3
Ferramentas de busca

A *Internet* se transformou em um vasto repositório de informações. Podemos encontrar *sites* sobre qualquer assunto, de futebol a religião. O difícil, porém é conseguir encontrar a informação certa no momento desejado. Desta forma, para auxiliar na busca de conteúdo dentro da *Internet* foram criadas as ferramentas de busca.

As ferramentas de busca são *sites* especiais da *Web* que têm por função ajudar as pessoas na busca por informação armazenada em outros *sites*. Elas têm um papel importante dentro do ambiente da WWW. Mais de um em cada quatro usuários da Internet nos Estados Unidos, algo em torno de 33 milhões de pessoas, fazem buscas em máquinas de busca em um dia típico (Pew).

As ferramentas de busca podem ser divididas da seguinte forma (UC Berkeley):

- Diretórios por assunto.
- Máquinas de busca.
- Meta máquinas de busca.

Os diretórios são organizados em hierarquias de assuntos. São produzidos de forma manual, e por isto abrangem uma parte muito pequena da *Web*. O mais conhecido dos diretórios é o Yahoo! (http://www.yahoo.com.br/) com aproximadamente 2 milhões de páginas categorizadas (UC Berkeley). Esse tipo de ferramenta de busca não será visto aqui.

A máquina de busca é uma ferramenta onde o usuário entra com uma ou mais palavras-chave e a máquina retorna páginas que tenham palavras que casem com a palavra-chave. Diferentemente dos diretórios por assunto, nas máquinas de busca todo o processo é feito de forma automática, o que possibilita a abrangência de uma parcela muito maior da *Web*. Enquanto um diretório de busca grande como o Yahoo! (http://www.yahoo.com.br/) tem 2 milhões de páginas da *Web* categorizadas, máquinas de busca como o Google (http://www.google.com/) conseguem abranger 1.5 bilhões de páginas da *Web* (UC Berkeley).

Assim como na máquina de busca, na meta máquina de busca o usuário também entra com uma ou mais palavras-chave e a máquina retorna páginas que tenham palavras que casem com a palavra-chave, porém o que difere a máquina de busca da meta máquina de busca é como a procura pelas páginas é feita internamente. Na primeira existe um banco de dados indexado pelas palavras encontradas nas diversas páginas, enquanto na última não existe esse banco de dados, sendo utilizadas para a procura outras máquinas de busca.

A seguir fala-se sobre o funcionamento das máquinas de busca e das meta máquinas de busca.

3.1. Máquina de busca genérica

Existe uma gama enorme de máquinas de busca, cada qual com sua forma diferente de fazer o serviço de busca, porém todas executam as seguintes tarefas básicas:

- Percorrem a *Internet*. Devido ao volume de informação existente na mesma, cada máquina de busca só consegue percorrer uma parte da *Internet*.
- Mantêm um índice das palavras encontradas nos diversos *site*, com a URL dos mesmos e outras informações pertinentes.
- Permitem aos usuários procurar por palavras ou combinações encontradas no seu índice.

Apesar de todas as máquinas de busca cumprirem essas três tarefas, o que difere umas das outras é o modo como a tarefa é executada. É por isso que uma mesma busca em máquinas de busca diferentes normalmente produz resultados diferentes.

A seguir será apresentada uma descrição mais detalhada de como são executadas as tarefas acima.

3.1.1. Percorrendo a Internet

Para que uma máquina de busca possa dizer onde um documento HTML está, ela deve antes de tudo achar o mesmo. Para fazer o serviço de "varrer" as páginas da *Internet*, a máquina de busca emprega um tipo especial de agente de software, chamado *spider* ou *crawler*.

Esse tipo de agente tem por função percorrer páginas da *Web* obtendo informações relevantes para a formação e expansão do banco de dados interno da máquina de busca (índice de páginas). Além de percorrer páginas da *Web* que não estejam no banco de dados interno para a expansão deste, eles periodicamente voltam a páginas antigas para verificar se ainda estão ativas ou ficaram obsoletas.

Os pontos de início usuais desse tipo de agente são as páginas mais populares da *Web*. A partir dessas páginas ele vai percorrendo outras, conforme a estratégia de varredura do mesmo. Para cada página o agente monta uma lista de palavras e outras informações pertinentes. Por exemplo, o agente do Google (http://www.google.com/) também leva em consideração a localização da palavra dentro da página como uma informação relevante.

Apesar de esse processo ser executado em todas as máquinas de busca, ele pode variar nos seguintes aspectos:

- Escopo do agente de busca – Todas as máquinas de busca cobrem uma parte diferente da *Web*.
- Profundidade do agente – Uma vez entrando em um *site*, cada agente pode ter uma restrição de profundidade diferente.
- Freqüência de atualização – O agente retorna ao *site* para verificar se houve alguma alteração relevante no mesmo.

Apesar de a maioria dos *spiders* das máquinas de busca varrerem somente o documento HTML, que não fornece nenhuma descrição formal sobre a que o documento realmente diz respeito, algumas dessas máquinas de busca possuem também a capacidade de interpretar *meta tags* (Raggett et al., 1999a) colocadas dentro da página.

Meta tags permitem ao dono do documento especificar palavras-chave e conceitos com os quais o documento está relacionado, de maneira que o mesmo seja indexado de forma mais eficiente. Isso pode ser útil, especialmente em casos em que as palavras no documento podem ter duplo sentido. Com a utilização dessas *tags* o dono do documento pode guiar a máquina de busca na escolha de qual dos possíveis significados é o correto.

Atualmente a utilização dessas *meta tags* pode ser encarada como a única forma de descrição formal dentro da página HTML. Apesar de ser uma descrição formal, a capacidade das *meta tags* (Raggett et al., 1999a) em representar o

conteúdo da página é extremamente pobre em comparação com a capacidade da descrição formal de uma ontologia.

Após esse primeiro momento em que os *spiders* recolhem informações de um conjunto de páginas, passa-se para um segundo momento em que essas são organizadas de modo a facilitar a sua procura.

3.1.2. Mantendo um índice

Uma vez que os *spiders* retornaram informações sobre as páginas, a máquina de busca deve armazenar essas informações de modo a utilizá-las para responder posteriores perguntas dos usuários. Esse processo chama-se indexação.

Esse processo pode variar de uma máquina de busca para outra da seguinte forma:

- Características da indexação – Caso a indexação seja feita de forma automática, ela pode variar em sua sofisticação de forma a melhorar a precisão da resposta. Documentos podem ser indexados por freqüência de palavras e frases, pode-se atribuir pesos para as posições onde aparecem as palavras (por ex.: uma palavra no título da página tem maior peso que uma palavra no texto), ou até por uma análise mais detalhada do documento (Invention Machine Corp., 2000). A utilização das *meta tags* supracitadas também podem ser usadas nessa fase.
- Velocidade de indexação – O processo de indexação consome tempo. Por exemplo, o Altavista (www.altavista.com.br) demora em média 6 semanas até uma URL achada pelo *spider* ser listada em sua base de dados indexada e, portanto, ser passível de ser encontrada (Müller, 1999).

Ao final desse processo teremos uma base de dados indexada com as informações dos diversos *sites* percorridos pelos *spiders* na *Web*.

3.1.3. Construindo a busca

Após os passos anteriores, a máquina de busca é capaz de receber pedidos de busca. Quando um pedido é requerido, a máquina de busca procura no índice

entradas que casem com o pedido de busca e ordena as respostas pelo o que acredita ser mais relevante.

A forma de execução dessa tarefa dentro de uma máquina de busca pode variar nos seguintes aspectos:

- Flexibilidade de formulação da questão pesquisada – A questão a ser pesquisada pode ser desde uma só palavra a até uma combinação de palavras. Para fazer essa combinação um ou mais operadores lógicos podem ser utilizados. Entre os operadores mais utilizados nós temos: AND, OR, NOT, " " (as palavras entre " " são tratadas como frases que devem estar presentes no documento).
- A forma de determinar a relevância das páginas que compõem a resposta ao usuário – Cada máquina utiliza uma técnica diferente para descobrir a relevância de cada página que casa com a pesquisa requerida. Uma técnica comumente utilizada para determinar a relevância do documento é a associação de um peso à página devido à freqüência / localização da palavra procurada dentro da mesma. Técnicas mais avançadas podem ser utilizadas. Por exemplo, o Google (http://www.google.com/) utiliza uma técnica chamada PageRank™. Nessa técnica leva-se em consideração na formação do peso de relevância de cada página a quantidade de *links* de outras páginas que apontam para ela.

A figura 8 apresenta um diagrama de uma máquina de busca genérica.

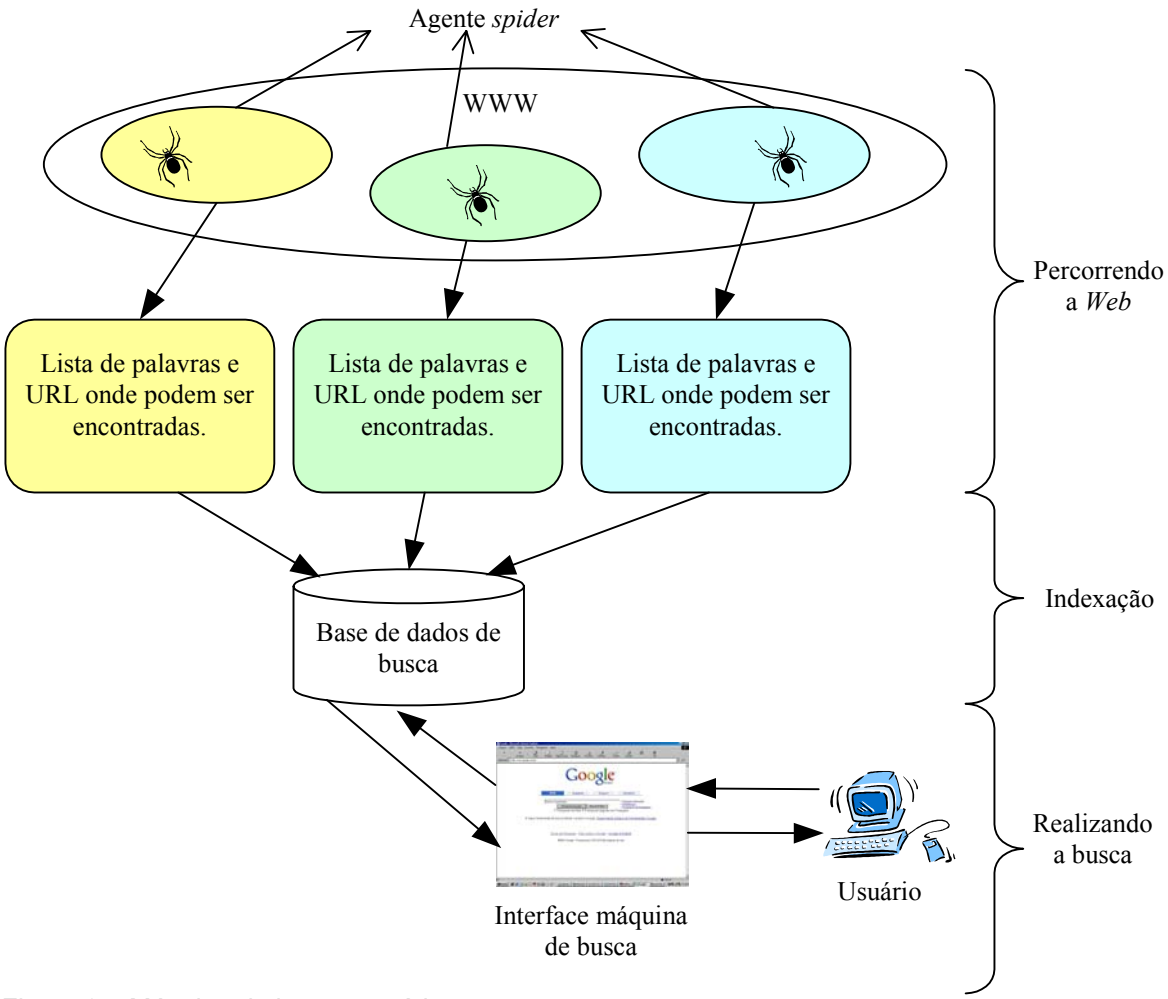

Figura 8 – Máquina de busca genérica.

Todos os 3 passos acima citados podem ser feitos de forma independente pela máquina de busca, pois enquanto os agentes estão recolhendo novas informações na *Internet*, há páginas já percorridas sendo indexadas para o banco de dados, e ao mesmo tempo a máquina de busca está respondendo a pedidos dos usuários.

3.2. Meta máquina de busca

Outro tipo de ferramenta de busca na *Internet* é a meta máquina de busca. Do ponto de vista do usuário ela tem a mesma interface da máquina de busca convencional, porém ela difere do modo como a busca é executada dentro da ferramenta.

Essa ferramenta não possui qualquer banco de dados próprio com *Web sites* indexados. Para responder uma busca, ela utiliza diversas máquinas de busca convencionais existentes.

O funcionamento da meta máquina pode ser dividido em 2 passos principais a serem executados pela mesma:

- Enviar a busca do usuário às máquinas de busca convencionais – Uma vez que o usuário entra com o pedido de busca, este é enviado para várias máquinas de busca convencionais. Nesse passo é necessário que sejam feitas conversões entre o pedido de busca feito na sintaxe da meta máquina para a sintaxe da máquina convencional onde será executada realmente a busca. Por exemplo, se a sintaxe da meta máquina para E lógico for "+" e de uma máquina convencional for "AND", recebendo a busca "futebol + Brasil" esta deve ser convertida para "futebol AND Brasil" a fim de ser enviada a essa máquina de busca convencional.
- Integrar as respostas das diversas máquinas de busca convencionais – Após a execução das buscas nas diversas máquinas convencionais, a meta máquina precisa integrar as diversas respostas retornadas por estas. Para isso são utilizados conversores específicos para cada resposta proveniente de uma máquina convencional diferente. Esses conversores são também chamados de *wrappers*.

A implementação do *wrapper* é a principal dificuldade na implementação desse tipo de ferramenta. Isso porque para cada uma das máquinas convencionais há um *wrapper* diferente. Caso a máquina de busca convencional mude o *layout* HTML de sua resposta, todo o *wrapper* desta terá que ser refeito.

Outro problema que dificulta a implementação dos conversores é que cada um desses tem que converter a resposta em uma resposta canônica padrão da meta

máquina de busca para que seja feita a integração das respostas. Como ele está lidando com um documento HTML que é estruturado orientado a *layout* a extração da informação relevante fica muito vinculada ao *layout* do HTML da resposta.

Apesar de existirem técnicas que diminuem um pouco o problema da implementação dos *wrappers* (Kushmerick et al., 1997), esses ainda apresentam grandes dificuldades.

A figura 9 ilustra o funcionamento de uma meta máquina de busca genérica.

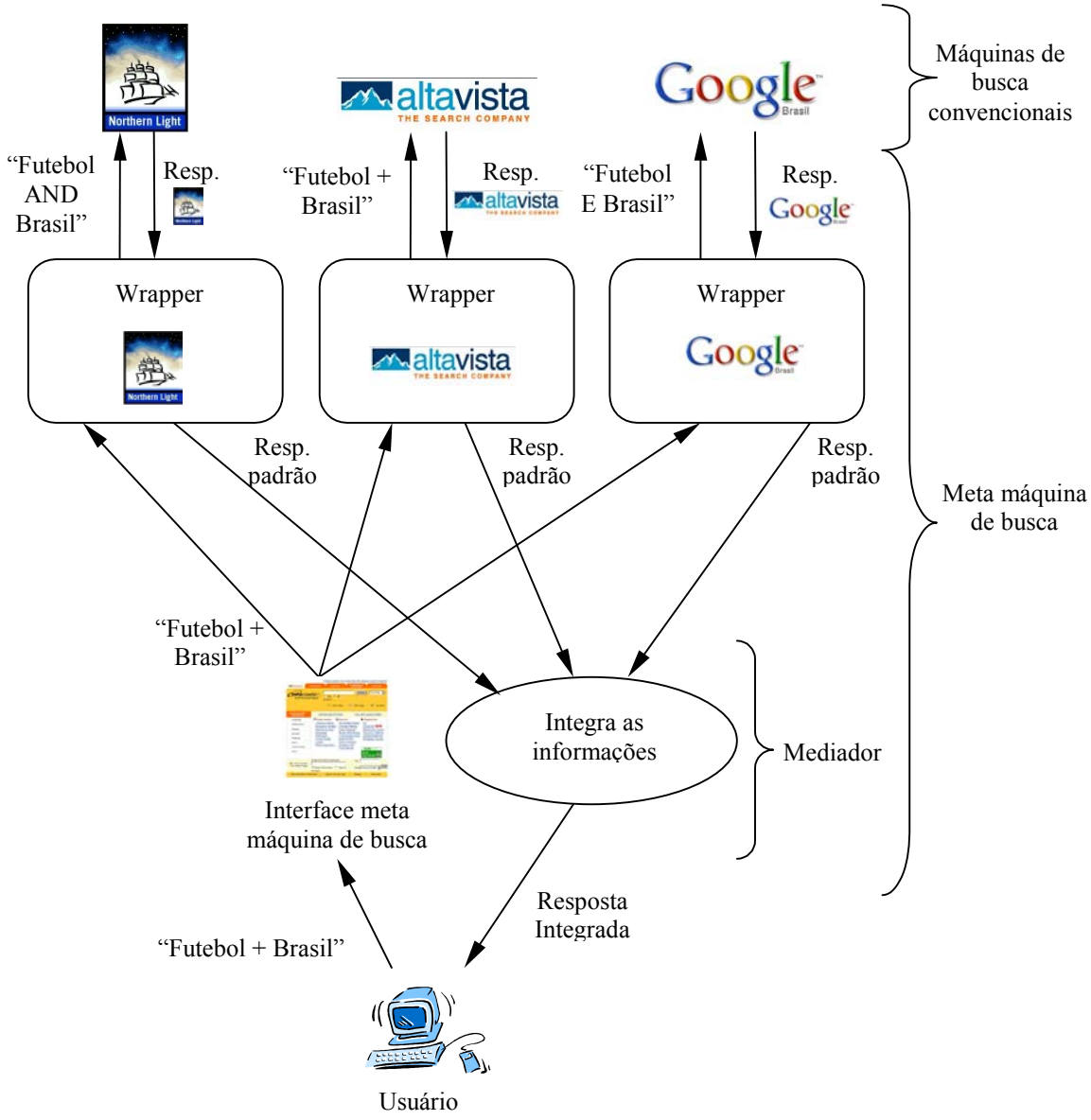

Figura 9 – Meta máquina de busca genérica.

Um detalhe interessante desse tipo de arquitetura é que há um retardo maior na resposta para o usuário do que na busca direta em uma máquina convencional. Isto acontece, pois além de esperar o resultado de várias pesquisas em máquinas

de busca distintas, ainda há a tarefa de converter e integrar as respostas em uma única para o usuário.

3.3.
Softbots

Softbots (robôs de *software*) são agentes inteligentes que usam ferramentas de *software* e serviços como representantes de pessoas (Etzioni, 1996). Em muitos casos os *softbots* utilizam-se das mesmas ferramentas que um usuário humano pode utilizar, por exemplo, ferramentas para enviar e-mail, máquinas de busca, etc.

Meta máquinas de busca podem ser vistas como estando dentro dessa categoria, pois fazem uso de máquinas de busca como representantes do usuário (ao invés do usuário ter que ir a cada máquina de busca a meta máquina faz esse serviço por ele).

Um outro tipo de *softbot* é o *shopbot* que será visto logo abaixo.

3.3.1.
Shopbots

Shopbot (Fensel, 2001; Doorenbos et al., 1997; Etzioni, 1996), ou agente de compra, é um tipo especial de ferramenta de busca voltada para pesquisas em um nicho específico da *Web*.

Eles são agentes que buscam em vários vendedores on-line informações sobre preços e outros atributos de bens de consumo e serviços de forma a facilitar a comparação de atributos na hora da decisão de compra. Os *shopbots* têm uma precisão muito maior que uma máquina de busca genérica, pois estão concentrados em um nicho especifico da *Web* (lojas virtuais de um determinado produto).

Eles conseguem uma extensa cobertura de produtos em poucos segundos, cobertura essa muito maior que um comprador humano paciente e determinado poderia alcançar, mesmos após horas de busca manual.

Através do uso desse tipo de ferramenta consegue-se estabelecer uma interface muito mais amigável entre usuário e máquina para a execução da tarefa de comparação de preços e atributos de produtos que são vendidos pela *Web*.

Pesquisa realizada por R. B. Doorenbos, O. Etzioni, e D. S. Weld (Doorenbos et al., 1997) mostra que, através da utilização dessa ferramenta, o tempo gasto para realizar a tarefa de pesquisar o melhor preço de um produto é muito menor do que através da comparação *site* a *site*.

Podemos dividir os *shopbot* em 3 categorias diferentes conforme os serviços prestados (Fensel, 2001):

- Agentes de compra passivos que procuram informações de produtos baseados no pedido explícito do usuário. Podemos citar como exemplo o Miner (http://www.miner.com.br) que foi desenvolvido pela UFMG.
- Agentes de compra ativos que tentam antecipar os desejos do usuário propondo sugestões. Eles procuram por produtos que podem ser de interesse do usuário tendo por base um perfil do mesmo.
- Agentes de compra que tentam antecipar os desejos do usuário não somente levando em consideração o mesmo, mas também levando em consideração outros usuários.

Para exemplificar o funcionamento de um agente de compra passivo, considere um *shopbot* especializado no domínio de comparação de livros. Nesta ferramenta o usuário inicia a busca determinando as características do livro que procura. Diferentemente das máquinas de busca e das meta máquinas que possuem um campo único e genérico para a busca, nos *shopbots* há um conjunto de campos para indicar as características específicas do produto. No caso de livro características como título, ISBN, autor e preço são possíveis campos de procura. Após receber o pedido, o *shopbot* percorre um conjunto predefinido de lojas on-line fazendo o pedido de busca do livro com as características requisitadas. Após receber a resposta de cada loja on-line, é feita a integração das respostas de cada loja em uma única resposta que é apresentada ao usuário.

A figura 10 ilustra o funcionamento do *shopbot* passivo acima explicado.

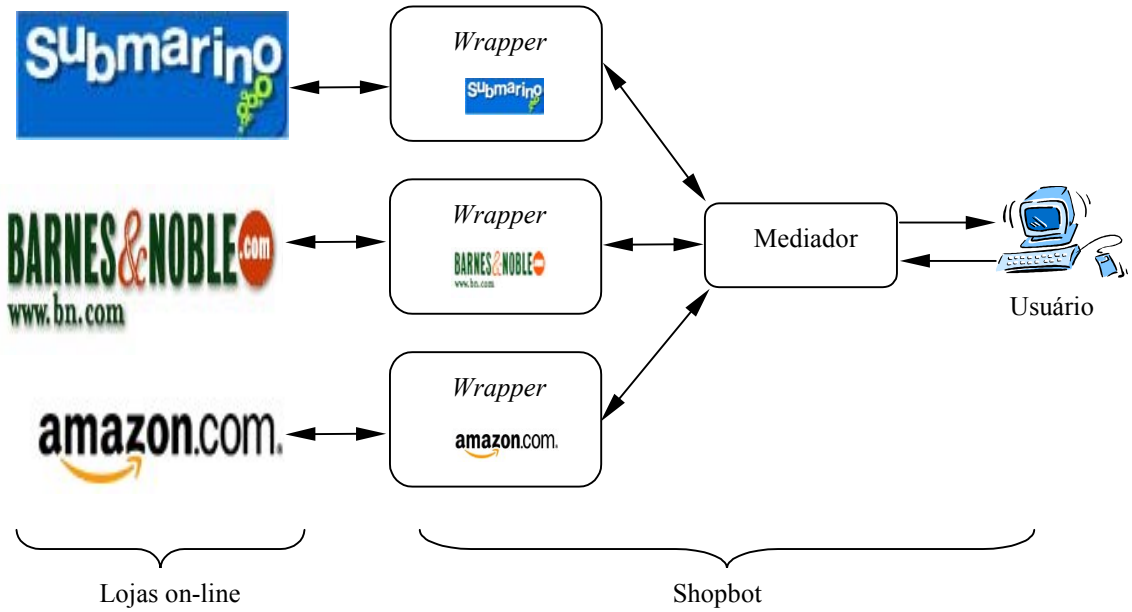

Figura 10 – Agente de compra (Shopbot) passivo de livros.

Podemos ver pela descrição acima que o agente de compra passivo tem uma arquitetura muito parecida com a meta máquina de busca, porém algumas diferenças importantes existem:

- Enquanto a meta máquina de busca tem como domínio todas as páginas da *Web*, o *shopbot* tem como domínio lojas on-line.
- Enquanto o *wrapper* da meta máquina de busca tem que converter uma lista de *sites* retornada por cada uma das máquinas de busca em uma lista canônica padrão, o *wrapper* do *shopbot* tem que converter a descrição de um produto retornada por cada uma das lojas on-line. Como a quantidade e a complexidade semântica dos atributos da descrição de um produto é maior que a complexidade de uma lista de *sites*, o *wrapper* do *shopbot* tende a ser mais complexo que o da meta máquina de busca.

Diferentemente do agente de compra passivo, o agente de compra ativo pode buscar por informações de produtos que podem ser do interesse do seu usuário. Para executar esse tipo de serviço é necessário que ele tenha o conhecimento sobre as preferências do usuário. A arquitetura mostrada na figura 10 também é válida para esse tipo de agente, sendo somente diferente a capacidade do mediador. O mediador deve ter uma implementação mais complexa, porém os *wrappers* continuam sendo os mesmos.

3.3.2.
Wrappers e a Web

Os *shopbots*, assim como todas as ferramentas de busca que lidam com conversores (*wrappers*), têm limitações devido às características impostas pelo ambiente *Web*. Essas limitações têm forte relação com a linguagem HTML, que é a linguagem utilizada neste ambiente. A figura 11 apresenta como essas ferramentas estão organizadas na *Web* (Etzioni, 1996).

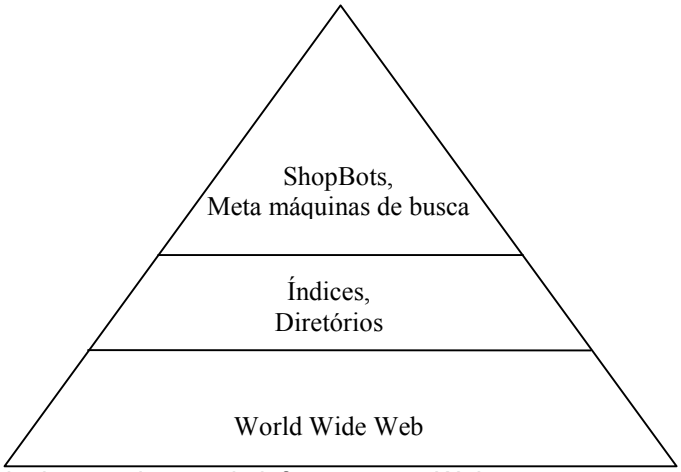

Figura 11 – Níveis de organização da informação na Web.

Para publicar informação com o fim de ter uma distribuição global, é necessário utilizar uma linguagem universal, um tipo de língua mãe que todos os computadores tem o potencial entender. A língua para publicação de documentos usada pela *Web* é o HTML (*Hyper Text Markup Language*) (Raggett et al., 1999b).

O HTML tem as seguintes características principais:

- Permite publicação de documentos com título, texto, tabelas, listas, fotos, etc.
- Recupera informação via *links*.
- Permite criação de formulários que possibilitam a realização de transações com serviços remotos. Por exemplo, possibilita a compra de produtos, busca por informação, etc.
- Permite incluir *video-clips*, som, e outras aplicações diretamente nos documentos.

Um mesmo documento HTML deve poder ser visualizado igualmente em duas máquinas diferentes, em sistemas operacionais diferentes, e em *browsers*

diferentes. A idéia central da linguagem HTML é que ela possa ser uma linguagem universal.

A HTML foi inventada essencialmente para ser uma linguagem de apresentação. Documentos escritos em HTML foram idealizados para serem interpretados pelos *browsers*. Estes têm por função apresentar a informação presente no documento da forma indicada pelas marcações HTML. Normalmente essa apresentação é feita de forma gráfica.

O *browser* sabe somente fazer a apresentação do documento. Todas as informações presentes no documento que não sejam marcações HTML não são entendidas pelo mesmo.

Pelo fato de ser uma linguagem de apresentação, o documento HTML é somente compreensível por seres humanos, pois a HTML foi desenvolvida para ter estes como seus principais consumidores.

Dessa forma, a criação de agentes de software que utilizam informação proveniente desse tipo de documento esbarra em dois problemas:

- Apesar de a informação contida no documento ter uma estrutura intrínseca, esta não é utilizada na estruturação do documento. Dessa forma para um agente de software se todos os documentos têm uma mesma estrutura e esta estrutura nada tem a ver com a estrutura da informação nela contida, é como se o documento não tivesse estrutura.
- Dado um documento HTML qualquer, um agente de software não tem a menor idéia da semântica da informação nele contida. Por exemplo, considere que dois documentos HTML descrevam CDs do Roberto Carlos. Nesses dois documentos o nome do cantor está descrito de forma diferente (por ex.: cantor e interprete). Mesmo que o agente consiga de alguma forma obter a estrutura do documento ele não conseguirá entender que o termo cantor é exatamente a mesma coisa que o termo interprete para o contexto de CDs.

Para poder extrair informação útil de um documento HTML, os agentes de software têm que utilizar conversores (*wrappers*) específicos para cada tipo de documento.

As técnicas atuais para extração de informação de documentos HTML utilizadas pelos *wrappers* (Kushmerick et al., 1997) estão fortemente vinculadas

ao *layout* do documento para obtenção da informação. Qualquer mudança neste *layout* exige que a forma de extração tenha que ser revista.

Outro problema cada vez mais enfrentado pelos *wrappers* na hora da extração da informação é que nem sempre ela está em HTML. Apesar de o documento HTML não ter estrutura e nem informação semântica associadas, através da utilização de certas heurísticas, como por exemplo, *Wrapper Induction* (Kushmerick et al., 1997), é possível conseguir extrair a estrutura da informação. Se a mesma informação fosse fornecida através de uma imagem ou som embutido no documento HTML, a extração desta seria muito mais difícil de ser realizada.

A questão de informações embutidas em imagens era um problema relativamente pequeno no início da *Web*. Isto acontecia devido às pequenas taxas de transmissão entre servidores e usuários existentes na época, que tornava proibitiva a inclusão de muitas imagens e recurso extras no *site*, pois este perderia capacidade de interatividade com o usuário. Dessa forma poucos eram os *sites*, e no nosso caso as lojas on-line, que utilizavam esse tipo de recurso para passar informações. A tendência porém é que isso não venha a ser um grande impeditivo no futuro devido ao aumento das taxas de transmissão dentre servidor e usuário.

Devido às dificuldades supracitadas para o desenvolvimento de agentes que utilizam documentos HTML, é que devemos utilizar uma outra abordagem para descrever informações para serem consumidas por agentes de software. Conforme visto acima dois requisitos são fundamentais:

- Fornecer informações em um formato estruturado.
- Fornecer uma descrição formal da semântica da informação.

Os dois requisitos acima são preenchidos com o uso de ontologias.