SISTEMA PARA RECUPERAÇÃO DE INFORMAÇÃO DE DOWNLOADS DE SOFTWARES

Elzo Soares Pereira Filho, Matheus Rossi de Oliveira Costa,
Vinícius Lobo Silva, Luciene Chagas de Oliveira,
UNIUBE – Universidade de Uberaba
Uberlândia – MG, Brasil
elzo.soares@yahoo.com.br, matheus-rossi07@hotmail.com,
loboo_u2@yahoo.com.br, luciene.oliveira@uniube.br

Resumo – A recuperação da informação é muito importante, pois sabemos que a cada dia que passa mais e mais informações vão sendo geradas e distribuídas pelo mundo. Com a invenção da Internet, ficou mais fácil e rápido disseminar essas informações entre as pessoas. Atualmente, muitos usuários precisam realizar a busca de downloads de software. Neste contexto, surge a necessidade de ferramentas para auxiliar em uma busca mais refinada de informações. O objetivo deste trabalho é realizar o desenvolvimento de um sistema Web que aplica os conceitos de Recuperação de Informação para realizar a busca de downloads de softwares que utiliza um motor de busca.

***Palavras-Chave* – desenvolvimento de software, recuperação de informação, site de busca, motor de busca.**

INFORMATION RETRIEVAL SYSTEM OF SOFTWARE DOWNLOADS

***Abstract* - Information retrieval is very important because we know that every day that passes more and more information are being generated and distributed worldwide. With the invention of the Internet has made it easier and faster to spread this information among people. Currently, many users need to perform a search for software downloads. In this context there is a need for tools to assist in a search for more refined information. The aim of this work is the development of a Web system that applies the concepts of Information Retrieval to perform a search for software downloads hat uses a search engine.**

***Keywords* - software development, information retrieval, search site, search engine.**

XI CEEL – ISSN 2178-8308
25 a 29 de novembro de 2013
Universidade Federal de Uberlândia – UFU
Uberlândia – Minas Gerais – Brasil

I. INTRODUÇÃO

Com o crescimento do volume de informações, ao longo dos anos, foram desenvolvidas técnicas de recuperação de informação para responder às necessidades dos usuários de bibliotecas, tradicionais ou digitais.

A Recuperação de Informação (RI) é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados geralmente texto. A área de Recuperação de Informação (RI) possui grande importância para a comunidade científica devido à grande disponibilidade de documentos existente hoje na forma digital, por exemplo, naWeb. O problema central em recuperação de informação é encontrar informações de interesse dos usuários. A principal ferramenta usada para resolver este problema é o emprego de sistemas de recuperação de informação (SRI).

Geralmente, a quantidade de downloads de software é grande, surgindo a necessidade de realizar a recuperação destas informações. RI é uma área que estuda o armazenamento, classificação, agrupamento e recuperação automática de documentos. A abundância de informações na Web é uma das principais razões para sua crescente popularidade. A facilidade do uso e acessibilidade da Web fez dela uma ferramenta muito importante, não somente para comunicação, mas também para armazenamento e compartilhamento de informação. Do ponto de vista de RI, a Web pode ser vista como grande repositório de dados contendo documentos, ou páginas Web, que são interconectados. [1]

Neste trabalho foi desenvolvido um sistema Web para recuperar informações relevantes de downloads de softwares para Engenharia. Neste sistema foi implementado o algoritmo vetorial no módulo de buscas dos downloads e desenvolvido um motor de busca para os links de downloads.

II. FUNDAMENTOS DA RECUPERAÇÃO DE INFORMAÇÃO

A. Sistema de Recuperação de Informação

Recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados

geralmente texto. Um sistema de Recuperação de Informação (SRI) pode ser estruturado conforme a Figura 1[2].

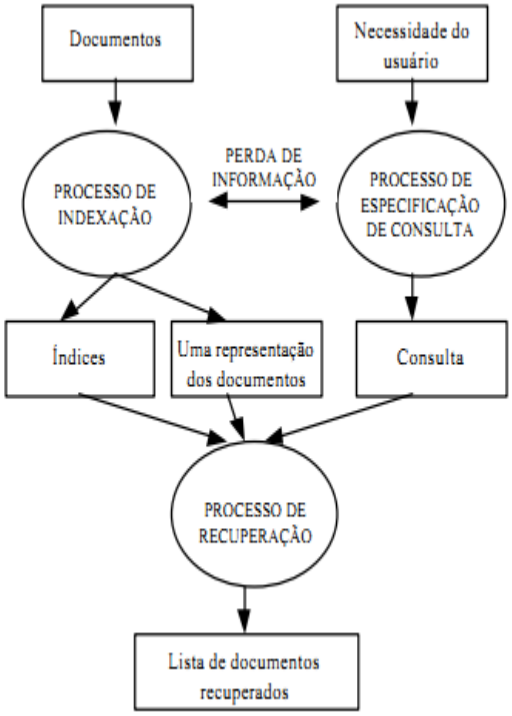

Fig. 1. Componentes de um Sistema de Recuperação de Informação [2]

Os componentes do sistema incluem documentos, necessidades do usuário, gera a consulta formulada, e finalmente o processo de recuperação que, à partir das estruturas de dados e a da consulta formulada, recupera uma lista de documentos considerados relevantes.

O processo de indexação envolve a criação de estruturas de dados associados à parte textual dos documentos, por exemplo, as estruturas de *arrays* e arquivos invertidos, discutidas em [3]. Estas estruturas podem conter dados sobre características dos termos na coleção de documentos, tais como a frequência de cada termo em um documento.

O processo de especificação da consulta geralmente é uma tarefa difícil. Há frequentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada. Essa distancia é gerada pelo limitado conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta.

O processo de recuperação consiste na geração de uma lista de documentos recuperados para responder consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos e são usados para acelerar esta tarefa. Além disso, a lista de documentos recuperados é classificada em ordem decrescente de um grau de similaridade entre o documento e a consulta.

O site da Google [4] é um exemplo de site de busca que utiliza técnicas e modelos de Recuperação de Informação.

B. Motores de Busca

Uma das principais funções de um Administrador de *Web Site* é saber qual o motivo de um usuário estar utilizando aquele serviço no qual ele oferece e o que o mesmo pode estar à procura naquele momento em seu site. Sabendo disso, o Administrador pode utilizar as ferramentas corretas e sistemas de pesquisa eficiente para que o usuário consiga localizar sua pesquisa com máxima eficiência e no menor tempo possível.

Visando em proporcionar ao usuário tais eficiências e agilidade, pode-se utilizar de um sistema chamado Motor de buscas no qual utiliza-se de palavras-chave para localizar assuntos relacionados na Web e trazer ao usuário uma lista de resultados que combinam com tais critérios. Estes motores de buscas utilizam algoritmos e modelos de Recuperação de Informação (RI).

Os motores de busca usam regularmente índices atualizados para funcionar de forma rápida e eficiente. Sem maior especificação, ele normalmente refere-se ao serviço de busca Web, que procura informações na rede pública da Internet. Alguns motores também extraem dados disponíveis em grupos de notícias, grandes bancos de dados ou diretórios abertos. Ao contrário dos diretórios Web, que são mantidos por editores humanos, os serviços de busca funcionam algoritmicamente. A maioria dos sites que chamam os motores de busca são, na verdade, uma "interface" (*front end*) para os sistemas de busca de outras empresas.

C. Modelos Clássicos de Recuperação de Informação

Um modelo de recuperação de informação representa documentos e consultas para predizer o que um usuário considera relevante para sua necessidade de informação. São três os modelos clássicos seguidos por sistemas de RI para determinar a relevância de documentos: booleano, vetorial e probabilístico [1].

Os modelos clássicos, utilizados no processo de recuperação de informação, apresentam estratégias de busca de documentos similares à consulta. Estes modelos consideram que cada documento é descrito por um conjunto de termos, considerados como mutuamente independentes.

Associa-se a cada termo k_i e um documento d_j um peso $w_{i,j} \geq 0$, que quantifica o peso do termo k_i no documento d_j . Este peso reflete a importância do termo k_i no documento d_j. Analogamente a cada par termo-consulta (k_i,q) associa-se o peso w_i,q.

Modelos mais avançados têm sido propostos, mas ainda existe uma grande necessidade por novos arcabouços que permitam a melhoria na qualidade das respostas [1].

Neste trabalho foi utilizado o modelo clássico vetorial, descrito a seguir. No modelo de espaço-vetorial, ou simplesmente modelo vetorial, os documentos e as consultas são representados por um vetor em um espaço de termos. O conjunto de termos de uma coleção de documentos é chamado de vocabulário. Cada termo possui um peso associado que indica seu grau de importância no documento. Em outras palavras, os documentos e as consultas possuem vetores associados a cada um [1].

Cada elemento do vetor de termos é considerado uma coordenada dimensional. Assim, os documentos e consultas do usuário são representados como vetores de termos em um espaço t-dimensional, onde t é o número de termos ou

tamanho do vocabulário. O j-ésimo documento em uma coleção de documentos é denotado por d_j. Um termo é uma palavra que semanticamente ajuda a lembrar o tema principal do documento. Um termo é denotado por k_i. Então o vetor associado ao documento d_j é dado por $\vec{j} = (w_{1,j} , w_{2,j} ,..., w_{t,j})$, onde $w_{i,j}$ é o peso associado ao termo ki no documento d_j . Consultas também são representadas por vetores. Uma consulta é um conjunto de termos que expressa a necessidade do usuário, e é denotada por q . O vetor associado à consulta q é $\vec{q} = (w_{1,q} , w_{2,q} ,..., w_{t,q})$, onde $w_{i,q}$ é o peso associado ao termo ki na consulta q .

Cada dimensão deste espaço é associada com um vetor de termos $\vec{k}_i$. Estes vetores de termos são ortogonais, ou seja, $i \neq j \Rightarrow \vec{k}_i * \vec{k}_j = 0$. Isto indica que assumimos que termos ocorrem independentemente dentro dos documentos e consultas. Além disso, $|\vec{k}_i| = 1$.

O modelo vetorial propõe avaliar o grau de similaridade entre um documento dj e uma consulta q como uma correlação entre vetores $\vec{d}_j$ e $\vec{q}$.

Esta correlação pode ser quantificada pelo cosseno do ângulo entre estes vetores. Então, a fórmula de similaridade é definida como:

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \qquad (1)$$

A Equação (1) representa a regra do produto para o cálculo da similaridade do modelo vetorial.Os pesos wi,j e wi,q quantificam a importância do termo ki para a consulta e para os documentos, respectivamente. Os pesos podem ser calculados da seguinte forma [6]: Seja N o número total de documentos na coleção e ni o número de documentos em que o termo ki aparece. Seja freqi,j a frequência natural do termo ki no documento dj , isto é, o número de vezes que o termo ki é mencionado no texto do documento dj . Se o termo ki não aparece no documento dj, então freqi,j = 0. Cada frequência do termo fornece uma medida de como o termo descreve o conteúdo do documento, denominada caracterização intra-documento.

Para cada termo é calculado também a frequência inversa dos documentos onde o termo aparece, IDF, que fornece uma caracterização inter-documento. A motivação para o seu uso é que termos que aparecem em muitos documentos não são úteis para distinguir um documento relevante de um documento não relevante. O IDF (ki), freqüência inversa de documentos do termo ki em uma coleção é dado por:

$$IDF(k_i) = \log \frac{N}{n_i} \qquad (2)$$

O peso do termo no documento é dado pela fórmula wi,j = freqi,j * IDF (ki) e o peso do termo na consulta é dado por wi,q = freqi,q * IDF(ki) [5]. A norma do documento dj de uma coleção é dada por:

$$|\vec{d_j}| = \sqrt{\sum_{i=1}^{t} w_{i,j}^2} = \sqrt{\sum_{i=1}^{t} (freq_{i,j} \cdot IDF(k_i))^2} \qquad (2)$$

Calculados os graus de similaridade pela Equação (1), é possível montar uma lista ordenada de todos os documentos ordenados por seus respectivos graus de relevância à consulta ou *ranking*. Um documento pode ser recuperado mesmo se ele satisfazer a consulta somente parcialmente. Assim, os documentos mais similares à consulta ficarão no topo desta ordenação.

Este é um modelo muito utilizado em sistemas de recuperação de informação. As principais razões para isto são a sua rapidez no processo de busca, a sua simplicidade, a flexível estratégia de agrupamento e a boa precisão na recuperação de documentos de coleções genéricas [5].

III. RESULTADOS E FUNCIONAMENTO DO SISTEMA

Ao acessar o site para recuperação de informações, denominado Eng. Downloads, o usuário visualizará a pagina inicial, nesse caso a Home Page, ela é composta por um menu horizontal, campo para imagem e um campo para texto no qual apresenta uma breve descrição do site e seu objetivos.

O menu é composto por 4 opções distintas, sendo elas: Home Page, News, Downloads e login.

Na Figura 2 é possível visualizar a página inicial do site.

Fig. 2. Página Inicial (Home Page) do Sistema Eng Downloads.

O usuário deve selecionar a opção desejada entre as citadas anteriormente para navegar no site e caso o mesmo selecione downloads ele será redirecionado para a página que executa o modelo de Recuperação de Informação para busca de downloads de softwares, que possui um campo para que

ou usuário possa informar palavras chave de pesquisa de softwares (Figura 3).

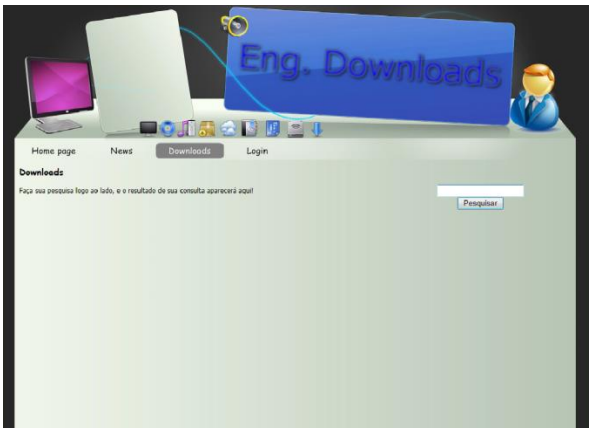

Fig. 3. Página de Downloads com área para Pesquisa

Para efetuar a pesquisa, o usuário deve inserir as palavras chave que deseja buscar e clicar em pesquisar. Assim, o sistema irá efetuar a recuperação de informações e mostrar os links para downloads em forma de uma lista ao usuário, com o nome do software, parte da descrição e tamanho do arquivo, conforme Figuras 4 e 5.

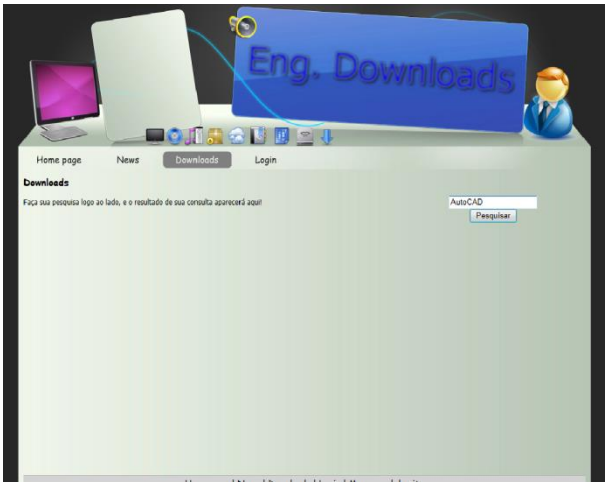

Fig. 4. Usuário inserindo no campo de busca software para pesquisa..

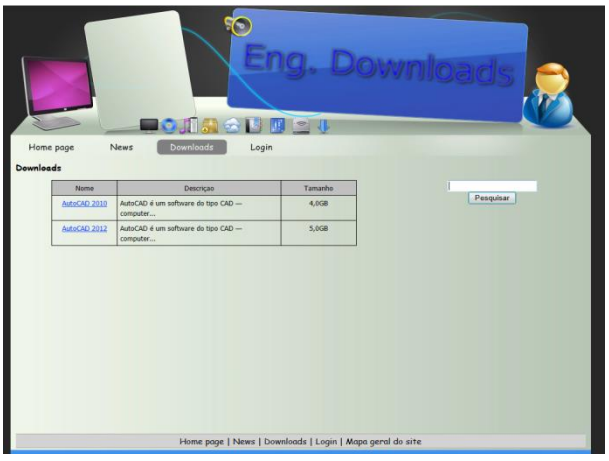

Fig. 5. – Tabela contendo resultado da pesquisa

Ao localizar o software que deseja, o usuário clicará em cima do nome arquivo e será redirecionado a uma nova página, na qual possuirá o nome do software, a sua descrição completa e imagem do produto, juntamente com o tamanho de arquivo e link para download, e caso seja necessário haverá também um breve tutorial de instalação junto à descrição do programa como mostra a Figura 6.

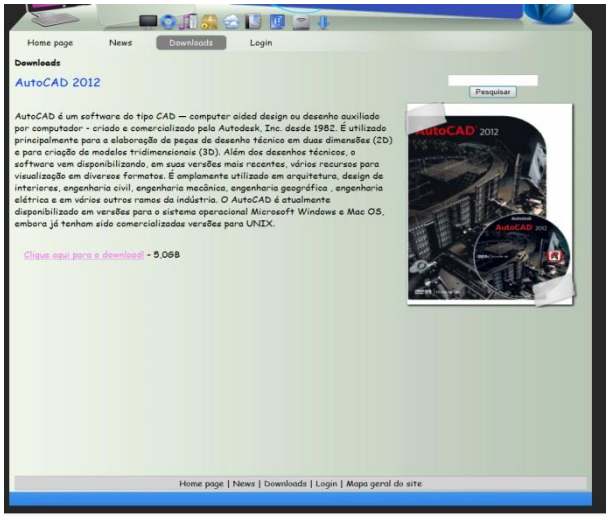

Fig. 6. Página mostrando detalhamento do aplicativo desejado pelo usuário.

Todas as informações estão armazenadas na Web e para o usuário efetuar qualquer tipo de download no site o mesmo deve estar cadastrado no site e efetuar o login.

O login é realizado através de outra página no site, cujo caminho pode ser acessado pelo menu de opções. A página de login é mostrada na Figura 7.

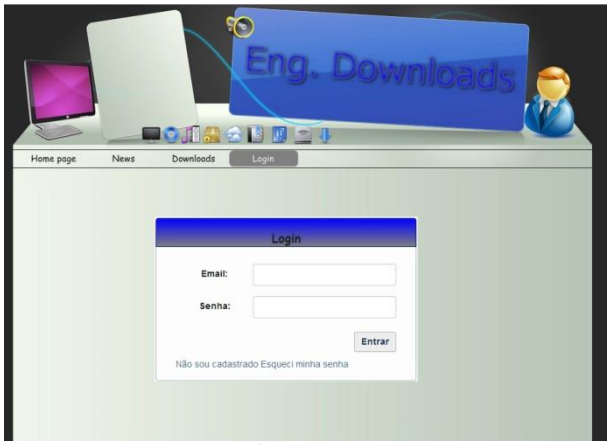

Fig. 7. Tela de Login do usuário

Caso o mesmo já esteja cadastrado poderá efetuar o login normalmente e prosseguir para área de downloads, caso contrário, haverá uma opção para cadastro no qual o usuário clica e é redirecionado a uma página de cadastro no qual ele insere todas as informações exigidas pelo site, como e-mail, senha, confirmação de senha, nome, pergunta secreta e resposta da pergunta secreta para a recuperação de senha caso o usuário esqueça-a (Figura 8).

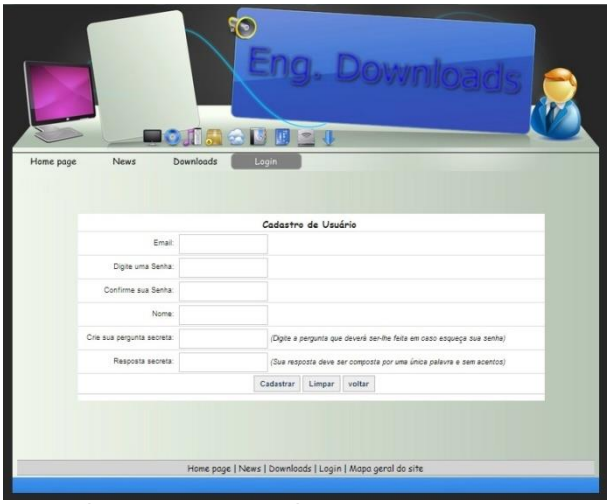

Fig. 8. Página de cadastro de usuários.

Se o usuário por algum motivo esqueceu sua senha ou a perdeu, o mesmo poderá acessar a área de login e selecionar a opção de "Esqueci minha senha" ao clicar no link, o usuário visualizará um campo para o mesmo inserir seu email e um botão para clicar chamado recuperar, ao efetuar isso será exigido do usuário a resposta da pergunta secreta que o mesmo inseriu durante seu cadastro no site, feito isso, será enviado um e-mail para o usuário com a senha cadastrada. Tal procedimento poderá ser visualizado nas Figuras 9 e 10.

Fig. 9. Página de recuperação de senhas.

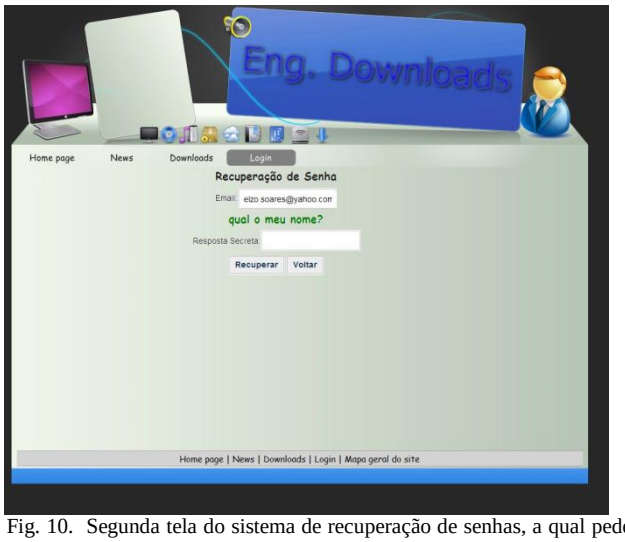

Fig. 10. Segunda tela do sistema de recuperação de senhas, a qual pede a resposta da pergunta secreta.

Além disso, o sistema possui uma página News que contem notícias de programas novos adicionados ao site e ou alguma outra informação relevante para o usuário como mostra a Figura 11.

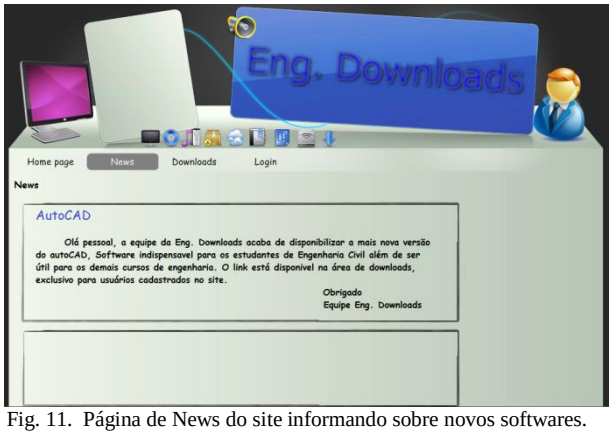

Fig. 11. Página de News do site informando sobre novos softwares.

A Recuperação de Informação no sistema é realizado através da utilização do motor de busca para localização de links para downloads de softwares, conforme mostrado na Figura 12.

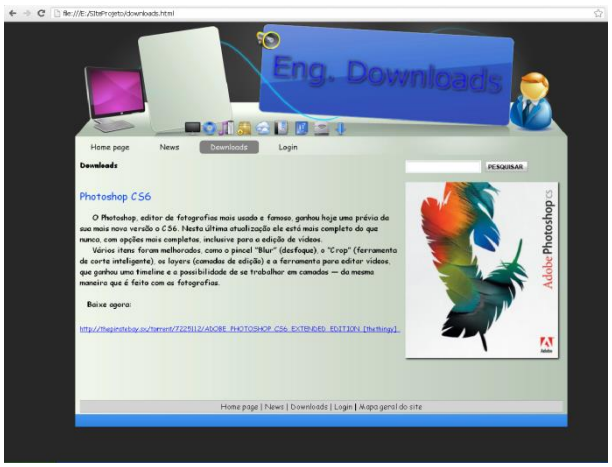

Fig. 12. Site de Download que utiliza motor de busca no banco de dados.

Estudantes de Engenharia e Tecnologia da Informação (TI) muitas vezes precisam utilizar algum tipo de software para utilização em algumas disciplinas, mas não sabem onde e como encontrar essas ferramentas, seja ela de desenvolvimento, simulação ou cálculos. Pensando nisso, foi desenvolvido o *web site* Eng Downlods cujo objetivo é facilitar o acesso do aluno a essas ferramentas.

Foi utilizado um sistema de Motor de buscas interno que utiliza a Web e um banco de dados para a localização de links para downloads desses aplicativos e os disponibiliza no site com as informações adicionais, para que o estudante, tenha acesso a esse conteúdo de maneira rápida sem a necessidade de acessar diversos site na procura de links que funcionem e sejam de confiança.

IV. CONCLUSÕES

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, recuperação de informação apresenta a cada dia, novos desafios e se configura como uma área de significância maior.

O estudo da área de Recuperação de Informação é de grande importância para a área de sistemas de informações em geral. Com a explosão do número de documentos e usuários na Web, modelos para RI passaram a se tornar relevantes.

Para demonstrar a utilização de um sistema de recuperação de informação, neste trabalho foi desenvolvimento um site de busca para recuperar informações relevantes para downloads de software para auxiliar profissionais da área de TI, tais como engenheiros e desenvolvedores.

Motores de buscas estão constantemente em melhorias em seu algoritmo para trazer aos usuários sempre resultados mais precisos e hoje é um componente essencial em grandes sites geralmente com grande quantidade de conteúdo e utiliza-se desse sistema para facilitar a pesquisa do usuário final.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] OLIVEIRA, L. C. Meta-Modelo funcional para recuperação de informaçãoTítulo: subtítulo. Tese Mestrado UFU, 2006.

[2] GEY, F. “Models in Information Retrieval”. Folders of Tutorial Presented at the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR), 1992.

[3] FRAKES, W. B. & Baeza-Yates, R. Information Retrieval. Data Structures & Algorithms, Prentice Hall, 1992.

[4] GOOGLE, “Google”. Disponível em: https://www.google.com.br/. Acesso em Junho de 2013.

[5] BAEZA-YATES & RIBEIRO-NETO. Modern Information Retrieval. Addison Wesley, 1999.

[6] ZOBEL J. ; MOFFAT A. Exploring the similarity space. SIGIR Forum, 32(1):18–34, 1998.