VERIFICADOR DE AUTENTICIDADE EM TRABALHOS ACADÊMICOS UTILIZANDO MÁQUINAS DE BUSCA

MOREIRA, Marcel Felipe Machado (1); REIS, José Cláudio de Souza (2)

(1) Acadêmico do Curso de Bacharelado em Ciência da Computação da UNIFENAS; (2) Orientador

RESUMO

Em constante evolução, a tecnologia da informação vem colaborando cada vez mais em práticas do dia-a-dia. O surgimento da Internet, e com ela os motores de busca, possibilitou maior ajuda na busca de informações, tornando mais fácil a obtenção de conhecimento. Este trabalho tem como objetivo estudar o funcionamento das máquinas de busca e compreender os detalhes deste funcionamento. Construiu-se um algoritmo que usa as máquinas de busca e as relacionam para que possam fazer massivas buscas na web, levando em consideração o tempo e a possibilidade de erro dessas máquinas. Tendo esta premissa como base, implementou-se um software usando o algoritmo de busca, de forma a encontrar ocorrências em textos acadêmicos que possam estar parcial ou totalmente dispostos na Internet. Com o uso do software, verificaram-se mais possibilidades de comparações, mais rapidez, mais eficiência e mais simplicidade na execução da tarefa de verificação de autenticidade. O software mostrou-se uma ferramenta simples na análise e verificação de trabalhos acadêmicos.

ABSTRACT

Information technology goes evolving, helping more and more the practical activities day to day. The advent of the internet and search engines has helped in finding information, facilitating the acquisition of knowledge. This work aims to study the operation of search engines as understand details of its operation. There has been built an algorithm that uses the search engines and so that can relate them in massive search on the web, taking into consideration the time and the possibility of error of these machines. Taking this as basis, one can implemented a software using the search algorithm in order to find occurrences in academic texts that can be partially or totally provided on the Internet. Using the software there are more possibilities for comparisons in a more efficiency, simply and faster way concerning the task of verification of authenticity. The software showed itself as a simple tool in the analysis and verification of academic works.

1 INTRODUÇÃO

1.1 Apresentação

No início da Era da Informática, computadores eram utilizados para cálculos balísticos de guerra; apenas efetuavam operações matemáticas, com um poder de processamento que superava as mentes dos grandes matemáticos. Porém, logo houve necessidade de armazenar os dados, uma vez que as respostas dessas grandes máquinas sobrecarregavam a mente dos operadores.

Em seguida, houve o desenvolvimento de formas e técnicas para armazenar dados em um sistema computacional. Ao mesmo tempo, foi preciso descobrir maneiras para gerenciar estes dados, uma vez que seu volume crescia cada vez mais e tornava-se impraticável gerenciá-los manualmente. Foi então que surgiram os sistemas de gerenciamento de dados. Para todo gerenciamento tem-se um acesso ao banco de dados.

Em se tratando de dados, o acesso a eles se torna muito importante. Por isso, as máquinas de busca são reconhecidas como um dos maiores inventos da modernidade, uma vez que dados públicos contidos em páginas de qualquer parte do mundo podem ser recuperadas por meio de simples palavras-chave.

As máquinas de busca evoluíram e tornaram-se acessíveis a todos que hoje utilizam computador com acesso à Internet. Entretanto, há necessidade de evoluir mais, de forma que palavras-chave sejam melhores compreendidas por essas máquinas, para que possam aperfeiçoar as buscas.

1.2 Objetivos

- Estudar o funcionamento das máquinas de busca e compreender detalhes deste funcionamento.

- Criação de um algoritmo que contenha as melhores máquinas de busca e que suporte massivas buscas.
- Utilizar algoritmo criado para desenvolvimento de uma aplicação que irá buscar textos acadêmicos pela web.

1.3 Justificativa

A motivação que levou a esse estudo é uma compreensão ampla das grandes máquinas de busca, sendo que as maiores empresas são suas detentoras. Além disso, hoje é o sistema de maior utilização pelos usuários.

Um dos grandes propósitos da computação é armazenar dados e descobrir como esses dados podem ajudar os usuários. Sendo assim, as máquinas de busca são tidas como maior banco de dados público existente. Então, procurar entender esses mecanismos se torna primordial para o desenvolvimento de uma aplicação que tenha como característica ajudar o usuário.

2. REFERENCIAL TEÓRICO

2.1 Máquinas de busca

2.1.1 História

Em 1993, com o crescimento acelerado da web, foi criado por Matthew Gray a primeira máquina de busca, denominada Wandex, específica para pesquisas de suas páginas. Utilizava um robô de computador, que contava os servidores ativos. No mesmo ano surgiu o Aliweb, que foi uma das primeiras tentativas de buscas por palavras-chaves, mas devido aos inúmeros acessos, causou um retardamento no sistema (COELHO, 2007).

O sistema pioneiro, que buscava por uma palavra em qualquer página, foi o WebCrawler, criado em 1994. Robôs localizavam documentos, indexavam e extraíam informações. Porém, no mesmo ano, o Lycos se tornou a máquina de busca mais popular da época (COELHO, 2007).

Com o sucesso do Lycos, que já tinha características dos buscadores que são usados atualmente, apareceram outras empresas com intuito de prover serviços eficientes, de forma organizada e de rápido acesso. Entre elas, Northern Light, Infoseek, Excite e AltaVista, que de uma certa forma competiram com diretórios populares como o Yahoo!, que fornecia indexação de páginas por sua categorização e descrição de cada endereço (COELHO, 2007).

Quatro anos depois, incentivados por David Filo, fundador do Yahoo!, dois jovens, chamados Larry Page e Sergey Brin, deram continuidade ao projeto de criar uma máquina de busca que tivesse habilidade de rastrear *links* da web, alcançando dados relevantes e grande volume de informação. Surgia o Google (COELHO, 2007).

2.1.2 Como funcionam as máquinas de busca

Estas máquinas operam sob a seguinte ordem: Web Crawling (percorrer *links*), Indexação e Busca Web.

2.1.2.1 Web Crawling

O processo de percorrer a web é chamado de web crawling ou spidering. As máquinas de busca utilizam o web crawler para varrer páginas de Internet, a fim de manter seu banco de dados sempre atualizado. As páginas são todas copiadas e é feita a indexação de todas as informações para que haja agilidade

nas buscas. Também pode ser usado para automatizar a manutenção de páginas web e validar conteúdo HTML (RICOTTA, 2007).

2.1.2.2 Indexação

Indexar é identificar um documento de acordo com o seu assunto. Pode-se fazer uma analogia do bibliotecário com um indexador, pois é feita por ele uma leitura técnica antes de indexar o livro (RICOTTA, 2007).

2.1.2.3 Busca Web

O usuário, ao escolher uma palavra-chave para a busca, está fazendo com que o sistema procure por índices e provenha uma lista de páginas que melhor combina com elas, normalmente seguido do título e uma breve descrição da página. As dificuldades de busca são a questão da ambiguidade e as estritas regras de sintaxe. A maior parte dos sistemas aceita expressões booleanas AND, OR e NOT para especificar a busca (RICOTTA, 2007).

2.1.3 Web Semântica

A Web Semântica tem como conceito introduzir estrutura e significado ao conteúdo web, transformando uma rede de documentos em uma rede de dados que seja compreensível tanto para humanos quanto para computadores, ou seja, uma desambiguação dos dados. O principal desafio da Web Semântica é expressar significado e ao mesmo tempo processar esses dados de forma a deduzir novos dados e regras, definindo as normas para gerir um significado, que devem ser transportadas para web, para que outros sistemas inteligentes possam interagir (MORAIS; SOARES, 2003).

2.1.4 Google

Atualmente o Google corresponde a aproximadamente 90% de toda busca via web. Desde seu surgimento, tomou proporções gigantescas, podendo ser comparado às revolução social ocorrida com a invenção da roda e, posteriormente, do celular (GRAF. 1, TECHNORECIPE, 2011).

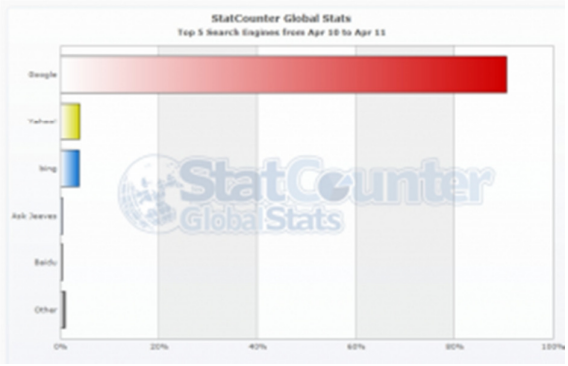


GRÁFICO 1 – Máquinas de busca mais utilizadas de abril de 2010 até abril de 2011 (TECHNORECIPE, 2011)

3 MATERIAL E MÉTODOS

3.1 Tecnologias utilizadas

No desenvolvimento da aplicação utilizou-se a UML (Unified Model Language) para modelagem e Photoshop para estruturação do *layout*. Para auxiliar no desenvolvimento foi utilizada a ferramenta Visual Studio na criação do projeto Web e SQL Server para modelagem e manipulação do banco de dados.

4 DESENVOLVIMENTO

4.1 Plágio em trabalhos acadêmicos

4.1.1 Informações sobre plágio

“O plágio acadêmico se configura quando um aluno retira, seja de livros ou da Internet, ideias, conceitos ou frases de outro autor (que as formulou e as publicou), sem lhe dar o devido crédito, sem citá-lo como fonte de pesquisa” (NERY et al., 2010).

4.1.2 Notícias sobre descoberta de plágio

Plágio pode trazer punições tanto para aluno quanto para o orientador, como no caso que ocorreu em fevereiro de 2011, em que um professor da USP foi demitido após 15 anos de carreira, por liderar uma pesquisa que plagiou trabalho de outros pesquisadores. O docente Andreimar Soares, da Faculdade de Ciências Farmacêuticas de Ribeirão Preto, foi demitido por ser o autor principal da pesquisa. Outra pesquisadora perdeu o título de doutorado por ser responsável pelas partes contestadas (USP…, 2011).

4.2 Modelagem do sistema

Verificar se trabalhos acadêmicos são autênticos é possível por meio da procura de trechos do trabalho em máquinas de busca. Com a intenção de facilitar e agilizar, no processo foi desenvolvido o sistema **Verificador de Autenticidade**, que tem como entrada um texto fornecido pelo usuário.

4.2.1 Orientações de uso do sistema

O usuário assim que entrar no sistema tem como página principal a tela de *login*. Após sua entrada, o usuário pode alterar suas configurações quanto à quantidade de sites a serem mostrados para cada trecho e a velocidade da análise, ver os últimos 10 trabalhos analisados para usuários que já utilizaram o sistema ou entrar com um novo trabalho para que seja analisado quanto a sua autenticidade.

4.2.2 Processo de preparação do texto para análise

Após o usuário entrar com o texto a ser analisado, a primeira etapa do sistema é dividir o texto em trechos de aproximadamente 100 caracteres. A utilização de 100 caracteres se da pelo motivo em que, se esse número for menor, poderá encontrar sites nos quais os trechos não necessariamente foram retirados do mesmo, e para mais, poderia nem ser encontrado mesmo existindo o trecho pela web. O sistema considera 100 caracteres mais o número de caracteres que existirem até que ele encontre um espaço. Os blocos serão armazenados em uma lista.

4.2.3 Colocação dos trechos para processamento paralelo

Com base na lista de trechos, o sistema irá colocar os 30 primeiros para serem analisadas no algoritmo de busca, de forma a serem executadas paralelamente, após o termino da analise desses trechos serão colocados os próximos 30 e assim sucessivamente até o término da lista de trechos. Foram escolhidos 30, por ser o maior número de trechos executados simultaneamente que não faz uma sobrecarga no algoritmo, uma vez que, quanto maior o número de trechos sendo analisado no mesmo instante, mais rápido o sistema será.

4.2.4 Desenvolvimento do algoritmo de busca

Para verificar a autenticidade de um texto é necessário analisar se os trechos do mesmo estão dispostos na web, ou seja, se foram copiados. Para fazer essas buscas, foram implementadas 12 máquinas de busca, para que retornem os sites resultantes. São necessárias duas etapas para isso: a etapa de extração dos sites resultantes dos buscadores e a etapa do processo de escolha do buscador.

4.2.4.1 Etapa de extração de sites resultantes dos buscadores

Na extração de sites que contêm ocorrência do trecho é necessário colocar junto com o endereço (URL) da máquina de busca o trecho, que será analisado entre aspas. Tudo concatenado, fazendo um acesso à máquina de busca. Com isso, será copiado o HTML da página, que irá conter os sites formadores do resultado da busca.

Foi implementado um método genérico de retorno de HTML usado em todos os buscadores em que serão passados o trecho a ser analisado, a URL do buscador, o texto que o buscador mostra quando não houver resposta e o conjunto de caracteres que indicam que, a partir dele, tudo na string do HTML que for “http” é resultado de busca. Será retornado o índice deste início e também o HTML da página.

Com base nos métodos implementados para cada um dos 12 buscadores métodos de busca, que terá denominação de método HtmlBusca retornando o seu HTML, quando houver resultados e não houver nenhum erro. Se for vazio será colocado na lista de resultados apenas a palavra “vazio” e retornar; se for erro será colocado “erro” e retornar. Caso não ocorra nenhuma dessas duas situações, ele passará para o métodos ExtrairUrl, em que será retornada uma lista com os sites de resultados do trecho para o algoritmo.

4.2.4.2 Processo de escolha do buscador

O algoritmo buscador tem como entrada no seu método construtor o trecho a ser analisado e a quantidade de sites resultantes, que será determinada pelo usuário na parte de configurações do sistema.

O trecho será submetido a um sorteio, selecionando randomicamente um número e a máquina de busca que estiver naquele intervalo será executada. Cada máquina de busca foi estudada para que pudesse determinar um valor de pertinência para cada um deles, baseado na velocidade, tempo de erro, tempo de congelamento, entre outros elementos. Foram atribuídos aos melhores um maior intervalo de valores que podem ser sorteados. Caso o buscador não se porte muito bem com massivas buscas, ganharão um menor conjunto de números, sendo esta uma solução para que não haja sobrecarga em um buscador, pois se houvesse sido implementado somente um, este entraria em estado de congelamento e o software iria parar de funcionar.

As pertinências foram definidas conforme mostra o gráfico 2.

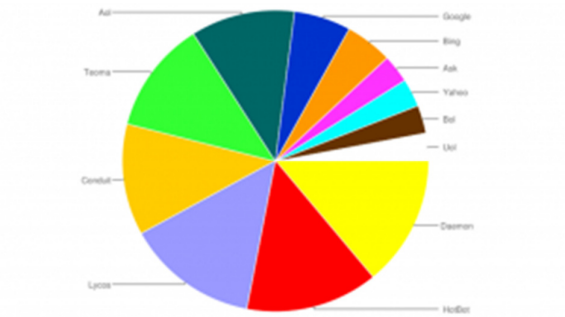


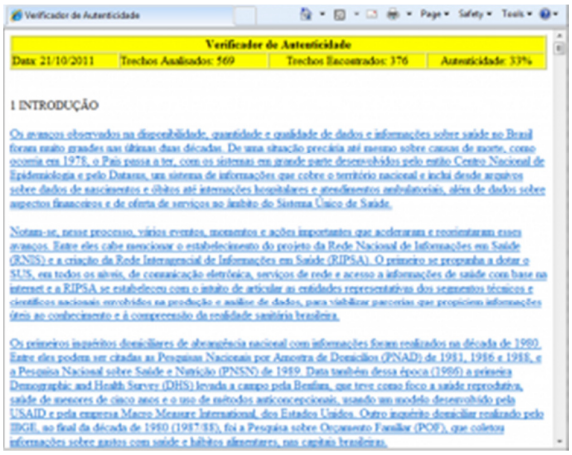
GRÁFICO 2 – Possibilidade de sorteio das máquinas de busca

Uma máquina de busca pode retornar um erro após muitas buscas feitas em um espaço de tempo curto, entrando em estado de congelamento. Se o sistema continuasse fazendo busca, a máquina continuaria retornando erro e o estado de congelamento continuaria infinitamente. Para resolver tal situação, o sistema grava no banco de dados data e hora de congelamento, bloqueando a

máquina de busca por um tempo determinado. Este tempo é baseado nos estudos feitos a cada buscador no seu tempo necessário para descongelamento, voltando a serem utilizados após passar o tempo determinado de bloqueio.

4.2.5 Apresentação da monografia verificada

Após etapas de verificação do texto, será apresentado outro texto para o usuário, de modo que mostre as estatísticas dos buscadores e os *links* para cada trecho analisado.


Verificador de Autenticidade

Verificador de Autenticidade

Data: 21/10/2011	Trechos Analisados: 569	Trechos Encontrados: 376	Autenticidade: 33%

1 INTRODUÇÃO

Os avanços observados na disponibilidade, quantidade e qualidade de dados e informações sobre saúde no Brasil foram muito grandes nas últimas duas décadas. De uma situação precária até mesmo sobre causas de morte, como ocorria em 1978, o País passa a ter, com os sistemas em grande parte desenvolvidos pelo então Centro Nacional de Epidemiologia e pelo Datasus, um sistema de informações que cobre o território nacional e inclui desde arquivos sobre dados de nascimentos e óbitos até internações hospitalares e atendimentos ambulatoriais, além de dados sobre aspectos financeiros e de oferta de serviços no âmbito do Sistema Único de Saúde.

Notam-se, nesse processo, vários eventos, momentos e ações importantes que aceleraram e reorientaram esses avanços. Entre eles cabe mencionar o estabelecimento do projeto da Rede Nacional de Informações em Saúde (RNIS) e a criação da Rede Interagencial de Informações em Saúde (RIPSA). O primeiro se propunha a dotar o SUS, em todos os níveis, de comunicação eletrônica, serviços de rede e acesso a informações de saúde com base na internet e a RIPSA se estabeleceu com o intuito de articular as entidades representativas dos segmentos técnicos e científicos nacionais envolvidos na produção e análise de dados, para viabilizar parcerias que propiciem informações úteis ao conhecimento e à compreensão da realidade sanitária brasileira.

Os primeiros inquéritos domiciliares de abrangência nacional com informações foram realizados na década de 1980. Entre eles podem ser citadas as Pesquisas Nacionais por Amostra de Domicílios (PNAD) de 1981, 1986 e 1988, e a Pesquisa Nacional sobre Saúde e Nutrição (PNSN) de 1989. Data também dessa época (1986) a primeira Demographic and Health Survey (DHS) levada a campo pela Bemfam, que teve como foco a saúde reprodutiva, saúde de menores de cinco anos e o uso de métodos anticoncepcionais, usando um modelo desenvolvido pela USAID e pela empresa Macro Measure International, dos Estados Unidos. Outro inquérito domiciliar realizado pelo IBGE, no final da década de 1980 (1987/88), foi a Pesquisa sobre Orçamento Familiar (POF), que coletou informações sobre gastos com saúde e hábitos alimentares, nas capitais brasileiras.

FIGURA 1 – Trabalho verificado

A figura 1 apresenta o trabalho verificado. Na parte superior estão às estatísticas de busca, mostrando a quantidade de trechos que foram analisados e a quantidade de trechos que foram encontrados na WEB. Com base nos dois realizam-se as estatísticas de busca.

FIGURA 2 – Site resultado

A partir de um clique em algum dos trechos com *link* na figura 1, será aberta uma nova página que contém exatamente as mesmas palavras do trecho clicado. Como pode ser percebido, a figura 2 apresenta vários fragmentos de trechos da figura 1.

4.3 Análise de eficiência do sistema

Programas que fazem análise de plágio em monografias são pouco conhecidos atualmente, porém existem algumas universidades que já os utilizam para verificar a autenticidade das monografias dos alunos. Em geral, esses programas são pagos; os gratuitos, em sua maioria, são limitados.

Foram utilizados os programas mais conhecidos da Internet que identificam plágio em monografias e comparou-se com o Verificador de Autenticidade. Os resultados estão demonstrados na tabela 1.

TABELA 1 – Comparação com outros sistemas

Nome	Sistema WEB	Tipo de Verificação	Porcentagem de Ocorrência de Plágio	Tempo Gasto Para Conclusão de Verificação
Plagium	Sim	Rápida	Não	5,3s
Plagius	Não	Rápida	63,80%	21,3s
Farejador de plagio	Não	Rápida	Não	98,6s
Verificador de Autenticidade	**Sim**	**Rápida**	**83,60%**	**1,3s**
Verificador de Autenticidade	**Sim**	**Normal**	**85,30%**	**2,6s**

A tabela 1 mostra a comparação de alguns sistemas mais conhecidos em relação ao sistema desenvolvido, feita por meio de três arquivos diferentes de uma página. O Tipo de Verificação indica qual velocidade para realização da tarefa de verificação. A Porcentagem de Ocorrência de Plágio mostra o valor percentual de plágio indicado nos sistemas para cada um dos três arquivos analisados somados e divididos por três. O Tempo Gasto Para Conclusão de Verificação é a quantidade em segundos da soma do tempo gasto para a execução dos arquivos dividida por três.

5 RESULTADOS E DISCUSSÃO

O software obteve como resultado um dos fatores mais importantes da computação, que refere-se ao seu desempenho. O software também apresenta como característica sua facilidade de uso e uma maior precisão quanto à análise do trabalho.

Para o desenvolvimento desse sistema foram inúmeros os problemas que se apresentaram ao longo do processo. O principal deles foi o congelamento dos buscadores, uma vez que inicialmente parecia ser impossível a implementação do sistema sem a obrigatoriedade de pagamento pelas máquinas de busca para cada resultado.

Para tanto, foi necessário o estudo minucioso de cada uma das máquinas de busca implementadas e colocá-las de forma a fazer revezamento durante o processo de recolhimento de sites resultados.

6 CONCLUSÃO

Com o uso do Verificador de Autenticidade, verificou-se que este aplicativo conseguiu mais comparações, mais rapidez, mais eficiência e mais simplicidade na execução da tarefa de verificação de autenticidade, conforme explicitado pela TAB. 1. O aplicativo mostrou-se uma ferramenta simples na análise e na verificação de trabalhos acadêmicos.

O sistema desenvolvido atendeu às necessidades de análise de uma monografia quanto a sua autenticidade, de maneira a apresentar os resultados ao usuário de forma explícita.

Por estes resultados, conclui-se que a ferramenta poderá auxiliar os professores na correção de monografias e outros trabalhos acadêmicos, indicando as que sejam realmente escritas pelo próprio conhecimento do aluno.

Sugere-se como implementações futuras neste aplicativo o uso de novas máquinas de busca, verificação da ocorrência do trabalho a ser analisado escrito em outras línguas e o uso de técnicas de inteligência artificial na escolha da máquina de busca, que deve ser executada por período de tempo e na escolha dos tamanhos do texto para análise.

REFERÊNCIAS

COELHO, Ed. **A história do tempo dos buscadores.** 2007 Disponível em: <http://imasters.com.br/artigo/5444/webmarketing/a_historia_do_tempo_dos_buscadores/> Acesso em: 05 Junho 2011.

MORAIS, Erikson Freitas; SOARES, Marcelo Borghetti. **Web Semântica para Máquinas de Busca.** 2003. 6f. Artigo. Departamento de Ciência da Computação. Universidade Federal de Minas Gerais, Belo Horizonte.
NERY, Guilherme et al . **Nem tudo que parece é: entenda o que é plágio.** Rio de Janeiro, 2010. Disponível em: <http://www.noticias.uff.br/arquivos/cartilha-sobre-plagio-academico.pdf> Acesso em: 06 Novembro 2011.

OLIVEIRA, Rômulo Silva de Oliveira; CARISSIMI, Alexandre da Silva; TOSCANI, Simão Sirineo. Sistemas Operacionais.**Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 6, n.3, p. 17, dez. 2001.

SILBERSCHATZ, Abraham; KORTH, Henry; SUDARSHAN, S. **Sistema de Banco de Dados.** São Paulo: Makron Books do Brasil., 3. ed., 1999.

RICOTTA, Fábio Carvalho Motta. **"Como os search engines funcionam?"** 2007 176 f. Trabalho de Graduação de Curso em Matemática e Computação – Instituto de Ciências Exatas da Universidade Federal de Itajubá, Minas Gerais.

TECHNORECIPE. **Bing & Baidu, vs The Google ??.** 2011 Disponível em: <http://www.technorecipe.info/2011/05/bing-baidu-vs-google.html> Acesso em: 06 Junho 2011.

USP demite professor por plágio em pesquisa. **FOLHA.COM.** 20 Fevereiro 2011. Disponível em: <http://www1.folha.uol.com.br/saber/878368-usp-demite-professor-por-plagio-em-pesquisa.shtml>. Acesso em: 6 Novembro 2011.