parte ou o texto completo, sem citar a fonte;
- **Mudança na Frase/Paráfrase:** reordenação das palavras mantendo o sentido;
- **Estilo:** basear em trabalho de outros autores mantendo a mesma estrutura;
- **Metáforas:** metáforas são usadas para fazer uma ideia mais clara e dar ao leitor uma analogia que toca os sentidos ou emoções melhor do que uma simples descrição do objeto ou processo. Caso utilize metáforas de outros autores para ilustrar uma ideia importante, atribuir os créditos ao autor original.
- **Ideia:** se o autor do artigo fonte expressa uma ideia, solução criativa ou sugere uma solução para um problema, ela deve ser claramente atribuída ao autor.

B. Detectores automáticos de plágio

Detectores de plágio são ferramentas automatizadas que têm por finalidade identificar similaridade entre dois ou mais documentos de texto. Segundo McKeever (2006) em [8], o processo de detectar plágio possui um grande custo computacional, sendo necessário tomar medidas para reduzir o domínio de comparação e reduzir o tempo das análises.

Existem no mercado diversas ferramentas automatizadas de detecção de plágio, algumas são gratuitas e outras pagas. Em sua essência as ferramentas são similares, pois possuem o mesmo propósito, identificar o plágio, mas cada uma delas implementa técnicas e algoritmos diferentes.

Para Clough (2000) em [2], alguns fatores podem ser utilizados pelas ferramentas para detectar o plágio, como:

- **Uso de vocabulário:** o uso de vocabulários avançados que não são frequentes em textos do mesmo autor podem ser indícios de plágio;
- **Mudança de vocabulário:** mudanças de vocabulário no decorrer do texto podem ser indícios de plágio;
- **Quantidade de similaridade entre textos:** sempre existirá uma similaridade entre os textos escritos sobre o mesmo tema. Mas é improvável o compartilhamento de grande quantidade de texto;
- **Erros de gramática comum:** a ocorrência de erros gramaticais idênticos em dois textos distintos pode indicar o plágio.

Em seu trabalho Kang, Gelbukh e Han (2006) em [5] classificam o grau de dificuldade de detecção de plágio em documentos de texto segundo a característica do tipo de plágio. Na Figura 1 nota-se que é mais fácil identificar plágio em cópia exata de documentos do que em documentos que tiveram alteração em sua estrutura.

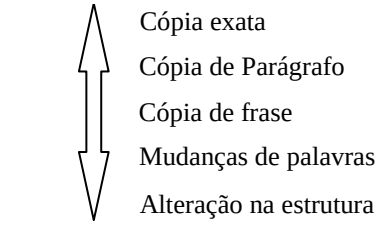

Figura 1: Padrões de plágio e seus níveis de sofisticação.
Fonte: Adaptado de Kang, Gelbukh e Han (2006) em [5].

Para Mussini (2008) em [9], o plágio em documentos de texto é mais difícil de detectar do que em código fonte, pois a gramática completa da linguagem de programação é limitada e pode ser definida e especificada. Quando se trata de documento de texto, além de ser difícil de impor limites, pode haver ambiguidade.

III. SELEÇÃO DAS FERRAMENTAS

Para determinar o conjunto de ferramentas a serem analisadas neste trabalho de pesquisa, fizeram-se diversas pesquisas no *Google*. Foram utilizadas as palavras- chave: "Detector de plágios", "*Plagiarism detection*", "*Plagiarism detection software*", *"detect copying in text documents"* e "*automated document comparison*". Para cada consulta foram verificadas as 10 primeiras páginas de resposta com 10 ocorrências cada. As ferramentas encontradas nessas buscas estão listadas no Quadro 1.

Após buscas, estabelecem-se critérios para reduzir a quantidade de ferramentas e selecionar algumas para análise. Os critérios adotados foram:

- ser gratuita;
- permitir análises em documentos de texto;
- ser uma ferramenta para *desktop*.

Das ferramentas acima foram selecionadas: *Ferret 4.0, Sherlock, Viper, CopyCath Gold, WCopyFind 2.7, Pl@giarism.* No entanto, duas ferramentas não foram consideradas: *Pl@giarism e Viper*. A primeira encontra-se em fase de testes e a segunda apresentou problemas em suas análises, os quais não foram solucionados.

QUADRO 1
FERRAMENTAS DE DETECÇÃO DE PLÁGIO

Ferramentas	**Gratuita?**	**Permite análise em documentos de texto?**	***Desktop*?**
Plagius	Não	Sim	Sim
Farejador de Plágio	Não	Sim	Sim
Ferret 4.0	Sim	Sim	Sim
*Yap*3	Sim	Não	Sim
Sherlock	Sim	Sim	Sim
Moss	Sim	Não	Sim
EVE2	Não	Sim	Sim
CROT Antiplagiarism	Não	Sim	Sim
iPlagiarism Check	Não	Sim	Não
Viper	Sim	Sim	Sim
Plagiarism Detector	Não	Sim	Sim
AntiPlagiarist	Não	Sim	Sim
CopyCatch Gold	Sim	Sim	Sim
Turnitin	Não	Sim	Não
Plagiarized.org	Sim	Não	Não
CrossRefme	Sim	Não	Não
WCopyFind 2.7	Sim	Sim	Sim
Grammarly	Não	Sim	Não
Pl@giarism	Sim	Sim	Sim
Dupli Checker	Sim	Não	Não
Check for plagiarism	Não	Sim	Não
Academic Plagiarism	Não	Sim	Não
Plagiarism Search	Não	Sim	Não
Plagiarisma.net	Não	Sim	Não
Plagiarism Checker	Sim	Não	Não
CopyScape	Sim	Não	Não
The Plagiarism	Não	Sim	Não
DOC Cop	Sim	Sim	Não
Plagiarism detection	Não	Sim	Não
SIM	Sim	Não	Sim

IV. ANÁLISE COMPARATIVA E DISCUSSÃO DOS DADOS

A análise comparativa foi dividida em: a) análise qualitativa, cujo objetivo é comparar os principais recursos oferecidos pelas ferramentas; e b) análise quantitativa, cujo objetivo é aplicar testes de eficácia, sensibilidade e desempenho nas ferramentas selecionadas.

A. Análise qualitativa das ferramentas

Para análise qualitativa, os seguintes critérios foram adotados:

- ***WEB*:** a ferramenta possui versão *WEB*?
- **Análise em código fonte:** as ferramentas oferecem análise em código fonte?
- **Apresentação dos resultados:** para apresentação dos resultados foram adotados:
  (*) – Ruim: relatório que omite informações relevantes e mal estruturadas;
  (**) – Razoável: relatório de difícil interpretação e informações irrelevantes;
  (***) – Boa: relatório simples e organizado, com informações relevantes.
- **Extensões suportadas:** quais extensões de documentos de texto são suportadas?
- ***FAQ*:** a ferramenta possui *FAQ* (*Frequently Asked Questions)*?
- **Multiplataforma:** pode ser usada em vários sistemas operacionais?
- **Usabilidade:** para usabilidade foram adotados:
  (*) – Ruim: para as ferramentas que proporcionam pouca facilidade de uso;
  (**) – Razoável: para as ferramentas que proporcionam facilidade de uso razoável;
  (***) – Boa: para as ferramentas que proporcionam boa facilidade de uso.
- **Análise na *WEB*:** a ferramenta estende sua capacidade de análise aos motores de busca da *WEB*?
- **Cadastro:** para adquirir a ferramenta é necessário cadastrar?
- **Manual:** a ferramenta possui manual ou instruções de uso?

No Quadro 2, apresenta-se um quadro resumo da análise qualitativa propiciando uma comparação direta entre as ferramentas, seguindo os critérios estabelecidos a *priori*.

QUADRO 2
COMPARATIVO ENTRE AS FERRAMENTAS ESTUDADAS

Critérios de Comparação	***Ferret***	***Sherlock***	***CopyCatch Gold***	***WCopyFind***
Versão *WEB*	Não	Não	Não	Não
Análise em código fonte	Sim	Sim	Não	Não
Apresentação dos resultados	**	*	**	**
Extensões suportadas	DOC, RTF, TXT, PDF	TXT	RTF, DOC, TXT	TXT, DOC
Manual	Sim	Não	Sim	Sim
FAQ	Não	Não	Não	Sim
Multiplataforma	Sim	Sim	Sim	Sim
Usabilidade	***	*	**	***
Análise na *WEB*	Não	Não	Não	Não
Cadastro	Não	Não	Não	Não

Percebe-se através do Quadro 2, que as ferramentas selecionadas não necessitam de cadastro, não possuem versão oferecida como serviço *WEB*, e também não estendem suas análises a documentos presente na *WEB*.

As ferramentas *Ferret*, *CopyCatch Gold* e *WCopyFind* possuem manual detalhado de suas funcionalidades. No entanto, apenas a ferramenta *WCopyFind* possui *FAQ* em seu *site* oficial.

Todas as ferramentas apresentadas são multiplataforma e permitem análises de documentos no formato TXT. Das ferramentas, apenas o *Ferret* 4.0 e *Sherlock* detecta plágio em códigos fontes.

A ferramenta *Sherlock* foi a única classificada como ruim no critério usabilidade, a falta de uma interface gráfica tornou-a de difícil utilização. Já as ferramentas *WCopyFind* e *Ferret* foram as que apresentaram boa usabilidade. Em relação à apresentação dos resultados, nenhuma foi classificada como boa, em geral, as ferramentas apresentam relatórios mal organizados, com informações irrelevantes ou poucas informações.

B. Análise quantitativa das ferramentas

Para compor a base de testes, foram utilizados 47 livros retirados do Projeto Gutenberg (2011) em [11], onde encontra-se uma grande quantidade de livros digitalizados e disponíveis. Os testes foram realizados cinco vezes com documentos diferentes, e cada documento sofreu três execuções. Os valores apresentados nas tabelas seguintes são as médias dos resultados obtidos através das execuções nos testes realizados.

1) Teste de eficácia

O objetivo do teste de eficácia nas ferramentas de detecção de plágio é verificar se as similaridades apresentadas pelas ferramentas correspondem com as similaridades esperadas.

Para realizar esse teste, foram criados dez documentos a partir de um documento original de 500 palavras. Os documentos possuíam similaridades diferentes e conhecidas.

Os testes foram conduzidos da seguinte maneira: os documentos de entrada 01 foram submetidos a análises com os documentos de entrada 02 para obter a similaridade esperada (ver Tabela 1).

TABELA 1
SIMILARIDADES ESPERADAS

Documento de entrada 01	**Documento de entrada 02**	**Similaridades esperadas**
Documento Original	Documento diferente	0%
Documento Original	Documento 1	10%
Documento Original	Documento 2	20%
Documento Original	Documento 3	30%
Documento Original	Documento 4	40%
Documento Original	Documento 5	50%
Documento Original	Documento 6	60%
Documento Original	Documento 7	70%
Documento Original	Documento 8	80%
Documento Original	Documento 9	90%
Documento Original	Documento 10	100%

O Documento diferente trata-se de um texto distinto sem qualquer relação com o documento original; o Documento 1 possui 10% (50 palavras) do documento original; o Documento 2 possui 20% (100 palavras) do documento original; o Documento 3 possui 30% (150 palavras) do documento original; e assim por diante, até o Documento 10 possui 100% do documento original (trata-se de uma cópia exata).

Os resultados obtidos podem ser vistos na Tabela 2.

TABELA 2
TESTE DE EFICÁCIA DAS FERRAMENTAS SELECIONADAS

Documentos / % esperada	***Ferret***	***Sherlock***	***CopyCatch Gold***	***WcopyFind***
	Média das similaridades obtidas			
Doc. 0 / 0%	0%	0%	21%	15%
Doc. 1 / 10%	9%	8%	23%	10%
Doc. 2 / 20%	18%	22%	39%	20%
Doc. 3 / 30%	27%	33%	52%	30%
Doc. 4 / 40%	39%	48%	63%	40%
Doc. 5 / 50%	48%	55%	72%	50%
Doc. 6 / 60%	59%	66%	80%	60%
Doc. 7 / 70%	69%	74%	85%	70%
Doc. 8 / 80%	78%	82%	90%	79%
Doc. 9 / 90%	87%	88%	95%	90%
Doc. 10 / 100%	100%	100%	100%	100%

Para obter uma pontuação que indique a ferramenta mais adequada neste teste, as similaridades obtidas na Tabela 2 foram subtraídas das similaridades esperadas, em seguida, o módulo dos valores foram tirados e, por fim, realizou-se o somatório dos valores das colunas da Tabela 3. Nesse teste, considera-se a ferramenta mais adequada a que possuir o menor somatório.

TABELA 3
|SIMILARIDADE OBTIDA – SIMILARIDADE ESPERADA|

Documentos / % esperada	***Ferret***	***Sherlock***	***CopyCatch Gold***	***WcopyFind***
Doc. 0 / 0%	0	0	21	15
Doc. 1 / 10%	1	2	13	0
Doc. 2 / 20%	2	2	19	0
Doc. 3 / 30%	3	3	22	0
Doc. 4 / 40%	1	8	23	0
Doc. 5 / 50%	2	5	22	0
Doc. 6 / 60%	1	6	20	0
Doc. 7 / 70%	1	4	15	0
Doc. 8 / 80%	2	2	10	1
Doc. 9 / 90%	3	2	05	0
Doc. 10 / 100%	0	0	0	0
Somatório das colunas:	**17**	**35**	**169**	**16**

Verifica-se que as ferramentas não identificam o plágio como esperado. Os motivos podem ser vários, como: técnicas utilizadas, os algoritmos de comparação ou os métodos utilizados para calcular a similaridade entre os documentos.

Dentre os resultados obtidos, a ferramenta *CopyCatch Gold* apresentou os piores resultados, consequentemente, o maior somatório. As similaridades obtidas foram muito acima das similaridades esperadas.

A ferramenta *WCopyFind* obteve os melhores resultados, pois apresentou as menores variações nas similaridades encontradas, consequentemente, o menor somatório.

As ferramentas *CopyCatch Gold e WCopyFind* foram as únicas ferramentas que apresentaram similaridades diferentes

de 0% nos testes realizados com documentos diferentes.

As ferramentas comparadas mostraram-se eficazes na análise de cópias exatas.

2) Teste de Sensibilidade

O plágio pode variar do mais simples ao mais complexo. Sendo assim, realizou-se o teste de sensibilidade, que tem por objetivo verificar o quanto as ferramentas são sensíveis a alterações nos documentos. Para compor o teste de sensibilidade, foram usados 10 documentos modificados a partir de um original de 500 palavras.

O teste de sensibilidade foi dividido em três grupos:

- **Deslocamento:** parágrafos e frases foram deslocados aleatóriamente;
- **Inserção:** foram inseridas no documento original novas palavras. Essas palavras foram organizadas em parágrafos;
- **Paráfrase:** foram alteradas palavras e frases mantendo o sentido original e a mesma quantidade de palavras do documento original.

Com esse teste é possível apontar os pontos fortes e fracos das ferramentas e também identificar em quais condições as ferramentas apresentam os melhores resultados. Variando os documentos de entrada e efetuando comparações com os documentos originais, espera-se que as medidas exibidas pelas ferramentas sejam sempre de 100%, ou seja, independente das alterações nos documentos, as ferramentas devem ser capazes de constatar o plágio. As similaridades obtidas nos testes realizados podem ser conferidas na Tabela 4.

TABELA 4
TESTE DE SENSIBILIDADE DAS FERRAMENTAS SELECIONADAS

		Ferret	*Sherlock*	*CopyCatch Gold*	*WCopyFind*
		Média das similaridades obtidas			
Deslocamento	**Parágrafos.**	96%	96%	100%	100%
	Frases.	90%	90%	100%	100%
	Parágrafos e frases.	90%	81%	100%	100%
Inserção	**100 novas palavras.**	85%	82%	92%	100%
	300 novas palavras.	65%	57%	82%	100%
	500 novas palavras.	56%	43%	74%	100%
Paráfrase	**de 05% do texto.**	93%	93%	99%	98%
	de 15% do texto.	84%	80%	97%	95%
	de 30% do texto.	73%	71%	95%	91%
	de 50% do texto.	60%	54%	92%	86%

Para obter uma pontuação que indique a melhor ferramenta, as similaridades obtidas na Tabela 4 foram subtraídas das similaridades esperadas (neste caso sempre será 100%). Em seguida, foram calculados os somatórios dos valores das colunas da Tabela 5. Considera-se a ferramenta mais adequada neste teste a que possuir o somatório mais próximo de zero.

TABELA 5
SIMILARIDADE OBTIDA – SIMILARIDADE ESPERADA

		Ferret	*Sherlock*	*CopyCatch Gold*	*WcopyFind*
Deslocamento	**Parágrafos.**	4	4	0	0
	Frases.	10	10	0	0
	Parágrafos e frases.	10	19	0	0
Inserção	**100 novas palavras.**	15	18	8	0
	300 novas palavras.	35	43	18	0
	500 novas palavras.	44	57	26	0
Paráfrase	**de 05% do texto.**	7	7	1	2
	de 15% do texto.	16	20	3	5
	de 30% do texto.	27	29	5	9
	de 50% do texto.	40	46	8	14
	Somatório das colunas	**207**	**254**	**69**	**30**

Verifica-se que todas as ferramentas são sensíveis a determinadas alterações, ou seja, documentos plagiados que sofrem alterações podem não ser identificados como plágio em algumas ferramentas ou podem apresentar similaridade inadequada.

As ferramentas *WCopyFind e CopyCatch Gold* mostraram-se insensíveis ao deslocamento de texto, no entanto, somente a ferramenta *WCopyFind* mostrou-se insensível a inserção de texto. No que se refere à paráfrase, *CopyCatch Gold* mostrou-se pouco sensível.

As ferramentas *Sherlock* e *Ferret* apresentaram os maiores somatórios, pois se mostraram sensíveis nos três grupos analisados, já a ferramenta *WCopyFind* apresentou o menor somatório. Para esse teste foi considerada a melhor ferramenta.

3) *Teste de desempenho*

O ideal para esse teste seria realizar uma análise baseada nos algoritmos, porém, os fabricantes das ferramentas não disponibilizam informações a respeito dos algoritmos ou técnicas utilizadas. Também, poderia ser examinado baseando no tempo do sistema, mas, algumas ferramentas não fornecem o tempo de análise de plágio. Sendo assim, os testes foram

realizados manualmente com auxílio de um cronômetro. Logo, o teste realizado neste trabalho não é preciso, mas sugere um tempo de execução aproximado de cada ferramenta.

Para realizar os testes, utilizou-se um Computador Core 2 *Quad* com 4 *Gbytes* de memória *RAM* e um disco rígido *Serial ATA II* de *7200 RPM*, rodando *Windows* 7 *Ultimate*.

Para o teste de desempenho, foram utilizados 10 documentos que variam de 10.000 palavras (25 páginas) a 100.000 palavras (180 páginas) com intervalo de 10.000 palavras como representado na Tabela 6.

TABELA 6
DOCUMENTOS TESTE DE DESEMPENHO

Documentos	Constituído de:	Total de Palavras
Documento 01 =	10.000 palavras	10.000
Documento 02 =	Documento 01 + 10.000 palavras	20.000
Documento 03 =	Documento 02 + 10.000 palavras	30.000
Documento 04 =	Documento 03 + 10.000 palavras	40.000
Documento 05 =	Documento 04 + 10.000 palavras	50.000
Documento 06 =	Documento 05 + 10.000 palavras	60.000
Documento 07 =	Documento 06 + 10.000 palavras	70.000
Documento 08 =	Documento 07 + 10.000 palavras	80.000
Documento 09 =	Documento 08 + 10.000 palavras	90.000
Documento 10 =	Documento 09 + 10.000 palavras	100.000

Para cada ferramenta, os testes foram realizados com três grupos de documentos, são eles:

- **Grupo A** - Documentos idênticos (100% plágio);
- **Grupo B** - Documentos semelhantes (50% plágio);
- **Grupo C** - Documentos distintos (0% plágio).

As análises foram realizadas da seguinte forma: o Documento 1 (original de 10.000 palavras) foi comparado com o Documento 1 do grupo A (10.000 palavras - cópia exata), Documento 1 do grupo B (10.000 palavras – 50% plagiado) e Documento 1 do grupo C (10.000 palavras - diferente). O Documento 2 (original de 20.000 palavras) foi comparado com Documento 2 do grupo A (20.000 palavras - cópia exata), Documento 2 do grupo B (20.000 palavras – 50% plagiado) e Documento 2 do grupo C (20.000 palavras - diferente). O Documento 3 (original de 30.000 palavras) foi comparado com Documento 3 do grupo A (30.000 palavras - cópia exata), Documento 3 do grupo B (30.000 palavras – 50% plagiado) e Documento 3 do grupo C (30.000 palavras - diferente). Esse procedimento foi repetido em todos os Documentos.

Com o teste de desempenho é possível observar o comportamento da ferramenta mediante aos três grupos analisados. Também é possível identificar para qual tipo de plágio a ferramenta é mais eficiente e apresenta os menores tempos. Os tempos das análises nos três grupos podem ser vistos na Tabela 7.

O tempo demonstrado na Tabela 7, não é somente o tempo gasto nas comparações, mas desde o momento que o documento foi submetido até o momento em que os resultados foram exibidos, ou seja, inclui o tempo de processamento do documento, de comparação e processamento dos resultados. Em algumas ferramentas o tempo obtido foi muito pequeno e colocou-se a média do menor tempo medido manualmente.

A ferramenta Sherlock apresentou tempo de análise extremamente baixo quando foram testados os documentos de 10.000 a 60.000 palavras. A ferramenta respondia imediatamente e em tempo menor do que clicar para disparar e clicar para parar o cronômetro. Ao fazer o disparar e parar o

TABELA 7
TESTE DE DESEMPENHO DAS FERRAMENTAS SELECIONADAS

	Ferret			***Sherlock***			***CopyCatch Gold***			***WCopyFind***		
	A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Documento 1	0.34s	0.38s	0.37s	-	-	-	2.27s	1.84s	1.35s	0.70s	0.95s	0.94s
Documento 2	0.38s	0.45s	0.43s	-	-	-	13.37s	7.34s	5.14s	0.98s	1.13s	1.03s
Documento 3	0.45s	0.50s	0.48s	-	-	-	28.56s	16.40s	11.58s	1.12s	1.22s	1.23s
Documento 4	0.51s	0.57s	0.57s	-	-	-	48.89s	30.50s	21.16s	1.17s	1.34s	1.32s
Documento 5	0.61s	0.62s	0.64s	-	-	-	77.23s	46.22s	34.15s	1.27s	1.40s	1.44s
Documento 6	0.66s	0.72s	0.70s	-	-	-	112.47s	68.22s	45.81s	1.32s	1.48s	1.51s
Documento 7	0.72s	0.79s	0.76s	0.19s	0.20s	0.19s	147.84s	94.23s	60.23s	1.41s	1.63s	1.57s
Documento 8	0.78s	0.84s	0.81s	0.22s	0.23s	0.22s	190.82s	127.70s	91.51s	1.49s	1.67s	1.66s
Documento 9	0.88s	0.90s	0.89s	0.27s	0.28s	0.27s	247.51s	165.06s	122.56s	1.55s	1.76s	1.73s
Documento 10	1.01s	1.02s	1.06s	0.32s	0.32s	0.31s	317.07s	216.38s	175.83s	1.62s	1.85s	1.79s
∑ da coluna	6.36s	6.78s	6.69s	1.00s	1.02s	1.00s	1186.03s	773.89s	569.33s	12.64s	14.43s	14.21s
∑ Total:	19.83s			3.02s			2529.25s			41.28s		

cronômetro atingiu-se 0.19s, e o sistema processou antes disto.

Das ferramentas comparadas, Sherlock apresentou os menores tempos de execução. Os tempos de análise nos três grupos testados tiveram pouca variação.

Nos testes realizados, as ferramentas *WCopyFind* e *Ferret* 4.0 mostraram-se mais rápidas nas comparações dos documentos do grupo A, ou seja, quando há 100% de plágio. No entanto, mostraram-se lentas nas análises do grupo B, quando há 50% de similaridade. Entre essas duas ferramentas, a *Ferret* 4.0 apresentou os melhores tempos.

A ferramenta *CopyCatch Gold*, diferente das outras ferramentas mostrou-se mais lenta nas comparações de documentos idênticos e mais rápida em documentos não plagiados.

A Tabela 8 elucida uma classificação geral das ferramentas baseando-se nos resultados dos testes obtidos neste trabalho. Para determinar a melhor ferramenta, foram calculados os somatórios e os produtórios das posições obtidas pelas ferramentas em cada teste. A ferramenta que apresentar o menor somatório é a mais adequada. Em caso de empates nos somatórios, os produtórios podem ser utilizados como critério de desempate.

TABELA 8
TESTE DE SENSIBILIDADE DAS FERRAMENTAS SELECIONADAS

	Ferret	***Sherlock***	***CopyCatch Gold***	***WCopyFind***
Teste de eficácia	2º	3º	4º	1º
Teste de sensibilidade	3º	4º	2º	1º
Teste de desempenho	2º	1º	4º	3º
∑	7	8	10	5
Π	12	12	32	3

Baseando-se nos valores da Tabela 10, verifica-se que a ferramenta WCopyFind apresentou os melhores resultados e se sobressaiu em detrimento das outras.

V. CONCLUSÕES E TRABALHOS FUTUROS

O plágio pode apresentar vários tipos e níveis de complexidade. Por isso, é difícil de identificar. Em alguns casos, uma análise manual torna-se impraticável. Diante disso, surgem diversas ferramentas automatizadas de detecção de plágio, algumas focadas na detecção de plágio em código fonte e outras em documento de texto. Entretanto, selecionar qual ferramenta utilizar tornou-se uma tarefa árdua.

Os professores devem utilizar de ferramentas de plágio para evitar comportamentos antiéticos e garantir a originalidade dos trabalhos. Dessa forma, visando facilitar e reduzir o esforço empregado na escolha das ferramentas, este trabalho propôs a elaboração de um quadro comparativo gerado a partir dos testes de eficácia, sensibilidade e desempenho realizado na comparação quantitativa.

Ao longo do trabalho foram realizadas comparações, qualitativa e quantitativa entre ferramentas de detecção de plágio gratuitas. Na comparação qualitativa, as ferramentas tiveram seus recursos comparados de acordo com os critérios estabelecidos. Entretanto por ser puramente descritiva, a comparação qualitativa não permite apontar qual é a melhor ferramenta. Na comparação quantitativa observou-se que as ferramentas de detecção de plágio podem fornecer medidas precisas em suas análises. Considerando-se os três testes realizados, eficácia, sensibilidade e desempenho, verificou-se que a ferramenta *WCopyFind* se sobressaiu em detrimento das outras.

Como trabalhos futuros, propõem-se:

- aplicação dos testes em ferramentas de detecção de plágio em código fonte;
- análise estatística dos resultados obtidos nos testes realizados visando melhorar as conclusões;
- modelar uma curva que norteie a relação, precisão e tempo de processamento para cada ferramenta.

REFERÊNCIAS

[1] BARNBAUM, C. *PLAGIARISM: A Student's Guide to Recognizing It and Avoiding It.* 2002. [Online]. Disponível em: <http://www.valdosta.edu/~cbarnbau/personal/teaching_MISC/plagiarism.htm>. Acesso em: 12 Outubro 2010.

[2] CLOUGH, P. *Plagiarism in natural and programming languages: an overview of current tools and technologies*. Department of Computer Science, University of Sheffield. p. 1-31. 2000.

[3] HAGE, J.; RADEMAKER, P.; VUGT, N. V. *A comparison of plagiarism detection tools*. Utrecht University. Utrecht, The Netherlands, p. 28. 2010.

[4] HARTMANN, E. *Variações sobre plágio. Confraria - Arte e Literatura*, 2006. Disponivel em: <http://acd.ufrj.br/~confrariadovento/numero8/ensaio03.htm>. Acesso em: 12 Outubro 2010.

[5] KANG, N.; GELBUKH, E.; HAN, S. *PPChecker: Plagiarism Pattern Checker in Document Copy Detection. In: SOJKA, P.; KOPECEK, I.; PALA, K. Text, Speech and Dialogue*. Springer Berlin / Heidelberg, v. 4188, 2006. p. 661-667.

[6] KLEIMAN, A. B. *Análise e comparação qualitativa de sistemas de detecção de plagio em tarefas de programação*. Dissertação (mestrado), Universidade Estadual de Campinas. Campinas, SP, p. 94. 2007.

[7] LANCASTER, T. *Effective and Efficient Plagiarism Detection*. 368 f. Tese (Phd) - South Bank University, London, Uk, 2003.

[8] MCKEEVER, L. *Online plagiarism detection services – Saviour or scourge? Assessment & Evaluation in Higher Education*, v. 31, n. 2, p. 155-165, 2006.

[9] MUSSINI, J. A. *Novas arquiteturas Para detecção de plágio baseadas em redes P2P*. PUC Paraná. Curitiba PR, p. 107. 2008.

[10] PLAGIARISM.ORG. *What is plagiarism? Plagiarism.org*. [Online]. Disponivel em: <http://www.plagiarism.org/plag_article_what_is_plagiarism.html>. Acesso em: 14 outubro 2011.

[11] PROJECT GUTENBERG, 2011. [Online]. Disponivel em: <http://www.gutenberg.org/browse/scores/top>. Acesso em: 26 Maio 2011.

[12] SMITH, N.; WREN, K. R. *Ethical and legal aspects part 2: plagiarism-"what is it and how do I avoid it?"* American Society of PeriAnesthesia Nurses, v. 25, n. 5, p. 327-330, 2010.

[13] TEDFORD, R. *Plagiarism detection programs: A comparative evaluation.* College & University Media Review, v. 9, n. 2, p. 111-118, 2003.