CENTRO UNIVERSITÁRIO UNIVATES
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
CURSO DE ENGENHARIA DA COMPUTAÇÃO

COPYTOPASTE - DESENVOLVIMENTO DE SISTEMA DE DETECÇÃO DE PLÁGIO

JOSÉ FERNANDO ECKERT

Lajeado
2012

JOSÉ FERNANDO ECKERT

COPYTOPASTE - PROPOSTA DE SISTEMA DE DETECÇÃO DE PLÁGIO

Monografia apresentada ao Centro de Ciências Exatas e Tecnológicas do Centro Universitário UNIVATES, Disciplina de Trabalho de Conclusão de Curso II, do curso de Engenharia da Computação, como parte dos requisitos para a obtenção do título de Bacharel em Engenharia da Computação.

Orientador: Fabrício Pretto

Lajeado, julho de 2012

JOSÉ FERNANDO ECKERT

COPYTOPASTE - DESENVOLVIMENTO DE SISTEMA DE DETECÇÃO DE PLÁGIO

Este trabalho foi julgado adequado para a obtenção do título de Bacharel em Engenharia da Computação e aprovado em sua forma final pelo Orientador e pela Banca Examinadora.

Orientador: _____

Prof. Fabrício Pretto, UNIVATES

Mestre pela PUC RS – Porto Alegre, Brasil

Banca Examinadora:

Prof. Fabrício Pretto, UNIVATES

Mestre pela PUC RS – Porto Alegre, Brasil

Prof. Marcelo de Gomensoro Malheiros, UNIVATES

Mestre pela UNICAMP – Campinas, Brasil

Prof. Luis Antônio Schneiders, UNIVATES

Mestre pela UFRGS – Porto Alegre, Brasil

Lajeado, julho de 2012

RESUMO

Com o surgimento da Internet houve muitas mudanças quanto às relações humanas, disposições de obras protegidas pelos direitos autorais e forma de ensino adotada pelas escolas e demais centros educacionais. As mudanças ocasionadas por redes sociais, centros de armazenamento de arquivos, bibliotecas digitais e o compartilhamento direto facilitaram o acesso a obras e induziram o plágio devido ao sentimento de impunidade no mundo virtual. Para mudar este paradigma o presente trabalho tem como principal objetivo orientar professores, alunos e demais cidadãos quanto à prática do plágio, demonstrando que é possível determinar valores referentes à similaridade entre documentos e identificar o plágio. Para possibilitar a identificação de plágio, este trabalho propõe vários métodos de tratamento de texto e comparação de similaridade, trazendo um resultado rico em informações e indicando pontos do texto que possuem similaridade com documentos disponíveis no mundo virtual. Com acesso livre à solução mediante cadastro, o ambiente de pesquisa proporciona a escolha do grau de comparação que determina as ferramentas que o sistema irá utilizar para realizar a busca e o tempo da análise. O resultado exibido é produzido pelo uso em conjunto de métodos de limpeza textual e comparação de similaridade.

Palavras-chave: Plágio, Processamento de texto, Detecção de similaridade.

ABSTRACT

With the advent of the Internet there were many changes with human relationships, arrangements of works protected by copyright and the teaching methods adopted by schools and other educational centers. The changes brought about by social networks, storage centers, digital libraries and direct sharing facilitated the access to works and induced plagiarism because of the feeling of impunity in the virtual world. To change this paradigm this work aims to guide teachers, students and other citizens about the practice of plagiarism, demonstrating that it is possible to determine values for the similarity between documents and identify plagiarism. To enable the plagiarism detection, this work proposes several methods of word processing and comparison of similarity, bringing a wealth of information and results indicating points of the text that have a high degree of similarity with other published documents. With free access to the solution, the research environment provides the choice of the degree of comparison determines which tools are used by the system to perform the search and the time of analysis. The result displayed is produced by use of text cleaning methods and comparison of textual similarity.

Palavras-chave: Plagiarism, Text processing, Similarity comparison.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ABDR: Associação Brasileira de Direitos Reprográficos

ABNT: Associação Brasileira de Normas Técnicas

AJAX: Asynchronous JavaScript and XML

API: Application Programming Interface

ASF: Apache Software Foundation

BSD: Berkeley Software Distribution

CDNA: Conselho Nacional de Direito Autoral

CERN: European Particle Physics Laboratory

CORDS: Copyright Office Eletronic Registration, Recordation and Deposit System

CSS: Cascading Style Sheet

CTP: Copy To Paste

ECAD: Escritório Central de Arrecadação e Distribuição

FSF: Free Software Foundation

GPL: General Public License

HTML: Hyper Text Markup Language

INPI: Instituto Nacional da Propriedade Intelectual

JSON: JavaScript Object Notation

MIT: Massachusetts Institute of Technology

MPL: Mozilla Public License

PDF: Portable Document Format

PERL: Practical Extraction and Report Language

PHP: Personal Home Page

SGBD: Sistema Gerenciador de Banco de Dados

SICAM: Sociedade Independente de Compositores e Autores Musicais

SQL: Structured Query Language

SSH: Secure Shell

URL: Uniform Resource Locator

VPS: Virtual Private Server

W3C: World Wide Web Consortium

XML: Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

O sistema de ensino vem evoluindo juntamente com a progressão da tecnologia, devido às inovações na área de interações humanas, compartilhamento de conhecimento e velocidade da informação. Antigamente a única ferramenta que o aluno possuía para o estudo era o livro, sendo que em muitos casos era uma coleção restrita de livros contidos em uma biblioteca. Estas bibliotecas localizavam-se nos centros urbanos e necessitavam de investimento proveniente dos centros administrativos dos municípios e doações da comunidade (CABRAL, 1996).

Aos poucos este paradigma foi mudando e fez-se necessário ampliar as unidades existentes, construindo novas bibliotecas em bairros menores. Neste mesmo período, as escolas e centros de formação que não eram administradas pelo Estado, começaram a criar bibliotecas internas ou privadas para atender a um público específico. Devido ao acréscimo no investimento do setor bibliotecário, os alunos tiveram maior acesso à informação e literatura existente (CABRAL, 1996).

Outro fator importante que trouxe maior informação e liberdade de acesso à literatura foi a Internet. Desde seu surgimento, são criados sítios, portais e bibliotecas digitais que alimentam esta rede e fazem com que esta seja a maior fonte de pesquisa dos estudantes na atualidade. A facilidade e rapidez de acesso à informação na rede só foi possível graças ao surgimento de centros de buscas, que auxiliam os internautas a encontrar endereços de sítios que contenham informações referentes aos dados concedidos pelo usuário (CABRAL, 1996).

Na Internet é possível realizar uma pesquisa e obter o resultado em menos de um segundo e este é o grande trunfo em relação a outros meios, tornando esta ferramenta muito poderosa. Além da informação textual o usuário encontra muitos conteúdos no formato de áudio, vídeo e imagens. Isto faz com que se estimule o plagiador a cometer a cópia do conteúdo, devido a seu fácil acesso e também por se tratar de um meio digital que ainda não possui legislação específica que contemple o todo (SANTOS, 2009).

Esta liberdade induziu e ainda induz plagiadores a cometer crimes, pois a legislação não estava preparada e não acompanhou o crescimento da Internet. Desta forma, os órgãos fiscalizadores estão utilizando o conceito de propriedade intelectual que rege os direitos autorais e industriais, para reprimir e punir os responsáveis por estes crimes, além de evitar a

disseminação de cópias indevidas e assegurar os direitos à pessoa que criou a obra. (SANTOS, 2009).

A propriedade intelectual busca assegurar os direitos referentes à criação e exploração financeira de obras intelectuais, resultando em grande incentivo aos criadores a produzir mais sem se preocupar com o uso indevido de sua obra. Com este incentivo houve um grande impulso no desenvolvimento, pois aumentou a quantidade de invenções que trouxeram grandes benefícios à população (SANTOS, 2009).

A propriedade intelectual possui várias definições de vários autores, porém seu objetivo principal é proteger o autor e garantir o direito sobre a obra. Este direito vem sendo desenvolvido há muito tempo e possui atualizações periódicas, pois deve acompanhar o desenvolvimento tecnológico agregando melhorias.

Sem leis que definem as restrições perante a utilização de material alheio, sendo este através do meio digital ou impresso, causa a impunidade e incentiva a prática de plágio. Ainda hoje as leis não estão plenamente adequadas ao ambiente virtual, porém, para possibilitar o controle do plágio estão sendo adotadas as leis criadas antes da disseminação da Internet. Este é um caso onde a legislação não acompanhou o desenvolvimento tecnológico que abriu um novo caminho para a prática do plágio (SANTOS, 2009).

Mesmo sem acompanhar a velocidade de crescimento da Internet a justiça está conseguindo concluir casos de plágio praticados através da Internet, pois estão sendo aplicadas leis existentes mesmo antes da revolução que a Internet vem produzindo. No mundo virtual há muito material com direitos de autor e proteção da lei como músicas, filmes, livros, monografias, teses e fotografias que possuem milhares de cópias espalhadas nesta rede com acesso livre para o *download* (SANTOS, 2009).

Este crime de reproduzir conteúdo criado por outra pessoa sem autorização, sendo por meio físico ou virtual é definido como plágio. O plágio também é definido como obra copiada ou utilizada como base para novos trabalhos sem dar os respectivos créditos ao autor da obra utilizada como referência. Este tipo de contrafação acontece muito, principalmente nos colégios e faculdades. Muitos dos alunos que utilizam obras alheias também fazem uso de trabalhos de terceiros para reproduzir ou criar o seu com base em outro (Barbosa, 2003).

Segundo Oliveira et al. (2007), para a prática do plágio no meio virtual muitas vezes são utilizadas as chamadas bibliotecas digitais. Estas bibliotecas estão surgindo e crescendo conforme o crescimento da Internet, e possuem conteúdo acadêmico, livros e demais obras.

Muitas destas bibliotecas têm acesso livre para qualquer pessoa e acaba sendo um amplo campo para os plagiadores.

Para frear a ação de plagiadores e minimizar o impacto que as bibliotecas digitais vêm causando, estão surgindo softwares e sistemas de combate ao plágio. A proposta desta monografia vem ao encontro deste tema que cada vez mais faz parte do dia a dia das pessoas (OLIVEIRA et al., 2007).

Com a utilização de um software que ajude aos professores detectar casos de plágio é possível reprimir este tipo de ação e desta forma, diminuir casos de plágio e trazer benefícios para a comunidade acadêmica. Para informar o usuário, o trabalho demonstra o nível de similaridade e a localização dos documentos encontrados na Internet para que o usuário possa julgar a autenticidade dos quais foram comparados.

1.1 Objetivos

Este trabalho busca informar usuários quanto a semelhanças entre documentos e possíveis casos de plágio. Para isto é necessário o desenvolvimento de uma ferramenta que utilize métodos computacionais para comparação entre documentos. Esta comparação deve demonstrar a similaridade entre os textos envolvidos. Além de demonstrar a similaridade, a ferramenta proveniente deste trabalho deve possibilitar a detecção de similaridade nas diversas formas de plágio mais utilizadas.

Para o desenvolvimento serão adotadas ferramentas livres que possibilitarão a redução dos custos do projeto. Além disso, o acesso à ferramenta deve ser de forma gratuita e possibilitar a utilização através de diversos navegadores e sistemas operacionais.

1.2 Estrutura do trabalho

O presente trabalho apresenta no Capítulo 2 as leis para a proteção das obras e como estas protegem os autores e suas obras no Brasil e no mundo. Também está definido o conceito de plágio e o histórico deste tipo de crime, traçando um comparativo entre o antes e o depois da Internet. O Capítulo 3 descreve métodos de comparação de similaridade,

tratamento de texto e ferramentas existentes que comparam documentos e retornam o nível de similaridade. Após, no Capítulo 4 está definida a implementação da ferramenta.

2 REFERENCIAL TEÓRICO

Através das criações, o ser humano pode obter novos caminhos e progredir, porém antigamente a humanidade não tinha direitos que protegiam suas criações e obras. Com o desenvolvimento foi necessário desenvolver uma proteção que garantisse ao criador os direitos de propriedade, ou seja, que o retorno financeiro sobre o objeto em questão fosse para o respectivo criador. Além desta garantia, também era necessário que se punisse os contraventores que utilizassem da pirataria e plágio para beneficiar a si mesmos através do trabalho alheio.

Para definir a proteção ao autor foi criada a Propriedade Intelectual que engloba os direitos autorais e industriais, e é descrita nas próximas seções. Com a criação destes direitos, foi notável a aceleração do desenvolvimento, pois desde então os criadores sentiram-se incentivados a produzir e reproduzir suas obras.

A seguir é abordado o conceito de propriedade intelectual juntamente com os direitos autoral e industrial. Também são descritos os sistemas *Copyright, Copyleft e Droit d´ auteur* criados internacionalmente para proteção das obras. É feita uma abordagem específica para a propriedade intelectual no Brasil, uma breve descrição da Internet e sua ligação com os direitos autorais. Após está definido reparações judiciais para o caso de plágio e é descrito também a relação entre bibliotecas digitais e o plágio. Por fim, é definido o plágio e a sua relação com o crescimento da Internet.

2.1 Propriedade intelectual

A Propriedade Intelectual é o direito referente às criações e invenções do ser humano em todos os campos de atividade e garante a proteção pública sobre o trabalho resultante de sua invenção. Este direito abrange os direitos autorais de trabalho ou criação imaterial e os direitos de propriedade industrial sobre o trabalho material (SANTOS, 2009).

Para Barbosa (2003) a Propriedade Intelectual possui os direitos referentes a obras literárias, artísticas e científicas, aos trabalhos dos intérpretes, execuções dos artistas, emissões de radiodifusão, invenções de qualquer atividade, descobertas, desenhos e modelos

industriais, marcas, proteção contra concorrência desleal e todos os direitos referentes às atividades intelectuais na indústria, ciência, literatura e arte.

Basso (2000) cita que há dois modelos conceituais para produção intelectual, o tradicional (histórico) e o atual, que possui caráter imaterial e internacional, independentemente do modelo conceitual. A Propriedade Intelectual nacional tem seu reconhecimento dificultado devido sua proteção ser insuficiente, porém juntamente com a Propriedade Intelectual internacional fica completa na defesa dos direitos e mantém maior proteção ao autor.

Habitualmente as leis que vigoram no Brasil são aplicáveis somente dentro de seu território, porém no campo da Propriedade Intelectual não há como haver somente legislação interna, portanto através de tratados e convenções são definidas leis internacionais nas quais todos países que fazem parte concordam em aplicá-las dentro de suas fronteiras.

Não há limite territorial para a Propriedade Intelectual, pois invenções de todos os cantos do mundo estão por toda parte melhorando a qualidade de vida do ser humano e gerando *royalties*[1] aos criadores das obras e editores (BASSO, 2000).

2.1.1 Direito autoral

Direito autoral define os direitos que o autor possui sobre as criações de obras intelectuais, como textos, monografias, livros e artigos. Regulamentada pela Lei nº 9.610/98, promulgada pelo presidente Fernando Henrique Cardoso e chamada Lei de Direitos Autorais, define a proteção sobre as criações do autor. Esta lei também regulamenta os direitos conexos que são chamados de vizinhos ou análogos, e protege os direitos dos artistas e demais trabalhadores de radiodifusão, televisão, teatro e produção cinematográfica (SANTOS, 2009).

Este direito resguarda, sob aspecto moral e pecuniário, a proteção somente para criações produzidas dentro do limite territorial do país de origem, porém para ampliar os limites foram criados tratados e acordos internacionais que garantem a aplicação dos direitos fora dos limites do país, desde que as condições sejam recíprocas (PAESANI, 2009).

1 Na antiguidade, Royalties eram chamados os valores que eram remetidos ao rei pelo uso de recursos naturais ou obras que o pertenciam. Atualmente são denominados desta maneira os valores pagos ao detentor de uma marca, produto ou patente.

Segundo Bittar (1999), o direito do autor é a proteção de criação que obtém maior sucesso a nível universal devido à evolução intelectual e tecnológica. Incluído nos direitos fundamentais do ser humano é lentamente regulado por leis especiais nos países civilizados. Estas leis têm por base o passado, pois possuem diretrizes básicas definidas na convenção de Berna em 1886, e atualmente são aperfeiçoadas e periodicamente revisadas.

Devido ao direito do autor ser difundido em diversos países, isso possibilitou a circulação de obras intelectuais por todo o globo, contribuindo para a aproximação entre países sem deixar de proteger os direitos dos autores. Isto ocorre porque os países devem conferir aos estrangeiros os mesmos direitos que os cidadãos nacionais possuem (BITTAR, 1999).

Em alguns países, os direitos do autor sofrem algumas alterações para possibilitar o uso da obra não licenciada em alguns casos. A legislação americana assegura o direito de uso de obra protegida pelos direitos autorais a uso acadêmico, para a crítica, a notícia e a pesquisa. Este é um conceito chamado de Fair Use, utilizado também por outros países como Israel e Coreia do Sul com um conceito similar chamado Fair Dealing (QUEIROZ, 2009).

2.1.2 Teorias do direito do autor

Para melhor entendimento foram definidas sete teorias que compõe o direito autoral e cada uma abrange parte deste direito. Estas teorias completam a estrutura, compreensão e definição do direito autoral, representadas na Tabela 1.

Tabela 1: Teorias do direito do autor

Teoria	Descrição
Teoria da natureza do direito do autor	O direito do autor compõe o direito real que envolve autor e obra, patrimonial vinculado a direito de propriedade, personalidade vinculado a direitos morais do autor, pessoal vinculado a direito privado e por último o direito especial que está em uma categoria diferente, pois possui teoria própria.
Teoria da natureza	O direito do autor compõe o direito real que envolve autor e obra

do direito do autor	patrimonial vinculado a direito de propriedade, personalidade vinculado a direitos morais do autor, pessoal vinculado a direito privado e por último o direito especial que está em uma categoria diferente, pois possui teoria própria.
Teoria do sujeito	O sujeito define o criador da obra intelectual e que produz o que era inexistente. Também é considerado criador o sujeito que coordena e dirige produções de terceiros concluídos em produto único. Na criação colaborativa o direito possui regras comuns para ambos e na derivação de obras, é necessário o consentimento do titular na obra resultante.
Teoria do objeto	A teoria do objeto caracteriza obras de caráter estético que podem ser de literatura, artes e ciência. Estas obras são protegidas pelo direito de autor mesmo que sejam de esteticidade.
Teoria do conteúdo	Os direitos da propriedade intelectual se completam na defesa dos direitos do criador e estão diretamente ligados a vínculos pessoais e patrimoniais em relação a autor e obra, mesmo após sua morte. Dentre os direitos da propriedade intelectual, o direito moral é o que mantém vinculado autor e obra. Já o direito patrimonial permite ao autor participar dos resultados financeiros de sua criação.
Teoria da circulação	Para possibilitar o uso público da obra e lograr-se diretamente ou indiretamente, é necessário fixar contratos, definindo a exclusividade de exploração ao autor e a dependência para o uso. Fora os direitos de exploração definidos no contrato, permanecem os demais direitos do autor.
Teoria da administração	A administração dos direitos juntamente com a reunião dos titulares das entidades de defesa e intervenção estatal, constitui o mecanismo

	público de apoio ao autor. O conselho Nacional de Direito Autoral (CNDA) exercia intervenção estatal, porém foi abolido e em seu lugar foram criadas associações ou fundações como a Associação Brasileira de Direitos Reprográficos (ABDR), Sociedade Independente de Compositores e Autores Musicais (SICAM) entre outros.
Teoria da tutela	A tutela é um direito que se confere a alguém, neste caso o autor ou criador da obra, para representar, usufruir e administrar. Nos direitos autorais, esta possui diferentes níveis que propiciam ao criador da obra a justa proteção no campo civil, penal e administrativo. Isto traz ao autor recorrer a medidas extrajudiciais e judiciais, de caráter civil ou penal, podendo gerar busca e apreensão do material em desacordo com as leis, reparação de danos, ação penal e outras.

Fonte: BITTAR (1999, p. 26 - 30).

2.1.3 Direito industrial

Para o direito de propriedade à criação material foi criado o direito industrial que é regulamentado pela Lei nº 9.279/96, chamada de Lei da Propriedade Industrial. Devido à necessidade de uma proteção para o criador de artigos industriais, os primeiros manuscritos sobre o direito industrial foram criados em meados de 1474 em Veneza, Itália, e definia o direito sobre a criação durante o período de dez anos, desde que o invento fosse novo, genial e que pudesse ser utilizado (SANTOS, 2009).

Para Barbosa (2003), propriedade industrial é o direito que engloba patente de invenção, modelos de utilidade, desenhos ou modelos industriais, marcas, nome comercial e repressão à concorrência desleal. Este direito não se aplica somente à indústria e comércio, mas também a todos os produtos manufaturados.

2.1.4 Direito Autoral X Direito Industrial

O direito autoral foi elaborado para atender interesses do autor de obra intelectual, independente da forma que foi documentada, executada ou recitada. Este direto não necessita

de registro para nomear o autor, ficando a critério deste, registrar ou não. Já o direito industrial, juntamente com interesses técnicos, econômicos e políticos engloba produtos industriais e impede o monopólio, como acontece com as marcas. Para o direito industrial é necessário registro no INPI – Instituto Nacional da Propriedade Intelectual para ter a proteção sobre a invenção (SANTOS, 2009).

2.1.5 *Copyright, Copyleft e Droit d´ auteur*

Por haver a necessidade de exclusividade na publicação de livros, foi criado na Inglaterra e Europa continental o sistema *Copyright* em 1557. Neste regime é assegurado o direito de cópia, ou seja, o direito do editor.

Para assegurar os direitos morais do criador da obra foi criado em meados de 1791 o regime *Droit d´ auteur*, ou seja, direito do autor que possui características inversas as do *Copyright*, pois aboliu o privilégio dos editores e atendeu as necessidades do autor (SANTOS, 2009).

O *Copyright* protegia por 21 anos obras impressas que foram formalizadas, e 14 anos para as não impressas. O prazo de 21 anos começava a contar a partir da data de impressão da obra. Já o *Droit d´ auteur* concedia a proteção referente ao autor e obra durante toda vida do criador e até mesmo após sua morte, sendo transferidos os direitos sobre a criação para os herdeiros (GANDELMAN, 2001).

Ao contrário do Copyright, o Copyleft assegura a liberdade de cópia da obra. Neste caso, desde o primeiro licenciamento o criador da obra autoriza os direitos de uso, reprodução, distribuição e alteração da obra por qualquer pessoa.

Um projeto que exemplifica o Copyleft é o software livre, que possui livre direito de uso, alteração e reprodução. O software livre possui um desenvolvimento colaborativo com a adesão de milhares de voluntários (GANDELMAN, 2001).

Outros tipos de licença existentes atualmente são utilizadas por projetos *Open Source* (código aberto) que tem como principal propósito não fornecer direito exclusivo de posse a uma pessoa. Estes projetos possuem diferentes tipos de licenças como MIT, BSD, GPL, LGPL, MPL, Creative Commons Licenses e Microsoft Shared Source Initiative (LAURENT, 2004).

2.1.6 MIT, BSD, GPL, MPL, Creative Commons

Os tipos de licenças Massachusetts Institute of Technology (MIT) e Berkeley Software Distribution (BSD) foram as pioneiras em projetos *Open Source* e definem os princípios deste modelo colaborativo. Além de estarem presentes em muitos projetos desta área, também regulam os direitos de uso em dois grandes projetos existentes atualmente que denominam-se BSDNet e FreeBSD (LAURENT, 2004).

2.1.6.1 MIT

A licença *Open Source* MIT tem as garantias definidas como segue:

> The software is provided “as is”, without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings in the software (LAURENT, 2004, p.15).

Como pode-se observar na citação anterior, esta licença exime as autores de qualquer responsabilidade perante a utilização do software. Além disso, deixa bem claro que em hipótese alguma é fornecida garantia ou reparações perante qualquer dano que o objeto em questão possa causar.

2.1.6.2 BSD

Não muito diferente da MIT a licença BSD tem garantias que igualam em seu conteúdo com a MIT, porém define que é necessária a permissão para utilizar o nome do criador do software. Esta definição protege a imagem do autor perante as próximas versões do software, não ligando seu nome à possíveis erros no software ou desenvolvimento de baixa qualidade causados por terceiros (LAURENT, 2004).

2.1.6.3 GPL

Para garantir que o software livre continue livre, é elaborado um contrato GNU-GPL (GNU General Public License) criando redes de contratos. Desta forma uma pessoa não pode alterar o software livre e vender licenças como um produto de sua criação. Criada pela Free Software Foundation FSF, o GPL forneceu grande contribuição aos projetos *Open Source* sendo o tipo de licença preferido pelos que possuem autorização da FSF.

A licença GPL garante o direito de continuação do desenvolvimento do código fonte do software que a possui. Porém descreve claramente que não é permitida a alteração da licença (LAURENT, 2004).

2.1.6.4 MPL

A Mozilla Public License (MPL) foi desenvolvida pela advogada Winifred Mitchell Baker na empresa Netscape Communications Corporation. Com grande parte herdada do Copyleft, a MPL possui características que não a deixam tão rígida, pois permite a utilização do código em conjunto com outros que podem ser proprietários. Além de possibilitar o uso do código proprietário, também é possível criar versão proprietária através do código que possui licença MPL (LAURENT, 2004).

2.1.6.5 Creative commons

Juntamente com o Copyleft que trata-se de um movimento baseado no compartilhamento de conhecimento, o projeto Creative Commons criado pelo Prof. Lawrence Lessig, tem o objetivo de expandir obras criativas. Com o uso deste sistema os autores podem licenciá-los através de licença pública, autorizando qualquer pessoa a utilizar a obra dentro dos limites da licença (PARANAGUÁ ET AL., 2009).

2.1.7 Propriedade Intelectual no Brasil

O Brasil possui legislação sobre proteção intelectual desde 28 de abril de 1809, isto o torna uma das quatro primeiras nações a desenvolver algo sobre o tema. Desde então foram

criadas novas leis que sucederam a de 1809. Uma delas é a lei de propriedade industrial de 1945 e a lei de Medeiros e Albuquerque de 1898 sobre os direitos do autor. Estas leis vigoraram até a criação do novo código penal em 1996 (BARBOSA, 2003).

Segundo Santos (2009), após a independência o Brasil ainda utilizava o sistema de privilégios, ou seja, somente editores e impressores possuíam direitos sobre as obras. No Brasil, o desenvolvimento de leis e proteções legais para propriedade intelectual é ainda muito recente, mas há cerca de trezentos projetos de lei em tramitação na Câmara dos Deputados que envolvem direitos autorais.

No plano nacional de propriedade intelectual e direito autoral, é concedido o princípio da exclusividade a favor do autor na exploração econômica, de propriedade e a transmissão dos direitos a terceiros por sucessão. Neste plano está definido em lei o tempo limite de exclusividade e após, a obra possuirá domínio público (BITTAR, 1999).

Com a criação de uma obra, seja um livro, música, peça teatral dentre outros, muitos autores desejam que o público tome conhecimento da obra e passe a utilizar, ver e escutar. Um exemplo disso é a música, pois normalmente uma pessoa compra um CD ou DVD de um artista quando já tenha conhecimento de algumas destas que fazem parte do álbum. Para tornar estas obras públicas são utilizadas como meio de transmissão a radiodifusão, os impressos ou a Internet.

Para uma obra ser transmitida através de radiodifusão é necessário que o autor se associe a uma das associações que compões o ECAD (Escritório Central de Arrecadação e Distribuição) e após estes meios de comunicação poderão reproduzi-la. Com isso, as empresas de radiodifusão devem pagar à ECAD um valor que varia conforme a quantidade de população atingida e o valor socioeconômico. Estes valores pagos são repassados pelas entidades aos compositores e artistas (PARANAGUÁ ET AL., 2009).

As associações que fazem parte da ECAD são:

- ABRAMUS (Associação Brasileira de Música e Artes).
- AMAR (Associação de Músicos, Arranjadores e Regentes).
- ASSIM (Associação de Intérpretes e Músicos).
- SBACEM (Sociedade Brasileira de Autores, Compositores e Escritores de Música).
- SICAM (Sociedade Independente de Compositores e Autores Musicais).
- SOCINPRO (Sociedade Brasileira de Administração e Proteção de Direitos Intelectuais).

- UBC (União Brasileira de Compositores).

- ABRAC (Associação Brasileira de Autores, Compositores, Intérpretes e Músicos).

- SADEMBRA (Sociedade Administradora de Direitos de Execução Musical do Brasil).

As associações, como citado anteriormente, recebem os valores correspondente a todas as emissoras de rádios e televisão do Brasil, e após repassam os valores aos compositores, músicos e intérpretes (PARANAGUÁ ET AL., 2009).

2.1.8 Internet

Quando houve o surgimento da Internet, a humanidade possuía livros, jornais e impressos como fonte de conhecimento e materialização de obras. Porém com o avanço do mundo digital, as obras começaram a ser publicadas e veiculadas através da Internet. Atualmente esta grande rede virtual possui um público maior que muitos outros tipos de mídia, como é o caso do jornal, cada vez mais comum ser encontrado em meio digital (CARDOSO, 2007).

Além de incorporar os jornais, televisão e rádio, a Internet trouxe novas formas de comunicação entre as pessoas e novos recursos. Estas ferramentas e novos recursos são redes sociais, centros de compartilhamento e armazenamento de vídeos e fotos, fóruns dentre outros (CARDOSO, 2007).

2.1.9 Direitos Autorais e a Internet

Para Santos (2009), a Internet não proveu mudanças nos direitos autorais para o judiciário, sendo que desta forma o autor continua com todos os direitos sobre suas obras. O arquivo digital de um livro, filme, música ou qualquer outro material que possua direitos autorais fora do mundo virtual continuam com suas proteções legais, porém muitas pessoas não observam tal proteção e acabam usufruindo do material como se não houvesse autor e proteção legal. O material postado na Internet possui visibilidade global, ou seja, qualquer pessoa do mundo pode acessá-lo e copiá-lo, isto pode produzir uma grande quantidade de

acessos e cópias com facilidade. Por ser difícil impor regras no mundo virtual, acabou por ser denominada “uma terra sem leis” mesmo que não seja.

Para reprodução de uma obra é necessário um contrato com o autor que possui direito exclusivo de reprodução. Mesmo que seja em formato digital a proteção é a mesma e para cada cópia, mesmo que seja no mesmo computador deve ter autorização (GANDELMAN, 2001).

Para a Internet ajudar a proteger as obras e colaborar com os direitos autorais, está sendo testado um novo sistema chamado CORDS (Copyright Office Eletronic Registration, Recordation and Deposit System) que irá permitir o registro de obras através da Internet.

2.1.10 Reparações judiciais a direitos autorais violados

Para reparação dos danos causados sobre a violação dos direitos autorais cabe a indenização por parte do réu ao autor. O cálculo da indenização compreende verba correspondente aos danos morais e patrimoniais sobre cada direito violado. O valor é estipulado de forma que desestimule uma suposta nova violação dos direitos. O valor deve ser bem acima dos valores de mercado do direito de cópia, além de ser uma reparação a altura dos danos causados ao autor (BITTAR, 1999).

2.2 Plágio e Bibliotecas Digitais

Nos anos 90 começou uma nova era para a humanidade, com mais interações e velocidade na informação, acesso ao conhecimento e ampliação dos limites que possuíam as tecnologias anteriores. Juntamente com a disseminação dos computadores, no final da década de 90 houve a popularização da Internet que continua crescendo de uma forma gigantesca (PAESANI, 2009).

A grande expansão da Internet se dá pelo fato de que as empresas estão investindo cada vez mais neste segmento. Para obter vantagem, elas procuram estarem atualizadas, e desta forma a Internet é uma ferramenta excelente para obter informações, pois nesta rede a informação chega mais rápida e privilegia os que a possuem antecipadamente. A vantagem

disto reflete principalmente nos negócios que lutam constantemente para se manterem atualizados e a frente dos concorrentes.

Para os jovens, viver sem acesso à Internet é estar fora da realidade, viver em um mundo à parte. Com a facilidade do acesso a informação e obras, muitas pessoas utilizam destas sem a devida referência, caracterizando o plágio (OLIVEIRA ET AL., 2007).

Com esta ascensão da Internet, as instituições de ensino começaram a criar bibliotecas digitais que segundo Valmorbida (2011), consistem em centros de armazenamento de conteúdo digital que garantem acesso, integridade e longo tempo de armazenamento aos dados. Este conteúdo pode ser disponibilizado a um público específico ou comunidades selecionadas.

O fácil acesso ao conteúdo acadêmico através de bibliotecas digitais e obras disponibilizadas e compartilhadas na Internet, acabou incentivando o crime do plágio que é praticado principalmente na comunidade acadêmica (MORAES, 2007).

2.2.1 Plágio

O plágio não é algo novo, desde a antiguidade têm-se registros de sanções e penalidades impostas a plagiadores, como desonra e repúdio público. Plágio vem do latim *plagiarius* que eram chamados os que roubavam escravos e após revendiam como livres, dessa forma o plágio começou a ser considerada apropriação indébita, até os dias atuais (MORAES, 2007).

Para ser considerado plágio, é necessário cópia parcial ou total de uma obra alheia sem divulgar sua autoria, como se fosse um documento autêntico. Para Cotta (1999), o significado de plágio mudou com o tempo, pois antigamente significava roubo e atualmente o conceito de plágio mudou um pouco, sendo definido como captura de total ou parte de obra intelectual.

O plágio possui alguns graus como a cópia literal que representa uma cópia idêntica da original, definido como *plagiarius simplex*. Outro grau seria o *plagiarius artifex* que é uma obra baseada em outra, porém com cópia ocultada por reescrita e reformulação do texto.

Segundo Moraes (2007), na antiguidade existia indícios de compra dos direitos de autoria para evitar o plágio, atualmente isso é proibido, pois o direito de autoria é intransferível e somente pode ser negociado o direito patrimonial sobre a obra. Isso acontece nas editoras, onde o autor pode negociar e elaborar um contrato de edição e venda de

exemplares da obra, retornando lucros para ambos sobre as vendas. Para definir um plágio não há número de linhas e palavras, pois se houvesse poderia ocorrer casos onde burlasse estas regras mesmo cometendo o plágio.

Para o desenvolvimento de um trabalho escolar, monografia, tese ou qualquer obra que utilize parte de outra, deve-se utilizar um dos tipos de descrição referenciada que estão explicadas na Tabela 2.

Tabela 2: Tipos de descrição referenciada

Tipo	Descrição
Citação	Na citação deve-se descrever exatamente como é texto original e estar devidamente referenciado. Além disso, também deve possuir aspas (quando possuir menos de 4 linhas) no início e fim da descrição ou pode estar em um parágrafo separado conforme padrões ABNT (Associação Brasileira de Normas Técnicas).
Paráfrase	A paráfrase é um texto reformulado com as próprias palavras do escritor sobre outra obra, além de ser direcionado ao público alvo. Neste caso também deve-se referenciar o autor da obra original, pois segundo Kirkpatrick (2007) mesmo que escreva com suas palavras, não torna seu este trecho descrito.
Resumo	Parecido com a paráfrase, o resumo é um texto feito com as próprias palavras e também deve referenciar o autor da obra original, porém normalmente mais curto e

	mais abrangente.
Referência	A referência é a identificação do autor da obra original que está sendo citada, resumida ou utilizada na paráfrase. Atualmente os textos jornalísticos, monografias ou teses exigem a referência completa.

Fonte: KIRKPATRICK (2007, texto digital).

Os textos plagiados são caracterizados por ser cópia ou baseado em obra alheia, por não estar devidamente referenciado e autor divergente do original. Na tabelas 3 e 4 estão nove tipos de plágio que são praticados atualmente.

Tabela 3: Tipos de plágio

Tipos	Descrição
Plágio Direto	O plágio direto é caracterizado por ser uma cópia idêntica à fonte original conforme a citação, porém o criador da obra original não é referenciado fazendo com que o leitor pense que o texto é íntegro.
Receber emprestado trabalhos de outros estudantes	Nas escolas normalmente colegas de aula emprestam seus trabalhos para outros que podem utilizar para desenvolver os próprios, porém muitas vezes acabam por copiá-lo e isto é considerado um plágio direto.
Referência incorreta ou incompleta	Isto acontece quando o leitor não identifica o texto como sendo uma obra alheia, devido a referência estar de uma forma incorreta ou somente em parte do texto. Para evitar este

	caso o escritor deve indicar quando uma paráfrase, resumo ou citação começa, termina ou é interrompida.
Plágio mosaico	Considerado o tipo de plágio mais comum, não é uma cópia idêntica a obra alheia e não identifica o autor da obra original. Caracteriza-se por palavras trocadas em cada sentença e com frases alteradas e reformuladas de um texto original. Muitas vezes estes textos podem possuir características de citações ou paráfrases sem referências.

Fonte: KIRKPATRICK (2007, texto digital).

Tabela 4: Tipos de plágio

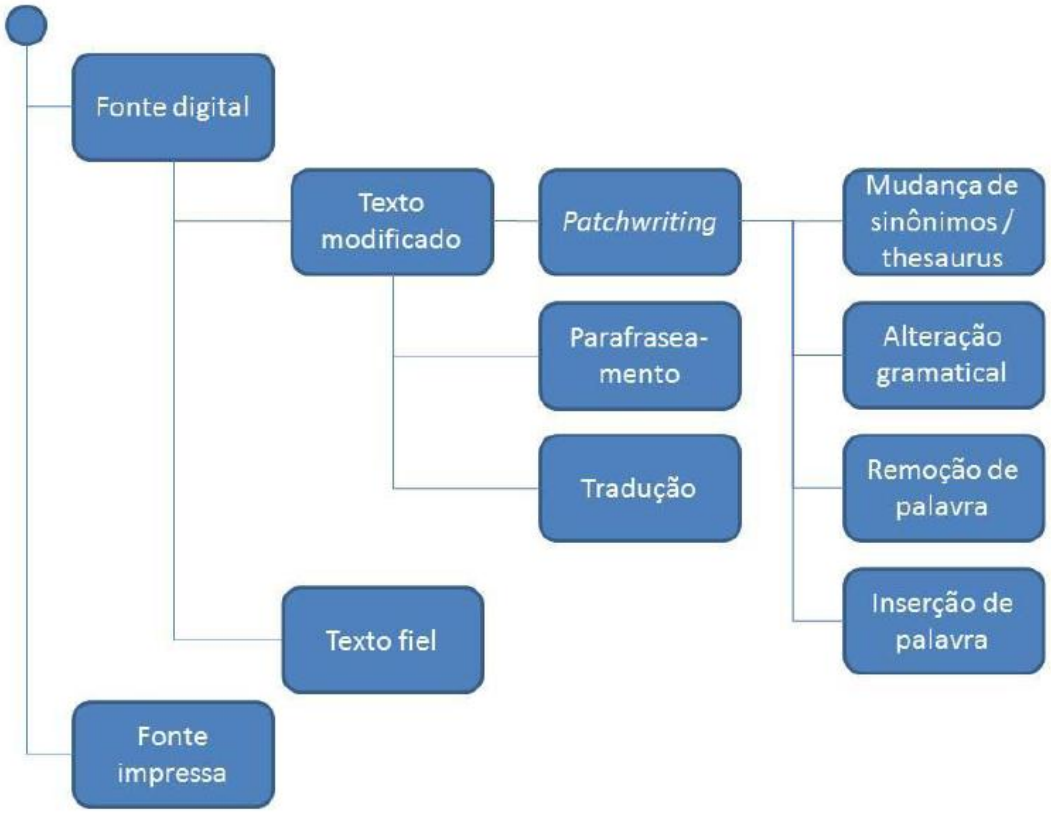
Tipos	Descrição
Translations	O tipo de plágio *Translations* é simplesmente uma tradução do texto original por uma ferramenta ou feito manualmente. Este tipo de plágio assemelha-se ao plágio direto, porém com uma tradução.
Shake & Paste Collections	Tipo de plágio que procura utilizar textos de várias fontes, após é feito uma ordenação aleatória de cada trecho extraído criando um novo documento. Este tipo de plágio normalmente é identificado em uma leitura simples que possa identificar diversas formatações, estilos de escritas e termos

	utilizados no mesmo texto.
Structural Plagiarism	Texto que utiliza paráfrase, porém sem identificar a fonte original. Nesta forma, é utilizada a mesma estrutura, argumentos, desenvolvimento, resultados e fontes do texto original como se fosse autêntico.
Patchwriting	Semelhante ao plágio mosaico, o *patchwriting* é a troca de palavras por sinônimos, alterar a estrutura e subtrair palavras do texto a ser plagiado.
Collusion	Tipo de plágio que é produzido por vários autores e no final quando concluído somente mencionar um destes. Desta forma fica mais definido o plágio devido ao menor campo de busca por plágio, pois normalmente foi plagiado na mesma turma de uma instituição de ensino e pelo fato de haver mais formas de escrita no mesmo texto.

Fonte: ABREU (2011, p. 6 - 7).

Na Figura 1, pode-se observar um diagrama que apresenta tipos de plágio e a derivação destas técnicas sobre um texto. Nesta figura é demonstrado que um texto plagiado pode possuir mais de uma técnica e desta forma quanto mais técnicas de plágio são utilizadas, mais distante fica o texto plagiado do texto original. Desta forma, quanto mais técnicas são utilizadas, mais distante estará a possibilidade de detectar o plágio.

Figura 1: Diagrama de técnicas de plágios aplicadas a um texto (Abreu, 2011).



Fonte: Abreu (2011, p. 9).

2.2.2 O Plágio e a Internet

Antigamente os professores não tinham acesso a uma rede como a Internet que possui diversos buscadores, e devido a isto não identificavam ou relutavam em divulgar casos de plágio de alunos. Isto era e ainda é uma das motivações para o plágio e desta forma poderia ter uma possível impunidade. Pelo fato do conteúdo disponibilizado na Internet ter fácil acesso, muitas vezes o estudante não reconhece que a obra possui direitos de propriedade (KIRKPATRICK, 2007).

A Internet é uma grande ferramenta de comunicação que contribui para a socialização e maior interação entre as pessoas. Além disso, há pouco tempo atrás foram criadas poderosas ferramentas de busca que, com seus robôs digitais, indexam milhares de páginas por dia. Para fazer a busca, o usuário somente digita a palavra ou assunto sobre qual quer retorno de endereços de sítios *online* que possuem conteúdo relacionado (OLIVEIRA et al., 2007).

A Internet, disseminada na década de 90 é uma grande contribuição para o plágio, sendo uma imensa área onde não há um controle para este problema. Uma obra disponibilizada na Internet logo está disponível a milhares de internautas que por sua vez, em alguns casos não valorizam os esforços do autor e utilizam o material para o plágio (MORAES, 2007).

Para Silva (2011) historicamente os alunos de ensino fundamental, médio e universitário possuem constante contato com o plágio. Por consequência disso e com a facilidade que a Internet impõe, os alunos acabam utilizando produções textuais alheias cometendo o plágio.

Segundo Kirkpatrick (2007), atualmente há sites que produzem e vendem trabalhos através da Internet que podem ser enviados digitalmente na hora da compra. Muitos estudantes que estão com curto prazo de entrega ou atrasados são os que mais utilizam deste tipo de serviço. Estes sites, que possuem baixo custo de criação e manutenção, oferecem trabalhos acadêmicos que são produzidos por universitários, tornando mais difícil a definição de integridade dos textos. Outra forma que acaba sendo disponibilizada aos contraventores são *sites* de professores, que acaba por disponibilizar, mesmo que de forma não intencional trabalhos de alunos.

Com a expansão da Internet foram criados cursos à distância, que estão em grande expansão atualmente. Estes cursos podem ser técnicos, graduação ou até mesmo de pós-graduação que são parcial ou totalmente através da Internet, armazenando conteúdo de livros e trabalhos de professores e alunos. A multiplicação destes conteúdos, íntegros ou não, pela Internet e muitas vezes com livre acesso, acaba sendo um facilitador para a prática do plágio sendo mais rápido e fácil que o conteúdo impresso (OLIVEIRA et al., 2008).

Para armazenar arquivos de qualquer conteúdo, existem vários *data centers* que possuem um grande espaço de armazenamento e, na maioria, não cobram taxas para isto. Uma vez feito o envio do arquivo para o servidor do *Data Center*, este estará disponível para milhares de internautas que por sua vez, podem ou não utilizar este serviço para cometer o plágio. Nestes servidores pode-se encontrar diversos tipos de arquivos como músicas, fotos, vídeos, livros e muitos outros que podem possuir proteção por direitos autorais. Estes *data centers* possuem grande volume de arquivos que se multiplicam diariamente pela Internet, devido a isso é quase impossível a missão de extinguir um arquivo da Internet.

Para o armazenamento e consulta de trabalhos acadêmicos foram criadas as bibliotecas digitais, que possuem grande acervo de livros, dissertações, teses, monografias e artigos. Devido a falta de segurança para coibir o plágio, muitos autores deixam de postar suas obras e desta forma acabam prejudicando o acesso ao material (OLIVEIRA et al., 2007).

Com a crescente preocupação perante a cópia de obra alheia, foram sendo criadas soluções que procuram identificar o plágio. Atualmente há várias formas de identificar a similaridade entre documentos. O capítulo 3 descreve algumas técnicas utilizadas pelos softwares existentes.

3 TÉCNICAS PARA DETECÇÃO DE PLÁGIO E SOLUÇÕES EXISTENTES

Para conter o avanço de plagiadores foram criadas técnicas que possibilitam rastrear e detectar similaridade entre documentos. Estas técnicas utilizam algoritmos capazes de verificar palavra por palavra e quantificar o quão é similar um documento ao outro. Desta forma busca-se diminuir este tipo de ação que ocorre através da Internet e propiciar principalmente aos profissionais da educação uma forma de detectar plágio. Nas seções a seguir são apresentados métodos que comparam textos e verificam similaridades, como também os métodos de limpeza textual que complementam estes métodos de detecção de similaridade.

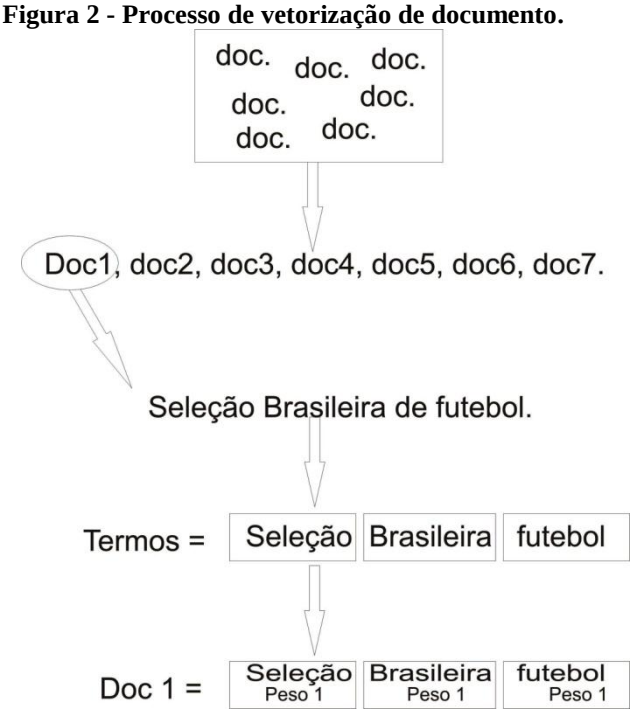
3.1 Representação vetorial de documentos

A técnica de representação vetorial utiliza o computador como ferramenta de cálculo e de forma automática busca a similaridade entre documentos. Conforme Oliveira et. al. (2007) para o desenvolvimento dessa técnica, é utilizada a representação na forma vetorial de um documento, onde os termos são extraídos e alocados em um vetor.

Estes vetores podem ser formados por termos de um documento, parágrafo ou oração. Para uma busca mais detalhada é necessário extrair termos e criar vetores por parágrafo ou oração, já que em um documento completo o número de termos seria muito abrangente.

Na Figura 2 está representado o processo de vetorização de documento, onde demonstra o procedimento com um documento. Pode-se notar que no primeiro momento os documentos estão apresentados de uma forma desorganizada e não numerada. Após ser feita a ordenação, são colocados em uma fila e seus termos são extraídos e colocados em um vetor. Ao possuir lugar no vetor, o termo recebe um peso que é determinado pelo número de ocorrências deste no documento.

Figura 2 - Processo de vetorização de documento.



Fonte: Do autor.

Nesta representação cada texto é um vetor de termos R e cada um destes vetores poderá possuir R^n termos, onde n é o número de palavras extraídas do texto.

Para melhor entender este procedimento considera-se uma lista de n documentos e cada posição representa um destes documentos DOC = $\{d_1, d_2, d_3, \ldots, d_n\}$. Cada uma dessas posições (documentos) terá um vetor representando os pesos $w_i = [w_1, w_2, w_3, \ldots, w_k, w_{k+1}, w_{k+2}, \ldots, w_n]$, onde a letra k representa todas as palavras distintas do documento d_i. As posições do vetor representadas por p_{k+n} definem os termos que aparecem em mais documentos. Assim representados ficam claro os termos que possuem maior número de repetições e a partir disso pode-se atribuir pesos para todos os termos. Para cada documento são definidos pesos para os termos e cada peso é referente à ocorrência do termo neste documento (OLIVEIRA ET. AL., 2007).

Para fazer a análise de quão similar um documento d_1 é em relação à d_2 pode-se calcular da seguinte forma:

$$sim(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i| x |d_j|} = \cos(\theta) \qquad (1)$$

O resultado desta fórmula pode ter valores entre 0 e 1, ou seja, quanto mais próximo de 0, mais distante está a similaridade entre os documentos e quanto mais perto de 1, mais similar será. Este resultado representa o ângulo entre os vetores que representam os documentos conforme explanado a seguir (OLIVEIRA ET. AL., 2007).

Para representar melhor a explicação anterior, foi elaborado um texto que é posto a prova de similaridade com outros documentos. O documento d_1 representa o original e d_2 e d3 serão somente termos no vetor que os representarão, como segue:

d_1- "*O campeonato brasileiro está próximo do fim. Tal campeonato foi muito prejudicado pela desorganização e times famosos poderão ser rebaixados. Alguns times estão entrando na Justiça para pedir a anulação do campeonato* " Fonte: OLIVEIRA ET. AL. (2007, p. 7).

O vetor de termos do documento d_1 possui 17 termos distintos e cada um ocupa uma posição no vetor de termos, portanto ficam agrupados conforme demonstrado a seguir:

d_1 = [*campeonato, brasileiro, próximo, fim, foi, prejudicado, desorganização, times, famosos, poderão, rebaixados, entrando, justiça, pedir, anulação*].

Como pode-se observar no vetor d_1, foram excluídos os chamados *stopwords* que são termos e preposições existentes no documento. Segundo Oliveira et al. (2007), são termos que aparecem em qualquer tipo de documento e não beneficiam o resultado da similaridade. Desta forma ficam menos termos para serem utilizados nos cálculos e verificações, tornando estas operações mais leves e consequentemente mais ágeis. O conceito de *stopwords* é descrito na seção 3.4.1 deste trabalho.

Após são adicionados ao vetor de termos os pesos que cada termo possui neste documento. A Tabela 5 apresentada a seguir demonstra como ficam os pesos de cada palavra, que variam entre um e três. Pode-se observar que muitas delas possuem peso um e somente duas tem pesos diferentes, sendo peso dois e três para as palavras "times" e "campeonato" respectivamente.

Tabela 5: Representação vetorial de um documento.

Índice i	Peso w_i	Termo t_i d_1
1	**3**	***campeonato***
2	1	*brasileiro*
3	1	*próximo*
4	1	*fim*
5	1	*foi*
6	1	*prejudicado*
7	1	*desorganização*
8	**2**	***times***
9	1	*famosos*
10	1	*poderão*
11	1	*rebaixados*
12	1	*entrando*
13	1	*justiça*
14	1	*pedir*
15	1	*anulação*

Fonte: Oliveira Et. Al. (2007, p. 8)

Para melhor visualizar e anular a influência de palavras com pesos menores, Oliveira et. al. (2007) adotou uma estratégia diferente, eliminando os termos que possuem peso menor que a metade do maior peso existente no vetor. Desta forma verifica-se que o maior peso é 3 e sua metade, ou seja, 3/2 é o peso de 1,5 e conforme a tabela pode-se observar que somente as palavras times e campeonato possuem peso acima deste limite. O vetor resultante deste documento é demonstrado a seguir:

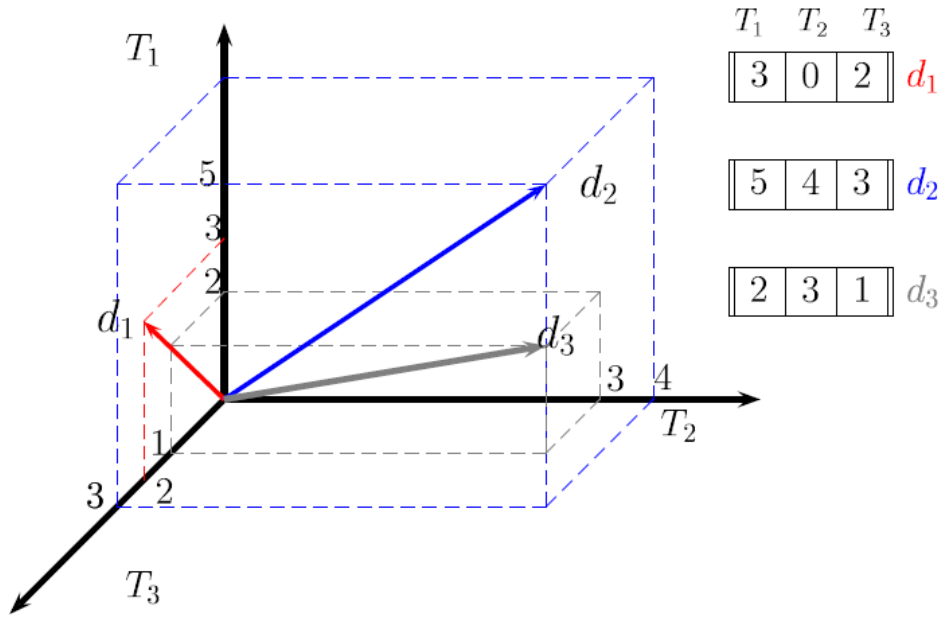
$$d_1 = [campeonato_3, times_2]$$

Para proceder são necessários mais documentos e conforme Oliveira et.al. (2007) considera-se os dois documentos que já estão representados por vetores de termos juntamente com o peso de cada palavra conforme demonstrado a seguir:

$$d_2 = [campeonato_5, brasileiro_4, times_3].$$
$$d_3 = [campeonato_2, brasileiro_3, times_1].$$

Após possuir os termos dos documentos nos respectivos vetores é traçado o gráfico onde pode-se observar as linhas traçadas sobre os eixos X, Y, Z representados pelos termos T_1, T_2, T_3 respectivamente. Desta forma fica claro a distância entre um documento e outro, sendo que os que estão mais próximos são mais similares.

Figura 3: Termos dos documentos d1, d2, d3 representados graficamente.



Fonte: Oliveira ET. AL. (2007, p. 9).

Na Figura 3 pode-se observar que os documentos mais próximos representados vetorialmente são d_2 e d_3, além disso, também é visível que possuem um ângulo menor se comparado com o documento d_1.

Os eixos X, Y, Z são representados respectivamente pelo termo T_1 campeonato, T_2 brasileiro e T_3 times que aparecem uma ou mais vezes nos documentos d_1, d_2, d_3 representados por linhas vetoriais vermelha, azul e cinza. Somente através do gráfico é notável que o peso referente ao termo T_1 é cinco no documento d_2 e peso dois para o mesmo termo no documento d_3. Desta forma percebe-se que para o documento d_2 este termo tem maior importância que para os outros dois, sendo que para o documento d_3 tem menor importância se comparando com outros dois textos analisados. A linha vetorial de d_1 está somente entre os eixos T_1 e T_3, pois para este documento o peso do termo T_2 é nulo (OLIVEIRA ET. AL., 2007).

É necessário representar os documentos desta forma porque os computadores são digitais e necessitam de uma estrutura lógica para executar o teste. Para continuar o método de detecção de similaridade, deve-se calcular *sim*(d_i, d_j) dentre os documentos existentes na comparação. Desenvolvendo-se este cálculo, ficará da seguinte forma:

$$sim(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i| x |d_j|} = \frac{\sum_{k=1}^{n} w_k^i \, x \, w_k^j}{\sqrt{\sum_{k=1}^{n} \{w_k^i\}^2} \; x \; \sqrt{\sum_{k=1}^{n} \{w_k^j\}^2}} = \cos(\theta) \qquad (2)$$

Através do desenvolvimento do cálculo na fórmula 2 pode-se observar que o cos (θ) representa o ângulo resultante entre os vetores dos dois documentos d_i e d_j. O valor resultante dos cos (θ) sempre será um valor entre 0 e 1 que representa o ângulo que está entre 0^o e 90^0. Quanto à representação de similaridade, estes resultados definem quanto similar um documento é do outro, sendo que quanto mais próximo a 0 maior é a distância de similaridade e quanto mais próximo a 1 maior é a similaridade entre eles (OLIVEIRA ET. AL. (2007)).

Colocando em prática o cálculo com os três documentos citados anteriormente apresentará o resultado conforme a seguir:

$$sim(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| x |d_2|} \; \frac{3 \, x \, 5 + 0 \, x \, 4 + 2 \, x \, 3}{\sqrt{3^2+0^2+2^2} \, x \, \sqrt{5^2+4^2+3^2}} = \frac{21}{25{,}49} = 0{,}82 = \cos(\theta_{1,2}) \qquad (3)$$

$$sim(d_1, d_3) = \frac{d_1 \cdot d_3}{|d_1| x |d_3|} \; \frac{3 \, x \, 2 + 0 \, x \, 3 + 2 \, x \, 1}{\sqrt{3^2+0^2+2^2} \, x \, \sqrt{2^2+3^2+1^2}} = \frac{8}{13{,}49} = 0{,}59 = \cos(\theta_{1,3}) \qquad (4)$$

$$sim(d_2, d_3) = \frac{d_2 \cdot d_3}{|d_2| x |d_3|} \; \frac{5 \, x \, 2 + 4 \, x \, 3 + 3 \, x \, 1}{\sqrt{5^2+4^2+3^2} \, x \, \sqrt{2^2+3^2+1^2}} = \frac{25}{24{,}49} = 0{,}94 = \cos(\theta_{2,3}) \; (5)$$

Pode-se observar nas fórmulas 3, 4 e 5 que no primeiro cálculo é comparado o documento 1 e 2, obtendo o resultado de 0,82 que representa o ângulo de 35^o. O segundo cálculo obteve o resultado de 0,59 com a comparação entre os documentos 1 e 3, representando o ângulo de $53{,}8^o$. Este resultado significa que os documentos d_2 e d_3 são menos similares que d_1 e d_2, porém verificando o resultado do terceiro cálculo pode-se notar que os documentos d_2 e d_3 são os mais similares e pode ser um caso de plágio. O resultado do cálculo 3 é 0,94 que representa o ângulo de 20^o (OLIVEIRA ET. AL. (2007)).

3.2 Sumarização de textos

Segundo Luhn (1958), a técnica de sumarização de textos foi inicialmente criada para identificar o conteúdo relevante em um artigo ou relatório. Este conteúdo relevante seria utilizado para criar resumos automaticamente sem intervenção humana e desta forma agilizar este processo e eliminar um possível direcionamento por parte do autor na interpretação do texto pelo leitor.

Para criar um resumo conforme foi citado anteriormente, o algoritmo busca o que é mais relevante no texto e retira frases que serão pontos chaves para criar este resumo. Já quando o autor cria o resumo do artigo ou relatório normalmente escolhe assuntos de forma aleatória, podendo assim criar um resumo que não contemple todo o texto e oculte possíveis pontos que possuem grande importância neste.

3.2.1 Definição de frases

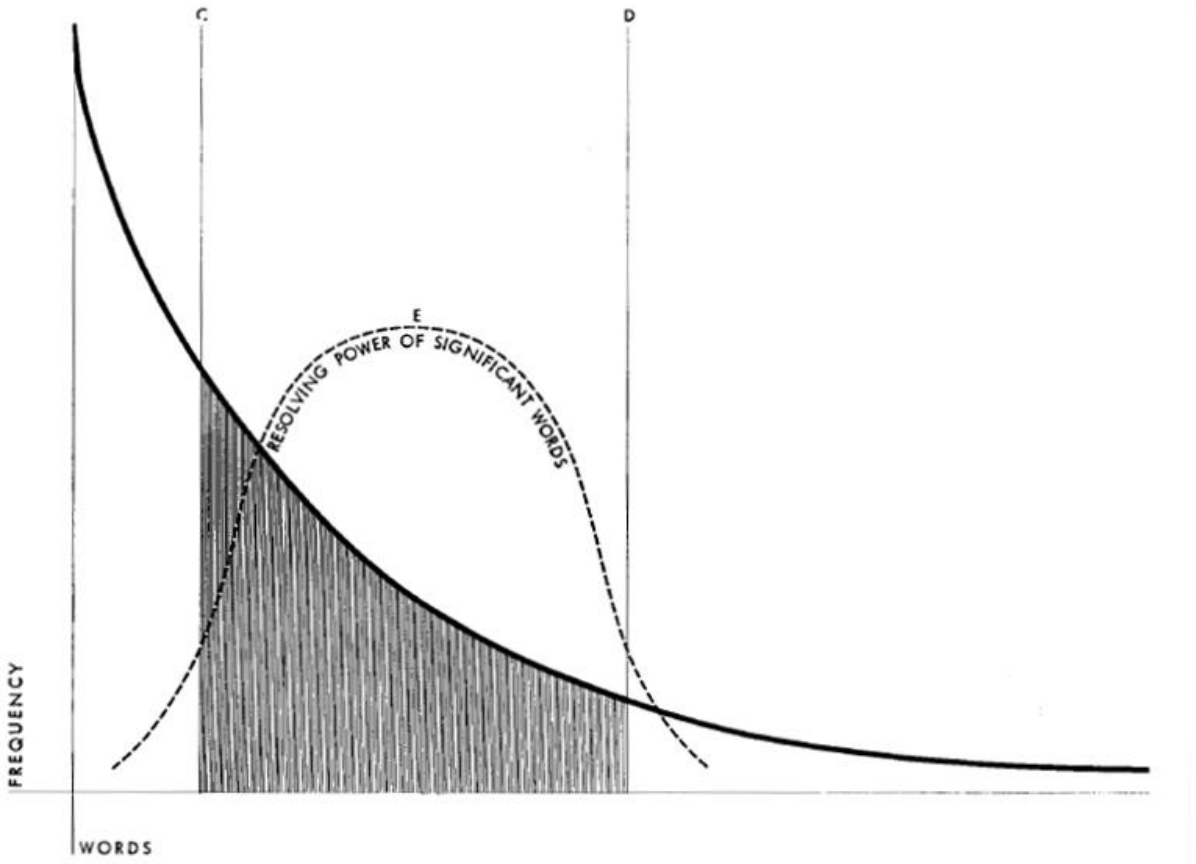
Para determinar quais frases que melhor representarão o texto no resumo é necessário ter uma medida que possa ser utilizada para a comparação entre frases. Para cada frase é definido um peso que após será utilizado para mensurar sua relevância no texto e desta forma que possa produzir melhores resultados para compor o resumo.

Para definir o peso de cada frase selecionada é necessário verificar as palavras que o compõe e a partir disso classificar estas frases em ordem de relevância. Além das frases, as palavras também possuem sua relevância no texto e é dada pela frequência que ocorre neste. Outra medida que também é utilizada é a distância dentro da frase entre palavras que possuem maior relevância, sendo que o número de palavras não significativas entre estas também é determinante no peso da frase.

Esta forma de classificação de peso se justifica devido que os autores repetem os assuntos principais várias vezes no texto e desta forma acabam repetindo palavras, mesmo que se esforcem para utilizar sinônimos, acabam caindo na repetição de palavras. Para definir melhor estas palavras foi utilizado um filtro que elimina as que possuem maior repetição, porque artigos e preposições certamente são os termos mais repetidos do texto e não devem interferir no resultado da seleção de frases, sendo eliminados das palavras mais relevantes.

A Figura 4 demonstra a frequência de termos em um texto qualquer e como pode-se observar, as palavras estão ordenadas de forma decrescente em relação a frequência definindo a linha exponencial demonstrada na figura. Os termos que estão fora do campo entre as barras verticais C e D são definidas como ruído, ou seja, palavras com baixa frequência não representam relevância no texto e palavras com muita frequência da mesma forma, conforme foi descrito no parágrafo anterior.

Figura 4: Diagrama de frequência de termos.



Fonte: Luhn (1958, p. 161).

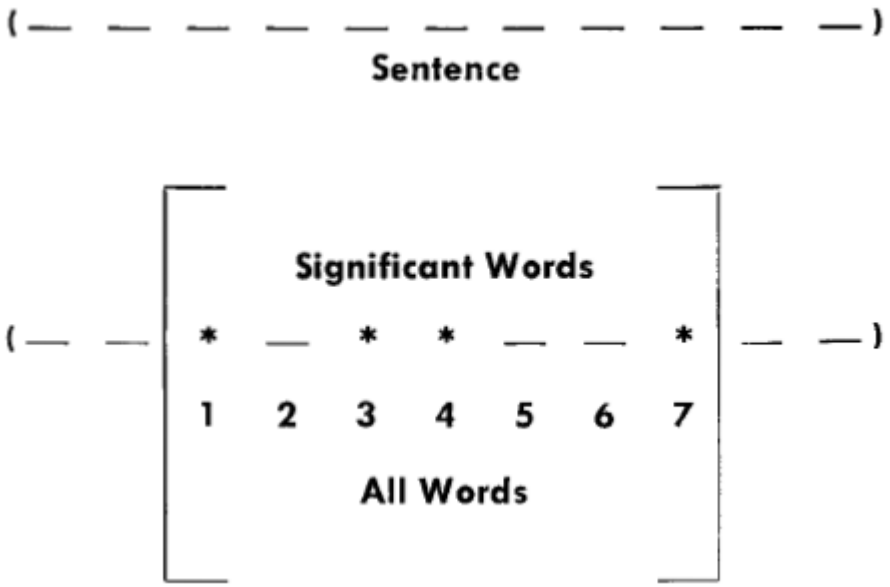
Luhn (1958) cita que quanto mais palavras distintas que possuem alta frequência estão na mesma frase e que não sejam muito distantes umas das outras, a probabilidade é alta desta frase ser do assunto principal do texto.

Para determinar a relevância da frase é necessário selecionar uma frase conforme Figura 5, onde entre parênteses há 11 termos sendo analisados. O cálculo é feito conforme segue:

$$S = \frac{\sqrt{|Termos\ relevantes|}}{|Total\ de\ termos|} \qquad (6)$$

Para calcular é necessário contar o número de palavras significativas dentro dos parênteses e após fazer a raiz quadrada deste valor e por fim, dividir pelo número total de termos que ali contém. Na Figura 5 está sendo demonstrando o trecho (entre colchetes) que foi selecionado e as palavras mais relevantes (asterisco).

Figura 5: Análise de frase.



Fonte: Luhn (1958, p. 162).

Luhn (1958) utilizou vários documentos para analisar através do algoritmo em questão e os resultados obtidos, demonstraram que entre palavras que são mais significativas não pode ter mais de 4 ou 5 não significativas. Desta forma muitas frases serão selecionadas como relevantes para compor o resumo, e a classificação ou peso destas frases é determinado pelo número de palavras relevantes que possui. Ao retirar as frases relevantes do texto é necessário colocar em ordem decrescente de peso e após selecionar as que possuem maior peso para compor o resumo.

Este método favorece a detecção do plágio porque agrega muito na seleção de frases e trechos de texto que podem servir como busca de documentos que são plágios. Ao invés de utilizar o texto na íntegra, o que poderia resultar em um grande custo computacional, esta

ferramenta traz a solução que diminui a quantidade de texto a ser comparado, sendo que o texto é quebrado em várias partes relevantes.

Um exemplo prático desta ferramenta demonstra o quanto agrega para a detecção de similaridade entre documentos. Na Figura 6 está sendo apresentado um texto retirado do site Wikipédia.org e será utilizado para demonstrar como este método atua.

Figura 6: Texto extraído da Internet.

O **futebol**,[1] (do inglês ***association football*** ou simplesmente ***football***) é um desporto de equipe jogado entre dois times de 11 jogadores cada um e um árbitro que se ocupa da correta aplicação das normas. É considerado o desporto mais popular do mundo, pois cerca de 270 milhões de pessoas participam das suas várias competições.[2] É jogado num campo retangular gramado, com uma baliza em cada lado do campo. O objetivo do jogo é deslocar uma bola através do campo para colocá-la dentro da baliza adversária, ação que se denomina golo (português europeu) ou gol (português brasileiro). A equipe que marca mais gols ao término da partida é a vencedora.[3]

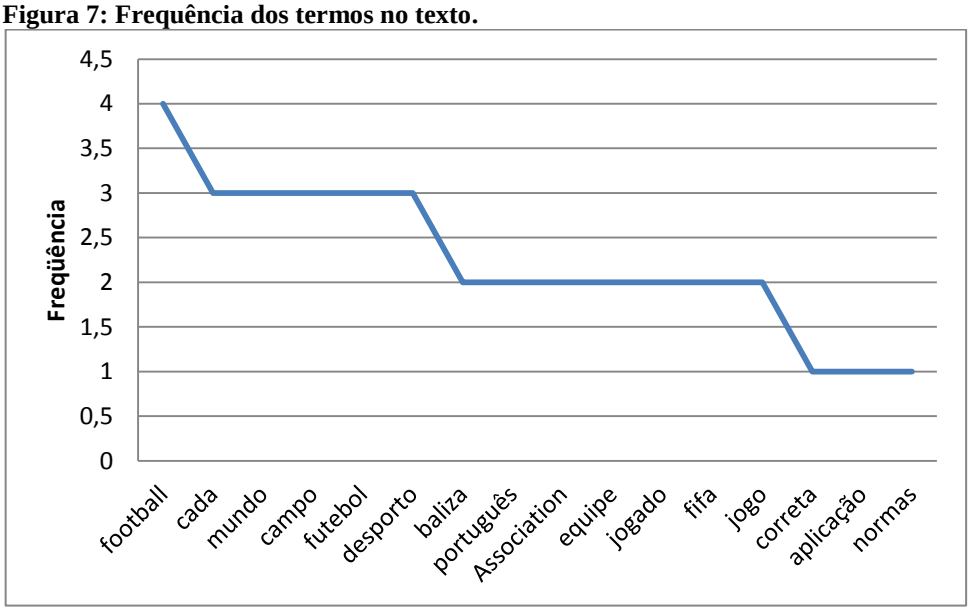
O jogo moderno foi criado na Inglaterra com a formação da Football Association, cujas regras de 1863 são a base do desporto na atualidade. O órgão regente do futebol é a *Fédération Internationale de Football Association*, mais conhecida pela sigla *FIFA*. A principal competição internacional de futebol é a Copa do Mundo FIFA, realizada a cada quatro anos. Este evento é o mais famoso e com maior quantidade de espectadores do mundo, o dobro da audiência dos Jogos Olímpicos.[4]

Fonte: http://pt.wikipedia.org/wiki/Futebol (2012, texto digital).

Do texto original foram retirados dois parágrafos conforme apresentado na Figura 6. Este texto possui 188 palavras e para ter um melhor resultado, é feita a limpeza que extrai as *stopwords*, resultando em 109 termos. Após feita a limpeza, este texto foi vetorizado para então analisar a frequência destes termos.

Na Figura 7 estão demonstrados os termos e frequências, neste gráfico foram apresentados apenas 3 termos com 1 frequência, deixando os demais ausente do gráfico para não prejudicar a visualização.

Figura 7: Frequência dos termos no texto.



Fonte: Do autor.

Após ter as frequências dos termos, conforme pode-se analisar na Figura 7 e segundo Luhn (1958) deve-se desconsiderar o ruído, que são os termos com baixa frequência e os termos que possuem alta frequência. Desta forma foram selecionados os termos com frequência entre 2 e 3.

As duas frases relevantes que foram selecionadas através deste método são:

I. É **jogado** num **campo** retangular gramado, com uma **baliza** em **cada** lado do **campo**.

II. A principal competição internacional de **futebol** é a Copa do **Mundo FIFA**, realizada a **cada** quatro anos.

As palavras que estão em negrito são as mais relevantes e na frase I, possui cinco termos relevantes que possuem entre si menos de cinco termos não relevantes. Já na segunda frase, existem quatro palavras relevantes, porém somente três delas possuem menos de cinco palavras que as distanciam.

3.3 Distância de Levenshtein

Em 1966 foi desenvolvida por Levenshtein (1966) a distância de edição de uma palavra para que se transformasse em outra. Para uma palavra ser transformada em outra é necessário fazer inserções, substituições e deleções, e este método visa identificar a quantidade mínima dessas alterações para que se chegue ao resultado pretendido. Este método é aplicado em várias soluções como a ortografia, recuperação de informação e classificação.

Para chegar ao número mínimo de alterações de um termo, Duarte (2011) apresentou os seguintes passos que deve-se seguir:

> "I. Preenche-se uma matriz bidimensional D, onde D[i][j] representa a distância entre o prefixo de tamanho i da primeira string S1 (de tamanho m) e o prefixo de tamanho j da segunda string S2 (de tamanho n);
>
> II. Para i de 1 até m e j de 1 até n calcula-se D[i][j] da seguinte forma:
>
> III. Se o caractere i de S1 é igual ao caractere j de S2 então D[i][j] = D[i – 1][j – 1];
>
> IV. Caso contrário é atribuído a D[i][j] o mínimo entre:
>
> D[i – 1][j – 1] + custo de substituição;
>
> D[i][j – 1] + custo de adição;
>
> D[i – 1][j] + custo de remoção;
>
> V. A distância entre S1 e S2 será o valor na posição D[m][n] ao termino da execução do algoritmo;" (DUARTE, 2011, p. 76).

Conforme citação anterior de Duarte (2011), desta forma pode-se chegar ao número mínimo de alterações de um termo ao outro na posição D[i][j] da matriz. Este algoritmo possui a complexidade de O(m,n) em tempo e espaço.

A seguir foi desenvolvido um exemplo prático para demonstrar como este método atua utilizando os termos "forte" e "mostre" conforme segue:

I. *Forte*;

II. *Morte* (a letra "F" é substituída pela letra "M");

III. *Moste* (a letra "r" dá lugar a letra "s");

IV. *Mostre* (é inserida a letra "r" entre "t" e "e").

Como pode-se observar no exemplo anterior, a palavra "Forte" foi transformada em "Mostre" e para isto foi preciso 4 passos e 3 alterações. O primeiro passo é a palavra inicial,

neste caso “Forte”, e a primeira alteração feita foi uma substituição da letra “F” pela letra “M”. Do segundo para o terceiro passo foi necessário fazer uma nova substituição, desta vez foi substituída a letra “r” pela letra “s” finalizando o passo 3 com a palavra “Moste”. Para finalizar o processo foi necessário fazer uma inserção da letra “r” entre as letras “t” e “e”, chegando na palavra final “Mostre”.

3.4 Limpeza textual

Para obter melhores resultados nas consultas e comparações de texto é necessário fazer uma limpeza textual. Muitos métodos fazem este tipo de limpeza, sendo que possuem níveis e critérios diferentes para a seleção do que será excluído. Desta forma serão eliminados termos que não teriam significado ou poderiam alterar o resultado dos testes de similaridade de textos.

Alguns destes termos possuem uma frequência alta na maioria dos textos, ou seja, se for comparado dois textos, a chance dos dois textos possuírem várias vezes estes termos é muita alta, podendo produzir ruído em uma análise de similaridade.

3.5 Stopwords

Stopwords são palavras que aparecem frequentemente em todos os textos e não é possível redigir sem utilizar *stopwords*. *Stopwords* se encaixam no que foi citado na seção anterior, como estão em todos os textos e possuem grande frequência, não é considerado relevante na comparação de textos.

As palavras que compõem *stopwords* são artigos, preposições, pronomes e outros. Segundo Abreu (2011), o projeto *Open Source* Apache Lucene possui bases de dados com *stopwords* em várias línguas. O projeto Lucene está na versão 3.3.0 e a mais de dez anos vem aprimorando suas listas de *stopwords*. Nas tabelas a seguir estão disponibilizadas algumas *stopwords* da língua Inglesa e Portuguesa:

Tabela 6: *Stopwords* em Inglês

but	***be***	***with***	***such***	***then***	***for***	***no***	***will***	***not***	***are***	***and***
their	***if***	***this***	***on***	***into***	***a***	***or***	***there***	***in***	***that***	***they***
was	***is***	***it***	***an***	***the***	***as***	***at***	***these***	***by***	***to***	***of***

Fonte: Do autor.

Tabela 7: *Stopwords* em Português

a	**ainda**	**alem**	**ambas**	**ambos**	**antes**	**ao**	**aonde**	**aos**	**apos**	**aquele**
aqueles	**as**	**assim**	**com**	**como**	**contra**	**contudo**	**cuja**	**cujas**	**cujo**	**cujos**
da	**das**	**de**	**dela**	**dele**	**deles**	**demais**	**depois**	**desde**	**desta**	**deste**
dispõe	**dispõem**	**diversa**	**diversas**	**diversos**	**do**	**dos**	**durante**	**e**	**ela**	**elas**
ele	**eles**	**em**	**então**	**entre**	**essa**	**essas**	**esse**	**esses**	**esta**	**estas**
este	**estes**	**ha**	**isso**	**isto**	**logo**	**mais**	**mas**	**mediante**	**menos**	**mesma**
mesmas	**mesmo**	**mesmos**	**na**	**nas**	**não**	**nas**	**nem**	**nesse**	**neste**	**nos**
o	**os**	**ou**	**outra**	**outras**	**outro**	**outros**	**pelas**	**pelas**	**pelo**	**pelos**
perante	**pois**	**por**	**porque**	**portanto**	**próprio**	**próprios**	**quais**	**qual**	**qualquer**	**quando**
quanto	**que**	**quem**	**quer**	**se**	**seja**	**sem**	**sendo**	**seu**	**seus**	**sob**
sobre	**sua**	**suas**	**tal**	**também**	**teu**	**teus**	**toda**	**todas**	**todo**	**todos**
tua	**tuas**	**tudo**	**um**	**uma**	**umas**	**uns**				

Fonte: Do autor.

Nas Tabelas 6 e 7 pode-se observar que o número de termos é maior na língua portuguesa em comparação com a língua inglesa. Na tradução do texto original para realizar buscas em espanhol, também pode ser utilizada uma tabela com *stopwords* nesta outra linguagem. A tradução pode ser utilizada em casos onde necessita-se de uma busca mais abrangente podendo assim, encontrar indícios de plágio nestes campos que normalmente é utilizado por estudantes em trabalhos acadêmicos.

3.6 Radicalização

A radicalização também é utilizada na limpeza textual, pois consiste em retirar o radical de um termo e desta forma resultar em tornar palavras equivalentes. Extraindo o mesmo radical de dois termos é grande a chance de estas possuírem o mesmo sentido. Radicalização também é conhecida com *stemming* e uma aplicação prática é demonstrada a seguir:

I. termo 1: *comprar*
II. termo 2: *comprou*
III. radical: *compr*

Neste exemplo é utilizado os termos “comprar” e “comprou” que possuem mesmo radical e equivalência. Este exemplo serve para ilustrar uma palavra que teve sua conjugação alterada em um texto plagiado e desta forma pode-se confirmar isto.

Além do prefixo, este algoritmo também pode remover o sufixo e até mesmo a sua forma no infinitivo. Para ser possível utilizar este método na língua portuguesa foi necessário adaptar o algoritmo de *stemming* (MATSUBARA ET AL., 2003).

3.7 Bag of words

A tradução de *Bag of words* é bolsa de palavras e este método consiste em comparar textos através dos termos e de suas respectivas frequências, diferentemente do algoritmo de Luhn (1958), este método não considera a ordem dos termos no texto analisado.

Para utilizar este método, usualmente é utilizada a limpeza de texto antes da análise e são utilizados *stopwords*, Radicalização (*stemming*) após os termos estarem dispostos em um vetor. Após, os termos de sentenças ou documentos são alocados em um vetor juntamente com o sua frequência para então ser analisados os resultados obtidos juntamente com os resultados de outros documentos (MATSUBARA ET AL., 2003).

3.8 MySQL Fulltext

MySQL é um sistema gerenciador de banco de dados que possui também, agregado a suas funcionalidades, várias ferramentas para a comparação de textos. Entre as comparações mais utilizadas estão o " = " que retorna na consulta uma *string* igual a que foi comparada, o "like" que pode trazer resultados como:

I. like '%busca%' - Retorna uma *string* que contenha a palavra "busca";
II. like 'busca%' - Retorna uma *string* que inicie com "busca";
III. like '%busca' - Retorna uma *string* que termine com "busca".

A função Fulltext do MySQL permite a criação de índices que tem a função de agilizar o processo de busca e agregar mais campos de texto a uma tabela. Para ser utilizada é necessária a criação de índices nos campos textos que será feita a busca. Este método possui alguns parâmetros pré-configurados conforme listado a seguir:

I. Palavras que tiverem menos de 5 letras não são considerados na busca;
II. Palavras que estão na lista de *stopwords* do MySQL também serão descartadas.
III. Palavras que estão em mais de 50% dos documentos não serão consideradas na busca.

Segundo Varella (2007) a função do MySQL Fulltext leva em consideração na busca o somatório dos pesos dos termos multiplicado pela frequência conforme calculado pela fórmula:

$$r = w * qf \qquad (7)$$

Onde r é a relevância do termo para o documento que iguala ao peso multiplicado pelo pela frequência que o termo aparece na consulta. Desta forma é calculada a relevância para cada termo da consulta:

$$R = w_1 * qf_1 + w_2 * qf_2 + w_3 * qf_3 + \ldots + w_n * qf_n \qquad (8)$$

Como pode-se observar, quanto maior for a frequência do termo na consulta, maior será seu peso e consequentemente sua relevância na pesquisa.

Quanto maior é a quantidade de documentos maior é o custo computacional, e para diminuir este custo, pode-se realizar a consulta pelo MySQL Fulltext para descobrir quais são os documentos mais relevantes.

A seguir é demonstrado um exemplo onde se aplica a busca com MySQL Fulltext. Considerando que em uma tabela de banco de dados chamada Livros estejam cadastrados 5 livros com os dados demonstrados na Figura 8.

Figura 8: Tabela exemplo.

Livros	
livroId	**nome**
1	JavaScript: The Definitive Guide
2	Cascading Style Sheets: The Definitive Guide
3	A Mídia na Sociedade em Rede
4	Understanding Open Source And Free Software Licensing
5	Html & Xhtml: The Definitive Guide

Fonte: Do autor.

O comando SQL a seguir é executado sobre a tabela demonstrada na Figura 8:

“SELECT nome, (MATCH (nome) AGAINST(‘free > the > definitive’ IN BOOLEAN MODE)) AS valor FROM livros WHERE MATCH (nome) AGAINST(‘free > the > definitive’ IN BOOLEAN MODE) ORDER BY valor DESC”

Este comando SQL é uma forma de representar os benefícios do MySQL Fulltext. Os termos que se encontram dentro das aspas simples serão procurados na coluna nome da tabela livros. O operador “>” determina que os termos “the” e “definitive” terão uma contribuição maior no resultado.

O resultado exibido na consulta é demonstrado na Figura 9. Como pode-se observar, os livros que possuem dois termos da busca receberam o valor 1.5, já o livro que possui somente um termo da busca recebeu o valor 1. Desta forma é possível classificar os resultados e definir facilmente os melhores.

Figura 9: MySQL Fulltext - Resultado de consulta

Resultado	
nome	**valor**
JavaScript: The Definitive Guide	1 . 5
Cascading Style Sheets: The Definitive Guide	1 . 5
Html & Xhtml: The Definitive Guide	1 . 5
Understanding Open Source And Free Software Licensing	1

Fonte: Do autor.

Na próxima seção são apresentadas soluções já desenvolvidas na busca de documentos similares e plágio.

3.9 Soluções existentes

Existem hoje três soluções que buscam atender os usuários na busca por documentos similares e plágio. Estas soluções possuem diversas características e podem ser gratuitas ou pagas. Nas próximas seções estão apresentadas algumas ferramentas existentes.

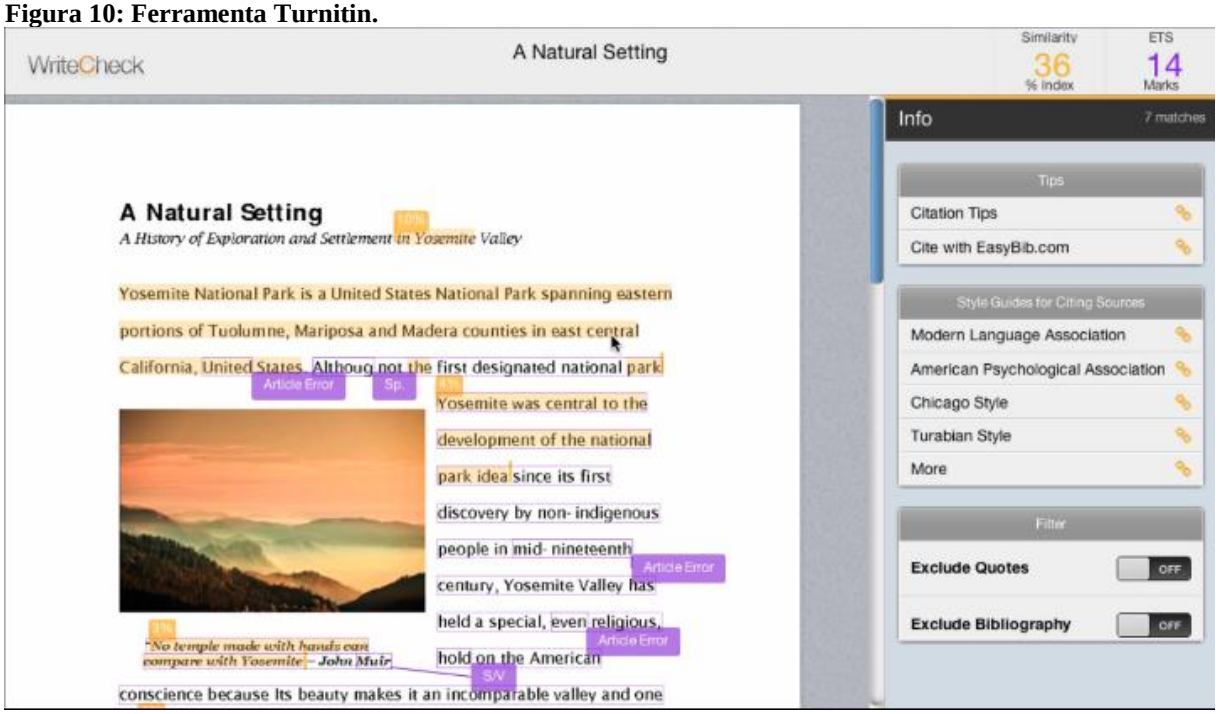
3.9.1 Turnitin (iParadigms, *LLC*)

Ferramenta com plataforma *online* paga, que possibilita ao usuário realizar buscas para determinar se o documento que postou é original ou não. Além de demonstrar a similaridade do documento, a plataforma emite um relatório apresentando todos os dados referentes a documentos similares.

A busca que esta ferramenta utiliza é baseada em mecanismos de busca que possuem grandes bases de dados com documentos e artigos. Além de produzir comparações entre documentos, o Turnitin também produz comentários e correções através de avaliação *online*.

Atualmente a base de dados da ferramenta possui mais de 220 milhões de documentos armazenados, 90 mil periódicos acadêmicos e livros e 20 bilhões de páginas web rastreadas. Além deste conteúdo, estão cadastrados mais de 1 milhão de professores, 10 mil instituições de ensino, 20 milhões de alunos e 126 países (https://turnitin.com/pt_br/home, acessado em 11/2011).

Figura 10: Ferramenta Turnitin.



Fonte: Do autor, adaptado de https://www.writecheck.com/static/demo.html (2012).

Na Figura 10 é possível visualizar a interface da ferramenta Turnitin e algumas opções que são disponibilizadas ao usuário. No canto superior direito há o nível de similaridade que este documento possui com outros armazenados na base de dados do Turnitin. Além da similaridade, esta ferramenta coloca marcações no texto demonstrando similaridade e correção gramatical.

Na parte inferior direita da Figura 10 há os filtros que o Turnitin possui e que tem como ação de retirar citações (partes do texto que estejam entre aspas) e bibliografia da análise. Como é demonstrado na Figura 10, o documento é visualizado na mesma formatação facilitando assim o entendimento.

Além de comparações de similaridade e correção gramatical, o Turnitin possui 3 tipos de níveis cadastrais:

I. O nível Professor pode criar classes, aulas, definir datas de entrega de trabalhos, comparar similaridade, fazer correção *online*, enviar documentos.

II. O nível Professor Assistente pode alterar definições criadas pelo professor, enviar documentos, verificar a similaridade dos trabalhos e fazer a correção on-line.

III. Aluno: enviar documentos, ver comparação de similaridade, corrigir o documento on-line.

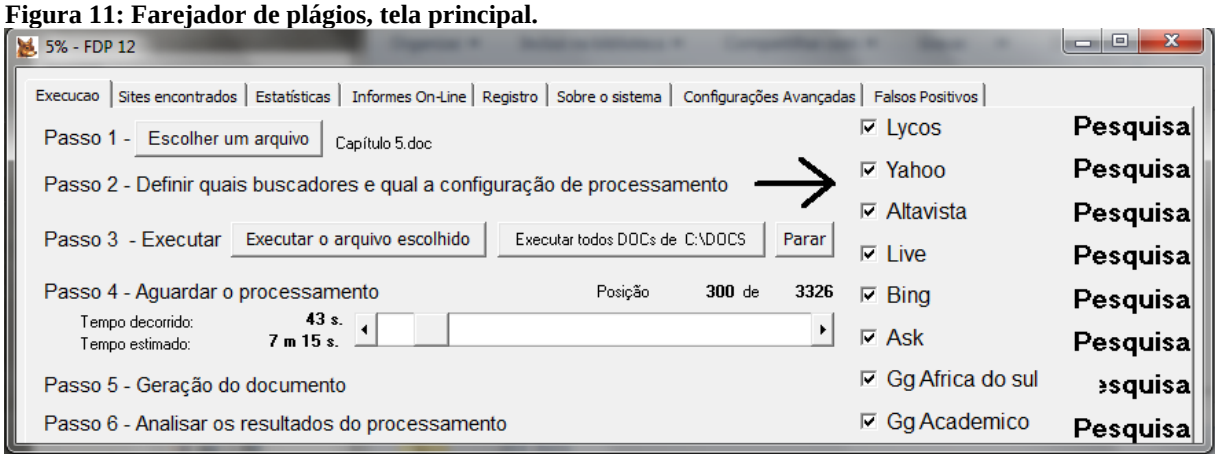
3.9.2 Farejador de Plágios

A ferramenta Farejador de Plágios possui duas versões, sendo uma gratuita e outra paga. Na versão gratuita o usuário poderá enviar documentos com até 300 KB de tamanho e somente a análise de metade do arquivo é apresentada ao usuário. A versão paga tem um custo de R$ 19,90 para uso pessoal e de R$ 59,90 para uso empresarial/institucional.

Este software necessita de instalação no computador do usuário e aceita somente dois tipos de formatos de arquivos: Word ou RTF. Para encontrar os documentos similares, os mecanismos de busca utilizados são Google, Yahoo, Altheweb entre outros.

Esta ferramenta faz um rastreamento em todo o documento e retira trechos que variam de 4 a 10 palavras. Estes trechos do texto são utilizados para fazer a busca de similaridade do documento. Ao finalizar a busca de similaridade, a ferramenta cria um documento com o conteúdo do documento fornecido pelo usuário, porém com os endereços das páginas e arquivos onde foram encontradas as similaridades e coloca estes endereços logo após o trecho similar. Além de inserir os endereços, o Farejador de plágios também altera o trecho similar deixando-o sublinhado (http://www.farejadordeplagio.com.br/, 11/2011).

Figura 11: Farejador de plágios, tela principal.



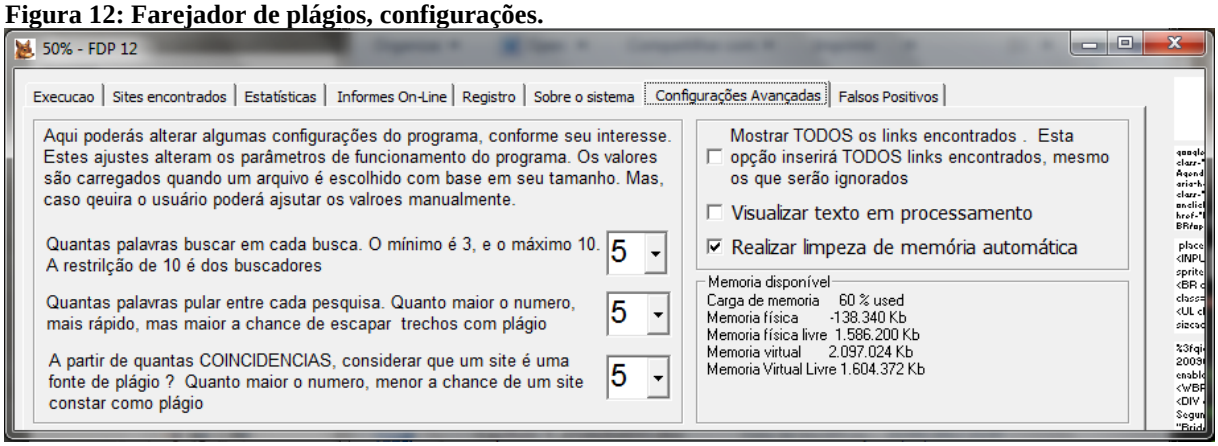
Fonte: Do autor.

Na Figura 11 está a interface principal o software Farejador de plágios. Como pode-se observar, no lado direito estão os mecanismos de busca utilizados, onde o usuário pode selecionar qual deseja utilizar.

Para utilizar a ferramenta é necessário selecionar um arquivo e após clicar em executar o arquivo escolhido. Também é possível executar vários arquivos juntos clicando em "executar todos DOCs de C:\DOCS" como demonstrado na Figura 11.

Na Figura 12, são demonstrados os tipos de configuração da ferramenta. Nesta tela do sistema é possível alterar parâmetros de busca como quantidade de palavras para cada busca, quantidade de palavras que são descartadas após cada busca dentre outros.

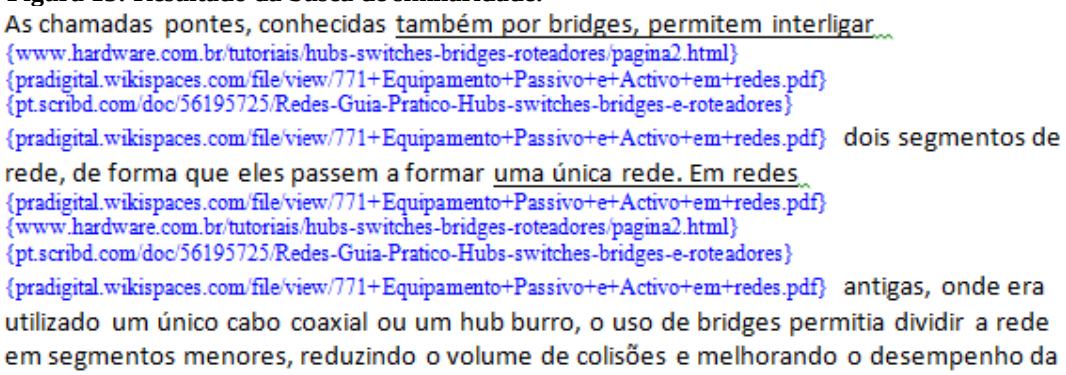
Figura 12: Farejador de plágios, configurações.



Fonte: Do autor.

Para visualizar melhor como o Farejador de Plágios apresenta os resultados, a Figura 13 demostra o resultado de um teste realizado selecionando todos os buscadores na execução da busca desta ferramenta.

Figura 13: Resultado da busca de similaridade.



Fonte: Do autor.

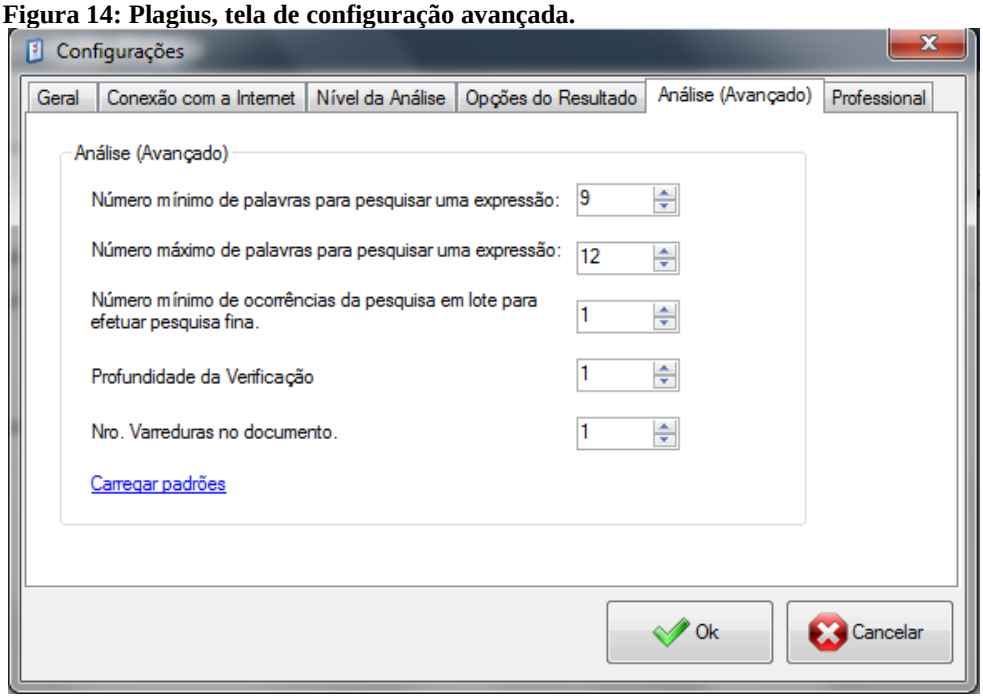
Como é demonstrado na Figura 13, os conteúdos entre chaves são os endereços onde foram encontradas similaridades com o texto que esta sublinhado. Segundo o site desenvolvedor da ferramenta, quanto mais vezes o mesmo endereço aparecer, maiores são as chances de haver indícios de plágio.

3.9.3 *Plagius*

Este software é encontrado no endereço http://www.plagius.com e possui versão *Personal* e *Professional*, ambas no modo gratuito e pago. A diferença entre os modos pago e gratuito é que na versão sem custos o tamanho do documento enviado é limitado a 500 palavras, já no pago não há limitações. A diferença entre as versões *Personal* e *Professional* é a busca através de uma página web, análise de documentos em lote, maior tempo de suporte grátis e comparar arquivos com outros existentes na máquina local.

O custo para adquirir o software é de R$ 29,90 para a versão *Personal* e de R$ 59,90 para a versão *Professional*. A versão do software que é demonstrada na Figura 14 é a Professional no modo gratuito.

Figura 14: Plagius, tela de configuração avançada.

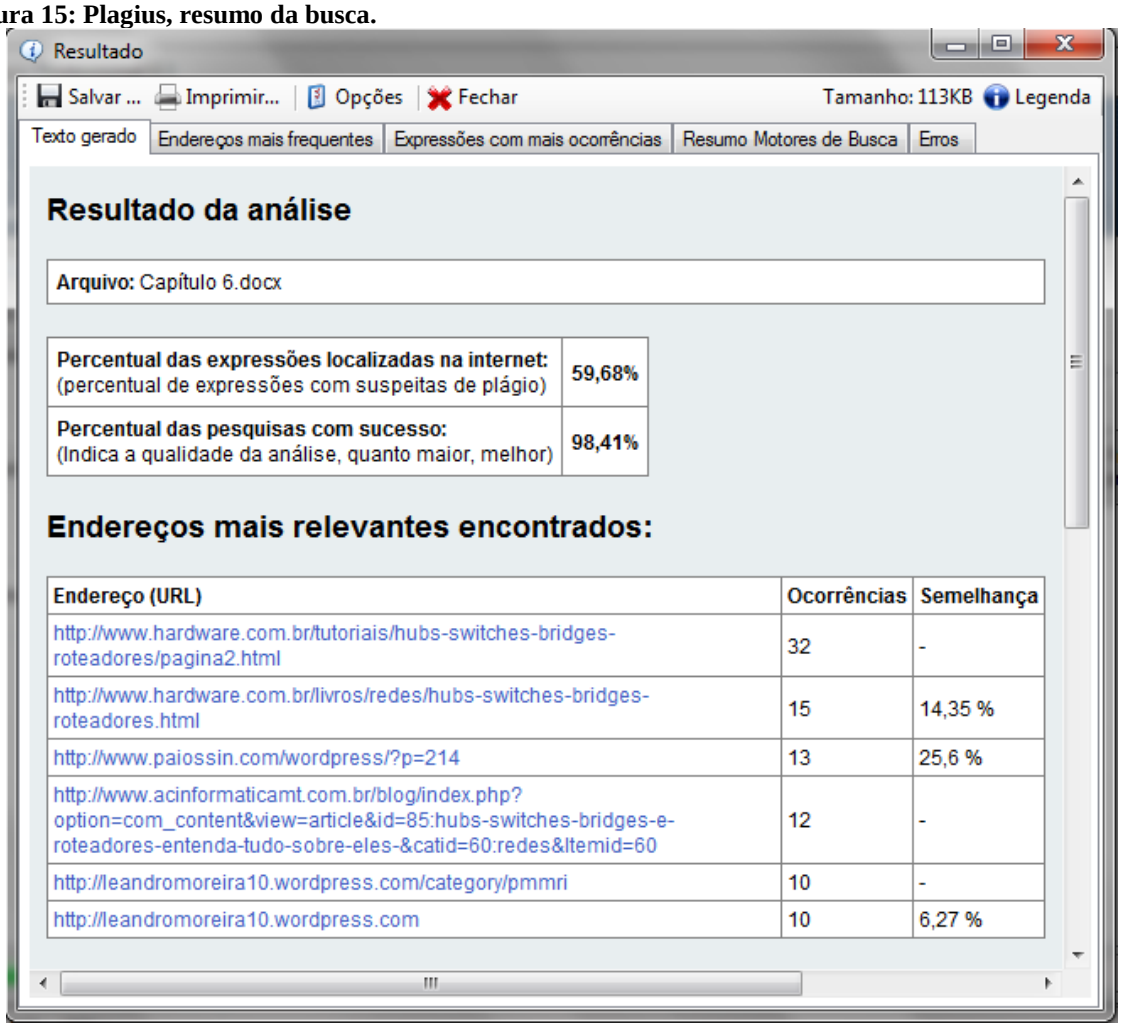


Fonte: Do autor.

Na Figura 14 é demonstrada a tela de configuração avançada do software. Como pode-se observar, as opções de ajustes são número mínimo e máximo de palavras por expressão, profundidade da verificação dentre outros. Nesta mesma tela do sistema é possível configurar a conexão com a Internet, alterar nível da análise e opções de resultados.

O resultado que a ferramenta apresenta é um resumo da análise contendo o percentual de expressões localizadas na Internet, percentual de pesquisas realizadas com sucesso e endereços com maior número de ocorrências e semelhanças. Na Figura 15 é demonstrada a tela de resultado da busca de similaridade de um documento nesta ferramenta.

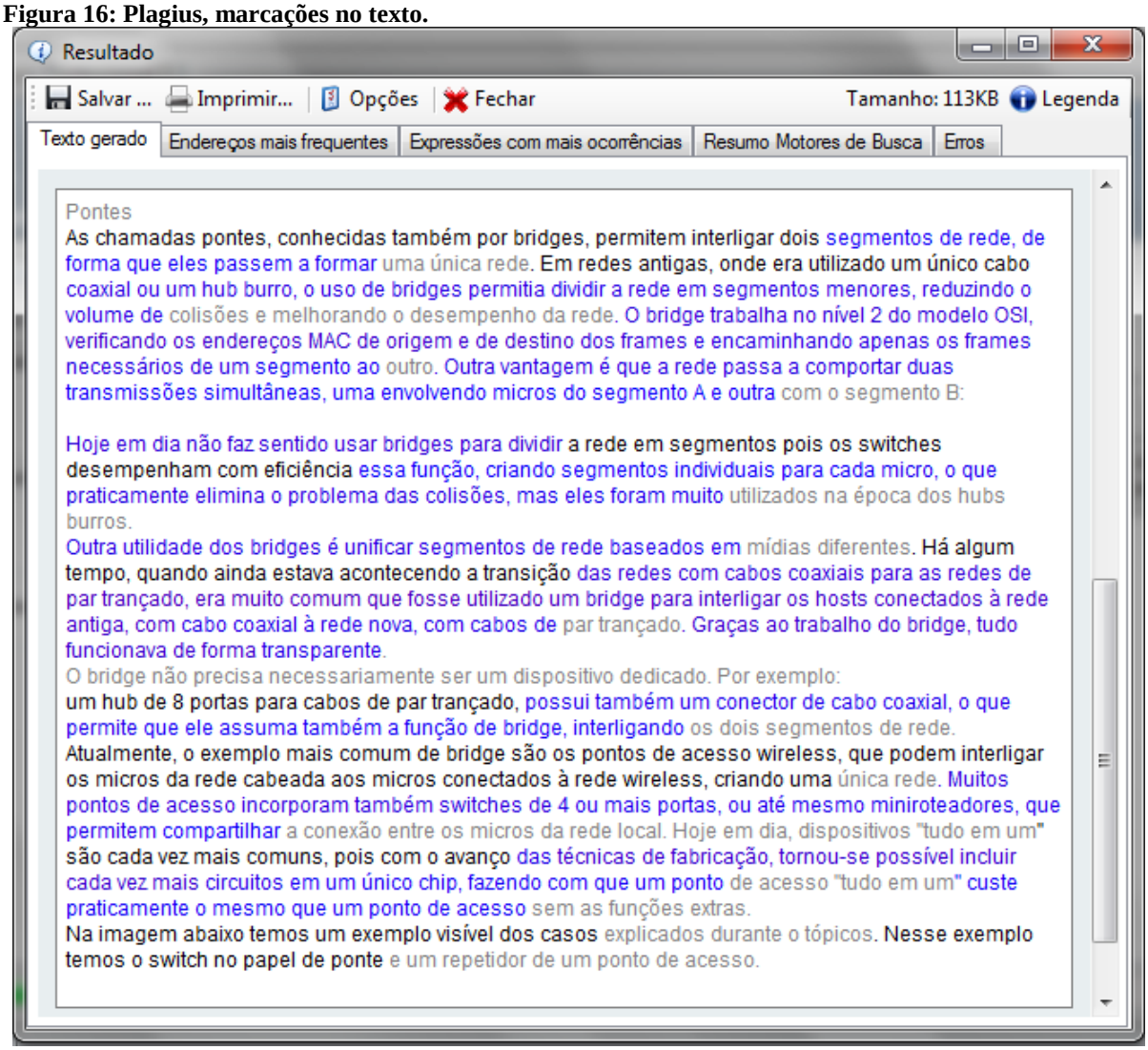
Figura 15: Plagius, resumo da busca.



Fonte: Do autor.

Além deste resumo, o software também exibe o texto de busca com diversas marcações. Estas marcações demonstram em preto as expressões que não obtiveram índices de similaridade, em azul as partes de texto que tiveram algum índice de similaridade e o texto que está em cinza não foi utilizado na busca. Esta parte da tela de resultado pode ser observada na Figura 16.

Figura 16: Plagius, marcações no texto.



Fonte: Do autor.

Além do resumo e das marcações no texto como demonstrado na Figura 16, ao passar o mouse pelas sentenças em azul, o software demonstra os endereços das páginas onde foram encontradas similaridades. Nas outras abas desta mesma tela é possível visualizar dados estatísticos referente a endereços encontrados, ocorrência das expressões e um resumo de dados referente às buscas em cada um dos 31 motores de buscas utilizados pelo sistema (http://www.plagius.com/s/br/default.aspx, 11/2011).

Dentre as três ferramentas analisadas somente o Turnitin se destaca por possuir funções mais avançadas. Os softwares Plagius e Farejador de plágios utilizam a mesma forma de busca, ou seja, analisam o texto que o motor de buscas retorna, não analisando os documentos que estão nos endereços encontrados. Por outro lado, o Turnitin possui uma

grande base de dados com as principais bibliotecas digitais indexadas, trazendo um melhor resultado na análise.

4 IMPLEMENTAÇÃO DA FERRAMENTA WEB PARA DETECÇÃO DE PLÁGIO

Nesta monografia foi desenvolvida uma ferramenta para a detecção de similaridades e plágio em documentos, a qual é acessada através da Internet. Desta forma qualquer pessoa pode acessá-la de qualquer lugar do mundo, desde que possua acesso à Internet. Além de facilitar o acesso, a ferramenta pode ser acessada através de qualquer browser disponível atualmente e de qualquer sistema operacional de computadores.

Para prover a busca por plágio a ferramenta CTP (Copy To Paste) possui três formas de busca por documentos além de também possuir uma plataforma que suporta o envio de mais de um tipo de documento. Isto traz uma gama maior de possibilidades para o usuário que além das citadas anteriormente, é demonstrado de forma detalhada ao usuário a localização do trecho similar e endereço do documento.

As formas de envio irão possibilitar o envio de texto puro, que o usuário poderá digitar em um campo texto da aplicação. A outra forma de disponibilização do texto à aplicação é o envio de arquivo PDF sem limite de tamanho e páginas. Por último, há a possibilidade de postar uma URL para que a aplicação extraia o HTML da mesma para fazer a análise de similaridade.

Este capítulo descreve a implementação da ferramenta CTP que busca atacar o problema principal, a busca por documentos plagiados e consequentemente por similares. Este capítulo também demonstra como esta ferramenta faz a busca e detecção de similaridade entre os documentos e como os resultados são exibidos para o usuário.

Através dos estudos realizados, observou-se que para realizar uma busca por documentos similares, primeiramente seria necessário tratar o texto que é disponibilizado pelo usuário e o texto que é capturado na Internet pelo sistema. Este texto pode ser proveniente de várias fontes e possuir diversos tipos de *layout*. Para isto, foi necessário buscar métodos de limpeza textual que retira elementos que são irrelevantes ou prejudicam a comparação de similaridade.

Para o desenvolvimento buscou-se utilizar ferramentas livres, para diminuir os custos de implementação. Nas próximas seções estão descritas as ferramentas utilizadas, API´s[2], bem como o hardware de testes.

4.1 Estrutura da ferramenta

Para o desenvolvimento da CTP buscou-se utilizar ferramentas que são muito utilizadas atualmente e que possuem suporte e desenvolvimento contínuo. As linguagens de programação utilizadas são PHP, Javascript, SQL e de marcação de texto HTML.

4.1.1 PHP (Personal Home Page)

A linguagem de programação Hypertext Processor (PHP), também conhecida como Personal Home Page, é uma linguagem de programação para servidores WEB que atualmente é a mais utilizada pelos programadores e software *houses*. O PHP possibilita uma infinidade de ações como utilizar banco de dados, compartilhar o mesmo arquivo com outras linguagens de programação, ser relativamente simples de utilizar e isto tudo a torna muito popular.

O PHP possui licença Copyright e é mantida pelo Grupo PHP. Conforme Licença disponibilizada no endereço http://www.php.net/license/3_01.txt, os usuários podem utilizar livremente o PHP desde que não infrinjam as restrições impostas nesta licença.

Atualmente o PHP encontra-se na versão 5.4.3 e pode ser encontrado para download no site www.php.net, endereço oficial desta linguagem. Além do download dos arquivos binários, este site disponibiliza a documentação completa juntamente com fórum e diversas fontes de ajuda (HOLZNER, 2008).

Para ser possível analisar um documento encontrado na Internet foi necessário buscar ferramentas que possibilitassem fazer o download do texto. Para isto foram utilizadas a função “file_get_contents”, “shell_exec” e a biblioteca “CURL”, todas suportadas pelo PHP. A função “file_get_contents” busca todo o conteúdo HTML do endereço e retorna como String para a variável definida. A biblioteca CURL tem função semelhante, porém com mais opções suportadas como vários tipos de servidores e protocolos de comunicação. Já a função

[2] API são rotinas e regras impostas pelo software para que possa utilizar seus serviços e funcionalidades por outros softwares. Através da programação pode-se ter acesso às funções disponibilizadas pelo software.

"shell_exec" tem característica distinta das outras descritas, pois possibilita a execução de comando no shell[3] através do código PHP. Esta função foi muito utilizada para descarregar o arquivo PDF encontrado na Internet para o servidor (PHP, 2012).

Ao descarregar o arquivo PDF é necessário ler e disponibilizar o texto do mesmo para o PHP. Isto é feito através da função "SHELL_EXEC" do PHP em conjunto com a função "PDFTOTEXT" que o software XPF[4] disponibiliza. Desta forma é possível carregar o conteúdo do arquivo PDF para uma variável no PHP para que possa ser utilizado pelo CTP.

Para processar o PHP foi utilizada uma ferramenta livre, que possui grande aceitação pelos desenvolvedores e que provem um bom desempenho. O Apache, como descrito a seguir, é muito utilizado juntamente com PHP para o desenvolvimento de páginas WEB e sistemas, por possibilitar a utilização de vários recursos e possuir versões para os sistemas operacionais mais utilizados atualmente.

4.1.2 Web Service Apache

O Apache é um WEB *Service* mantido pela Apache Software Foundation (ASF) muito utilizado atualmente em conjunto com a linguagem de programação PHP. Criado em 1995 pela National Center for Supercomputing Applications foi baseado no NCSA server que era o WEB server mais utilizado da época.

Após a incorporação do módulo mod_rewrite[5], o primeiro módulo de linguagem de programação a ser criado foi o mod_perl para a linguagem PERL[6]. Com o grande sucesso e crescimento que teve a utilização do Apache e a linguagem PERL, logo foram criados os módulos para outras linguagens, inclusive PHP (KEW, 2007).

Atualmente o Apache encontra-se na versão 2.4 e mantido pela ASF, está em constante evolução. Assim como o PHP, o apache pode ser utilizado e copiado conforme

[3] Linguagem utilizada por sistemas operacionais para iteração entre usuário e computador.
[4] Software que possui como principal função a visualização de documentos PDF. Possui diversas funções e entre elas a "pdftotext", que converter o arquivo PDF em texto.
[5] O mod_rewrite trouxe ao Apache a funcionalidade de reescrever de forma instantânea URLs solicitadas. Este módulo suporta uma quantidade ilimitada de regras e cada regra suporta um número infinito de condições para fornecer uma grande flexibilidade na manipulação de URLs (APACHE SOFTWARE FUNDATION, 2012).
[6] A linguagem de programação Pratical Extraction and Report Language (PERL) surgiu em 1987 criada por Larry Wall e liberada com código-fonte aberto. PERL foi criada para ser uma linguagem flexível, portável e eficiente e juntamente com Apache e Linux formaram um movimento pelo código aberto. Esta linguagem teve seu auge na década de 90 e é utilizada até hoje para o desenvolvimento de portais na WEB (DEITEL, 2001).

informa a licença disponibilizada no site oficial da ASF no endereço http://www.apache.org/licenses/LICENSE-2.0.

Para o desenvolvimento da interface e iteração com o usuário foi utilizada a linguagem de programação JavaScript juntamente com a linguagem de marcação de texto HTML . Estas linguagens possibilitam criar a interface de comunicação entre o usuário e a ferramenta.

4.1.3 HTML (Hypertext Markup Language)

A linguagem de marcação de textos Hypertext Markup Language (HTML) foi criada quando a Internet havia sido liberada para uso comercial. Criada por Tim Berners-Lee e Robert Caillau em 1989 quando trabalhavam no European Particle Physics Laboratory (CERN), o HTML provia som, imagem e texto, se tornando a base de comunicação na Internet da época.

Na Figura 17 está ilustrada de forma simplificada a estrutura básica de um documento HTML.

Figura 17: Estrutura HTML.

```
<html>
    <head>
        <title>Olá Mundo</title>
    </head>
    <body>
        <h1>Olá mundo!</h1>
        <!-- Comentário em HTML -->
    </body>
</html>
```

Fonte: Do autor.

No entanto, HTML que é uma linguagem de marcação, não continha um modelo rápido e organizado de criação e alteração do visual de páginas WEB. Devido a essa necessidade, foi desenvolvida a ferramenta descrita na próxima seção.

4.1.4 CSS (Cascading Style Sheet)

O Cascading Style Sheet (CSS) foi criado em 1995 pelo World Wide Web Consortium (W3C). Nesta época os documentos não possuíam muitas opções de edição de estilo, porém quando foi criado o CSS houve uma grande aceitação e agora era possível editar a cor da letra, cor de fundo, bordas dentre outros. Em 1996 já era utilizado em grande escala agilizando o processo de customização do HTML, tornando a interface mais elegante.

Além do HTML e CSS houve uma nova necessidade que era a execução de determinadas ações no computador do usuário, então foi desenvolvida uma nova linguagem de programação que atendesse as carências da época.

4.1.5 Linguagem JavaScript

A linguagem de programação JavaScript trouxe novos recursos para as páginas WEB, principalmente na parte do cliente. Esta linguagem possui uma programação leve e recursos de orientação a objetos e seu núcleo assemelha-se às linguagens C, C++ e Java. Após a chegada do JavaScript, houve uma mudança na forma com que as páginas WEB eram apresentadas, pois agora era possível tornar as páginas HTML dinâmicas no lado do cliente.

Na implementação da ferramenta desta monografia, foram utilizadas diversas vezes esta linguagem para validações e interação entre o usuário e a solução. Na Figura 18 há um exemplo de utilização do JavaScript em uma ação de clique em um botão. Como pode-se observar, o código que é colocado dentro da ação do botão, não necessita das *tags* citadas anteriormente.

Figura 18: Javascript.

```
<textarea id="textBusca" name="textBusca" onfocus="if(this.value == 'Digite seu texto aqui.'){this.value='';
this.style.color='black';}" onblur="if(this.value.length == 0){ this.style.color='#DCDCDC'; this.value='Digite seu
texto aqui.';}" style="color: rgb(220, 220, 220); ">Digite seu texto aqui.</textarea>
```

Fonte: Do autor.

Na Figura 18 há um elemento HTML chamado *textarea* com duas ações chamadas *onfocus* e *onblur*. No *onfocus* a ação será executada quando o elemento ganhar foco e o JavaScript contido ali executará as ações de troca de cor do texto e alteração no conteúdo HTML do elemento *textarea*, caso a condição do *if* ocorra. A ação *onblur* será executada

quando o elemento *textarea* perder o foco e o JavaScript também irá trocar a cor do texto e seu conteúdo, conforme condição do *if*.

Outra funcionalidade que o JavaScript proporciona e que foi muito utilizada na implementação do CTP é o Asynchronous JavaScript and XML (AJAX). Esta funcionalidade torna possível a comunicação cliente-servidor de forma assíncrona utilizando objeto XMLHttpRequest e todos os navegadores atuais a suportam.

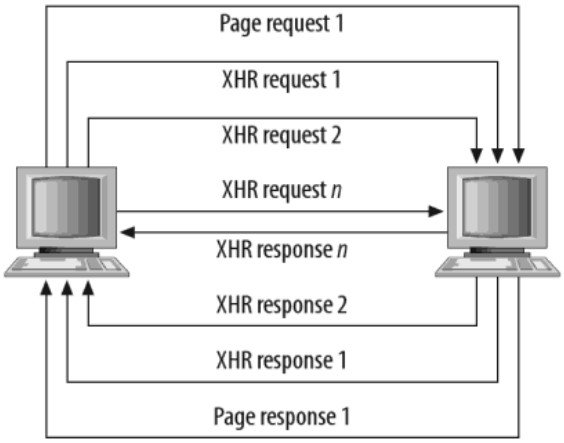
> XML is the metalanguage the W3C created and that developers use to define markup languages such as XHTML. Browser developers rely on XML´s metalanguage rules to create automated processes that read the language definition of XHTML and implement the processes that ultimately display or otherwise process XHTML documents (MUSCIANO ET AL., 2007, p.472).

> XHTML is a Family of current and future document types and modules that reproduces, subset, and extend HTML 4. XHTML family document types are XML based, and ultimately are designed to work in conjunction with XML-based user agents. (W3C, 2012, texto digital).

O AJAX trouxe vários benefícios para o desenvolvimento WEB, tornando as páginas e requisições mais rápidas. Isso acontece devido ao menor número de requisições, sendo que em muitos casos somente é necessário atualizar uma pequena parte da aplicação. Reduzindo a carga ao servidor a aplicação terá uma resposta mais rápida a cada nova requisição e, além disso, recarregando somente uma parte da página o tamanho e quantidade de tráfego na rede do navegador cliente diminuem.

Além da diminuição da carga ao servidor, outro benefício que o AJAX proporciona é aproximar a aplicação WEB de uma aplicação desktop. Com a utilização do AJAX é possível fazer *n* requisições ao servidor sem necessitar carregar a página inteira como mostra a ilustração da Figura 19. Isto trouxe um grande benefício às aplicações WEB proporcionando maior aproximação de uma aplicação WEB a uma aplicação desktop (HOLDENER, 2008).

Figura 19: Requisições AJAX.



Fonte: Holdener (2008, p.9).

No desenvolvimento do CTP o AJAX foi muito utilizado para enviar os dados dos documentos, URLs, textos que o usuário envia para a análise e atualização de conteúdo na visualização da busca. Juntamente com o JavaScript, é utilizada a biblioteca jQuery, descrita na próxima seção.

4.1.6 Biblioteca jQuery

jQuery é uma biblioteca criada para facilitar e agilizar a interação entre o JavaScript e o HTML e além disso, trouxe inovações na iteração do conteúdo HTML e o navegador. Esta biblioteca é muito utilizada atualmente pelos desenvolvedores WEB porque possui muitas vantagens como possuir um tamanho reduzido, ser *Open Source*, diminui as diferenças entre os navegadores atuais, possui documentação completa e é relativamente simples de programar.

Além de possuir as ações que o JavaScript proporciona, o jQuery trouxe animações para as páginas WEB, coisas que antes somente era possível com ferramentas como o FLASH. As opções de animação *slide, toggle, animate, hide, show* dentre outros, transformam páginas WEB e definem um novo conceito de iteração máquina usuário. Esta biblioteca possui uma grande vantagem frente ao principal concorrente FLASH sendo muito mais leve, porém perde em quantidade de efeitos e ferramentas que facilitem sua programação (JQuery Community, 2010).

Para a utilização do banco de dados foi utilizada a linguagem SQL[7] juntamente com PHP e o gerenciador de banco de dados MySQL para tornar possível o armazenamento de dados e análise dos documentos.

4.1.7 Banco de dados

Para armazenar informações do sistema e de dados que são inseridos foram criados os Sistemas Gerenciadores de Banco de Dados (SGBD). Devido a sua funcionalidade, os SGBD tornaram-se essenciais na criação de software e permitiram criar novas funcionalidades e gama de opções de software (COSTA, 2007).

Segundo Costa (2007), o acesso simultâneo ao SGBD é necessário para prover maior velocidade de acesso. Isso somente é possível porque o Banco de dados possui vários mecanismos que possibilitam e gerenciam este acontecimento. Para possibilitar o acesso simultâneo existe um complexo sistema de bloqueio que deixa o usuário aguardando sem realizar transações[8] enquanto este bloqueio durar.

4.1.8 Gerenciador de banco de dados MySQL

O MySQL é o gerenciador de banco de dados *Open Source* mais utilizado atualmente e a cada dia milhares de desenvolvedores iniciam atividades utilizando MySQL. Este SGBD entrou recentemente na área dos bancos de dados relacionais, conceito criado pela IBM. Além dos aplicativos modestos, o MySQL também está presente em grandes sistemas que utilizam grande quantidade de dados (TAHAGHOGHI, 2007).

Em 2004 o MySQL já possuía mais de 4 milhões de usuários, este sucesso só foi possível porque o MySQL possui versões para os sistemas operacionais mais utilizados. Criado especificamente para possuir alto desempenho e estabilidade, teve o desenvolvimento

[7] Para padronizar o acesso a SGBD foi criada em 1986 a primeira versão da linguagem de programação Structured Query Language (SQL) e logo após, os fornecedores de SGBD a adotaram. Após a adoção da linguagem, os fornecedores de SGBD criaram funções e comandos que trouxeram grandes benefícios e desta forma, acabaram sendo incorporadas na linguagem SQL (COSTA, 2007).

[8] Transações são unidades lógicas de processamento do banco de dados. Elas possuem quatro propriedades desejáveis (propriedades ACID): atomicidade, consistência, isolamento e durabilidade. Para manter essas propriedades, o SGBD utiliza-se de diversos mecanismos como áreas diferenciadas de buffers e logs, além dos bloqueios. *(COSTA, 2007, p. 4)*.

de sua primeira versão em 1979 por Michael Widenius e se chamava UNIREG. Em 1996, Widenius decidiu criou uma nova versão do banco de dados para utilizar a Linguagem SQL e que desempenhar funções específicas. Este novo banco de dados ganhou o nome MySQL e poucos meses depois teve seu desenvolvimento produzido pela Solaris que começou a distribuir a versão 3.11 (Tahaghoghi, 2007).

Após escolher esta ferramenta como gerenciadora do banco de dados, foi necessário criar um planejamento para a construção do banco. Para isto utilizou-se o modelo ER juntamente com o software MySQL Workbench[9] para construir a modelagem dos dados.

4.1.9 Diagrama ER do banco de dados

O Diagrama Entidade Relacionamento possibilita o planejamento e esquematização do banco de dados. Este modelo é uma das principais ferramentas utilizadas por profissionais para representar o modelo de forma conceitual do negócio em questão.

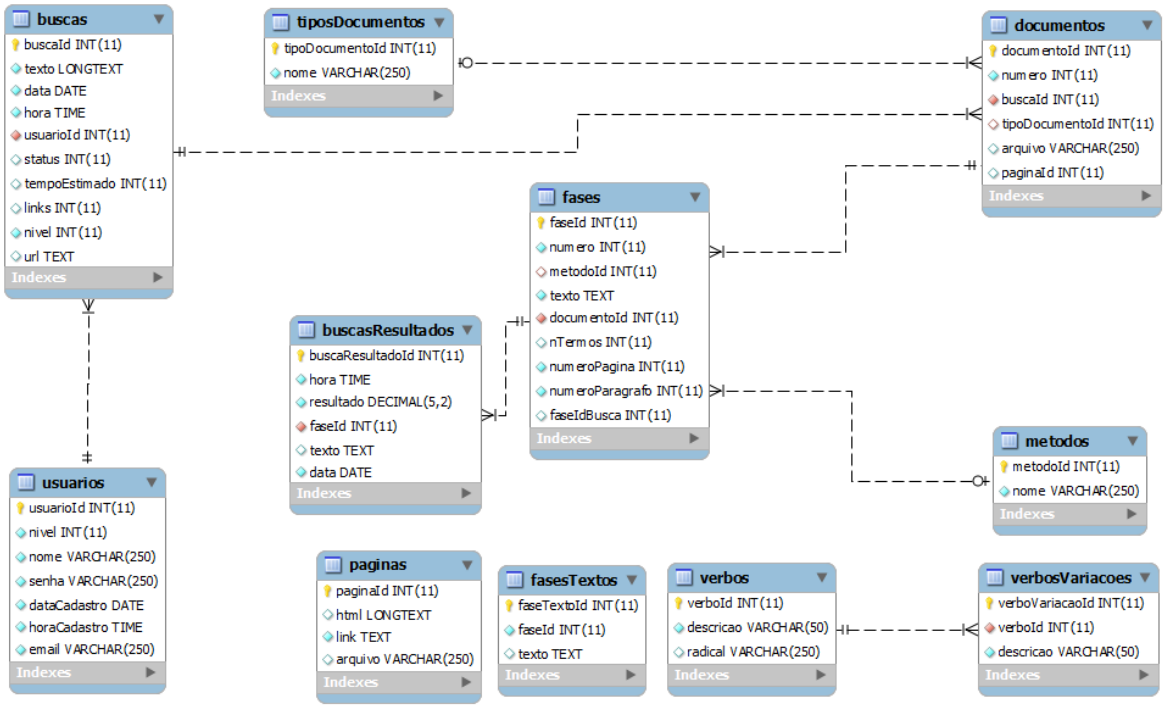
A Figura 20 demonstra modelo conceitual desenvolvido para o CTP, desta forma fica simplificada a visualização do banco de dados com os campos, chaves primárias e chaves estrangeiras.

No diagrama demonstrado na Figura 20 nota-se que há 11 tabelas sendo que duas delas não possuem ligações de chaves estrangeiras com outras tabelas, são elas a tabela “pagina” e “fasesTextos”. Estas tabelas não possuem ligações devido ao fato de possuir *engine* (motor) MyISAM que é necessária para utilizar a busca com MySQL FullText.

A *engine* MyISAM é o padrão utilizado pelo MySQL e possui algumas características como a criação de três arquivos para cada tabela que possui este padrão. Um destes arquivos armazena o formato da tabela, outro armazena os dados e o terceiro possui os índices (MYSQL, 2012).

[9] MySQL Workbench é uma ferramenta que possibilita a visualização do banco de dados através da modelagem de diagrama ER. Além da modelagem do banco de daods, o MySQL Workbench também disponibiliza o desenvolvimento SQL e ferramentas administrativas para configuração de servidores. Atualmente este software está disponível nas plataformas Windows, Linux e Mac OS (<http://www.mysql.com/products/workbench/>).

Figura 20: Diagrama ER do banco de dados.



Fonte: Do autor.

As tabelas que possuem alguma ligação de chave estrangeira possuem *engine* InnoDB que é muito utilizado atualmente. Este padrão permite maior segurança em transações do banco de dados. Esta segurança está diretamente relacionada aos mecanismos de transação “begin”, “commit” e “rollback”. Estes mecanismos possibilitam executar diversas transações e se houver erro, pode ser executado “rollback” para restaurar as informações como estavam antes do comando “begin”, caso ocorra o contrário, é executado o comando “commit” para finalizar as transações. Outra característica da *engine* InnoDB é o bloqueio por chaves estrangeiras que permitem manter a integridade dos dados (MYSQL, 2012).

Como pode-se observar na Figura 20, na tabela “usuarios” serão armazenados os dados referentes aos usuários cadastrados no sistema, são dados básicos para *login*, contato e nível de acesso. As tabelas “tiposDocumentos” e “metodos” possuem dados armazenados referente à extensão de documento e método de busca respectivamente. Estes dados são úteis para consulta e para demonstrar a informação ao usuário.

Outras tabelas que possuem uma grande contribuição para o método de radicalização são “verbos” e “verbosVariacoes”. Possuindo as informações dos verbos e suas conjugações

foi possível identificar o radical do mesmo. Desta forma pode-se realizar consultas com partes do termo até que o radical seja encontrado.

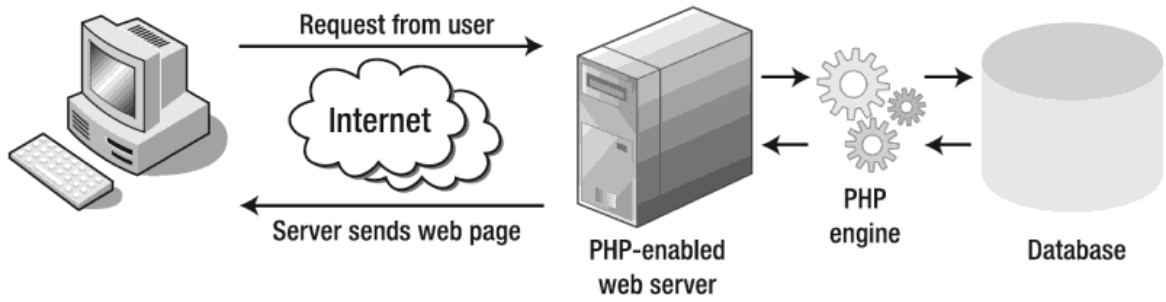
A tabela “buscas” guarda informações de cada busca que o usuário fizer. Além da data e hora inicial, esta tabela guarda informações de URL, texto, status, usuário, tempo estimado e nível da busca. Após gravar a busca, o sistema CTP grava o documento de número 1 na tabela “documentos”. Esta tabela registra os documentos encontrados e o documento que o usuário forneceu. Juntamente com a tabela “documentos”, a tabela “paginas” irá armazenar o texto ou HTML que é resgatado do documento proveniente da Internet.

Estes textos ou HTML são utilizados nas tabelas “fases” e “fasesTextos”, que armazenam as informações referente às frases ou partes dos textos. Estes dados gravados são os textos originais e textos vetorizados que são utilizados na comparação de similaridade.

Por fim, a tabela “buscasResultados” armazena os dados da comparação entre os documentos e também data e hora de término do processamento.

Para entender melhor como funciona a relação HTML, PHP, servidor WEB e banco de dados, a Figura 21 demonstra como acontecem as requisições, como são processadas e devolvidas ao navegador.

Figura 21: Requisições WEB ao PHP.



Fonte: Powers (2010, p. 3).

Como pode ser observado na Figura 21, quando um usuário acessa uma página PHP e faz requisições, acontecem as seguintes ações:

I. O navegador envia a requisição ao servidor WEB.

II. O servidor passa a solicitação para o PHP.

III. O PHP processa o código e se houver alguma requisição ao banco de dados, envia a requisição, recebe os dados e constrói a página.

IV. Retorna a requisição ao navegador do usuário com os dados processados.

Este processo de requisição e resposta leva frações de segundos, o que pode variar dependendo da Internet do usuário e da lógica utilizada no desenvolvimento (POWERS, 2010).

Além das linguagens de programação e banco de dados, também foram necessárias algumas API´s que são disponibilizadas gratuitamente. Estas API´s tornaram possível a tradução do documento fornecido pelo usuário e também a utilização dos gráficos na demonstração dos resultados.

4.1.10 API de tradução de texto

A API de tradução disponibilizada através do produto Bing da empresa Microsoft possibilita a tradução instantânea do texto enviado. Através desta informação foi possível inserir juntamente a busca de documentos na ferramenta CTP.

Para utilizar a API de tradução foi necessário criar uma Application ID no endereço https://br.ssl.bing.com/webmaster/Developers/CreateAppId.

Figura 22: Bing - Application ID


APPLICATION

Developers with existing AppIDs can continue using Bing Search API 2.0 until August 1, 2012. On and after this date, Bing Search API 2.0 AppIDs will no longer return results. Developers can continue using the API by signing up for it in the Windows Azure Marketplace. Read the Migration Guide and FAQs to get started.

Application name	Application ID	Website	Status
copytopaste	848D733080CA41C47A9816EDB6A607504195BD87	www.copytopaste.com.br	Enabled

Fonte: Do autor.

Para a criação deste ID foi necessário fornecer alguns dados como conta no Windows Live, website, e-mail dentre outros. O ID gerado pelo Bing Search API 2.0 foi 848D733080CA41C47A9816EDB6A607504195BD87 para a aplicação no domínio www.copytopaste.com.br conforme Figura 22.

A partir deste ID foi possível criar o código que envia a requisição para os servidores da API de tradução do Bing. Para realizar o envio foi necessário encapsular os dados na URL de envio utilizando a função do PHP chamada “urlencode”. A Figura 23 demonstra um exemplo de requisição à ferramenta de tradução.

Figura 23: Requisição à API de tradução.

```
http://api.bing.net/json.aspx?AppId=848D733080CA41C47A9816EDB6A607504195BD87&Version=2.2&Query=Ol%C3%A1+Mundo%21
&Sources=Translation&Translation.SourceLanguage=pt&Translation.TargetLanguage=en
```

Fonte: Do autor.

Na url de requisição exposta na Figura 23 possui alguns parâmetros que devem ser observados. Um destes parâmetros é o Application ID, uma sequência de caracteres que fica entre "?AppId=" e "&Version=". Outro parâmetro é a própria frase que se deseja traduzir, neste caso "Olá Mundo!" que com urlencode[10] resultou em "ol%c3%a1+Mundo%21" e na URL fica entre "&Query=" e "&Sources=". Os dois últimos parâmetros são as linguagens de origem e destino localizadas na URL logo após "&SourceLanguage=" e "TargetLanguage=" respectivamente.

A resposta que a ferramenta de tradução retorna é uma sequência de caracteres no formato JSON[11] como demonstrado na Figura 24.

Figura 24: Resposta de API de tradução.

```
{"SearchResponse":{"Version":"2.2","Query":{"SearchTerms":"olÃ¡ Mundo!"},"Translation":{"Results":[{"TranslatedTerm":"Hello world!"}]}}}
```

Fonte: Do autor.

Através da resposta da API de tradução é possível verificar os termos de pesquisa e resposta localizados logo após "SearchTerms" e "TranslatedTerm" respectivamente conforme Figura 24.

Além da API de tradução, foi utilizada uma API disponibilizada pela empresa Google.

4.1.11 Google Charts

Esta API denominada Google Charts é disponibilizada gratuitamente pela empresa Google[12] mediante cadastro em um de seus produtos. Esta ferramenta possui uma grande

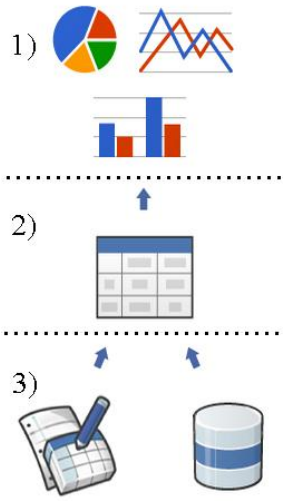
[10] Função de codificação de URL encontrada nas versos 4 e 5 do PHP, retorna os caracteres não alfanuméricos substituídos pelo caractere "%" seguido de dois dígitos exceto o espaço que é substituído com sinal de +. Os caracteres não alfanuméricos que não sofrem codificação são "-", "_" e "." (http://php.net/manual/pt_BR/function.urlencode.php).

[11] O JavaScript Object Notation (JSON) é um formato para troca de dados. Esta representação dos dados é ideal para ser utilizado por seres humanos e o computador. Seu uso não necessita a utilização de JavaScript especificamente (www.json.org).

variedade de gráficos customizáveis desde o mais simples até o complexo e sua utilização é simples.

Para a criação do gráfico são necessários três elementos conforme demonstra a Figura 25.

Figura 25: Elementos de criação do Google Charts.



Fonte: Google Charts (2012, texto digital).

O primeiro elemento é a biblioteca Chart que possui classes em JavaScript. Esta biblioteca permite que o desenvolvedor utilize a API e faça requisições para criar gráficos. O segundo elemento são as tabelas de dados que possuem uma estrutura de dados e classes. Esta estrutura proporciona a troca do gráfico sem a necessidade de alterar a forma que os dados são fornecidos. As classes possibilitam filtrar, modificar e classificar dados. O terceiro elemento são os dados fornecidos que terão a função de popular os gráficos e tornar sua criação possível (GOOGLE CHARTS, 2012).

Para possibilitar a utilização da API é necessário carregar as classes utilizando o código demonstrado na Figura 26.

Figura 26: Google Charts - Carregar classes.

```
<script type="text/javascript" src="https://www.google.com/jsapi"></script>
```

Fonte: Do autor.

[12] Empresa criada por Larry Page e Sergey Brin em 1998 nos Estados Unidos da América, possuía um mecanismo de buscas chamado Google como seu único produto. Em poucos anos a empresa cresceu exponencialmente e alcançou milhões de usuários com seus diversos produtos.

Carregando a classe no software em desenvolvimento é possível começar a utilizar a API conforme exemplos demonstrados no site da ferramenta que se encontra no endereço “https://developers.google.com/chart/interactive/docs/examples”. Na Figura 27 é demonstrada a sequência de códigos para a requisição do gráfico de barras.

Figura 27: Google Charts - Código de requisição do gráfico.

```
var chart = new google.visualization.ColumnChart(document.getElementById('chart_div'));
chart.draw(data, {width: '100%', height: 266,
           colors: ['#c7cfc7', '#b2c8b2', '#d9e0de', '#cdded1'],
           chartArea: {left:38,top:30, width:"75%",height:"70%"},
           legendTextStyle: {color:'#666666'},
           hAxis: {title: 'Year',
           titleTextStyle: {color: '#5c5c5c'},
           titlePosition: 'out'}
          });
```

Fonte: Google Charts (2012, texto digital).

Como pode-se observar, é chamada a classe google.visualization.ColumnChart que irá criar um gráfico de barras no elemento HTML com o ID “chart_div”. Na função draw são passados vários parâmetros como largura e altura do gráfico, cores das barras, cor da legenda, título e cor do texto. Existem outros parâmetros além dos já citados anteriormente, desta forma é possível personalizar totalmente o gráfico que será exibido.

Outra ferramenta muito utilizada no CTP é a barra de progresso que está descrita na próxima subseção.

4.1.12 jQuery progressBar

Esta biblioteca proporciona, dentre outras muitas funções, a criação de barras de progresso. Desenvolvida através de jQuery, esta ferramenta possui características como animação e customizações de cores. Atualmente esta ferramenta encontra-se na versão 2.01 e pode ser encontrada no endereço “http://t.wits.sg/jquery-progress-bar”.

A Figura 28 exibe os códigos de criação da barra de progresso. Na ilustração, há dois códigos sendo exibidos, o primeiro é responsável pela criação da barra utilizando a imagem da cor laranja no elemento HTML com ID progressBar1. O segundo código define a porcentagem do progresso, neste caso 100%.

Figura 28: jQuery progressBar - Código de criação.

```
$(document).ready(function(){$('#progressBar1').progressBar({barImage:'lib/progressBar/images/progressbg_orange.gif', showText:true});});
$('#progressBar1').val(100);
```

Fonte: Do autor.

A biblioteca jQuery também é utilizada no plugin jTruncate descrita na próxima seção.

4.1.13 jTruncate

Este plugin[13] do jQuery proporciona ao desenvolvedor uma forma fácil e rápida de ocultar parte do texto. Além desta função, também proporciona uma melhora na estética do software. Esta ferramenta possui versão única e fornece diversos parâmetros de configuração:

I. length: Tamanho do texto que ficará em exibição.
II. minTrail: Tamanho mínimo do texto a ser ocultado.
III. moreText: Texto exibido no botão que irá mostrar o texto ocultado.
IV. lessText: Texto exibido no botão que irá ocultar o texto.
V. ellipsisText: Texto exibido após o texto que não será ocultado.
VI. moreAni: Animação do texto ocultado quando for exibido.
VII. lessAni: inverso ao moreAni.

Esta ferramenta é muito utilizada em notícias, dicas entre outros e pode ser encontrada no endereço “http://www.jeremymartin.name/projects.php?project=jTruncate”. A ferramenta CTP utilizou este *plugin* ao exibir o resultado com os textos dos documentos que obtiveram maior nível de similaridade.

4.2 Funcionalidades do sistema

Para validar proposta da presente monografia foi desenvolvido o sistema CTP que analisa os documentos enviados pelo usuário e encontrados na Internet e apresenta resultados de similaridade.

[13] Software de computador que trabalha como um módulo ou extensão de outro software.

Com a exposição de resultados e diversas ferramentas e métodos implementados, é possível visualizar várias análises e resultados obtidos com documentos encontrados na Internet. Cada ferramenta utilizada possui características próprias e atuam em pontos específicos do software como descrito mais adiante.

Para possibilitar o acesso de qualquer lugar do mundo o CTP foi desenvolvido para estar disponível para ser acessado através da Internet. Outro benefício que o sistema proporciona é o acesso através de diversos sistemas operacionais através dos navegadores e inclusive dispositivos como smartphones e tablets.

A ferramenta de inibição de plágio CTP possui várias telas que formam seu conteúdo e demonstram de forma intuitiva suas funcionalidades. Estas telas juntamente com suas funções são descritas a seguir.

4.2.1 Página inicial e *login*

Ao acessar o sistema CTP no endereço “www.copytopaste.com.br” é demonstrada a tela de *login*, onde o usuário deve fornecer as informações necessárias para ter acesso às outras partes do mesmo.

Foi necessário o desenvolvimento de restrição de acesso através de *login* porque há a necessidade de restrição de buscas e privilégios de processamento para não sobrecarregar o sistema e não privilegiar a somente uma pessoa.

Na Figura 29 é demonstrada a tela de *login* e como pode-se observar, há o formulário do acesso com e-mail e senha e duas informações importantes sobre o sistema.

Figura 29: CTP, Tela de login.

Fonte: Do autor.

Estas informações demonstram que o sistema pode ser utilizado gratuitamente e o que ele oferece. Além das informações descritas, há também uma imagem que passa a informação que os documentos enviados são analisados com outros documentos de qualquer lugar do mundo.

Para acessar, o usuário utiliza o e-mail e senha de cadastro e caso não possua, deve acessar o “Cadastre-se” ou “Recuperar senha”.

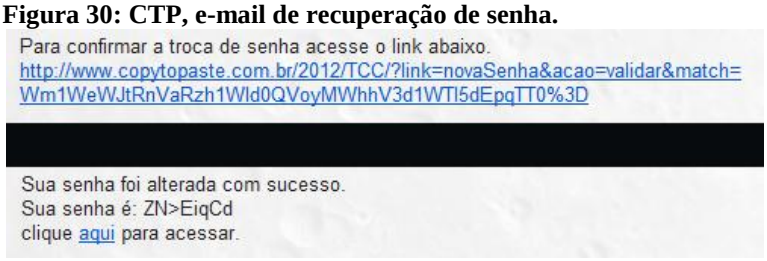
4.2.2 Cadastro e recuperação de senha

No cadastro de usuário são requisitadas somente informações básicas como “nome”, “e-mail”, “senha” e “confirmação da senha”. Com este cadastro é possível ter informações para contato e para acesso ao CTP. Após preenchidas os dados requisitados, é possível enviar e acessar o sistema na tela de *login*.

Caso o usuário possua cadastro e tenha extraviado a senha, será necessário recuperar a senha deste cadastro. Para isto o usuário terá que acessar a tela de recuperação de senha clicando em “Recuperar senha” na tela de *login*. Após clicar o usuário será redirecionado ao formulário que solicita somente a inclusão do e-mail. Ao enviar a informação de e-mail o

CTP envia um e-mail ao usuário através da função PHPMailer[14], solicitando o acesso ao *link* gerado para realizar a troca da senha. Na Figura 30, acima da faixa preta é possível observar a informação enviada ao e-mail do usuário e ao clicar no endereço, é enviado um novo e-mail com as informações de senha e endereço para o *login* da ferramenta, demonstrados nesta figura na parte abaixo da faixa preta.

Figura 30: CTP, e-mail de recuperação de senha.



Fonte: Do autor.

Após efetuado o *login*, o usuário será redirecionado para a página principal da ferramenta. Esta página é a responsável por iniciar as buscas dos usuários e também fornece as opções de configuração das mesmas. Nesta tela é possível alterar as informações de nome e senha do cadastro de usuário.

4.2.3 Envio do texto

Para possibilitar uma gama maior de formas de envio de documentos para a ferramenta CTP, foram desenvolvidas três que são descritas a seguir.

Na tela demonstrada na Figura 31, é possível identificar os campos de envio de arquivo, texto e endereço de página WEB. Através dos campos citados, o sistema possibilita o envio de arquivos no formato PDF, que é um dos formatos de arquivo mais utilizados em trabalhos acadêmicos e para troca de informações. Também é possível enviar um texto puro através do campo texto localizado logo abaixo do menu superior do CTP. Por último, é possível informar o endereço de uma página WEB para servir como documento enviado pelo usuário.

[14] Função do PHP que possibilita a transferência de e-mail através de POP3 e SMTP. Atualmente encontra-se na versão 5.2.1 no endereço “http://phpmailer.worxware.com/”.

Figura 31: - CTP, tela de buscas.

Fonte: Do autor.

Enviando um documento no formato PDF pelo campo “arquivo” a ferramenta faz o *upload* deste e armazena em uma pasta específica para o armazenamento. Após armazenar, o software utiliza a função descrita anteriormente chamada “pdftotext” que irá extrair o texto deste arquivo e armazenar em uma variável para fazer o tratamento.

Se o usuário utilizar o campo de texto puro para iniciar a busca, o CTP irá armazenar o texto na tabela buscas do banco de dados e iniciará o tratamento deste. Diferentemente do arquivo PDF enviado, o texto puro não necessita de função para ser armazenado no banco de dados, porém este requer maior atenção na segurança para evitar intrusão por meio de SQL injection[15] e demais técnicas.

A terceira opção de iniciar a busca é fornecendo um endereço WEB através do campo URL que por sua vez irá requisitar o HTML do mesmo. Após receber o HTML, o CTP faz um tratamento para eliminar conteúdo menos importante. Este conteúdo inclui partes de script JavaScript, menus, botões, imagens em HTML e o script CSS. Após feito o filtro dos dados recebidos, o restante destes é gravado na tabela buscas, como acontece com o texto puro.

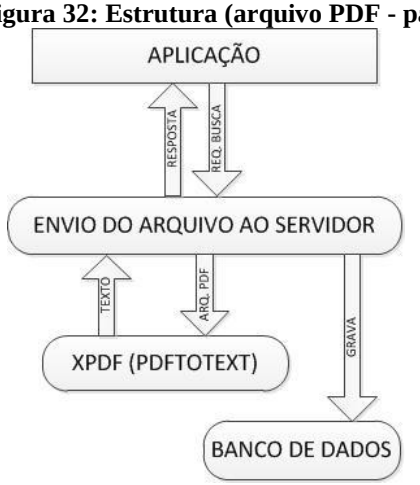
4.2.3.1 Envio de arquivo PDF

Quando o usuário utiliza o envio do texto em arquivo PDF, o CTP realiza várias ações de tratamento do texto e por fim a vetorização deste que é definido como “Documento 1”. A partir do “Documento 1” o CTP realizará buscas para encontrar os documentos similares.

[15] Técnica utilizada por invasores através de inserções de comandos SQL em sistemas com fraca proteção. Esta técnica é utilizada em páginas que não possuem validação de inserções como as aspas (CARR J., 2012).

A Figura 32 demonstra o processo de tratamento do texto quando o usuário utilizar a opção de envio de arquivo PDF. Este arquivo pode possuir diversos formatos de *layout*, possuindo tabelas, figuras, códigos de programação entre outros. Desta forma é necessário tratar este texto que o CTP recebe para que não haja problemas de segurança e também não ocorram falhas no processamento da comparação de similaridade. Devido a estes percalços, foi necessária a adição de funções que auxiliam no tratamento do texto.

Figura 32: Estrutura (arquivo PDF - parte 1).



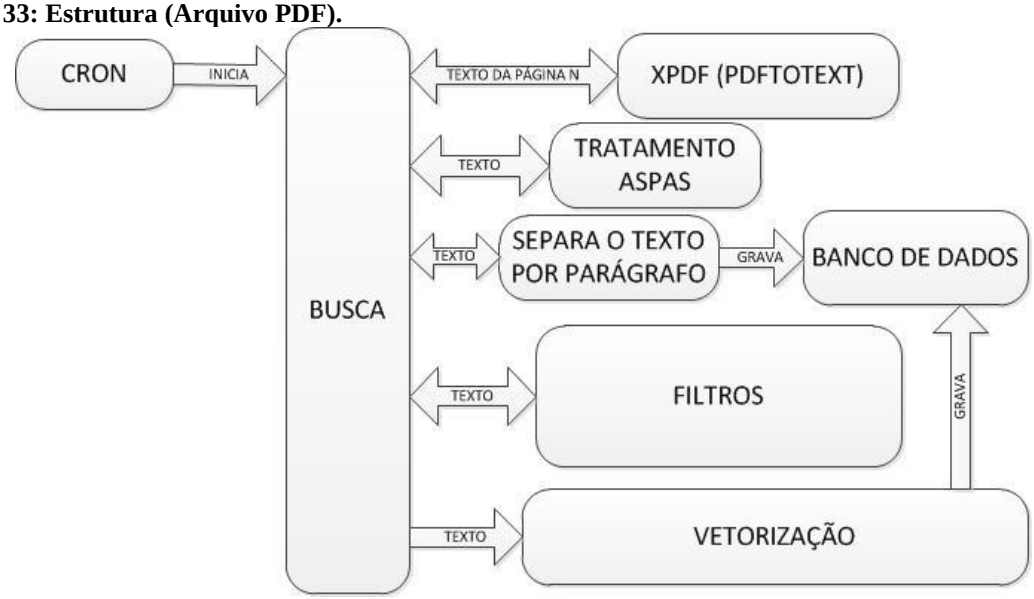
Fonte: Do autor.

Como pode-se observar na Figura 32, o arquivo PDF é enviado ao servidor e transformado em texto pela função “pdftotext” do software XPDF. Após feita a transformação de todo o documento, é armazenado na tabela buscas como texto original.

Na Figura 33 é demonstrada a próxima parte que será executada. Quando o CTP inicia o processo de busca através da CRONTAB[16], este arquivo PDF é lido página por página, sendo que cada página é quebrada em parágrafos e o texto resultante é gravado na tabela fases juntamente com o número da página e parágrafo. Para não haver conflitos na inserção dos dados, é feito um tratamento de aspas, eliminando assim possíveis erros e problemas de segurança neste processo.

[16] Software da plataforma Unix que possibilita a execução de comandos SHELL podendo determinar data e hora da execução e também possibilita a opção de execução a cada faixa de tempo determinada.

Figura 33: Estrutura (Arquivo PDF).



Fonte: Do autor.

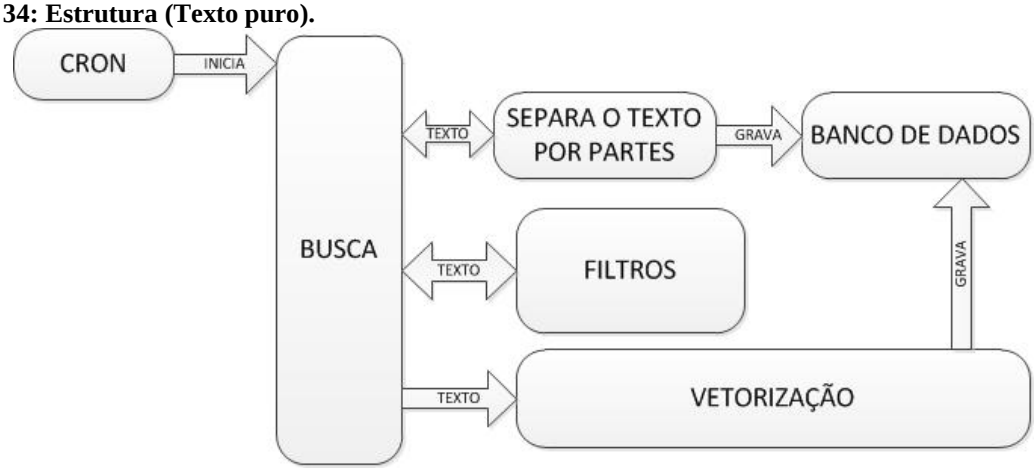
Na continuação do processo, cada um dos parágrafos é tratado de forma isolada. O texto passa por diversos filtros como eliminação de quebras de linha, pontuação, elementos HTML, *stopwords*, acentuação e espaços concatenados. Após passar por esta bateria de filtros o texto está pronto para iniciar o processo de vetorização, onde são eliminados os termos com menor relevância e é definida a quantidade que cada termo possui no texto. Concluída a vetorização, é gravado na tabela "fases" o texto resultante deste processo para posteriormente ser utilizado pelo MySQL Fulltext.

4.2.3.2 Envio de texto puro

No envio de texto puro, o processo de tratamento do texto é semelhante ao de arquivo exceto na utilização de ferramenta externa. Ao enviar o texto, o CTP grava este texto na tabela "buscas" e aguarda o início da busca executada através da cron.

Quando é iniciada a busca pela cron, o CTP resgata o texto gravado na tabela "buscas" e quebra por partes que possuem aproximadamente 200 caracteres. Após quebrar por partes, o texto destas é gravado na tabela fases ainda na forma original. Ao finalizar o processo de gravação, o CTP inicia o tratamento do texto e vetorização de forma semelhante ao realizado no texto do arquivo PDF, como demonstra a Figura 34.

Figura 34: Estrutura (Texto puro).



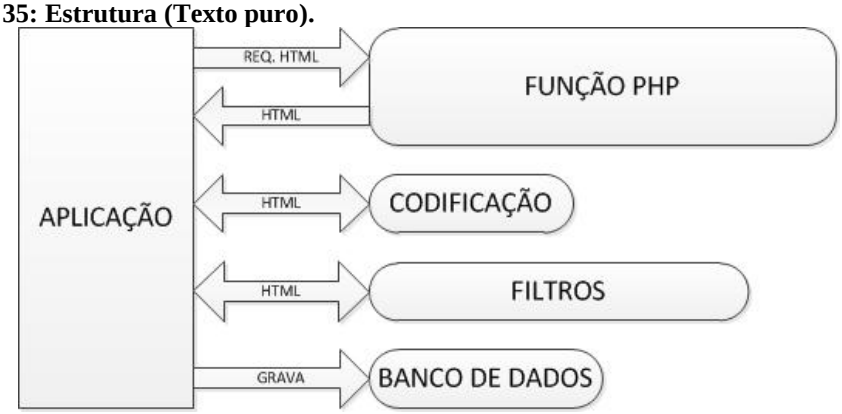
Fonte: Do autor.

Como pode-se observar na Figura 34, o texto passa por diversos filtros como tratamento de aspas, eliminação de caracteres não alfanuméricos, pontuação, stopwords e espaços concatenados. Após esta limpeza do texto é iniciada a vetorização do mesmo, passando pelo mesmo processo da vetorização descrito na subseção anterior. Com o texto vetorizado, é gravado na tabela “fases” para ser utilizado pelo MySQL Fultext posteriormente.

4.2.3.3 Envio de URL

A rotina efetuada quando o usuário envia um endereço WEB é semelhante ao de texto puro, porém possui um tratamento do texto antes da gravação na tabela buscas, diferentemente das outras duas formas, conforme Figura 35.

Figura 35: Estrutura (Texto puro).



Fonte: Do autor.

A Figura 35 demonstra que quando o usuário envia um endereço WEB, o CTP inicia o processo de busca do HTML através da função "file_get_contents" do PHP que resgata este HTML e o coloca na variável que armazena temporariamente o texto. Com posse do texto, é verificada a codificação utilizada e se for necessário, é feita a conversão para o formato ISO-8859-1[17], padrão do CTP. Além disso, o texto também passa por diversos filtros como a decodificação de elementos HTML utilizando as funções "htmlspecialchars_decode[18]" e "html_entity_decode[19]", eliminação dos scripts JavaScript e de estilo que estiverem presentes no texto, juntamente com a eliminação de elementos HTML e espaços concatenados. Por fim, é feito o tratamento de aspas e o texto é gravado na tabela buscas.

Ao iniciar a busca, o CTP irá tratar o texto resultante do tratamento descrito no parágrafo anterior da mesma forma que o texto puro. Com o tratamento do texto fornecido pelo usuário realizado, é possível iniciar a busca por documentos similares na Internet.

Nesta tela, além das opções de envio dos dados também é possível escolher entre três tipos de personalização da busca. Cada um destes possui características distintas e tempo de processamento.

[17] Codificação de caracteres também conhecida como latin1, é utilizada como padrão em diversos navegadores.
[18] Função do PHP de decodificação de caracteres HTML chamados especiais.
[19] Função do PHP que decodifica entidades HTML para os caracteres que o representam.

4.2.4 Tipos de buscas

O software CTP utiliza três padrões de buscas que utilizam diversas ferramentas. Estas buscas requisitam diferentes níveis de custo computacional, sendo que conforme este custo computacional aumenta o tempo que o CTP utilizará para finalizar a comparação também aumentará. Antes de enviar o texto, o usuário pode escolher entre estes níveis de buscas descritos a seguir.

4.2.4.1 Busca básica

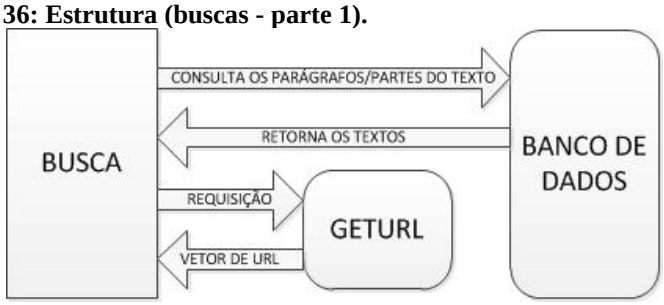
Na busca básica, o CTP realiza a comparação de similaridade utilizando MySQL Fulltext, que disponibiliza um resultado de similaridade com menos exatidão, dentre os níveis disponíveis. Neste nível, não é executado o processamento de vetorização dos documentos encontrados, melhorando o desempenho da busca em relação ao tempo de processamento.

O documento fornecido pelo usuário é definido como “Documento 1” e passa por uma limpeza de texto e vetorização como descrito nas subseções anteriores. A rotina em que o texto será submetido está relacionada com a forma que foi utilizada para enviá-lo.

Quando a CRONTAB inicia a busca, o CTP faz uma consulta para resgatar os parágrafos ou parte de texto que foram armazenados na tabela “fases”. Estes textos já estão vetorizados e desta forma é possível fazer as requisições as ferramentas de buscas *online* que são mais utilizadas atualmente.

Para demonstrar melhor o funcionamento do nível de busca básico, a Figura 36 demonstra a estrutura e ações que são realizadas pelo software até que sejam retornados os endereços dos documentos encontrados pela ferramenta de buscas. Este processo demonstrado na Figura 36 é idêntico nos três níveis e não somente na busca básica.

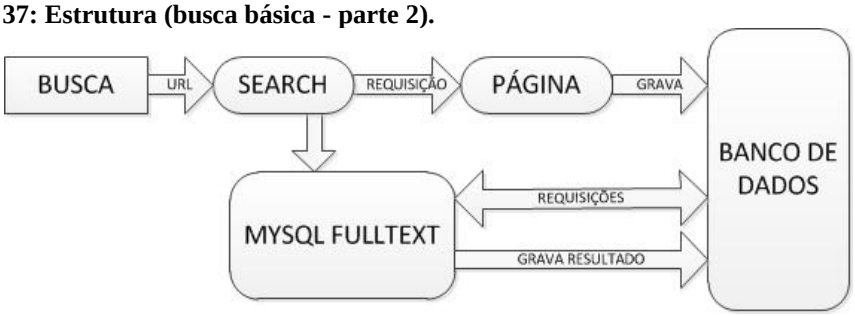
Figura 36: Estrutura (buscas - parte 1).



Fonte: Do autor.

Após o usuário enviar o texto, a CRONTAB inicia a busca conforme Figura 36 e a primeira ação a ser executada é a requisição de todos os parágrafos ou partes de textos do documento um que já estão vetorizados, ao banco de dados. Com os textos vetorizados é possível realizar as requisições de buscas que, neste nível de busca é executado na ferramenta Google. Esta ferramenta disponibiliza as URL referentes às buscas realizadas e a partir disso, o CTP extrai os endereços e armazena em um vetor. Estas buscas são realizadas a cada parágrafo ou parte de texto e após, os endereços são enviados a função "search" que faz as requisições do HTML ou arquivo PDF destes. A estrutura demonstrada nas Figuras 37 e 38 exemplifica o processo executado após a requisição destes endereços.

Figura 37: Estrutura (busca básica - parte 2).

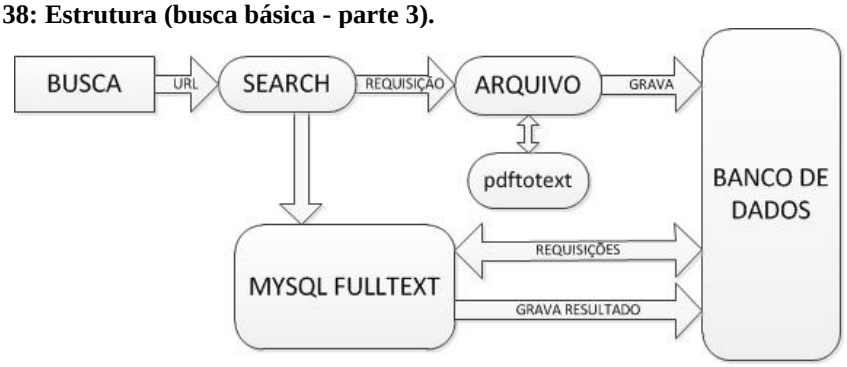


Fonte: Do autor.

De posse dos endereços o CTP faz a requisição do HTML ou arquivo através da função "search" e executa a função "pagina" quando a fonte for um HTML ou "arquivo" quando for arquivo PDF. Na Figura 37 está exposto o caminho que o CTP realiza quando o texto for proveniente de uma página HTML. A função "página" separa o texto por partes, executa filtros como o tratamento de aspas, retira espaços em branco, tags HTML entre outros e após grava no banco de dados o texto resultante.

A função “arquivo” é executada quando o endereço WEB for de um arquivo PDF e requisita através da função “pdftotext” o texto de cada parágrafo deste arquivo. Após, é gravado no banco de dados na tabela fases como demonstrado na Figura 38.

Figura 38: Estrutura (busca básica - parte 3).



Fonte: Do autor.

Com todos os dados no banco de dados, a função que é responsável por construir e executar a consulta MySQL Fulltext é executada. Esta função traz os parágrafos ou partes de texto que possuem a maior quantidade dos termos resultantes da vetorização destes. Este número é transformado em porcentagem referente à quantidade total dos termos e após é gravado no banco de dados na tabela de resultados.

4.2.4.2 Busca média

A busca média utiliza mais recursos para trazer o resultado de similaridade. Diferentemente da busca básica, esta necessita que os parágrafos ou partes de textos sejam vetorizados para que seja feita a comparação entre documentos. Além desta diferença, também altera a forma em que é feita a comparação, que neste caso é feita pelo método vetorização.

Para proceder a explicação do funcionamento deste tipo de busca pode-se utilizar a Figura 39 e 40 que demonstra o processo realizado após a requisição dos endereços. A primeira parte do processo até que o CTP possua os endereços WEB é idêntica ao executado na busca básica, como descrito anteriormente.

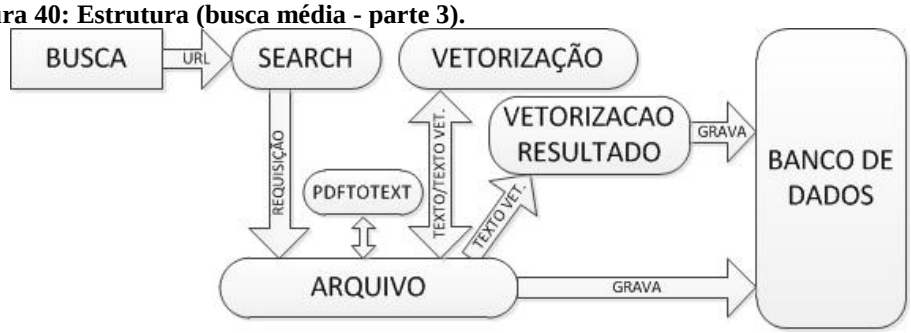
Figura 39: Estrutura (busca média - parte 2).

Fonte: Do autor.

possui a parte inicial idêntica a Figura 35, porém após a função “search” há diferenças no processo da busca média em relação a básica.

Como pode-se observar na Figura 39 e 40, as funções “arquivo” e “página” executam outras funções além do executado no tipo de busca básica. O caminho executado quando for texto HTML está demonstrado na Figura 39, já o caminho percorrido quando for um arquivo PDF está exibido na Figura 40. Na busca média estas funções fazem requisições à função vetorização e após gravam no banco de dados o resultado que retornou dessa função. Possuindo o texto vetorizado, é possível executar a função “vetorizaçãoResultado” que por sua vez irá realizar a comparação entre o documento atual e o fornecido pelo usuário, conforme o método de vetorização descrito nas seções anteriores. Esta comparação é feita com o parágrafo ou texto do documento fornecido pelo usuário denominado neste momento como uma fase, com o texto recuperado da Internet resultante da busca feita com os termos desta fase.

Figura 40: Estrutura (busca média - parte 3).



Fonte: Do autor.

O método vetorização fornece um resultado mais preciso na comparação de similaridade, devido ao fato do cálculo levar em consideração a importância do termo no texto. Outro fator que contribui para um resultado mais rico é a utilização de dois mecanismos de busca: o Google e o Bing, que desta forma trazem mais URLs distintas para a comparação.

O terceiro e último nível de busca é definida como completa, pois além de utilizar a mesma forma que a busca média, agrega mais elementos que aprimoram os resultados.

4.2.4.3 Busca completa

Este nível de busca traz o melhor resultado encontrado por este software, possui custo computacional maior em relação aos outros níveis e consequentemente necessita de uma fatia de tempo maior para apresentar os resultados da comparação de similaridade.

Este nível de busca utiliza juntamente com o nível médio a comparação de similaridade pelo método vetorização, porém com a utilização do método de radicalização e a tradução do texto fornecido pelo usuário. Após a função “search” é demonstrado na Figura 41 o processo realizado quando o texto for de uma página HTML. Já a Figura 42 exibe o processo realizado quando o endereço web for um documento PDF.

Figura 41: Estrutura (busca completa - parte 2).

Fonte: Do autor.

Como pode-se observar na Figura 42, as mudanças em relação a busca média ocorrem nas funções “arquivo” e “vetorizacaoResultado”. A função “arquivo” utiliza agora a tradução

do texto do documento fornecido pelo usuário, realizando as buscas por documentos similares nas linguagens português e inglês.

Figura 42: Estrutura (busca completa - parte 3).

Fonte: Do autor.

A função “pagina”, demonstrada na Figura 41 não possui a adição da função de tradução porque o texto enviado pelos modos texto puro e URL são traduzidos logo após o usuário enviar o texto, antes da gravação no banco de dados. O texto traduzido para inglês é adicionado ao texto em português.

Outra mudança significativa entre os níveis média e completa é a radicalização do texto que ocorre nas partes dos documentos que obtiverem maiores índices de similaridade. A radicalização busca melhorar este resultado tornando-o mais preciso e está disponível nas duas estruturas exibidas nas Figuras 41 e 42. Desta forma é feita a primeira comparação e se o nível de similaridade for maior que vinte por cento, os textos são submetidos à radicalização para após ser feita uma nova comparação.

As mudanças são ainda maiores, pois o nível completo utiliza também o buscador Yahoo, sendo que desta forma é resgatado um número maior de URLs para a requisição de documentos.

4.2.4.4 Ferramentas de busca e tradução

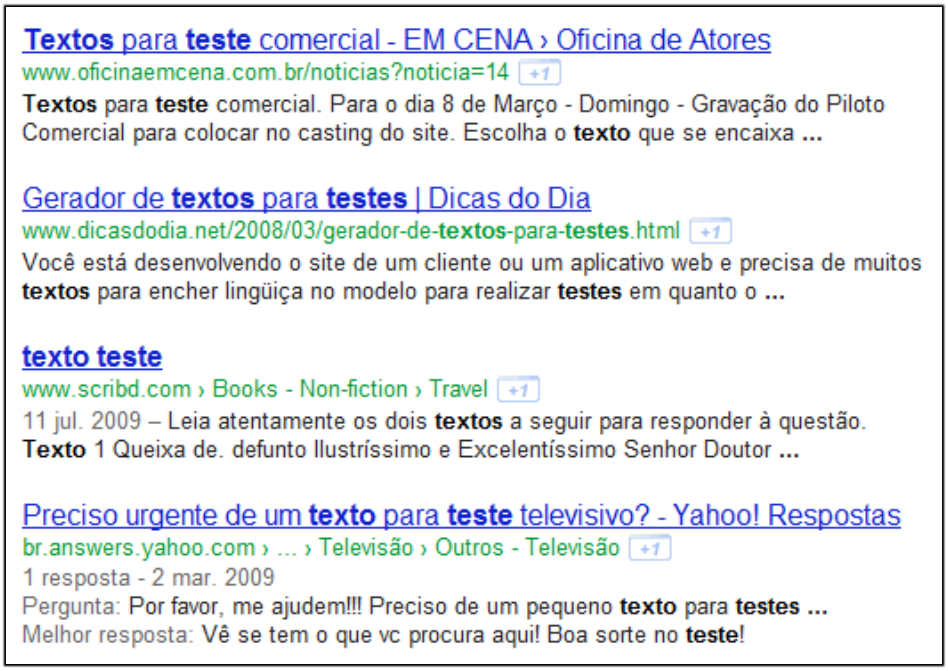
A ferramenta de busca do CTP proporciona o resgate de textos e documentos em qualquer parte do mundo, para isto são utilizadas três ferramentas pelos níveis de busca. Estes

buscadores possuem a maior parte do mercado de buscas *online*, resultado este devido a buscas mais eficientes, ferramentas agregadas e fácil utilização.

Uma destas ferramentas de busca é o Google, que foi fundada em 1998 pelos estudantes Larry Page e Sergey Brin da Universidade de Stanford, dois anos após ser desenvolvido o mecanismo de busca. Este mecanismo possui ótimos resultados porque o Google passa por toda web rastreando as páginas com rastreadores chamados de "Googlebots", além de utilizar vários parâmetros para classificar o conteúdo que é disponibilizado ao usuário (GOOGLE, 2011).

Com o texto ou documento disponibilizado pelo usuário já vetorizado, o CTP inicia a busca utilizando os buscadores. Ao fazer uma requisição à ferramenta de buscas Google, é apresentado os resultados conforme demonstra a Figura 38. Para este resultado demonstrado nesta figura, foi utilizado como texto de busca "texto teste".

Figura 43: Resultados buscador Google.



Fonte: Do autor.

O texto que está em azul na Figura 38 é a URL da página onde foram encontrados os dados que estão abaixo desta. Esta URL possui um texto como título e a seguir há o endereço da mesma na cor verde.

Outra ferramenta de buscas utilizada é o Bing que foi desenvolvida pela Microsoft, lançada em 28 de maio de 2009 é uma concorrência direta ao buscador do Google. Esta ferramenta possui foco em quatro áreas: decisão de compra, planejamento de viagens, saúde e negócios (MICROSOFT, 2011).

Como pode-se observar na Figura 39, os resultados que são exibidos ao usuário possuem forma semelhante ao buscador descrito anteriormente. Porém o HTML desta página possui características diferentes como a utilização de outros elementos. Para isto foi necessário fazer alterações na forma em que é retirado o endereço do documento destino e também adequar o código para descartar o conteúdo irrelevante.

Figura 44: Resultado buscador Bing.

Fonte: Do autor.

A terceira e última ferramenta utilizada é o buscador do Yahoo, companhia que possui diferentes soluções para atender necessidades dos usuários. Yahoo é uma empresa que possui anos de atuação e por muito tempo foi líder nas buscas *online*. Foi escolhida para compor as buscas por possuir bons resultados e uma grande base de páginas indexadas. A visualização retornada por esta ferramenta está disponibilizada na Figura 40.

Figura 45: Resultados buscador Yahoo.

Fonte: Do autor.

Como pode-se observar, os resultados das três ferramentas possuem uma visualização similar, porém em seu interior a construção do HTML é único e diferente em cada.

Para possibilitar a extração dos textos e documentos foi necessário isolar os *links* (endereços) disponibilizados pelo HTML dos buscadores. Estes endereços estão no atributo *href* da *tag* 'a' do HTML, a seguir está demonstrado como os três buscadores disponibilizam:

I – Google:

HTML - *<a class="l" onmousedown="return rwt(this,'','','1', 'AFQjCNEDiL7WKlheI-Nykr4aVw3N0sD8Mg', '','0CB0QFjAA')" href="http://www.oficinaemcena.com.br/noticias?noticia=14">*

Endereço - *http://www.oficinaemcena.com.br/noticias?noticia=14*

II – Bing:

HTML - <a onmousedown="return si_T('&ID=SERP,5044.1')" href="http://www.sato.adm.br/testes/teste_de_digitacao_ou_datilografia.htm">

Endereço - *http://www.sato.adm.br/testes/teste_de_digitacao_ou_datilografia.htm*

III – Yahoo:

HTML - *<a class="yschttl spt" data-bk="5064.1" href="http://br.wrs.yahoo.com/_ylt=A0geu8JSfMJOa2UA6WejIRh.;_ylu=X3oDMTByamlqaW9mBHNlYwNzcgRwb3MDMwRjb2xvA2FjMgR2dGlkAw--/SIG=14h31jn26/EXP=1321397458/**

*http%3a//educacao.centralblogs.com.br/post.php%3fhref=texto%2bnarrativo%2bteste%2bde%2bavaliacao%26KEYWORD=21799%26POST=3861922" dirtyhref="http://br.wrs.yahoo.com/_ylt=A0geu8JSfMJOa2UA6WejIRh.;_ylu=X3oDMTByamlqaW9mBHNlYwNzcgRwb3MDMwRjb2xvA2FjMgR2dGlkAw--/SIG=14h31jn26/EXP=1321397458/**http%3a//educacao.centralblogs.com.br/post.php%3fhref=texto%2bnarrativo%2bteste%2bde%2bavaliacao%26KEYWORD=21799%26POST=3861922">*

Endereço - *http://br.wrs.yahoo.com/_ylt=A0geu8JSfMJOa2UA6WejIRh.;_ylu=X3oDMTByamlqaW9mBHNlYwNzcgRwb3MDMwRjb2xvA2FjMgR2dGlkAw--/SIG=14h31jn26/EXP=1321397458/** http%3a//educacao.centralblogs.com.br/post.php%3fhref=texto%2bnarrativo%2bteste%2bde%2bavaliacao%26KEYWORD=21799%26POST=3861922*

O endereço que é extraído do HTML é o caminho para o acesso ao documento ou página, então a função "getUrl" do CTP consegue extrair o endereço exatamente como demonstrado nos exemplos anteriores.

Este endereço pode ser um documento ou uma página de Internet, no caso de um documento, o CTP extrai o texto de seu conteúdo interno através do "pdftotext" e após salva no banco de dados. Para páginas de Internet acontece da mesma forma, porém diferentemente do documento, não é necessário utilizar ferramenta externa para capturar HTML.

4.2.5 Apresentação dos resultados

Ainda na tela principal do software, onde são feitas as buscas, são exibidos os resultados parciais referentes às buscas realizadas. As buscas realizadas ficam listadas abaixo do painel de busca, onde é demonstrada a porcentagem de progresso e ao clicar na linha de cada busca é possível visualizar estes resultados.

A Figura 41 demonstra como estes resultados são exibidos. Como pode-se observar, no lado esquerdo está a porcentagem de similaridade entre os textos original e o recuperado na Internet, que possuem os textos expostos de forma resumida no centro. É possível expandir o texto clicando em "ver mais", isto fará com que o texto do parágrafo ou parte do texto seja exibido por completo.

Figura 46: Exibição dos resultados.

Fonte: Do autor.

Além das características descritas anteriormente, a Figura 41 demonstra que no resultado são exibidos as páginas e parágrafos dos textos envolvidos. Esta funcionalidade está presente somente quando o texto for proveniente de um arquivo PDF.

Na tela de buscas ainda é possível acessar a alteração de cadastro e a exibição completa dos resultados obtidos nas buscas.

4.2.6 Visualização completa dos resultados

A tela de visualização completa dos resultados demonstra as similaridades encontradas agrupadas por documento recuperado. Isto é muito importante porque o usuário pode visualizar como o documento que está sendo visualizado afetou na similaridade, além disso, também é possível visualizar o progresso da similaridade dentro do arquivo. A Figura 42 exibe a tela dos resultados de uma busca e as informações que o usuário poderá coletar.

Figura 47: Resultados completos.

Fonte: Do autor.

A Figura 42 demonstra o resultado de similaridade de dois documentos recuperados. No lado esquerdo é exibido um gráfico de linha que disponibiliza o progresso da similaridade entre o documento que o usuário disponibilizou e o que foi recuperado na Internet. Como pode-se observar no gráfico do documento superior, o usuário pode posicionar o mouse em cima da linha do gráfico e no mesmo momento são exibidas algumas informações. Estas informações são a similaridade naquele ponto e os números de página e parágrafo.

O endereço do documento encontra-se no lado direito do gráfico exibido e clicando sobre este, o usuário é direcionado a página de onde foi recuperado. Além das informações citadas anteriormente, esta tela de resultados também exibe os textos de parágrafos ou partes do texto que obtiveram mais de 20% de similaridade com o texto disponibilizado pelo usuário. Estes resultados são ordenados conforme o número do parágrafo e são exibidos os textos com termos similares grifados em negrito. Além disso, também é disponibilizado o número em porcentagem da similaridade encontrada.

Nesta tela ainda há as informações de data e hora em que o usuário iniciou a busca, o nível que foi selecionado e também há o endereço para acesso ao documento enviado pelo usuário.

4.2.7 Hardware utilizado

Para possibilitar a utilização do software por qualquer pessoa foi necessário buscar um serviço que provesse as ferramentas necessárias para a implantação do CTP. Inicialmente o

serviço contratado chamava-se VPS1 (Virtual Private Server) da empresa HostGator que possuía as características descritas na Tabela 8.

Tabela 8: Configuração VPS1.

Sistema operacional	**Linux CentOS**
Painel de configuração	**Virtuozzo Power Panel**
CPU	**0,56 GHZ**
Memória RAM	**384 MB**
Espaço em disco	**10 GB**
Largura de banda	**250 GB**
Custo	**R$ 69,99**

Fonte: Do autor.

Além das configurações descritas na tabela 8, há diversas ferramentas que são disponibilizadas neste serviço. Este serviço é uma máquina virtual[20] alocada em servidores que estão estabelecidos nos Estados Unidos da América e possuem ping[21] médio de 200ms.

Com a utilização do serviço VPS1 notou-se que não seria suficiente para a implantação do software. Desta forma foi feita a transferência diretamente para o plano VPS3 que possui as características descritas na tabela 9.

Tabela 9: Configuração VPS3.

Sistema operacional	**Linux CentOS**
Painel de configuração	**cPanel & WHM**
CPU	**1,13 GHZ**

[20] Sistema operacional executado através de um software que possui as mesmas características de um sistema operacional instalado no sistema.
[21] Ferramenta presente nos Sistemas operacionais que mede a latência entre o dispositivo local e o destino especificado.

Memória RAM	**768 MB**
Espaço em disco	**30 GB**
Largura de banda	**500 GB**
Custo	**R$ 139,99**

Fonte: Do autor.

O serviço VPS3 contempla todas as ferramentas necessárias para a implantação do CTP, além de possuir um desempenho superior ao outro plano. Este serviço está muito além do que é disponibilizado em hospedagens de sites, porque é possível utilizar diversas ferramentas de controle, configuração e estatísticas que auxiliam na tomada de decisão e planejamento.

Não foi possível contratar um serviço de hospedagem simples de sites porque não permite acesso ssh[22] ao sistema e desta forma não possibilita a instalação e utilização de ferramentas como XPDF e CURL. Outro fator importante é a não concorrência com softwares de terceiros implantados no mesmo hardware, o que poderia ocasionar uma perda de desempenho e instabilidade do software.

4.3 Testes da ferramenta

Para validar o software desenvolvido foi necessário fazer alguns testes que demonstrassem as características dos níveis de busca e também o desempenho destas. Desta forma foi utilizado um documento PDF que possuía as características demonstradas na tabela 10.

[22] Software e protocolo de rede que permite a utilização de outro computador de forma remota.

Tabela 10: Características do documento (1 página) de testes.

Páginas	**1**
Parágrafos	**7**
Palavras	**378**
Caracteres	**1.976**

Fonte: Do autor.

Com posse do documento foi possível iniciar os testes. Primeiramente foi utilizado o nível básico sendo que a forma de envio foi por arquivo PDF. Este teste consumiu 32 segundos de tempo para ser finalizado, retornando as similaridades entre as partes de textos do documento enviado e os documentos recuperados na Internet. Nesta busca foram recuperados 15 documentos em um total de 11 endereços WEB que resultaram em 11 resultados de similaridade acima de 20% entre os parágrafos dos textos. Estes resultados se concentram em três parágrafos como demonstrado na Tabela 11.

Tabela 11: Resultados pesquisa básica (documento de 1 página).

Parágrafo	Similaridade
2	**80%**
3	**80%**
3	**80%**
3	**80%**
3	**60%**
1	**51%**
1	**43%**
3	**40%**

1	**36%**
2	**33%**

Fonte: Do autor.

Como pode-se observar na Tabela 11, o parágrafo 3 foi o que mais possuiu similaridades encontradas, sendo que muitos destes são de comparações com fontes diferentes.

Para demonstrar a diferença de desempenho quando as fontes já estão indexadas pelo software, realizou-se novamente a mesma busca com o mesmo arquivo PDF. Nesta nova busca os resultados obtidos foram praticamente os mesmos, divergindo apenas em um resultado a mais que obteve 90% de similaridade da nova fonte com o parágrafo 1. Esta divergência aconteceu porque a ferramenta de buscas disponibilizou um novo endereço. Porém, a grande diferença desta nova consulta está no tempo que reduziu em quase 70% marcando 10 segundos de busca após iniciado pela CRONTAB.

Outro teste realizado ocorreu utilizando o nível de busca média com o mesmo arquivo que foi enviado na pesquisa básica. Neste nível foram recuperados mais documentos que o teste anterior. Nesse nível foi feita a vetorização dos dados, conforme seção 4.24.2.

O teste de busca no nível médio resultou em 38 documentos recuperados, 28 endereços WEB visitados e com 10 registros de similaridade maiores de 20% entre os parágrafos do texto enviado e os recuperados na Internet. Estes dados levaram 65 segundos para serem gerados e as similaridades encontradas são demonstradas na Tabela 12.

Tabela 12: Resultados pesquisa média (documento de 1 página).

Parágrafo	Similaridade
4	**88%**
2	**76%**
3	**71%**
3	**71%**

1	**58%**
3	**37%**
3	**25%**
2	**23%**
2	**22%**
2	**21%**

Fonte: Do autor.

Este nível trouxe a mesma quantidade de resultados, porém foram recuperados 17 textos a mais que o nível básico.

Ainda no nível médio foi feita uma segunda busca com o mesmo arquivo, demonstrando a diferença do desempenho quando o CTP já possui as páginas indexadas. Esta nova busca trouxe os mesmos resultados, porém levou apenas 17 segundos para concluir a análise. Comparando-se com a busca básica, o aumento de desempenho foi ainda maior na busca média, chegando a 74% de redução no tempo.

O terceiro nível chamado completo utilizou a radicalização para melhorar os resultados encontrados, isto fica bem visível no teste que foi feito com o documento PDF utilizado nos outros níveis. Este nível levou mais tempo para concluir a comparação, foram 6 minutos e 40 segundos. Porém foram recuperados 136 documentos, 92 páginas foram visitadas e foram encontradas 80 similaridades acima de 20% entre os parágrafos do texto enviado com os que foram encontrados na Internet.

A Tabela 13 demonstra os parágrafos e a similaridade mais alta encontrada na comparação.

Tabela 13: Resultados pesquisa completa (documento de 1 página).

Parágrafo	Similaridade
2	**100%**

2	**95%**
4	**95%**
4	**88%**
3	**77%**
2	**72%**
3	**71%**
3	**71%**
3	**71%**
1	**58%**

Fonte: Do autor.

Neste nível também foi realizado novamente a busca, que conforme testes anteriores também obteve uma melhora significativa no desempenho. O tempo de busca reduziu em mais de 90%, concluindo em 40 segundos.

Neste primeiro documento pode-se observar vários casos de alta similaridade que se concentraram principalmente nos quatro primeiros parágrafos. Os resultados dos três testes descritos anteriormente trouxeram endereços dos documentos onde foram encontradas similaridades. Estes endereços que possuíam maior similaridade permaneceram os mesmos nos três níveis de busca.

Na Figura 43 está sendo apresentada parte da tela dos resultados da busca completa, onde pode-se observar que através do mesmo parágrafo do texto enviado pelo usuário foram encontradas duas fontes diferentes com o mesmo conteúdo. Os textos deste resultado encontram-se expandidos para visualização completa e em negrito os termos que coincidem com o texto enviado. Este texto que o usuário enviou chama-se “texto de teste 1.pdf”, porém como demonstrado na Figura 43 o nome do arquivo no sistema é “texto de teste 123062012020902.pdf”, isto acontece porque o sistema renomeia o arquivo para que não haja sobreposição futura.

Figura 48: Resultados da busca completa.

Fonte: Do autor.

Como pode-se observar na Figura 43, foram encontrados 2 resultados que obtiveram mais de 90% de similaridade. O endereço do registro superior possui um arquivo PDF, já o endereço inferior possui somente página HTML. Nos resultados das buscas realizadas no documento “texto de teste 1.pdf” pode-se identificar um caso de plágio que segundo Abreu (2011) é classificado como plágio *Structural Plagiarism*. Este tipo de plágio acontece quando o usuário utiliza trechos de outras obras como paráfrase sem identificar o autor. Também ocorreu neste documento, especificamente no parágrafo 2 o Plágio Direto que segundo Kirkpatrick (2007) é caracterizado como cópia idêntica.

Para verificar como o sistema se comporta em arquivos maiores foram feitos outros testes com arquivos de 5 páginas, 10 páginas e 42 páginas. O arquivo de 5 páginas possui as características listadas na Tabela 14.

Tabela 14: Características do documento (5 páginas) de testes.

Páginas	**5**
Parágrafos	**17**
Palavras	**1.041**
Caracteres	**5.547**

Fonte: Do autor.

Este texto produzido em grupo foi fornecido por uma aluna de um curso de técnico em informática, que desconhecia estes plágios cometidos pelos colegas.

O documento de 5 páginas encontra-se em anexo nesta monografia, excluindo a página de identificação, foi produzido em grupo e fornecido por uma aluna de um curso de técnico em informática, que desconhecia estes plágios cometidos pelos colegas. Este arquivo obteve 43 documentos na busca de nível básica, além de 27 endereços de documentos recuperados e 40 registros de similaridade acima de 20% nos parágrafos. Esta busca estendeu-se por 6 minutos e 30 segundos. Os resultados encontrados encontram-se na Tabela 15.

Tabela 15: Resultados pesquisa básica (documento de 5 páginas).

Página	Parágrafo	Similaridade
3	**3**	**91%**
2	**4**	**90%**
2	**3**	**90%**
2	**3**	**90%**
4	**2**	**85%**
4	**2**	**85%**
2	**4**	**84%**

2	**4**	**84%**
2	**3**	**80%**
2	**4**	**78%**

Fonte: Do autor.

Na busca de nível médio deste mesmo documento o sistema resgatou 108 documentos, visitou 63 endereços WEB, obteve 47 registros de similaridade e utilizou 8 minutos e 44 segundos para a comparação. Os resultados obtidos estão exibidos na tabela 16.

Tabela 16: Resultados pesquisa média (documento de 5 páginas).

Página	Parágrafo	Similaridade
3	**3**	**98%**
4	**2**	**89%**
2	**4**	**83%**
2	**3**	**78%**
2	**3**	**69%**
4	**1**	**69%**
3	**5**	**62%**
2	**3**	**61%**
2	**3**	**61%**
3	**4**	**60%**

Fonte: Do autor.

O nível de busca completo utilizado na terceira comparação deste documento necessitou 83% a mais de tempo, totalizando 16 minutos e 3 segundos. Este tempo é necessário porque o documento dobra de tamanho quando submetido a tradução, sendo que é utilizada na busca a forma original e a traduzida. Esta busca resultou em 185 endereços WEB visitados, 268 documentos recuperados e 202 resultados de similaridade acima de 20%. Estes resultados estão descritos na tabela 17.

Tabela 17: Resultados pesquisa completa (documento de 5 páginas).

Página	Parágrafo	Similaridade
3	**3**	**98%**
4	**1**	**91%**
4	**2**	**89%**
2	**4**	**83%**
3	**4**	**79%**
2	**3**	**78%**
3	**4**	**71%**
2	**3**	**69%**
4	**1**	**69%**
4	**4**	**68%**

Fonte: Do autor.

O terceiro e quarto documento foram utilizados para ter uma noção de tempo que o CTP necessita para concluir a busca, desta forma não serão demonstrados os detalhes destes documentos e também os detalhes das buscas.

O terceiro documento utilizado possui 10 páginas que foram retiradas de um documento armazenado em um endereço WEB. Este documento trouxe grande similaridade

justamente porque o CTP encontrou o mesmo endereço de onde foi retirado. O tempo consumido na busca do nível básico foi de 10 minutos e 19 segundos, no nível médio o tempo consumido foi de 15minutos e 26 segundos e no nível completo o tempo decorrido na busca foi de 46 minutos e 45 segundos.

O documento de 10 páginas requisitou ao sistema um tempo maior de busca devido à quantidade de páginas, parágrafos e termos que este contém.

Para continuar os testes com um documento maior, foi selecionado um Trabalho de conclusão de curso contendo 42 páginas ao todo. Na busca básica, o sistema necessitou de 3 horas e 4 minutos e 9 segundos para concluir. Na busca média o sistema necessitou de 4 horas e 8 minutos e 14 segundos para concluir a busca. Na busca completa o sistema utilizou 11 horas e 45 minutos e 15 segundos para finalizar as comparações. Este documento que foi utilizado está armazenado digitalmente em um endereço WEB que foi inúmeras vezes encontrado pelo CTP nos testes realizados. Estas comparações demonstraram 100% de similaridade.

Após realizados os testes pode-se observar os resultados de similaridades que foram encontrados. Estes resultados demonstraram que o CTP possui precisão em buscas de nível médio e completo na comparação, principalmente em parágrafos ou partes de texto que resultem diversos termos na vetorização. Devido à baixa precisão do CTP quando havia menos de 4 termos vetorizados, foi necessário adequar a vetorização para aceitar termos com frequências menores quando haviam menos de 4 selecionados. Além desta adequação, também foi adicionado um filtro que não permite a execução da busca de um parágrafo ou parte de texto que possua menos de 4 termos no resultado da vetorização.

Observou-se que o método MySQL Fulltext traz bons resultados na comparação de documentos, porém o número da similaridade não possui exatidão, pois há uma variação ocasionada pelo fato de que o método não leva em consideração a frequência dos termos.

Os documentos selecionados tornaram possível verificar diversas situações como a recuperação de textos de páginas HTML e documentos PDF que possuíam imagens, códigos, gráficos, tabelas entre outros conteúdos. Nas diversas situações verificou-se que a comparação de similaridade do CTP corresponde com a similaridade dos documentos.

5 CONCLUSÃO

Esta monografia apresentou a implementação de uma ferramenta para a detecção de similaridade entre documentos. Através dessa ferramenta foi possível validar os métodos implementados e realizar os testes necessários.

Na utilização dos métodos, pode-se observar que o MySQL Fulltext é uma excelente ferramenta para busca de similaridade quando não deseja-se um resultado extremamente preciso. O grande trunfo desta função é sua velocidade e forma de utilização, podendo-se aplicar em diversas situações. Esta função é muito utilizada atualmente por mecanismos de buscas e em direcionamento de publicidade.

O método de vetorização de documentos trouxe bons resultados na comparação de similaridade com bom aproveitamento do processamento. Este método elimina parte do texto que não é significativo para o resultado da comparação e isto reduz consideravelmente a requisição de processamento para concluir as tarefas.

Ao contrário dos outros dois métodos utilizados, o método de radicalização requisita uma quantidade maior de processamento. Desta forma decidiu-se não utilizá-lo juntamente com a vetorização dos documentos encontrados na internet, devido ao consumo do processamento. Isto resultou em utilizá-lo como complemento do método de vetorização, melhorando o resultado da comparação de similaridade, sendo utilizado somente quando o método de vetorização encontrasse algum indício de similaridade na comparação.

Com os métodos e filtros utilizados, o software desenvolvido nesta monografia trouxe bons resultados de similaridade demonstrando que atende aos requisitos propostos. Através dos métodos implementados o sistema possibilita encontrar os tipos de plágio citados por Kirkpatrick (2007) como o plágio direto e plágio mosaico. Além destes tipos citados, o sistema também é capaz de encontrar na busca “completa” o tipo de plágio Translations que segundo Abreu (2011) é uma tradução do texto original.

A deficiência no software está na utilização de arquivos maiores no nível de busca “completo” devido ao fato da tradução não ter um tempo de resposta rápido, cerca de 7 segundos por requisição, e de não possuir as páginas e arquivos indexados. Verificou-se através dos testes realizados que cerca de 90% do tempo decorrido em uma busca está no download do HTML ou arquivo, na limpeza do texto recuperado e na sua vetorização. Outra limitação encontrada é a velocidade disponibilizada para download de arquivos ou HTML nos

endereços encontrados. Muitas vezes esta velocidade está limitada em 30 ou 40 kb/s, que em um documento de 2,5MB levaria cerca de 1 a 2 minutos para concluir o *download*.

Apesar das limitações do software, foram realizados testes utilizando diversos documentos e textos. Foi possível verificar que o software desenvolvido e os métodos nele implementados atendem a proposta inicial desta monografia. Além dos métodos, os filtros têm um papel fundamental nos resultados apresentados, além de proporcionar segurança e estabilidade do software.

5.1 Trabalhos futuros

Verificou-se que é possível implementar outros métodos para a comparação de similaridade que possibilitem utilizar outras funções disponibilizadas pelas ferramentas utilizadas. Estes outros métodos, além da melhora no desempenho e correções de limitações do sistema são metas a serem utilizadas em um trabalho futuro.

Uma das limitações que pode ser sanada em um trabalho futuro é a diversificação dos arquivos enviados pelos usuários. Atualmente o CopyToPaste aceita somente arquivos no formato PDF e isto restringe a utilização de demais arquivos como os “DOC” e “DOCX” que são muito utilizados atualmente.

Além da implementação citada anteriormente, também podem ser desenvolvidas novas formas de personalização da busca. Estas personalizações podem englobar a utilização de busca por conteúdo exato, exclusão de termos e domínios na busca.

Por fim, para melhorar consideravelmente o desempenho do software pode-se criar um robô podendo utilizar as funções desenvolvidas para indexar páginas e documentos da internet. Além da indexação seria necessário aumentar o desempenho do hardware podendo ainda utilizar o processamento paralelo.

REFERÊNCIAS

ABREU, R. M. **Proposta de Arquitetura para Um Sistema de Detecção de Plágio Multi-Algoritmo**, Rio de Janeiro, 2011, 105 p., Disponível em: <www.cos.ufrj.br/uploadfiles/1318599146.pdf> Acessado em: novembro de 2011.

BARBOSA, D. B. **Uma Introdução a Propriedade Intelectual**. Rio de Janeiro: Ed. Lumen Juris, 2003. p. 8-139. ISBN 85-7387-370-1.

BASSO, M. **O Direito Internacional da Propriedade Intelectual**. Porto Alegre: Ed. Livraria do Advogado, 2000. p. 19-39. ISBN 85-7348-152-8.

BITTAR, C. A. **Contornos Atuais do Direito do Autor**. São Paulo: Ed. Revista dos Tribunais LTDA., 1999. p. 19-48. ISBN 85-203-1686-7.

CABRAL, M. L. À mão ou por computador. In: CABRAL, M. L. **Bibliotecas: Acesso, Sempre**. Lisboa: Ed. Colibri, 1996. p. 53-65. ISBN 972-8288-16-6.

CARDOSO, G. **A Mídia na Sociedade em Rede**, Rio de Janeiro: Editora FVG, 2007, 15-146, 135-242. ISBN 978-85-225-0620-0.

CARR, J. **Inside Cyber Warfare**: Mapping the Cyber Underworld, Estados Unidos da América: O´Reilly Media, 2012, p. 141 – 143, ISBN 978-1-449-31004-2.

COSTA, R. L. C. **SQL**: Guia Prático, São Paulo: Brasport Livros e Multimídia Ltda, 2007, p. 1-5 , ISBN 978-85-7452-295-1.

COTTA, A. G. **O Palimpsesto de Aristarco**: considerações sobre plágio, originalidade e informação na musicologia histórica brasileira. Belo Horizonte, 1999, v4.0, n.2 p. 185-209, Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/567> Acessado em: novembro de 2011.

DEITEL, D. N. M. **PERL**: como programar, Porto Alegre: Artmed Editora S. A., 2001, p50-51, ISBN 0-13-028418-1.

DUARTE, F. R. **Identificação de Reuso em Documentos Digitais**, Rio de Janeiro, 2011, 131p., Disponível em: <http://www.cos.ufrj.br/uploadfiles/1310057262.pdf>, Acessado em: novembro de 2011.

FLANAGAN, D. **JavaScript**: The Definitive Guide, Estados Unidos da América: O´Reilly Media, 2006, p.19-28. ISBN 0-596-00048-0.

GANDELMAN, H. **De Gutenberg à Internet**: Direitos autorais na era digital. Rio de Janeiro: Ed. Record, 2001. p. 28-91. ISBN 85-01-04877-1.

Apache Software Fundation, **Apache Module mod_rewrite**, 2012, Disponível em: <http://httpd.apache.org/docs/2.0/mod/mod_rewrite.html>

GOOGLE, **Companhia**, *2011, Disponível em:* <http://www.google.com.br/intl/pt-BR/about/corporate/company/>, acessado em 11/2011.

GOOGLE, **Google Charts**, 2012, Disponível em: <https://developers.google.com/chart/>, acessado em 05/2012.

HOLDENER III, A.T. **AJAX**: The Definitive Guide, Estados Unidos da América: O´Reilly Media, 2008, p. 3-10. ISBN 978-0-596-52838-6.

HOLZNER, S. **PHP**: The complete reference, Estados Unidos da América: The McGraw-Hill Companies, 2008, p. 1 – 7. ISBN 0-07-150854-6.

jQuery Community Experts, **jQuery CookBook**, Estados Unidos da América: O´Reilly Media, 2010, p.1-7 . ISBN 978-0-596-15977-1.

KEN, N. The Apache Modules Book, Estados Unidos da América: Pearson Education, Inc., 2007, p.1-6. ISBN 0-13-240967-4.

KIRKPATRICK K. ***Avoiding Plagiarism***, Estados Unidos, 2001, 5p., Disponível em: <http://mundobr.pro.br/uneal/wp-content/uploads/2010/04/Evitando_o_plagio.pdf> Acessado em: novembro de 2011.

LAURENT, A. M. ST. **Understanding Open Source And Free Software Licensing**, Estados Unidos da América: O´Reilly Media, 2004, p. 1 -113. ISBN 978-0-596-00581-8.

LEVENSHTEIN, V. I. ***Binary Codes Capable of Correcting Deletions, Insertions, and Reversals***, União Soviética, 1966, 4., Disponível em: <http://www.freearchive.org/o/964e741877a3ffa8f2195ee253597fbaeed69a207e77df150439802fd370b2f8> Acessado em: novembro de 2011.

LUHN, H. P. ***The Automatic Creation of Literature Abstracts***, 1958, 7 p., Disponível em: <http://www.di.ubi.pt/~jpaulo/competence/general/(1958)Luhn.pdf> Acessado em: novembro de 2011.

MATSUBARA, E. T.; MARTINS C. A.; MONARD M. C. **PreTexT**: Uma ferramenta para pré-processamento de textos utilizando a abordagem *bag-of-words*, São Paulo, 2003, 57 p., Disponível em: <http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_209.pdf> Acessado em: novembro de 2011.

MEYER, E.A. **Cascading Style Sheets**: The Definitive Guide, Estados Unidos da América: O´Reilly Media, 2004, p.1-7. ISBN 0-596-00525-3.

MICROSOFT, **Companhia**, 2011, Disponível em: <http://www.microsoft.com/latam/presspass/brasil/2009/maio/bing.mspx>, acessado em 11/2011.

MORAES, R. **O Plágio na Pesquisa Acadêmica**: a proliferação da desonestidade intelectual, Bahia, 2007, 19p., disponível em: <http://faculdadesocial.edu.br/dialogospossiveis/artigos/4/06.pdf> Acessado em: novembro de 2011.

MUSCIANO C., KENNEDY B. **HTML & XHTML**: The Definitive Guide, Estados Unidos da América: O´Reilly Media, 2007, p1-12, ISBN-13: 978-0-596-52732-7.

MYSQL, **MySQL Reference Manual**, disponível em <http://dev.mysql.com/doc/refman/5.0/en/>, acessado em junho de 2012.

OLIVEIRA E.; OLIVEIRA M.; PEREIRA F.; CIARELLI M. P.; CARDOSO B; WALLACE F. H.; VERONESE L. **Bibliotecas Digitais Aliadas na Detecção Automática de Plágio**, Espírito Santo, 2007, 15p. Disponível em: <http://www.informarcia.pro.br/disciplinas/extras/plagio.pdf> Acesso em: outubro de 2011.

OLIVEIRA, M. G.; OLIVEIRA E. **Uma Metodologia para Detecção Automática de Plágios em Ambientes de Educação a Distância**, Espírito Santo, 2008, 9 p. Disponível em: < http://200.169.53.89/download/CD%20congressos/2008/V%20ESUD/trabs/t38670.pdf>. Acessado em: novembro de 2011.

PAESANI, M. L. **Direito de informática**: Comercialização e desenvolvimento internacional do *Software*. São Paulo: Ed. Atlas S.A., 2009. p. 5-20. ISBN 978-85-224-4827-2.

PARANAGUÁ, P.; BRANCO, S. **Direitos Autorais**, Rio de Janeiro: Editora FVG, 2009, p. 124 – 138. ISBN 978-85-225-0743-6.

PHP, **Documentação PHP**, disponível em <http://php.net>, acessada em junho de 2012.

POWERS, D. **PHP Solutions**: Dynamic Web Design Made Easy, Estados Unidos da América: Apress and friends of ED books, 2010, p. 1 - 4. ISBN – 13: 978-1-4302-3250-6.

QUEIROZ, R. **Propriedade Intelectual Digital e o Conceito de Uso Razoável**, 2009, Disponível em <http://www.ibdi.org.br/site/artigos.php?id=224>. Acessado em maio de 2012.

SANTOS, M. **Direito Autoral na Era Digital**: Impactos, controvérsias e possíveis soluções. São Paulo: Ed. Saraiva, 2009. p. 1-123. ISBN 978-85-02-08123-9.

SILVA, O. S. F. **Entre o Plágio e a Autoria**: qual o papel da universidade?, Bahia, 2008, 14p. Disponível em: < http://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=27503812>. Acessado em: novembro de 2011.

TAHAGHOGHI A. M. M.; Williams H. E. **Learning Mysql**, Estados Unidos da América: O´Reilly Media, 2007, p.3-8. ISBN 978-0-596-00864-2.

VALMORBIDA, W. **Análise e Implementação de Um Sistema Integrado de Busca Baseado nos Padrões de Metadados e Protocolos de Interoperabilidade Utilizados por Catálogos On-Line de Bibliotecas e Repositórios Digitais**, Rio Grande do Sul, 2011, p.22, Disponível em <http://www.univates.br/bdu/bitstream/10737/249/1/WillianValmorbida.pdf>. Acessado em maio de 2012.

VARELLA, A. N. ***Coopractice*** – Comunidades de Prática Virtuais Apoiadas por Ontologias, Rio de Janeiro, 2007, 136 p., Disponível em: <http://teses.ufrj.br/COPPE_M/AmandaNascimentoVarella.pdf> Acessado em: novembro de 2011.

W3C, **XHTML**, 2012, Disponível em: http://www.w3c.org/TR/xhtml1, acessado em maio de 2012.

6 ANEXO A – DOCUMENTO FORNECIDO PARA ANÁLISES

Repetidores

Os repetidores são utilizados, geralmente, para a interligação de duas ou mais redes idênticas. Atuando no nível físico, os repetidores simplesmente recebem todos os pacotes de cada uma das redes que interligam e os repetem nas demais redes sem realizar qualquer tipo de tratamento sobre os mesmos. Vários pontos são dignos de nota na utilização de repetidores para interconexão de redes locais.

Primeiramente, em redes em anel onde a estação é responsável pela retirada dos próprios quadros, caberá ao repetidor a retirada dos quadros nas redes em que atua como retransmissor. Em anéis onde cabe à estação de destino a retirada dos quadros, a situação se complica. Como pode haver mais de um repetidor em uma rede, o repetidor não pode agir como uma estação de destino intermediária e retirar o quadro do anel. A solução seria deixar tal tarefa para a estação monitora, o que diminui desempenho da rede.

Um segundo ponto vem da utilização de repetidores em redes que utilizam protocolos baseados em contenção. Nesse caso caberá ao repetidor também a função de detecção de colisão e retransmissão. Em redes que se utilizam de alguns protocolos, ao se calcular o tamanho mínimo do pacote, deve se levar em conta o retardo introduzido pelo repetidor. Isto vai limitar o número de repetidores em série em tais redes.

Um terceiro ponto vem da observação de que nada impede que tenhamos vários repetidores em uma mesma rede ou vários repetidores no caminho de um quadro desde a estação de origem até a estação de destino. Para isso alguns cuidados devem ser tomados. Não pode haver um caminho fechado entre dois repetidores quaisquer da rede, por isso implicará duplicações infinitas de quadros (um quadro repetido retornaria, devido a repetições em outros repetidores, voltaria a ser repetido, tornaria a retornar e assim indefinidamente), além de provocar outros efeitos colaterais, como por exemplo, a colisão dos quadros em redes baseadas em contenção, o que causa uma consequente diminuição do desempenho.

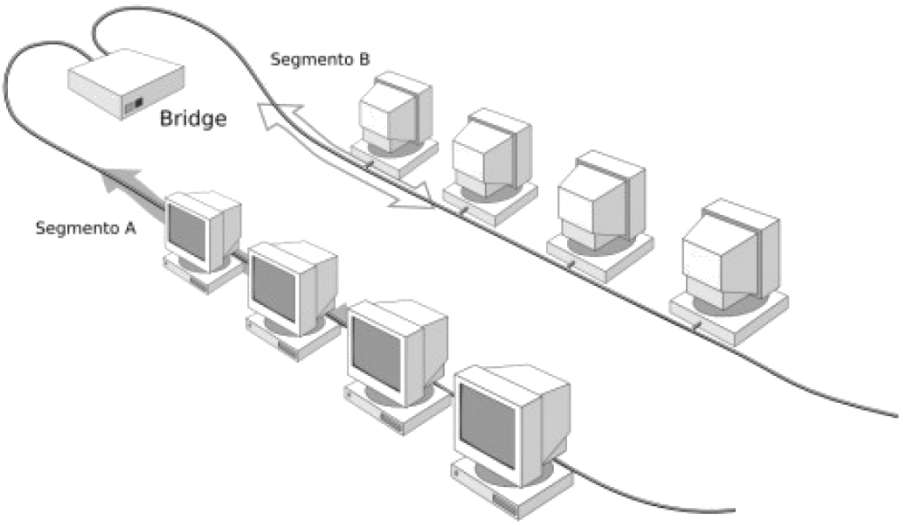
Outro ponto observado é que em protocolos onde reconhecimento do quadro é realizado automaticamente nos próprios quadros transmitidos, essa característica é perdida, pois existem dois motivos pelos quais não pode ser realizada pelos repetidores. Primeiro pelo fato de poder haver vários repetidores na rede. Nesse caso, a qual deles caberá a tarefa?

Segundo, mesmo que se pudesse decidir qual o repetidor teria a tarefa, como ele poderia saber da situação do quadro na estação de destino uma vez que ainda nem o retransmitiu?

Essa característica de alguns protocolos é irremediavelmente perdida. Ainda outro ponto a respeito dos repetidores deve ser mencionado, este ligado diretamente ao desempenho. Ao repetir todas as mensagens que recebe, um tráfego extra inútil é gerado pelo repetidor quando os pacotes repetidos não se destinam às redes que interligam. Uma solução para tal problema vem com a utilização de estações especiais denominadas pontes (bridges).

Pontes

As chamadas pontes, conhecidas também por bridges, permitem interligar dois segmentos de rede, de forma que eles passem a formar uma única rede. Em redes antigas, onde era utilizado um único cabo coaxial ou um hub burro, o uso de bridges permitia dividir a rede em segmentos menores, reduzindo o volume de colisões e melhorando o desempenho da rede. O bridge trabalha no nível 2 do modelo OSI, verificando os endereços MAC de origem e de destino dos frames e encaminhando apenas os frames necessários de um segmento ao outro. Outra vantagem é que a rede passa a comportar duas transmissões simultâneas, uma envolvendo micros do segmento A e outra com o segmento B:



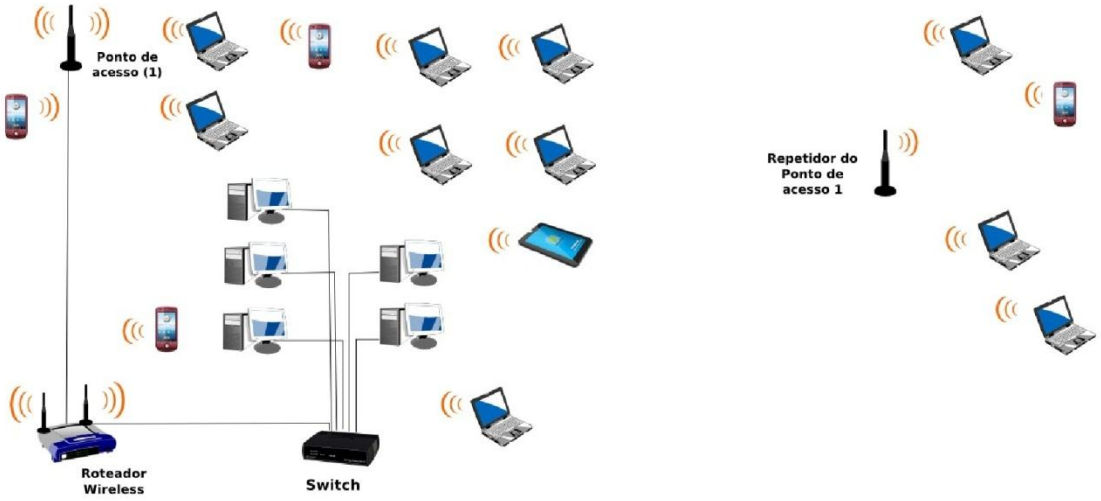
Hoje em dia não faz sentido usar bridges para dividir a rede em segmentos pois os switches desempenham com eficiência essa função, criando segmentos individuais para cada micro, o que praticamente elimina o problema das colisões, mas eles foram muito utilizados na época dos hubs burros.

Outra utilidade dos bridges é unificar segmentos de rede baseados em mídias diferentes. Há algum tempo, quando ainda estava acontecendo a transição das redes com cabos coaxiais para as redes de par trançado, era muito comum que fosse utilizado um bridge para interligar os hosts conectados à rede antiga, com cabo coaxial à rede nova, com cabos de par trançado. Graças ao trabalho do bridge, tudo funcionava de forma transparente.

O bridge não precisa necessariamente ser um dispositivo dedicado. Por exemplo: um hub de 8 portas para cabos de par trançado, possui também um conector de cabo coaxial, o que permite que ele assuma também a função de bridge, interligando os dois segmentos de rede.

Atualmente, o exemplo mais comum de bridge são os pontos de acesso wireless, que podem interligar os micros da rede cabeada aos micros conectados à rede wireless, criando uma única rede. Muitos pontos de acesso incorporam também switches de 4 ou mais portas, ou até mesmo miniroteadores, que permitem compartilhar a conexão entre os micros da rede local. Hoje em dia, dispositivos "tudo em um" são cada vez mais comuns, pois com o avanço das técnicas de fabricação, tornou-se possível incluir cada vez mais circuitos em um único chip, fazendo com que um ponto de acesso "tudo em um" custe praticamente o mesmo que um ponto de acesso sem as funções extras.

Na imagem abaixo temos um exemplo visível dos casos explicados durante o tópicos. Nesse exemplo temos o switch no papel de ponte e um repetidor de um ponto de acesso wireless:



Conclusão

Sabemos que a utilização de repetidores se faz necessária para repassar o sinal de rede em uma grande área, por exemplo. Ao mesmo tempo que é uma saída interessante para distribuir esse sinal, precisa-se tomar certos cuidados com o posicionamento desses aparelhos, pois conforme estão direcionados pode ocorrer de gerar um "looping" de envio de sinal desnecessário. Um outro problema se dá ao fato de um repetidor não saber identificar quadros e entregá-los ao respectivo já que o sinal ainda não fora enviado, logo o repetidor não saberá qual é. Esses sinais serão repetidos de forma gerar um tráfego inútl pela rede. Solução facilmente obtida com o uso de uma ponte que se controlará com transparência o tráfego da rede encarregará de encaminhar corretamente esses pacotes ao destinatário.

Referências bibliográficas

http://www.interacaovirtual.com/apostilas/equipamento_redes.pdf
http://www.hardware.com.br/livros/redes/hubs-switches-bridges-roteadores.html