XX Seminário Nacional de Distribuição de Energia Elétrica
SENDI 2012 - 22 a 26 de outubro
Rio de Janeiro - RJ - Brasil

Rosimeri Xavier de Oliveira	Light Serviços de Eletricidade S/A	rosimeri.oliveira@light.com.br
Carlos Jose Ribas Davila	Escola Politécnica da UFRJ - Universidade Federal do Rio de Janeiro	case@del.ufrj.br
Sergio Palma da Justa Medeiros	Escola Politécnica da UFRJ - Universidade Federal do Rio de Janeiro	sergio.mederios@del.ufrj.br
Ester Jose Casado de Lima	Escola Politécnica da UFRJ - Universidade Federal do Rio de Janeiro	esterlima@gmail.com
Andreia Luz Duarte	Light Serviços de Eletricidade S/A	andreia.duarte@light.com.br
Thalita Gaspar Telles	Light Serviços de Eletricidade S/A	thalita.telles@light.com.br
Marilia Nocchi	Light Serviços de Eletricidade S/A	marilia.nochhi@light.com.br

Busca Semântica no Atendimento Virtual

Palavras-chave

Busca Semântica
Indexação
Motor de busca
Web Crawling

Resumo

O advento da Internet democratizou o acesso às informações armazenadas, permitindo que a comunicação entre as pessoas se tornasse fácil e em tempo exíguo. Esta facilidade, porém trouxe um complicador: encontrar a informação correta no universo de textos armazenados. Para resolver este problema foram criados os ambientes de busca. Altavista, Google e Yahoo são exemplos de *sites* de busca que facilitam o acesso às informações armazenadas.

A Light, porém, inovou neste quesito ao implementar uma facilidade que não atendesse apenas aos requisitos tradicionais oferecidos pelos gigantes deste mercado de busca. A empresa resolveu endereçar o problema que existe quando as palavras usadas na busca não constam das páginas explicativas dos serviços. Para tal, foi desenvolvida uma solução de *software* denominada Máquina de Busca Semântica.

A diferença principal entre uma máquina de busca convencional e uma Semântica é que esta última permite ao cliente encontrar a página que responda aos seus anseios, mesmo que não utilize na busca palavras constantes dos textos descritos pela empresa.

Este diferencial permite que o cliente seja atendido no seu desejo de encontrar de forma rápida a informação precisa e que resolva o problema que o levou a procurar o atendimento Virtual da Light.

1. Introdução

A quantidade de páginas que contêm informação útil para o usuário tem crescido na mesma proporção que as empresas disponibilizam serviços para os seus clientes na Web. Este crescimento tem sido muito intenso nos últimos anos e os problemas decorrentes desta atividade também aumentam em tamanho e complexidade. A solução encontrada nos últimos anos para resolver esta questão foi a utilização de mecanismos de busca, que o mercado apelidou de Motores ou Máquinas de Busca. O Altavista, o Google, o Yahoo e o Bing são exemplos de Máquinas de Busca e que facilitam o acesso às informações armazenadas na Web.

A Light apresenta os mesmos problemas acima descritos, porém, em um escopo mais restrito. A empresa tem que criar mecanismos para que seus clientes encontrem os serviços disponibilizados por ela nas páginas constantes na sua Agência Virtual. Uma das soluções para este problema seria construir suas páginas de acordo com convenções estipuladas pelas Máquinas de busca públicas do tipo Google. O cliente entraria pelo Google, digitaria as palavras que representassem a sua necessidade de informação e o Google devolveria uma lista de páginas que tivesse relevância com relação aos termos fornecidos pelo usuário.

Este artifício, porém encontra a limitação de que é necessário colocar no texto das páginas todas as palavras passíveis de serem usadas pelos clientes e todos os seus sinônimos, bem como o entendimento da ideia que o cliente quer expressar quando ele digita um termo para a busca. Isto torna a tarefa de construção das páginas e sua atualização uma fonte constante de preocupação para as áreas responsáveis na empresa. O problema principal desta abordagem é que o uso das Máquinas de busca tradicionais só permite acesso às palavras usadas nas páginas e o seu significado literal, decorrente de um estudo inicial das palavras a serem usadas por ele.

Adicionalmente, quando o cliente está navegando na Agência Virtual e caso não encontre a informação desejada, ele tem que entrar em um buscador do tipo Google e só então retornar às páginas da Light. Esta situação acontece pela inexistência de um mecanismo de busca na atual estrutura web da empresa. É necessário, portanto, criar nas próprias páginas da empresa uma entrada para que o cliente possa fazer a busca especializada, onde na resposta da busca só retornem páginas da Agência Virtual, impedindo desta forma a dispersão da atenção do cliente.

O presente trabalho, desenvolvido no âmbito do Projeto de Pesquisa e Desenvolvimento, descreve a construção de um Mecanismo de Busca Semântica para a Agência Virtual da Light. No próximo item serão descritos os princípios que nortearam a solução, construída para resolver esta questão. A seguir, está resumida a arquitetura do mecanismo de busca semântica, com os seus componentes e seu detalhamento de funcionamento. Em sequência são mostrados alguns exemplos do uso da ferramenta com os resultados encontrados para a busca de palavras mais comuns. E na última seção são apresentadas as considerações finais deste trabalho com a análise das contribuições deste projeto e as perspectivas futuras.

2. Desenvolvimento

2.1. A busca da informação não estruturada

A questão central deste projeto foi fortemente motivada pelo desejo de se obter melhores resultados na busca de páginas da Agência Virtual visando o desenvolvimento de uma ferramenta para esse fim. O ponto focal do projeto foi o de encontrar um mecanismo simples e eficiente de busca de páginas Web sem que se tenha que modificar substancialmente o conteúdo das páginas. Foi importante considerar também a forma como os clientes utilizam a Agência Virtual na busca das informações que solucionem seus problemas, como pagar uma conta vencida e quaisquer outros problemas do cotidiano da relação com a empresa. Os estudos foram

conduzidos de forma a resolver a seguinte questão do ambiente pesquisado:

> *Como buscar páginas da Agência Virtual usando palavras chaves sem que haja necessidade de cadastramento dos termos a priori?*

Das perspectivas descritas neste trabalho, a busca por palavra chave é o principal assunto a ser estudado e precisa ser analisado cuidadosamente. Existem três razões que tornam esta atividade diferente de outras formas de pesquisas textuais. Em primeiro lugar, a busca exige que haja simbiose entre o usuário que redigiu o documento e o que vai buscar pelas palavras inseridas no texto. Em segundo lugar, os indivíduos possuem formas diferentes de interpretação semântica. Por último, a palavra procurada deve ter peso significativo nos documentos alvos. Normalmente, a atividade de busca de documentos não é vista como uma matéria importante até que sua ausência ou sua imperfeição interfira no trabalho das pessoas.

A questão da busca passou a ser a chave do problema dos usuários à medida que a quantidade de páginas cresceu e a sua qualidade se diversificou. Diversos produtos de busca passaram a tentar ocupar este espaço com dificuldades técnicas e comerciais para manter este serviço operativo com disponibilidade e atratividade compatíveis com as necessidades que se sedimentavam. Neste contexto o *site* Alta Vista se popularizou rapidamente criando um mercado e uma forma de financiamento para esta atividade que foi seguida por seus competidores. Nesta época, porém, a questão da abrangência e da precisão da busca passou a ser preponderante na escolha dos usuários por um produto de busca. Estes conceitos passaram a frequentar os estudos desta disciplina e passou a ser importante um *software* que conseguisse maximizar dois indicadores, descritos na figura a seguir.

FIGURA 1 – Conceitos de Precisão e Abrangência de Máquinas de Busca

Os conceitos de Precisão e Abrangência medem a capacidade de indexação do *site* de busca e sua competência para discernir o joio do trigo. Somente 10% (dez por cento) dos usuários de Máquina de Busca passam da terceira página em uma busca e se não encontrarem o que procuram, não voltam a usar o *site*. Estas características obrigam a uma especialização da busca de forma a tentar colocar no topo da lista das informações retornadas para o usuário os documentos mais relevantes para a busca efetuada. As Máquinas de Busca baseiam suas políticas de relevância em uma análise do conteúdo do documento, dentre as quais podemos citar as seguintes:

- Posição e frequências das palavras e frases contidas no texto;
- Sinônimos;
- Análise das URL;
- Data da última atualização;
- Verificação de ortografia;
- Análise do domínio.

Neste contexto, surgiram diversos *sites* que se propuseram a construir um ambiente de busca que alcançasse estas características, porém somente em 1998 foi lançado o produto da Google. Este *site* que rapidamente se tornou o mais popular, utilizando um conceito simples de classificar as páginas nomeado de *Page Rank*. A popularidade e consequentemente a sua relevância passou a ser medida pela quantidade de páginas que apontam para ela, repetindo o efeito de uma votação na Internet. Se o *site* é muito votado, então ele deve ser importante e provavelmente é relevante.

A questão chave do processo de busca está em encontrar páginas relevantes para os atributos fornecidos pelos usuários. Isto apresenta dois problemas principais. O primeiro é causado pela dificuldade de se especificar em palavras o pensamento do usuário. As pessoas têm dificuldades e a linguagem natural ajuda neste processo de colocar os vernáculos adequados e que exprimem o pensamento do interlocutor. O cliente pode digitar que a "luz pifou", mas isto pode não estar adequadamente previsto na construção das páginas. As dificuldades sintáticas e suas associações semânticas causam toda sorte de problemas na definição dos argumentos de pesquisa entregues à máquina encarregada de buscar as páginas armazenadas.

A segunda questão está associada à capacidade de indexação e recuperação do *software* de busca. Esta solução, normalmente por utilizar a existência pura e simples das ocorrências dos termos no corpo do documento, apresenta uma série de deficiências. Técnicas mais sofisticadas usam a posição do termo no texto e a freqüência com que ela ocorre. O Google, para páginas HTML, oferece uma facilidade adicional que é a capacidade de identificar a importância de um documento pela quantidade de acessos e de outras páginas que apontam para este texto.

2.2. A máquina de Busca

A arquitetura proposta e implementada para o Mecanismo de Busca deste trabalho pode ser visto na figura a seguir e é composta dos seguintes componentes:

- *Web Crawler*;
- Base de dados;
- Máquina de busca;
- Árvore similaridade.

O primeiro componente é o *Web Crawler*, ele atua na preparação do ambiente para uso da máquina de busca. O *Web Crawler* é um robô que através do acesso a *internet*, recupera e processa todas as páginas *Web* do ambiente que é alvo de indexação. No caso específico do sistema de busca proposto, o *Crawler* recuperará apenas as páginas *web* que compõem o *site* da Agência Virtual. O resultado desta atividade é uma lista de endereços de páginas contendo a URL (*Universal Resource Locator*) que remetem ao que foi buscado através da ferramenta. O *Crawler* recupera tanto as páginas visíveis pelo usuário quanto as que compõem dados de controle que não aparecem no *site*, tais como, comentários, meta *tags*, descrições, etc.

Estas informações obtidas pelo *Crawler* são armazenadas e indexadas em uma Base de Dados específica de forma a serem acessadas pela máquina de busca e apresentadas ao usuário quando do processamento do resultado final da busca. Estão sendo armazenadas todas as páginas obtidas da navegação a partir da URL https://agenciavirtual.light.com.br/LASView/av/home.do.

A Máquina de Busca é a responsável por tratar as palavras utilizadas em uma consulta do usuário e recuperar as páginas mais relevantes para tal consulta. Por ser um sistema de busca por similaridade essa máquina de busca conta com o apoio de uma árvore de similaridade. O esquemático da arquitetura proposta encontra-se na figura 2.

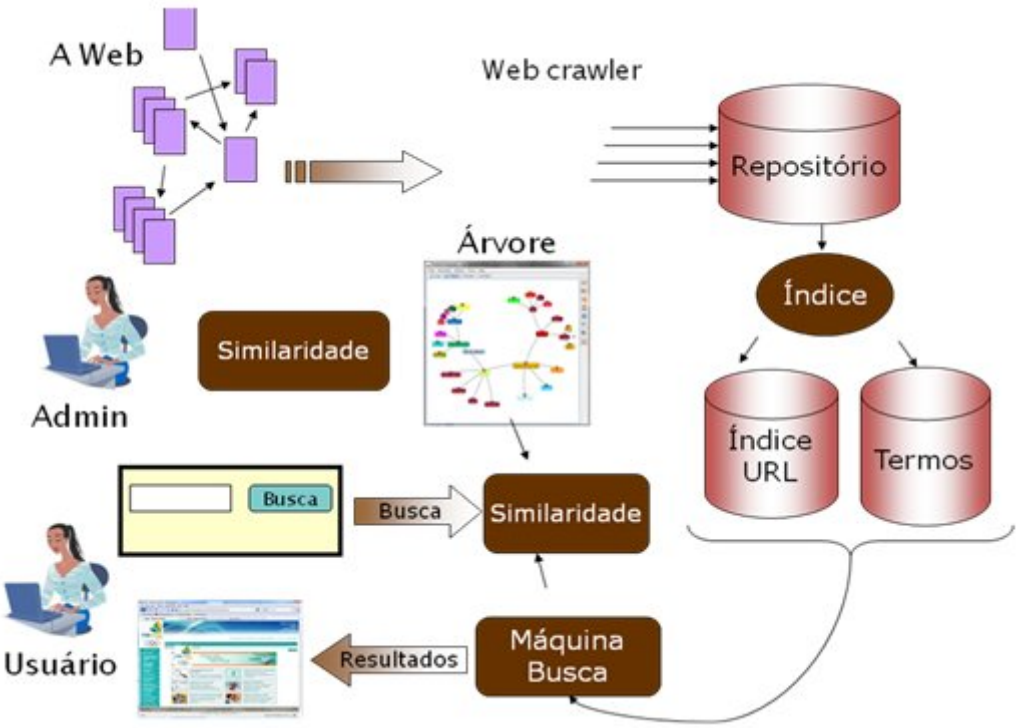

FIGURA 2 – Máquina de busca

2.3. A Árvore de Similaridade

A Árvore de Similaridade é uma estrutura de dados que contém para cada termo, outros termos semelhantes que podem ser utilizados por usuários que possuam diferentes culturas. Por exemplo, um usuário pode usar o termo “casa” para informar o seu interesse em uma busca nas páginas da Agência Virtual. Este termo pode não estar sendo utilizado de forma explícita no *site*, porém na árvore este termo pode estar associado ao termo “contrato”. Desta forma a Máquina de Busca vai encontrar as páginas que contém o termo “contrato” e devolverá para o usuário as páginas relevantes para a solução de seu problema.

A alimentação e a estruturação desta árvore é, nesta versão do produto, uma atividade semi-manual cujo exemplo pode ser visto a seguir. A equipe de administração do produto foi responsável por inserir as palavras similares associando-as às palavras que já constavam na árvore e por filtrar as palavras que não seriam utilizadas pela máquina de busca para as consultas dos usuários (palavras proibidas).Para esta atividade foi utilizado um produto de *software* livre denominado *Treebolic*. A equipe utiliza sua interface gráfica para a criação desta taxonomia e o resultado é um arquivo no formato XML que pode ser lido pelo sistema para o uso em tempo de busca. Alguns padrões para formalização da árvore no *Treebolic* foram utilizados, a saber:

- As palavras são inseridas na árvore no masculino e singular. Como a máquina de busca reduz todas as palavras dos textos rastreados para o radical, palavras como “medidores” e “medidor”, “religarão” e “religar” serão encontradas da mesma forma.
- É permitido inserir palavras compostas de mais de um termo, como por exemplo: “segunda via”, “abertura de contrato”, “interrupção programada”. Essas palavras serão utilizadas em conjunto no momento da busca.

Na literatura existem diversos algoritmos que calculam o peso das palavras em comparação com seus vizinhos. Medidas de distância ou similaridade são convenientes para diferentes tipos de análises. Existem diversas medidas padrões tais como Distância Euclidiana, Distância de *Manhattan*, Distância de Correlação e muitas outras. Neste trabalho foi eleito o método da distância Euclidiana baseada em uma árvore de um Thesaurus construído manualmente pelos usuários. O método usa a proximidade da palavra na árvore para calcular a sua importância e a sua relevância para a busca, conforme Figura 3.

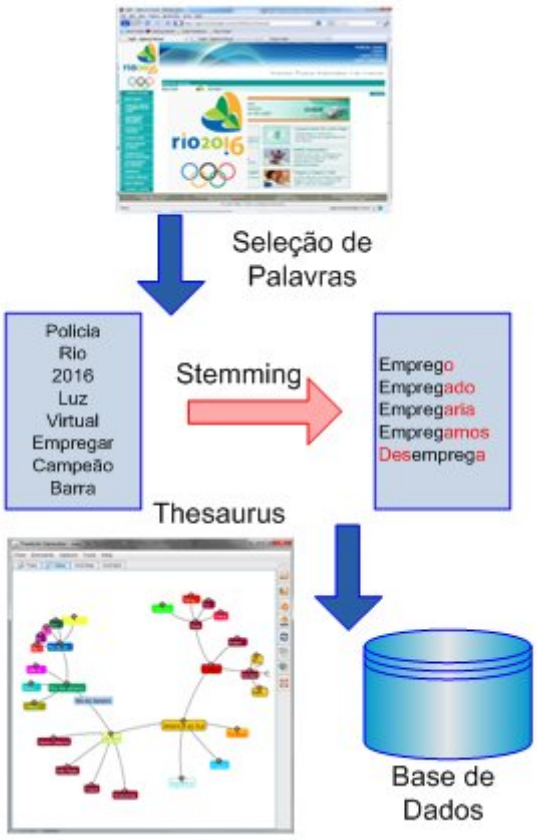

FIGURA 3 – Preparação da Base de Dados

O primeiro passo do processo é a Seleção das Palavras que serão utilizadas tanto na base de Dados quanto na construção do *Thesaurus* que comporão a infraestrutura da Máquina de Busca. As palavras encontradas na página se unirão às da Árvore de Similaridade estruturada no *Thesaurus* para compor a base para a operação da busca. Esta fase retira as palavras das páginas e filtra aquelas que são proibidas e que não são úteis em uma busca. Exemplos de palavras proibidas são os pronomes, preposições, adjuntos e artigos definidos. O próximo passo da construção da infraestrutura da máquina de busca é a fase do *stemming* onde é encontrado o radical das palavras, retirando os sufixos e os prefixos dos termos. Componentes de software especializados são capazes de tirar todas as terminações morfológicas e de flexão da língua portuguesa. Estes radicais encontrados são então armazenados na base de dados para a criação da rede semântica.

2.4. Busca no site da Agência Virtual

O mecanismo de busca está disponível na página principal do ambiente da Agência Virtual da Light na URL https://agenciavirtual.light.com.br/LASView/av/home.do.

O funcionamento é simples, o cliente deve digitar palavras na caixa de texto da busca para efetuar a consulta. Neste momento, o componente de busca é acionado e através da Base de Dados descrita no item anterior, procura as páginas ordenadas por relevância. O resultado da busca aparece para o cliente como uma lista de páginas, sendo que para cada item o sistema coloca o Título da página, sua descrição e a URL absoluta. Convém ressaltar que estas informações existem em cada página da Agência Virtual e foi alimentada pela equipe da Light especificamente para este fim. O cliente pode escolher o *link* que mais se aproxima de sua necessidade e seguir o *hyperlink* apresentado na lista, de acordo com a Figura 4.

FIGURA 4 – Lista do Resultado da Busca

A lista de páginas resultado da busca é apresentada com uma barra verde que representa um indicador visual do nível de relevância da página com relação às palavras fornecidas pelo cliente. Este indicador será mostrado em uma escala de 0 a 100%, sendo 100% o valor de maior relevância. Para facilitar o entendimento deste conceito, esta relevância é mostrada com uma barra verde onde o tamanho da barra indica a dimensão quantitativa da relevância.

3. Conclusões

Este trabalho teve como objetivo relatar a construção de uma Máquina de Busca Semântica e sua arquitetura no ambiente da Agência Virtual da Light. Este *software*, desenvolvido no âmbito do Programa de Pesquisa e Desenvolvimento, facilita o acesso rápido às informações contidas nas páginas do *site* da Agência Virtual, permitindo que o cliente encontre o serviço procurado mesmo que não consiga se expressar com a linguagem técnica utilizada nas páginas. Todas as palavras utilizadas pelos clientes na máquina de busca são armazenadas no *software*, permitindo a inserção de novos termos na Árvore de Similaridade, o que contribui para aproximar cada vez mais a Agência Virtual da linguagem cotidiana dos seus clientes.

Pretende-se, com essa nova funcionalidade, colaborar com o crescimento da participação da *Web*, canal de menor custo operacional, no *mix* do atendimento da empresa.

4. Referências bibliográficas

CARVALHO, G., SOUZA, J. M., MEDEIROS, S.P.J., SOUZA, J.M. (2009) “Collaboration Engineering, Philosophy, and Democracy with LaSca.” In: Proceedings of the 13th International Conference on Computer Supported Cooperative Work in Design - CSCWD 2009.

iProspect Search Engine User Behavior Study,
http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf, 2006.

MEDEIROS, S.P.J., CARVALHO, G., LIMA, E., SOUZA, J.M., (2011) “Locale Similarity Semantic Search in Large Groups Decision MUTIRÕ project for the Rio 2016 Olympic Games.” In: Proceedings of the 15th International Conference on Computer Supported Cooperative Work in Design - CSCWD 2011.

MANNING, C.D., RAGHVAN, P. & SCHÜTZE, H., 2009. Introduction to Information Retrieval Online ed., United Kingdom: Cambridge University Press.

WANGENHEIN, C. G.von, Raciocínio Baseado em Casos, Editora Manole, Barueri, SP, Brasil, 2003.