ILZA ALMEIDA DE ANDRADE

AS DIMENSÕES SEMÂNTICA E PRAGMÁTICA DA WEB E DOS MECANISMOS DE BUSCA NO CIBERESPAÇO

Londrina
2012

ILZA ALMEIDA DE ANDRADE

AS DIMENSÕES SEMÂNTICA E PRAGMÁTICA DA WEB E DOS MECANISMOS DE BUSCA NO CIBERESPAÇO

Dissertação apresentada ao Programa de Pós-Graduação em Gestão da Informação (Mestrado Profissional) da Universidade Estadual de Londrina como requisito parcial à obtenção do título de Mestre.

Área de Concentração: Gestão e Organização do Conhecimento.

Linha de Pesquisa: Organização e Representação da Informação e do Conhecimento.

Orientadora: Prof.ª Dr.ª Silvana Drumond Monteiro.

Londrina
2012

Catalogação elaborada pela Divisão de Processos Técnicos da Biblioteca Central da Universidade Estadual de Londrina

Dados Internacionais de Catalogação-na-Publicação (CIP)

A553d Andrade, Ilza Almeida de

As dimensões semântica e pragmática da Web e dos mecanismos de busca no ciberespaço / Ilza Almeida de Andrade. – Londrina, 2012.
121 f. : il.

Orientador: Silvana Drumond Monteiro.
Dissertação (Mestrado Profissional em Gestão da Informação) – Universidade Estadual de Londrina, Centro de Educação, Comunicação e Artes, Programa de Pós-Graduação em Gestão da Informação, 2012.
Inclui bibliografia.

1. Ferramentas de busca na Web – Teses. 2. Busca – Teses. 3. Web semântica – Teses. 4. Web pragmática – Teses. 5. World Wide Web (Sistema de recuperação da informação) – Teses. I. Monteiro, Silvana Drumond. II. Universidade Estadual de Londrina. Centro de Educação, Comunicação e Artes. Programa de Pós-Graduação em Gestão da Informação. III. Título.

CDU 025.4:519.68

ILZA ALMEIDA DE ANDRADE

AS DIMENSÕES SEMÂNTICA E PRAGMÁTICA DA WEB E DOS MECANISMOS DE BUSCA NO CIBERESPAÇO

Dissertação apresentada ao Programa de Pós-Graduação em Gestão da Informação (Mestrado Profissional) da Universidade Estadual de Londrina como requisito parcial à obtenção do título de Mestre.

Área de Concentração: Gestão e Organização do Conhecimento.

Linha de Pesquisa: Organização e Representação da Informação e do Conhecimento.

BANCA EXAMINADORA

_____

Profa. Dra. Silvana Drumond Monteiro
UEL - Londrina – PR
(Orientadora)

_____

Profa. Dra. Silvana Ap. Borsetti Gregório Vidotti
UNESP – Marília - SP

_____

Profa. Dra. Maria Elisabete Catarino
UEL - Londrina - PR

Londrina, 29 de novembro de 2012.

Dedico este trabalho aos meus pais,
Angelino e Maria (in memorian),
à minha amada filha Letícia e a
todos familiares e amigos.

AGRADECIMENTOS

Agradeço meus familiares e amigos que sempre me auxiliaram e incentivaram para que eu obtivesse êxito como pessoa e profissional.

À professora *Silvana Drumond Monteiro*, amiga querida e nesta oportunidade, orientadora, que com dedicação e conhecimento profundo do assunto me orientou e mostrou o caminho a trilhar no desenvolvimento desta pesquisa.

Aos membros da banca examinadora, professoras *Silvana Ap. Borsetti Gregório Vidotti* e *Maria Elisabete Catarino*, por aceitarem compartilhar seus conhecimentos, auxiliando no delineamento deste trabalho.

Aos amigos do mestrado profissional, turma 2010, pela amizade e troca de conhecimentos e experiências.

Aos amigos *Decio Wey Berti Junior* e *Geneviane Duarte Dias*, meu agradecimento especial, pela amizade e companheirismo em todos os momentos.

Aos amigos e profissionais da Biblioteca Central que me auxiliaram no desenvolvimento das atividades da Divisão de Circulação durante o período em que estive de licença para os estudos.

À magnífica reitora, professora *Nadina Aparecida Moreno*, que proporcionou com a abertura do mestrado a oportunidade de melhorar nosso nível acadêmico e profissional e, como amiga e comadre, sempre nos incentivou e compartilhou conosco seus conhecimentos e suas experiências profissionais.

Aos professores do Programa de Pós-Graduação em Gestão da Informação (Mestrado Profissional) por compartilharem seus conhecimentos e contribuírem para que obtivéssemos sucesso na conclusão de mais esta etapa do nosso desenvolvimento intelectual e profissional.

Os que se encantam com a prática sem a ciência são como os timoneiros que entram no navio sem timão nem bússola, nunca tendo certeza do seu destino.

Leonardo da Vinci

ANDRADE, Ilza Almeida de Andrade. **As dimensões semântica e pragmática da Web e dos mecanismos de busca no ciberespaço**. 2012. 121f. Dissertação (Mestrado Profissional em Gestão da Informação) – Universidade Estadual de Londrina, Londrina, 2012.

RESUMO

O ciberespaço é um espaço semântico/semiótico, desterritorializado, em constante modificação e a Web, seu principal constructo, tem crescido de forma vertiginosa ao ponto de ser fracionada para fins de estudo, dividindo-se em Web 1.0 ou Sintática, Web 2.0 ou Social, Web 3.0 ou Web Semântica, Web Pragmática etc. Assim, a Web Semântica (WS) é uma representação do conhecimento que tem na sua estrutura as tecnologias para atribuir semântica baseada nas linguagens de programação de modo geral. Porém, a semântica da WS é do tipo formal, mais ligada à sintaxe e a lógica, diferente da semântica da Linguística. A Web Pragmática está sendo construída a partir das experiências da Web 2.0 ou Social, de modo que as pessoas estão contribuindo para sua construção, fazendo o *upload* e uso da linguagem dentre de um contexto. Contudo, somente uma pequena parte da Web está representada por metadados, ontologias etc e, a outra parte, que é a grande maioria, ainda tem o problema da taxonomia do conhecimento e da multiplicidade dos signos. Dessa forma, os mecanismos de busca como tecnologias da informação fazem a indexação automática dos conteúdos da Web; no entanto, tradicionalmente, utilizam palavras-chave ou descrições textuais para processarem a busca. A grande riqueza da indexação maquínica são as múltiplas sintaxes, não obstante, o aspecto semântico é necessário para atribuir sentido a *query*, ou seja, a questão de busca. Nesse sentido, para entender as dimensões da Web e dos mecanismos de busca, utilizamos os conceitos de semântica e pragmática da Linguística e da Filosofia, relacionando o conceito de semântica ao sentido e o conceito de pragmática ao contexto de uso da linguagem. Trata-se de uma pesquisa teórico-informal que tem por objetivo estudar os mecanismos de busca que operam com semântica e a busca contextual. Para tanto, adotou-se a pesquisa documental com abordagem qualitativa e a análise documental como método e técnica tanto para construção do *corpus* teórico quanto para identificação, seleção, definição do *corpus* de análise e exemplificação da busca realizada pelos mecanismos de busca que operam com semântica. Foram analisadas a forma de organização (indexação) e o processo de busca dos mecanismos de busca selecionados. Os resultados demonstraram que a semântica e a pragmática são interdependentes quando se trata do estudo dos mecanismos de busca, porque não dá para desvincular a indexação e a busca, uma vez que o mecanismo é o interpretante da enunciação da busca e o leitor, o interpretante dos resultados. Os mecanismos de busca têm utilizado da colaboração dos leitores tanto na geração do índice quanto na definição de padrões de busca para que os resultados sejam obtidos em um contexto pragmático. Constatou-se que a pesquisa pode contribuir para a categorização ou tipologia dos mecanismos de busca, tanto que sugerimos a readigramação da categoria “apresentação dos resultados”. Por fim, acredita-se que o estudo possibilitará aos bibliotecários o conhecimento das funcionalidades dos mecanismos de busca, uma vez que a tendência atual das interfaces de busca dos catálogos online (OPACs) é facilitar o acesso às coleções em um ambiente similar ao dos *sites* de busca da Web.

Palavras-chave: Mecanismos de busca. Busca. Web semântica. Web pragmática. Ciberespaço.

ANDRADE, Ilza Almeida de Andrade. **The semantic dimensions and the pragmatics of the Web and search engines in cyberspace.** 2012. 121p. Dissertation (Professional Master's Degree in Information Management) – State University of Londrina, Londrina, 2012.

ABSTRACT

Cyberspace is a semantic/semiotic area, deterritorialized and in constant change. The Web, its principal construct, has been fractionated for study purposes because of its vertiginous development. It has been divided into Web 1.0 or Syntactic Web, Web 2.0 or Social Web, Web 3.0 or Semantic Web, Pragmatic Web and so on. Thus, Semantic Web (SW) is a representation of knowledge and holds in its structure the technologies which assign semantic based on programming languages in general. However, the SW semantic is formal. Unlike the Linguistic semantic, it has a stronger connection with syntax and logic. The Pragmatic Web has been built from the experiences of Web 2.0 or Social Web. Thus, people have been contributing to its construction, uploading and using language from a context. Nevertheless, only a small fraction of the web is represented by metadata, ontologies and etc., while its vast majority still has the problem of taxonomy of knowledge and multiplicity of the signs. Thus, search engines, as information technologies, make automatic indexing of the Web content. However, they traditionally use keywords or text descriptions to process the search. Multiple syntaxes are the machinic indexing great value; however, the semantic aspect is necessary to assign meaning to a query, that is, the search issue. In this regard, to understand the Web and search engines dimensions, we use the Semantic concepts and the Philosophy and Linguistics pragmatics, linking the concept of semantic to the meaning and the pragmatic concept to the context of language use. It is an informal and theoretical research which aims at studying the search engines which use semantic and the contextual search. Therefore, a documental research, a qualitative approach and a documental analysis were used as the method and technique both for building the theoretical corpus and for identification, selection, and definition of the analysis and exemplification corpus of the search through the search engines which operate with semantic. The form of organization (indexing) and the search process of the selected search engines were analyzed. The results showed that semantics and pragmatics are interdependent concerning to the study of the search engines due to the fact that both indexing and search are not dissociable. The search is the interpreter of the search enunciation and the reader is the interpreter of the results. Search engines have used readers' collaboration both in index generation and in the search patterns definition so that the results are obtained in a pragmatic context. It was observed that the search can contribute to the categorization, or typology of the search engines. That's why it was suggested the category new diagramming: "presented results". At last, this study will provide the knowledge of the search engines functions to the librarian, since the current trend of the Online public access catalogs (OPACs) is to facilitate the access to the collections in an environment similar to the web search engines.

Keywords: Search engines. Search. Semantic Web. Pragmatic Web. Cyberspace.

LISTA DE FIGURAS

Figura 1 - Evolução na Web 1.0, Web 2.0 e Web 3.0 32
Figura 2 - Processo de indexação realizado pelo mecanismo de busca 45
Figura 3 - Anatomia da busca 63
Figura 4 - Recurso *autocomplete* 71
Figura 5 - "Semantic Search Engines" - *Corpus* inicial 78
Figura 6 - Lista de verificação – Categorias de análise dos mecanismos de busca 79
Figura 7 - Representação do Gráfico do Conhecimento do Google 83
Figura 8 - Caixa do Lexxe para sugestão de chave semântica 85
Figura 9 - "Encontrar a coisa certa" - Gráfico do Conhecimento do Google 90
Figura 10 - "Obter o melhor resumo" – Gráfico do Conhecimento do Google 91
Figura 11 - "Ir mais profundo e mais amplo" – Gráfico do Conhecimento do Google 92
Figura 12 - *Autocomplete* e *autosuggest* do Google 93
Figura 13 - Buscas relacionadas sugeridas pelo Google 94
Figura 14 - Remoção do histórico de busca 94
Figura 15 - Pré-visualização do resultado com o *Google Instant* 95
Figura 16 - Recurso "Estou com sorte" do Google 95
Figura 17 - *Autocomplete* e *autosuggest* do Lexxe 96
Figura 18 - Remoção do histórico de busca do Lexxe 96
Figura 19 - Busca no Google por "indexing used google search engine" 97
Figura 20 - Busca no Hakia por "indexing used google search engine". 97
Figura 21 - Busca no Lexxe por "indexing used google search engine" 99
Figura 22 - Busca no Lexxe com a chave semântica "*search engine*:" 99
Figura 23 - Busca no Lexxe utilizando sugestão da lista "*Related info*" 100
Figura 24 - Busca no Lexxe com a adição da palavra *used* à chave semântica "*search engine*". 100

LISTA DE QUADROS

Quadro 1 - Resumo das abordagens relacionadas à semântica na Web Semântica 23
Quadro 2 - Oposição entre sistemas novos e antigos de recuperação da informação 40
Quadro 3 - Benefícios e problemas das *folksonomias* 42
Quadro 4 - Diferenças entre os mecanismos de busca tradicionais e os semânticos 57
Quadro 5 - Perfil cognitivo do leitor imersivo 61
Quadro 6 - Relação entre o leitor e a forma de elaboração da estratégia de busca 73
Quadro 7 - *Checklist* para construção da *query* 74
Quadro 8 - Mecanismos de busca selecionados e analisados 79

SUMÁRIO

1 INTRODUÇÃO 12

2 O QUE É O CIBERESPAÇO? 17

3 AS DIMENSÕES DA LINGUAGEM: SEMÂNTICA E PRAGMÁTICA 19
3.1 SEMÂNTICA E SENTIDO 19
3.2 PRAGMÁTICA E CONTEXTO 24

4 AS DIMENSÕES DA WEB E A INDEXAÇÃO 30
4.1 WEB SEMÂNTICA 33
4.2 WEB PRAGMÁTICA 37
4.3 INDEXAÇÃO NA WEB 39
4.3.1 Folksonomia 41
4.3.2 Search Engine Indexing 43

5 OS MECANISMOS DE BUSCA 48
5.1 CONCEITOS, DEFINIÇÕES E EVOLUÇÃO 48
5.2 MECANISMOS DE BUSCA SEMÂNTICA 56

6 BUSCA DE INFORMAÇÃO NA WEB 60
6.1 CLASSIFICAÇÃO DA BUSCA 65
6.1.1 Busca Sintática 66
6.1.2 Busca Semântica 69
6.1.3 Busca Pragmática 70
6.2 ESTRATÉGIA DE BUSCA 71

7 METODOLOGIA 76
7.1 ANÁLISE DOCUMENTAL 77
7.1.1 Corpus de Pesquisa 77
7.2 COLETA DE DADOS 79

8 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS 81

8.1 FORMA DE ORGANIZAÇÃO (INDEXING).....81
8.2 PROCESSO DE BUSCA (SEARCHING).....88

9 CONSIDERAÇÕES FINAIS 102

REFERÊNCIAS.....105

ANEXO 117

1 INTRODUÇÃO

O termo ciberespaço foi criado por Willian Gibson em seu livro *Neuromancer* (1984), e passou a ser utilizado para designar o espaço criado pelo computador e pelas redes de informação (LEÃO, 2005; SANTAELLA, 2011b).

O ciberespaço é um ambiente em evolução constante, e a este respeito, Monteiro (2006, p. 32) relata que

> Sendo o ciberespaço o espaço possível de criação de expressões culturais, ou seja, a cibercultura, de transações comerciais, econômicas e sociais, abordaremos o ciberespaço como um espaço semântico/semiótico, onde o signo se dá em várias semióticas, desterritorializado, nômade, em escrita espacializada e com a memória em constante modificação.

Contudo, é na Web, seu principal constructo, que a troca de informações e conhecimentos entre as pessoas acontece com maior intensidade, por ser um espaço social e cultural. Por isso a Web tem crescido de forma vertiginosa, ao ponto de ser fracionada para ser estudada melhor por pesquisadores tanto da Ciência da Computação e da Tecnologia da Informação, quanto por outras áreas, como a Ciência da Informação. No entanto, essa divisão acontece somente para fins didáticos e de estudo, pois a Web é uma só, mas os pesquisadores têm utilizado os termos Web 1.0 ou Sintática, Web 2.0 ou Social, Web 3.0 ou Web Semântica, Web Pragmática etc, para demonstrar seu desenvolvimento.

Nesse percurso, a Web Semântica (WS), que é uma fase da evolução da Web, está sendo estruturada no sentido de atribuir semântica aos conteúdos por meio de linguagens de representação capazes de possibilitar às máquinas (computadores) a busca do conhecimento disponível de forma legível/inteligível. Na Web há uma grande massa informacional estruturada e não-estruturada, bem como não existe uma forma padronizada de descrever todos seus conteúdos[1], o que possibilita o surgimento de inconsistências e ambiguidades durante

[1] Utilizaremos o termo **conteúdo** em detrimento a documentos ou recursos, porque “No ciberespaço, qualquer informação e dados podem se tornar arquitetônicos e habitáveis, de modo que o ciberespaço e a arquitetura do ciberespaço são uma só e mesma coisa.” (SANTAELLA, 2011b, p. 16) e, em razão disso, entende-se que o ciberespaço é povoado por signos, os quais podem se manifestar em uma multiplicidade de formas, uma vez que a linguagem hipermídia é uma “[...] linguagem polivalente que, a par das questões formais de justaposição e associação, também inclui a inter-relação ou colisão entre texto, imagem e som em camadas espaciais e temporais.” (SANTAELLA, 2011b, p. 385).

a recuperação da informação, as quais precisam ser solucionadas de alguma forma. Na Web Semântica, as especificações RDF (*Resource Description Framework*) e OWL (*Web Ontology Language*) são as únicas especificações da tecnologia "[...] construídas de propósito para uso como linguagem de metadados[2], inteiramente dedicadas a descrever e vincular dados de todos os tipos na escala Web." (POLLOCK, 2010, p. 63-64). Por outro lado, a busca e recuperação da informação na Web são aspectos importantes, uma vez que a atribuição de contexto à busca é fundamental, pois com as facilidades que a tecnologia oferece atualmente, os leitores[3] estão cada vez mais autônomos na busca de informação.

A Web Semântica preconizada na literatura, cunhada por Tim Berners-Lee, é uma representação do conhecimento que tem na sua estrutura as tecnologias para atribuir semântica; entretanto, essa semântica utilizada pelas tecnologias da WS, baseada nas linguagens de programação de modo geral, na teoria dos modelos etc., é um tipo de semântica formal[4], mais ligada à sintaxe e a lógica, diferente da semântica da Linguística. Na Linguística, a semântica inicialmente se constituiu como a ciência do significado, entretanto, "[...] a análise de que espécie de coisa é o significado não pode ser feita por meio de um prisma unicamente lingüístico." (FERRAREZI JR., 2010, p. 32), por isso, a semântica passou a caracterizar "[...] o sentido apreendido através das formas e estruturas significantes das línguas." (TAMBA, 2009, p. 10).

Já a Web Pragmática (WP) é uma Web que está sendo construída a partir das experiências da Web 2.0, também conhecida como Web Social. Na Web 2.0 a produção de conhecimento e informação se intensificou não só pela facilidade de

---

[2] "Os metadados são simplesmente formas de enriquecer os dados para que os sistemas de *software* possam interagir com a informação. Os metadados sobre os modelos, vocabulários, e até mesmo as linguagens de programação são simplesmente maneiras de fornecer 'dados sobre dados' para que um intérprete, processador ou algoritmo saiba o que fazer. Não há mágica com metadados." (POLLOCK, 2010, p. 118).

[3] Com o advento do ciberespaço, conforme explica Santaella (2011a, p. 18), "[...] fora e além do livro, há uma multiplicidade de tipos de leitores [...]" e, recentemente, "[...] o leitor das telas eletrônicas está transitando pelas infovias das redes, constituindo-se em um novo tipo de leitor que navega nas arquiteturas líquidas e alineares da hipermídia no ciberespaço." Santaella (2011a, p. 47) designa "leitor", "[...] todo aquele que desenvolve determinadas disposições e competências que o habilitam para a recepção e resposta à densa floresta de signos em que o crescimento das mídias vem convertendo o mundo." É um leitor imersivo que "[...] navega através de dados informacionais híbridos – sonoros, visuais e textuais – que são próprios da hipermídia [...].."; e é nessa linha de pensamento que entendemos que os usuários de outrora devem ser hoje tratados como "leitores".

[4] "[...] estabelecimento de condições de verdade de sentenças referenciais (semânticas extensionais) ou de proposições (semânticas intensionais) cujos sentidos [...] já estão construídos. Assim, ela não trata da questão da construção desses sentidos, mas da análise de sentidos já construídos." (FERRAREZI JR, 2010, p. 136).

uso das ferramentas e aplicações, mas principalmente pela colaboração entre as pessoas, promovendo o aumento da massa informacional no ciberespaço e, consequentemente, dificuldade na busca e recuperação da informação. Todavia, essa colaboração possibilitou a introdução de novas práticas em relação ao emprego da linguagem natural na indexação dos conteúdos e na busca e recuperação da informação. Denomina-se pragmática porque nessa Web as pessoas estão contribuindo para sua construção uma vez que fazem o *upload* e uso da linguagem dentro de um contexto. A pragmática na Linguística diz respeito às condições de uso da linguagem e na Filosofia, para Deleuze e Guattari (1997, v. 2), referem-se à linguagem como agenciamento ou acontecimento do corpo social.

Assim sendo, nessa pesquisa não abordaremos as Webs Semântica e Pragmática como representações conforme apresentadas na Ciência da Computação e na Tecnologia da Informação, porque somente uma pequena parte da Web está representada por metadados, ontologias etc., a outra parte, que é a grande maioria, “[...] a Web mundial sempre terá e trará o problema da taxonomia do conhecimento e da multiplicidade dos signos, seja em sua representação ou organização.” (MONTEIRO, 2008, p. 99). Por isso, utilizaremos os conceitos de semântica e de pragmática, da Linguística e da Filosofia, para estudar as dimensões da Web e dos mecanismos de busca no ciberespaço.

Conforme Santaella (2011b, p. 28), “Navegar pelas redes informacionais é se aventurar por territórios estrangeiros, travar contatos em busca de conhecimento e engendrar subjetividades.” Nesse aspecto, os mecanismos de busca se propõem a organizar o conhecimento e a informação no ciberespaço, facilitando a recuperação nesse ambiente aberto e imenso, onde encontramos conteúdos diversos dentro das áreas de conhecimento e em vários tipos de formatos. Porém, se de um lado temos um conteúdo imenso que pode ser “acessado” com apenas alguns comandos ou “cliques” no computador, do outro encontramos um conteúdo informacional de qualidade questionável recuperado pelos mecanismos de busca *on-line* tradicionais, se compararmos com os sistemas de informação formais, já que as “[...] linguagens antes consideradas do tempo – verbo, som, vídeo – espacializam-se nas cartografias líquidas e invisíveis do ciberespaço, assim como as linguagens tidas como espaciais – imagens, diagramas, fotos – fluidificam-se nas enxurradas e circunvoluções dos fluxos.” (SANTAELLA, 2011b, p. 24).

Os mecanismos de busca são os responsáveis pelo crescimento da Web, principal componente do ciberespaço, entretanto muitas informações não seriam recuperadas se eles não existissem, pois fazem a indexação automática do conteúdo, possibilitando o acesso à informação e ao conhecimento disponíveis, mesmo que a informação recuperada não tenha a precisão[5], a relevância[6] e a qualidade desejada pelo leitor; daí o investimento também em interfaces que proporcionem ao leitor a busca numa perspectiva mais pragmática (contexto).

Entretanto, embora os mecanismos de busca procurem melhorar seu desempenho constantemente, não conseguem indexar todas as páginas web[7], tendo em vista que na Web existem páginas web estáticas[8] produzidas manualmente e facilmente indexáveis e páginas web dinâmicas[9], geradas por computador, de indexação complexa ou não indexável. A indexação na Web é um trabalho gigantesco e interminável uma vez que conteúdos são inseridos e milhões de páginas são publicadas ou atualizadas diariamente. Desse modo, o acesso rápido e preciso à informação torna-se cada vez mais difícil mesmo com a evolução da tecnologia da informação.

Os mecanismos de busca tradicionais utilizam palavras-chave ou descrições textuais para processarem a busca no ciberespaço, e como os textos são descritos em linguagem natural, susceptíveis à ambiguidade e incompletude, essa técnica torna-se limitada (LEUF, 2006; BREITMAN, 2010). As buscas por palavras-

---

[5] Embora mencionados em vários pontos do estudo, nessa pesquisa não é nosso objetivo medir o coeficiente de precisão, bem como a relevância e a revocação. "**Precisão** – A extensão com que os itens recuperados durante uma busca numa base de dados são considerados relevantes ou pertinentes. Um busca que alcance uma precisão alta será aquela em que a maioria dos itens recuperados, se não todos, forem considerados relevantes ou pertinentes. O coeficiente de precisão – uma medida da extensão com que se alcança a precisão – é o número de itens relevantes (pertinentes) recuperados dividido pelo número total de itens recuperados." (LANCASTER, 1993, p. 305-6, grifo do autor).

[6] "**Relevância** – Refere-se à relação entre enunciados de necessidade de informação e fontes potenciais de informação. Por exemplo, considera-se que um artigo de periódico é relevante para um enunciado de necessidade se ele examina o problema ou a situação abrangida pelo enunciado. Essa relação é subjetiva, uma vez que diferentes pessoas tomarão diferentes decisões a respeito de quais itens são relevantes para quais enunciados ou em que medida eles são relevantes para esses enunciados." (LANCASTER, 1993, p. 306, grifo do autor).

[7] A palavra web será utilizada em minúsculo porque se trata de um adjetivo.

[8] **Páginas estáticas** - Uma página estática exibe apenas as informações escritas na página, é um arquivo pré-formatado que exibe somente as informações do arquivo em oposição a uma página dinâmica que pode exibir informações extraídas de uma base de dados. A maioria dos sites de iniciantes são estáticas. Páginas da web não são estáticas, se elas usam uma base de dados para exibir informações. (http://www.feedthebot.com/dynamicpages.html, tradução livre)

[9] **Páginas dinâmicas** - As páginas **geradas automaticamente** por ASP, PHP ou ColdFusion ou outra tecnologia. Bases de dados e muitas "lojas online" são dinâmicas. As páginas mais dinâmicas tem uma "?" na URL. (http://www.feedthebot.com/dynamicpages.html, tradução livre)

chave são sintáticas, ou seja, recuperam o termo exato sem uma análise semântica da palavra, uma vez que "[...] toda análise semântica pressupõe que sejam dadas de antemão informações sintáticas sobre as próprias expressões." (ILARI; GERALDI, 2006, p. 7), porque a semântica meramente da palavra não é capaz de alcançar o sentido, pois despreza o contexto e o cenário (FERRAREZI JR., 2010). A grande riqueza da indexação maquínica são as múltiplas sintaxes, não obstante, o aspecto semântico é necessário para atribuir sentido a "*query*", ou seja, a questão de busca.

Assim, a partir desse contexto, questionamos: Quais mecanismos de busca têm operado com a semântica? As interfaces de busca desses mecanismos possibilitam a busca em um contexto pragmático?

Essas questões conduziram a presente pesquisa, que teve por objetivo geral estudar os mecanismos de busca que operam com semântica e a busca contextual, e por objetivos específicos:

1) identificar e selecionar os mecanismos de busca que operam com semântica;
2) definir um *corpus* e analisá-lo a partir do constructo teórico; e
3) exemplificar como os mecanismos estão processando a busca (pragmática).

Por fim, considera-se que a pesquisa é relevante porque o estudo dos mecanismos de busca é um campo vasto de investigação, pouco explorado e de importante aplicação técnica para a área da Ciência da Informação, especificamente a Organização do Conhecimento, e porque acredita-se que trará contribuições tanto do ponto de vista prático como teórico. No campo prático, a principal contribuição da pesquisa é estimular os bibliotecários a utilizarem a Web como uma fonte de informação e, principalmente, comprovar que os mecanismos de busca são ferramentas auxiliares importantes na busca e recuperação da informação, uma vez que a busca contextual faz parte da sua *práxis*[10]. Para o campo teórico, espera-se que a pesquisa contribua para a área da Ciência da Informação uma vez que desenvolve e apresenta um *corpus* teórico no contexto da organização virtual do conhecimento e da busca e recuperação da informação no ciberespaço, ampliando, dessa maneira, a literatura na área sobre o assunto.

[10] A *práxis* no materialismo histórico significa "[...] conjunto de atividades humanas que engendram não só as condições de produção, mas, de um modo geral, as condições de existência de uma sociedade." (BLIKSTEIN, 1995, p. 54). Entretanto, nessa pesquisa, a *práxis* significa, tão somente, *prática* profissional.

2 O QUE É O CIBERESPAÇO?[11]

A palavra ciberespaço (*cyberspace*) conforme anunciada na introdução desse estudo, foi empregada pela primeira vez em 1984 no romance de ficção científica *Neuromancer* de autoria de William Gibson, e são vários os autores que utilizam o termo para se referir ao mundo digital.

Pierre Lévy (2000, p. 92) define o ciberespaço "[...] como o espaço de comunicação aberto pela interconexão mundial dos computadores e das memórias dos computadores." Ele inclui nessa definição "[...] o conjunto dos sistemas de comunicação eletrônicos (aí incluídos os conjuntos de redes hertzianas e telefônicas clássicas), na medida em que transmitem informações provenientes de fontes digitais ou destinadas à digitalização." Conclui que a codificação digital, "[...] condiciona o caráter plástico, fluido, calculável com precisão e tratável em tempo real, hipertextual, interativo e, resumindo, virtual da informação que é [segundo Lévy], a marca registrada do ciberespaço."

Porém há diversas outras definições, umas contrapondo a de Lévy (2000), como é a de Koepsell (2004), que acredita que o ciberespaço é físico, assim como seus componentes; e outros autores, como Silva e Tancman (1999), Rabaça e Barbosa (2001), e Ramal (2002) que, de certa forma, se assemelham a definição de Lévy.

Para Rabaça e Barbosa (2001, p. 130), o ciberespaço ou espaço cibernético, trata-se de

> Um universo virtual formado pelas informações que circulam e/ou estão armazenadas em todos os computadores ligados em rede, especialmente a Internet. Dimensão virtual da realidade, em que os indivíduos interagem através de computadores interligados.

E para Silva e Tancman (1999), como bem sintetiza Monteiro (2007, p. 6),

> [...] o ciberespaço é uma região abstrata invisível que permite a circulação de informações na forma de imagens, sons, textos, movimentos; um espaço virtual que está em vias de globalização planetária e já constitui um espaço social de trocas simbólicas entre pessoas dos mais diversos locais do planeta.

[11] O conceito de ciberespaço foi discutido por Monteiro (2007) no artigo "O ciberespaço: o termo a definição e o conceito".

Monteiro (2007, p. 14) define o ciberespaço como “[...] uma grande máquina abstrata, semiótica e social onde se realizam não somente trocas simbólicas, mas transações econômicas, comerciais, novas práticas comunicacionais, relações sociais, afetivas e sobretudo novos agenciamentos cognitivos.” A autora supõe que a “[...] compreensão do ciberespaço é mais ampla que a Web e a Internet [...]”, uma vez que entende que “[...] a Web é seu principal constructo, onde convergem as linguagens e a interoperabilidade necessária para efetuação das trocas simbólicas. Já a Internet é entendida [...] como a base técnica e operacional do ciberespaço.”

O ciberespaço, segundo Santaella (2011b, p. 177), é “O espaço que as redes fizeram nascer – espaço virtual, global, pluridimensional, sustentado e acessado pelos computadores [...].” Nesse espaço, um leitor, de qualquer terminal de computador pode acessar não só os fluxos ininterruptos e potencialmente infinitos de informação, mas, sobretudo, pode comunicar-se com qualquer outro leitor em outro ponto da esfera terrestre.

Santaella (2011b, p. 178-9) também afirma que

> O acesso ao ciberespaço se dá por meio de interfaces que nos permitem penetrar nos seus interiores e navegar a bel-prazer pela informação – consubstanciada em linguagens hipermidiáticas, linguagens mistas, híbridas, escorregadias, feitas de misturas de textos, linhas, sinais, gráficos, tabelas, imagens, ruídos, sons, músicas e vídeos – que esses interiores disponibilizam em arquiteturas de conteúdo organizado [...].

Assim sendo, o ciberespaço é “[...] todo e qualquer espaço informacional multidimensional que, dependente da interação do usuário, permite a este o acesso, a manipulação, a transformação e o intercâmbio de seus fluxos codificados de informação.” (SANTAELLA, 2011a, p. 45). Desse modo, tanto no ciberespaço quanto na Web, seu principal constructo, o acesso se dá por meio de interfaces, e a comunicação ocorre por meio de signos e linguagens.

3 AS DIMENSÕES DA LINGUAGEM: SEMÂNTICA E PRAGMÁTICA

Para o estudo das dimensões da Web e dos mecanismos de busca, sentimos a necessidade de conceituar inicialmente os termos, semântica e pragmática, dada a sua importância para alcançar o objetivo da pesquisa de estudar os mecanismos que operam com semântica, bem como os aspectos que envolvem a busca de informação com base no contexto. Assim sendo, buscamos na Linguística, na Ciência da Computação e na Tecnologia da Informação (TI) a distinção do emprego dos termos, embora estas duas últimas não sejam o campo conceitual mais adequado, mas aqui situadas para focos de inteligibilidade para a pesquisa.

Na Linguística encontramos o aporte teórico na Semântica de Contexto e Cenários (SCC), a qual tem por base a concepção de uma língua natural como um sistema de representação do mundo e de seus eventos (FERRAREZI JR., 2010). A abordagem SCC, de acordo com Ferrarezi Jr. (2010), parte do aproveitamento de elementos técnicos e teóricos de diferentes abordagens semânticas mais desenvolvidas para constituir-se em uma abordagem interfacial, semântico-pragmática.

Na Ciência da Computação e na Tecnologia da Informação, a partir do estudo de Almeida e Souza (2011), encontramos a distinção do emprego do termo em contextos diversos nas abordagens da semântica para sistemas computacionais.

No entanto, foi na Filosofia que buscamos a fundamentação necessária à formulação dos pressupostos teóricos básicos norteadores da pesquisa e dos seus resultados, uma vez que é a Filosofia que cria os conceitos (DELEUZE; GUATTARI, 1997, v. 2). Dessa forma, encontramos nas teorias filosóficas de Gilles Deleuze (1997, 2009), Félix Guattari (1997), Françoise Armengaud (2008), Lúcia Santaella (2009, 2011a, b) e Pierre Lévy (1997, 2000), dentre outros, essa fundamentação.

3.1 SEMÂNTICA E SENTIDO

O termo semântica é utilizado na Linguística, na Ciência da Computação, na Tecnologia da Informação e na Filosofia com diferentes acepções.

Semântica, no sentido etimológico, na Linguística, é o “[...] componente do sentido das palavras e da interpretação das sentenças e dos

enunciados [...].” (HOUAISS; VILLAR; FRANCO, 2004, p. 2540), e cabe a ela “[...] descrever a constituição dos sentidos, como passo inicial de seu processo descritivo, e, depois, dos fenômenos que decorrem do sentido e seu uso pelos sistemas linguísticos.” (FERRAREZI JR., 2010, p. 55). De acordo com Ferrarezi Jr. (2010, 133), “Uma semântica viva e reveladora deve ir além, buscando os sentidos da palavra em seu(s) contexto(s) e o(s) deste(s) em seu(s) cenário(s) [...].”

Na Filosofia, segundo Abbagnano (2007, p. 869, grifo do autor) a semântica é a “[...] doutrina que considera as relações dos signos com os objetos a que eles se referem, que é a relação de *designação.*” Na concepção de Deleuze (2009, p. 13, grifo do autor),

> A designação opera pela associação das próprias palavras com imagens particulares que devem “representar” o estado de coisas: entre todas aquelas que são associadas à palavra, tal ou tal palavra à proposição, é preciso escolher, selecionar as que correspondem ao complexo dado.

No âmbito da Ciência da Computação e da Tecnologia da Informação o termo, semântica, foi empregado para representar a evolução da Web e as limitações dos instrumentos de busca. Berners-Lee e Fischetti (2000) esclarecem que a palavra semântica foi utilizada para indicar um tipo de processamento pela máquina relativo à forma do significado e que a Web Semântica é a web das ligações entre as diferentes formas de dados que permitem uma máquina fazer algo que não era capaz de fazer de imediato. Uschold (2003, tradução livre) ao questionar *Onde estão as semânticas da Web Semântica?* concluiu que há muitas respostas para a pergunta:

a) as semânticas são muitas vezes apenas premissas humanas duvidosas derivadas do consenso implícito;
b) são especificações informais dos documentos (p. ex. as semânticas UML ou RDF SCHEMA);
c) são *hardwired* implementados no código (por ex. em ferramentas UML e RDF e agentes web comerciais);
d) estão em especificações formais para ajudar os humanos a compreender ou escrever código (p. ex. uma especificação lógica

modal do significado de *inform* na linguagem de comunicação do agente);

e) são formalmente codificadas para o processamento da máquina (p. ex., *fuel-pump* (*superclasses SHO: pump*));
f) estão na semântica axiomática e *modeltheoretic* da linguagem de representação (p. ex., a semântica formal do RDF).

Entretanto, Uschold (2003) esclarece que existem muitas outras questões importantes para a Web Semântica que não foram abordadas, dentre elas, os serviços web, a marcação semântica, a integração semântica e uso de técnicas de processamento da linguagem natural para recolher a semântica de documentos nessa linguagem.

Recentemente, Almeida e Souza (2011) analisaram e discutiram a semântica em diferentes contextos e avaliaram o espectro semântico de instrumentos para a organização da informação, e propuseram um novo espectro a partir das considerações apresentadas, o qual leva em conta o uso do instrumento por computadores e por pessoas. Entretanto, os autores não fizeram uma revisão exaustiva da literatura e salientaram que importantes pesquisadores da Filosofia e da Linguística com certeza não foram citados.

Almeida e Souza (2011) também destacaram a importância da abordagem linguística da Semântica Formal, uma vez que as abordagens relacionadas à Tecnologia da Informação são tipos de semântica formal; e, que no âmbito da Ciência da Informação é importante esse entendimento do uso do termo para evitar "[...] algum tipo de confusão entre a semântica usada em tesauros e a semântica usada em ontologias." (p. 46). Nesse caso, no entendimento de Miranda (2005, p. 153),

> Semântica é o estudo do significado de conceitos individuais utilizados na linguagem. É uma tentativa de descrever os significados das palavras e as condições sob as quais eles podem interagir para serem compatíveis com outros aspectos de uma linguagem.

Barreto (2001 apud MIRANDA, 2005, p. 186-7)[12] afirma que:

[12] Mensagem de email recebida pelo professor Marcos Luiz Cavalcanti de Miranda, da Universidade Federal do Rio de Janeiro.

> [...] a chave do problema no caso da semântica da Web é estabelecer, em diferentes níveis de qualidade e complexidade, as relações entre os conceitos ou, no caso, entre as páginas da Web que se quer mostrar aos usuários. Este tem sido um problema que os agentes inteligentes (humanos) ainda não resolveram, mas estão trabalhando com afinco nisso. Podemos dizer que na área de Ciência da Informação esta linha de investigação recebeu forte impulso com Farradane, com sua indexação relacional na década de 1960. A Internet com seu encantamento e a força de uma mídia intensa, renomeia problemas e re-inventa soluções, mas não coloca os agentes inteligentes (softwares) para evitar uma pretenciosa duplicação de pesquisa.

Na Tecnologia da Informação, conforme relatam Almeida e Souza (2011, p. 39),

> A maioria das interpretações para a semântica, descritas no âmbito da WS, são nada mais do que tipos de *Semântica Formal*, existindo algumas exceções. Tais exceções são indeterminadas, pois suas descrições não possibilitam verificar sua origem e classificá-las com os mesmos critérios.

Dessa forma, para ilustrar essas abordagens da semântica para sistemas computacionais, Almeida e Souza (2011) com base em diversos autores apresentam um quadro sinótico (Quadro 1) onde na última coluna (semântica linguística) classificam as abordagens da Web Semântica em relação ao tipo de semântica em seu campo de origem.

Observando-se o Quadro 1 percebe-se que as abordagens semântica do mundo real, semântica implícita e semântica informal, que ainda são indeterminadas para os sistemas computacionais, possivelmente só serão determinadas “[...] depois de compreendermos mais substancialmente o ‘bio-fisio-*modus operandi’* do cérebro humano.”, porque “A língua opera com sentidos para que o cérebro humano possa operar com o significado, de forma a haver compreensão.” (FERRAREZI JR., 2010, p. 13, 54, grifo do autor). O sentido, conforme explica Deleuze (2009, p. 18), “[...] não pode consistir naquilo que torna a proposição verdadeira ou falsa, nem na dimensão onde se efetuam estes valores.”, porque ele é o expresso da proposição e não existe fora dela.

Quadro 1 - Resumo das abordagens relacionadas à semântica na Web Semântica.

Abordagem	Breve descrição	Semântica linguística
Repr. do conhecimento	A semântica é formal e baseada em teorias lógico-filosóficas	formal
Repr. do conhecimento	A semântica é o significado de sentenças através de interpretação	formal
Semântica da Web	A semântica possibilita interpretação por um computador	formal
Semântica do mundo real	A semântica mapeia objetos do mundo para o sistema	indeterminada
Semântica axiomática	A semântica mapeia linguagens da WS para a Lógica	formal
Teoria dos Modelos	A semântica valida processos de inferência automáticos	formal
Semântica implícita	A semântica transmite o consenso obtido entre as pessoas	indeterminada
	Semântica inserida em padrões de dados não legível para máquinas	indeterminada
Semântica informal	A semântica é explícita e informal	indeterminada
Formal para humanos	Semântica explícita e expressa em linguagem formal, para pessoas	formal
Formal para máquinas	Semântica explícita e expressa em linguagem formal, para máquinas	formal
	Semântica definida por regras sintáticas mais interpretações	formal
Semântica nebulosa	Semântica baseada em estatística	formal

Fonte: Almeida e Souza (2011, p. 39).

Diante do exposto, nessa pesquisa, a semântica foi eleita como sentido e não significado porque os conteúdos do ciberespaço, gerados em todos os tipos de formatos, estruturas, estilos e linguagens, são atualizados constantemente e essa atualização dificulta a indexação da Web nos mesmos moldes da indexação manual, pois a cada segundo os dados, as informações e os conhecimentos se alteram. Por isso, para indexar esses conteúdos, os mecanismos de busca utilizam uma linguagem apropriada ao mundo digital, a linguagem artificial ou computacional, em oposição a linguagem documentária, que visa padronizar a representação do conhecimento dos documentos com base em uma cultura do impresso, de identidades fixas e sentido único.

No ciberespaço fica evidenciada essa ruptura com o sentido único e com as identidades fixas, ou seja, “[...] o bom senso e o senso comum, respectivamente, uma vez que o sentido é sempre um constructo, um

acontecimento.", porque o ciberespaço é uma máquina semiótica[13] com os signos em constante fluxo, em permanente desterritorialização (MONTEIRO, 2006, p. 32). É fato também que quando se trata do ciberespaço, devemos considerar as novas questões colocadas pela cultura digital voltando a nossa análise às linguagens das tecnologias utilizadas atualmente na representação do conhecimento, as quais possibilitam a organização virtual do conhecimento pelos mecanismos de busca "Uma vez que os algoritmos podem ser processados sem qualquer conhecimento sobre seus significados, eles podem ser processados por máquinas." (SANTAELLA, 2009, p. 58).

Desse modo, verifica-se que a semântica no ciberespaço pode ser analisada melhor sob uma perspectiva sígnica, por meio de "[...] uma estrutura complexa de três elementos íntima e inseparavelmente interconectados: (1.1) fundamento, (1.2) objeto e (1.3) interpretante."[14] (SANTELLA, 2009, p. 43). Assim, no nosso entendimento, um mecanismo de busca é o interpretante da enunciação da busca e o leitor, o interpretante dos resultados obtidos na busca.

3.2 Pragmática e Contexto

A palavra *pragmatismo* (do grego *pragma* = ação) foi introduzida pela primeira vez na Filosofia por Charles Peirce, em 1878, para apresentar uma nova teoria que reconhecia uma conexão inseparável entre a cognição racional e o propósito racional (PEIRCE, 2010; ABBAGNANO, 2007). Porém, para se diferenciar de seus contemporâneos, principalmente de William James e Ferdinand C. S. Schiller, preferiu utilizar o termo *pragmaticismo* para sua filosofia.

De acordo com Peirce (2010, p. 294), o pragmatismo

---

[13] Monteiro (2006) utilizou o termo "máquina semiótica" referindo-se ao ciberespaço para explicar que se trata de um espaço onde existem outras máquinas simbólicas (ou de linguagens) dentro dele.

[14] "(1.1) O fundamento é uma propriedade ou caráter ou aspecto do signo que o habilita a funcionar como tal. (1.2) O objeto é algo diferente do signo, algo que está fora do signo, um ausente que se torna mediatamente presente a um possível interprete graças à mediação do signo. (1.3) O interpretante é um signo adicional, resultado do efeito que o signo produz em uma mente interpretativa, não necessariamente humana, uma **máquina**, por exemplo, ou uma célula interpretam sinais. O interpretante não é qualquer signo, mas um signo que interpreta o fundamento. Através dessa interpretação, o fundamento revela algo sobre o objeto ausente, objeto que está fora e existe independente do signo." (SANTAELLA, 2009, p. 43-4, grifo nosso).

> [...] não pretende definir os equivalentes fenomenais das palavras e das idéias gerais, mas pelo contrário, elimina o elemento sensório destas e tenta definir o propósito racional, e isto ele descobre na conduta utilitária da palavra ou proposição em questão.

Peirce estabeleceu por meio do pragmaticismo a relação entre o signo e seu usuário. Ele afirma que nossas crenças nada mais são do que regras de ação, e que o importante é determinar que condutas o pensamento está apto a produzir, pois "[...] todo pensamento, seja qual for, é um signo, e é fundamentalmente da natureza da linguagem." (PEIRCE, 2010, p. 290).

Conforme Armengaud (2008, p. 28), "[...] Peirce é aquele que fez da vida dos signos e da troca de signos o ambiente vital do espírito e fez da semiótica o continente da linguística." Também possibilitou que a linguagem fosse compreendida sob o paradigma da comunicabilidade, e o sentido função do uso. "A *máxima pragmatista* de Peirce diz exatamente que a produção triádica do significado está orientada para a ação e que a idéia que temos das coisas é apenas a soma dos efeitos que concebemos como possíveis a partir delas." (ARMENGAUD, 2008, p. 10, grifo da autora).

Nessa direção, Bouyer (2010, p. 177) afirma que "Na vida cotidiana, não temos como usar a linguagem independentemente da ação. A linguagem está arraigada em contextos de interação." Para o autor

> No pragmatismo, a verdade de uma proposição depende de seus efeitos práticos e não se mostra completamente independente como na teoria da correspondência à realidade. Uma crença, por exemplo, é tida como pragmaticamente verdadeira quando suas consequências, na vida cotidiana, forem convenientes para aquele que crê. (BOUYER, 2010, p. 176).

Assim, a pragmática como uma das partes da Semiótica, conforme considera Morris (1938), "[...] é precisamente a que compreende o conjunto das investigações que tem por objeto a relação dos sinais com os interpretes, isto é, a situação em que o sinal é usado." (ABBAGNANO, 2007, p. 783). Na Linguística, a pragmática estuda a linguagem em situação de uso, tendo em conta a relação entre os interlocutores e a influência do contexto. Todavia, Deleuze e Guattari (1997, v. 2, p. 21) ressaltam que

> Enquanto a linguística se atém a constantes – fonológicas, morfológicas ou sintáticas – relaciona o enunciado a um significante e a enunciação a um sujeito, perdendo, assim, o agenciamento, remete as circunstâncias ao exterior, fecha a língua sobre si e faz da pragmática um resíduo. Ao contrário, a pragmática não recorre simplesmente às circunstâncias externas: destaca variáveis de expressão ou de enunciação que são para a língua razões internas suficientes para não se fechar sobre si.

Dessa forma, se a pragmática não recorre só às circunstâncias externas, conforme afirmam Deleuze e Guattari (1997, v. 2), o contexto abrange as variáveis externas e internas da produção de sentido uma vez que é o conjunto de pressupostos que tornam possível captar o sentido de um enunciado (ABBAGNANO, 2007). Assim, o sentido de um enunciado não é fixo e pré-determinado,

> [...] resulta de uma estreita negociação entre as várias possibilidades de significação que um certo enunciado possa assumir, quando mergulhado em um contexto histórico-cultural específico, no qual os interlocutores estão sensivelmente inseridos. (XAVIER, 2002, p. 145).

Nesse aspecto, como relatado, nessa pesquisa partimos do pressuposto que o **sentido** está presente na enunciação da busca e no processamento da questão (*query*), uma vez que o leitor estrutura sua estratégia de busca com base em um contexto. Esse contexto, no entanto, é um contexto de uso (pragmática) e não um contexto de condição de verdade do enunciado (semântica). O contexto é sempre pragmático e o sentido está nas duas dimensões da linguagem, ou seja, tanto na semântica quanto na pragmática; entretanto, Armengaud, Deleuze e Guattari e Peirce são teóricos da dimensão pragmática da linguagem, tendo, estes últimos estudado o sentido nesta dimensão.

A enunciação da busca, mesmo com as múltiplas sintaxes dos mecanismos de busca, ainda é realizada predominantemente por palavras pela maioria dos leitores, pois o enunciado, de acordo com Deleuze e Guattari (1997, v. 2), como unidade elementar da linguagem, é a palavra de ordem, e esta é a variável que faz da palavra como tal uma enunciação. Os autores chamam de palavra de ordem

> [...] não uma categoria particular de enunciados explícitos (por exemplo, no imperativo), mas a relação de qualquer palavra ou de qualquer enunciado com pressupostos implícitos, ou seja, com atos de fala que se realizam no enunciado, e que podem se realizar apenas nele. As palavras de ordem não remetem, então, somente a

> comandos, mas a todos os atos que estão ligados aos enunciados por uma ‘obrigação social’. (DELEUZE; GUATTARI, 1997, v. 2, p. 16).

Observa-se que a palavra de ordem, como apresentada pelos autores, pode ser na enunciação da busca, qualquer forma de entrada que o leitor utilize na busca, ou seja, pode ser expressa tanto por palavra-chave, descritor, frase, som, imagem etc. uma vez que há tanta diversidade na forma de expressão. Assim, a enunciação da busca pode ser formulada com base em diferentes contextos, “[...] fazendo variar não apenas o léxico, mas a estrutura e todos os elementos [...], ao mesmo tempo em que as palavras de ordem mudam.” (DELEUZE; GUATTARI, 1997, v. 2, p. 22-3). Nesse caso, na Web, esse modo de enunciação digital naturalmente híbrido, constituído por e com os demais modos de enunciação já existentes - o verbal, o visual e o sonoro - atua paralelamente a eles sem prejudicá-los (XAVIER, 2002, p. 10).

Nesse aspecto, verifica-se que no ambiente digital o conceito de contexto está relacionado a três sentidos diferentes:

a) o contexto como equivalente à situação: refere-se a uma especificação elaborada do ambiente no qual a busca de informação está inserida;
b) o contexto como contingência: refere-se a abordagens de contingência do contexto que estão preocupadas com a especificação chave dos principais fatores situacionais previsíveis do estado de busca de informação pelo produtor;
c) o contexto como estrutura: refere-se aos modos pelos quais o mesmo mundo pode ser visto de forma diferente diante das hipóteses interpretativas dadas. (JOHNSON, 2003).

O primeiro e o segundo sentidos sugerem que aí existem características objetivas de um ambiente que fornece um contexto real.

Contudo, na concepção de Armengaud (2008), o contexto é um conceito central e caracterizante para a pragmática e para distingui-lo a autora propõe a seguinte tipologia:

a) *contexto circunstancial, factual, existencial, referencial:* identidade dos interlocutores, seu ambiente físico, o lugar e o tempo em que suas sentenças são expressas – o contexto é aquele que contém os indivíduos existindo no mundo real;
b) *contexto situacional ou paradigmático:* é culturalmente mediado. A "situação" é qualificada e socialmente reconhecida como comportando uma ou várias finalidades e um sentido imanente partilhado pelos protagonistas pertencentes a mesma cultura. As práticas discursivas se inserem em situações definidas tacitamente ou por proclamação específica. As sentenças nelas proferidas fazem sentido e, transplantadas para outra situação, deixam de fazer sentido e parecem incongruentes;
c) *contexto interacional* – os interlocutores desempenham papéis propriamente pragmáticos: propor, objetar, retratar. Um ato de fala chama outro, mas especificado segundo determinada pressão seqüencial;
d) *contexto pressuposicional* – constituído por tudo o que é igualmente presumido pelos interlocutores: sejam pressuposições, crenças, expectativas ou intenções.

Kobashi e Fernandes (2009, v. 1, p. 5) apontam que a "Abordagem de Armengaud indica que é mais fácil reconhecer a existência de diferentes contextos do que delimitá-los operacionalmente para os fins da organização da informação.", bem como da organização do conhecimento. Nesse sentido, concordamos com a visão de Kobashi e Fernandes (2009, v. 1, p. 4) de que "A pragmática favorece as reflexões sobre os contextos em que ocorrem [...] a análise e a compreensão dos processos de busca e recepção de informação."

Vivemos em um momento de evidente saturação e sobrecarga de informação, e segundo Peterson (2003) exige que a informação seja por nós filtrada, administrada e manipulada para que ela se transforme em conhecimento, convertida em algo ajustado ao contexto em que estamos inseridos, porque como afirma Santaella (2011b, p. 306), "A estrutura do conhecimento de cada indivíduo é idiossincrática, de modo que cada qual deveria estruturar a informação de maneira que lhe faça sentido."

Por fim, nota-se que a busca no ciberespaço que é dispersa, alinear, fragmentada, é, entretanto, individualizada e adequada ao contexto do leitor.

4 AS DIMENSÕES DA WEB E A INDEXAÇÃO

Nosso objetivo não é apresentar um histórico detalhado da Web porque na literatura já existem estudos que abordam o tema com muita propriedade; no entanto, antes de prosseguirmos às questões relacionadas às dimensões da Web, a semântica e a pragmática, é necessário traçar um panorama geral para destacar alguns pontos importantes da sua evolução. Por uma questão meramente didática utilizaremos a divisão da Web em gerações: Web 1.0, Web. 2.0 e Web 3.0, etc.

A Web em um primeiro momento foi marcada pelo crescimento caótico das páginas (*home page*), pelo consumo de dados por humanos e pela falta de ordem e organização por se tratar de um portal de informação carente de contexto, interação e escalabilidade. Dentre os problemas destacamos a Web profunda (*Deep Web*), em que parte da Web não está acessível aos mecanismos de busca em oposição à Web visível, cujo conteúdo pode ser recuperado (SHERMAN; PRICE, 2001). Breitman (2010, p. 2-3), de uma forma pontual, destaca que os maiores problemas da Web são:

> - Grande número de páginas encontradas, porém com pouca precisão [...].
> - Resultados são muitos sensíveis ao vocabulário – em determinados casos, até a ordem em que as palavras são digitadas tem impacto nos resultados. [...]
> - Resultados são páginas individuais – em muitos casos temos um grande número de páginas no resultado que pertencem a um mesmo site. [...]

Entretanto, as principais vantagens da Web estão a descentralização, o compartilhamento de informação, o acesso fácil, a facilidade de contribuir com dados novos, a compatibilidade etc. Todavia, o uso da linguagem natural na representação da informação e do conhecimento é sua principal desvantagem, porque embora exista uma semântica implícita, ela é reconhecida e compreensível apenas por humanos, e complexa para ser processada automaticamente.

Em um segundo momento, a Web se constituiu em uma plataforma centralizada no poder de compartilhamento, no poder de um sistema de organização mais livre (*tagging*)[15] e no estabelecimento de conexões para a integração futura (RSS);

[15] "Tags são estruturas de linguagem de marcação que consistem em breves instruções, tendo uma marca de início e outra de fim. Há tendência, nos dias atuais, de usar as tags apenas como

porém, ainda falta-lhe personalização, portabilidade verdadeira e interoperabilidade. Essa segunda geração da Web, também chamada de Web 2.0 ou Web Social, tem o foco mais nas mudanças que ocorrem com as pessoas e a sociedade do que com a tecnologia. “Com a Web 2.0, as pessoas navegam procurando respostas para problemas complexos, para encontrar novas ideias que desafiem suas opiniões e para encontrar amizade e comunidade entre pessoas que compartilham seus valores.” (POLLOCK, 2010, p. 27) e, por isso, essa Web necessita de mecanismos de busca mais específicos para encontrar amigos, um lugar onde passar as férias ou fotos de outras pessoas na barra dos favoritos, etc.

Na Web 2.0, as pessoas passaram de usuários ou consumidores passivos para produtores ativos de conteúdos em escala global. Para Tredinnick (2006, p. 230, tradução livre), a

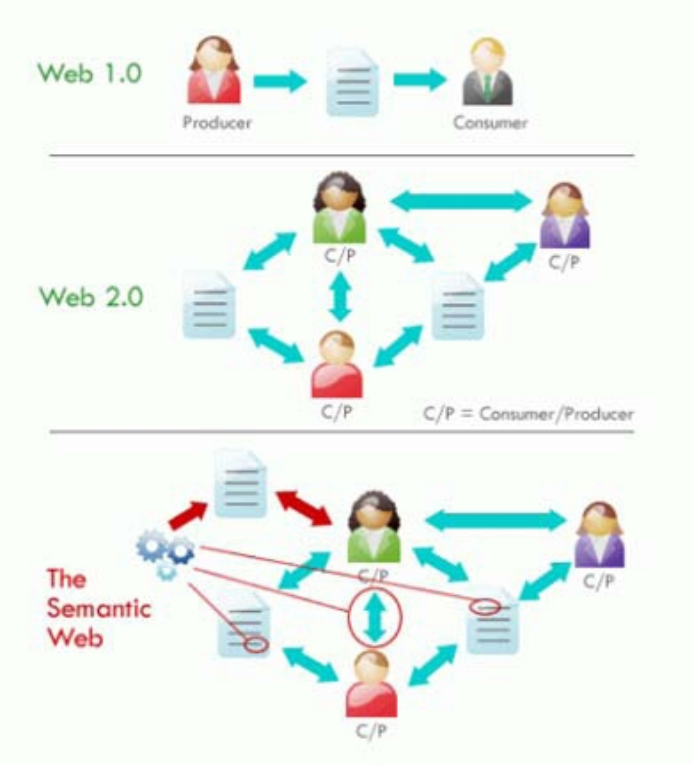
> Web 2.0 também é associada a abordagens novas para administrar a organização e recuperação da informação, como *folksonomies* e *bookmarking* social. Tais abordagens buscam construir estruturas de informação das contribuições de interações de usos.

De acordo com Pollock (2010, p. 27), a primeira geração da Web “[...] foi principalmente sobre publicação de páginas em HTML (Hypertext Markup Language) em um servidor.”, a segunda é ainda direcionada por páginas de documentos, mas o conteúdo é muito mais dinâmico e interativo do que antes que as “[...] páginas eram documentos estáticos que só podiam ser atualizados de formas rudimentares.”, e a terceira geração (atual) o conteúdo é colaborativo. (Figura 1).

---

delimitadores de estilo e/ou conteúdo, tanto em HTML quanto em XML.” (WEINBERGER, 2007, p. 92).

Figura 1 - Evolução na Web 1.0, Web 2.0 e Web 3.0.



Fonte: http://www.uv.es/bellochc/images/web123.jpg

Analisando esse panorama geral da Web, e a Figura 1, observa-se que realmente é preciso melhorar a Web, que ainda é sintática, dotando-a de máquinas capazes de ensinar os mecanismos de busca a compreenderem o sentido que envolve o processamento da linguagem natural, reconhecimento de imagens etc.. Também é necessário dotá-la de informação estruturada para que a representação da informação e do conhecimento sejam compreensíveis por máquinas, ou seja, que os conteúdos sejam expressos em um formato processável automaticamente, porque embora seja esse o objetivo da Web Semântica, ele ainda está longe de se concretizar pela complexidade da linguagem humana. Dessa forma, a Web Semântica, como a Web dos dados, tem seu foco na ligação entre dados para que os computadores façam coisas mais úteis e o desenvolvimento de sistemas possa oferecer suporte a interações na rede, de acordo com a visão da W3C.

Por isso, a Web Pragmática, que também não é uma Web nova, mas uma Web que emergiu a partir da experiência da Web 2.0 tem o propósito de adicionar contexto às informações dos leitores que navegam na rede, de acordo com o uso social da linguagem. Dessa forma, conclui-se que as Web Semântica e

Pragmática não são convergentes nem divergentes, ambas estão preocupadas em melhorar e/ou otimizar a recuperação da informação para o leitor.

4.1 WEB SEMÂNTICA

A World Wide Web Consortium (W3C), consórcio internacional que reúne organizações filiadas, uma equipe em tempo integral e o público, tem trabalhado em conjunto para o desenvolvimento de padrões e diretrizes para a Web. A W3C foi criada por Tim Berners-Lee e outros para ser um consórcio dedicado a construir consenso em torno das tecnologias da Web. As tecnologias da W3C são desenvolvidas para atender as questões de acessibilidade, internacionalização, independência de equipamentos, acesso móvel e garantia de qualidade. (ALESSO; SMITH, 2009; BADR, 2010; W3C, 2011; W3C BRASIL, 2011). A W3C, em 2000, tornou pública por meio de seu mentor Tim Berners-Lee, a primeira proposta sobre a arquitetura da Web Semântica.

Conforme afirmam Berners-Lee, Hendler e Lassila (2001) a “Web Semântica (SW) não é uma Web separada, mas uma extensão da atual, em que à informação é dada um significado bem definido, permitindo que computadores e pessoas trabalhem de forma cooperativa.” Trata-se de uma nova etapa da Web que tem por objetivo que os conteúdos adquiram estrutura e sejam enriquecidos com informação semântica explícita que permita organizar de forma global o conhecimento e que esta informação, independente da apresentação ao usuário, possa ser processada de forma automática por um programa (VIANELLO OSTI, 2004). “A idéia central é categorizar a informação de maneira padronizada, facilitando seu acesso.” (BREITMAN, 2010, p. 5).

Codina e Rovira (2006) apresentam uma definição separada da Web Semântica, porém complementares, uma do ponto de vista da Inteligência Artificial e a outra do processamento robusto. A Web Semântica na visão da Inteligência Artificial é um conjunto de iniciativas destinadas a promover uma Web futura cujas páginas estarão organizadas, estruturadas e codificadas de tal maneira que os computadores sejam capazes de efetuar inferências e raciocinar a partir de seus conteúdos e, na visão do processamento robusto, é um conjunto de iniciativas destinadas a converter a World Wide Web em uma grande base de dados capaz de suportar um processamento sistemático e consistente da informação.

No princípio das pesquisas sobre a Web Semântica, a preocupação era desenvolver linguagens computacionais para estruturar recursos informacionais e descrever aspectos semânticos inerentes a esses recursos. No entanto, após a padronização pelo W3C do XML como linguagem computacional padrão, os engenheiros de *software* começaram a perceber que não era suficiente apenas descrever os recursos informacionais sintaticamente, mas desenvolver tecnologias que permitissem descrever o significado das informações. (BERNERS-LEE; HENDLER; LASSILA, 2001; SHADBOLT; HALL; BERNERS-LEE, 2006; POLLOCK, 2010).

Nessa direção, o propósito da Web Semântica é estruturar o conteúdo significativo das páginas web, criando um ambiente em que agentes de *software* possam percorrer página por página para executar tarefas solicitadas pelos leitores. O objetivo é fazer com que os computadores entendam o conteúdo da Web. Por isso a W3C tem trabalhado no sentido de em um primeiro momento organizar e estruturar as informações e segundo adicionar semântica às informações na Web, de maneira que os agentes de *software* possam compreendê-las (WELLER, 2010).

A Web Semântica como extensão da Web atual possibilitará que os mecanismos e as pessoas trabalhem em cooperação. As características principais que a define, de acordo com Pollock (2010) são:

a) rede onipresente: requer que os dados estejam conectados e entrelaçados sem interesse em sua posição física (desenvolvimento da banda larga, acesso à Internet móvel);
b) abrir tudo: dados linkados (Linking Open Data), serviços, identidade aberta (Open ID), tecnologias – APIs e protocolos, formatos de dados abertos, plataformas de *software* de código aberto e dados abertos (*Creative Commons*, *Open Data License*), dados pessoais abertos (FOAF);
c) informação adaptável ou adaptativa: informações mais conectadas, de granulação mais fina e mais dinâmica;
d) nuvens de serviços adaptáveis: publicação e consumo de dados em nuvem (aplicações de *software* que estão hospedadas inteiramente via protocolos e serviços da Web);

e) dados federados: os dados são armazenados e recuperados a partir de locais diferentes durante uma única consulta;
f) inteligência simulada: introdução de algoritmos melhores para trabalhar com os dados.

Pode-se dizer que a Web Semântica é “muitas coisas” para “muita gente”, mas na realidade ela é um conjunto de tecnologias para a organização, representação e recuperação do conhecimento digital que adicionam semântica interpretável pelas máquinas, com o objetivo de proporcionar um acesso inteligente à informação heterogênica e distribuída da Web, possibilitando aos agentes de *software* fazerem a mediação entre as necessidades dos usuários e os conteúdos disponíveis. (RODRÍGUEZ PEROJO; RONDA LEÓN, 2005). Para isso é necessário construir uma Web de dados com semântica, que implica na adição de informação e conhecimento, para que um mecanismo de busca possa aprender tanto a respeito do que querem “dizer” os dados quanto acerca da informação necessária para processá-los.

Entretanto, Silva (2003) alerta que apesar da proposta da Web Semântica de compartilhar informação estruturada da Web, ela se esbarra nos seguintes problemas:

> - Muitas páginas na Web são desenvolvidas por pessoas não especializadas. De fato, o processo de disponibilizar uma página na Web é simples e rápido e não exige do autor conhecimentos especiais;
> - Não existem ferramentas de estruturação automática da informação que abranjam todos os domínios e assim possam ser usadas por qualquer editor de páginas HTML;
> - As páginas desenvolvidas ainda são direcionadas às pessoas. Mesmo que o autor disponha de conhecimento sobre Web semântica, será ele quem decidirá dispor ou não informações neste formato;
> - A Web semântica exige coerência do conhecimento disponibilizado. Frequentemente, as páginas contêm conteúdo vago e ambíguo não permitindo uma boa estruturação da informação;
> - Páginas publicadas dificilmente serão alteradas para prover informação semântica. (SILVA, 2003, p. 12).

Todavia, observa-se que ao longo dos anos esses problemas de falta de organização e dificuldade na recuperação das informações existentes na Web estão diminuindo em virtude da estruturação e criação de padrões para armazenamento das informações. Nessa direção, Lima (2006, p. 112) relata que

> [...] o processo de autoria de hiperdocumentos tem sido estudado por várias equipes de pesquisadores da informação, como o grupo Word Wide Web Consortium (W3C), com uma preocupação comum de inserir o conteúdo semântico nas páginas. Entre as soluções que surgiram para ajudar os autores de hipertextos a organizá-los semanticamente, estão duas propostas do W3C: o modelo Resource Description Framework (RDF) e a linguagem Extensible Markup Language (XML). [16]

De acordo com Catarino e Baptista (2008, p. 48), "A adição de 'semântica' aos conteúdos da Web permitirá que o intercâmbio e reuso das informações tenham maior qualidade, estejam elas disponíveis em quaisquer tipos de fontes de informações." Dessa maneira, as recomendações da W3C para a Web Semântica são fundamentais para a evolução da Web, não só em quantidade de sites e informações mas, principalmente, em qualidade das informações e serviços disponibilizados aos leitores.

Nesse sentido, deve-se considerar em primeira instância o RDF, porque "[...] foi concebido desde o início para o acesso e utilização ao longo da World Wide Web, e é projetado para proporcionar uma base fundamental para linguagens de dados mais avançadas com um propósito semelhante." (POLLOCK, 2010, p. 77-85). A semântica do RDF pode ser utilizada para especificar a semântica de outras linguagens de dados como por exemplo:

> *a)* *Really Simple Syndication* (RSS): [...] permite aos usuários da Web visualizar algum conteúdo da sua página sem ter, na verdade, que visitar o seu site diretamente. [...] Proporciona uma infraestrutura de distribuição de conteúdos para serem distribuídos e consumidos facilmente;
> *b)* *Friend of a Friend* (FOAF): [...] é um vocabulário legível por máquina para as pessoas descreverem um perfil online de si próprias para vincular em redes sociais, sem a necessidade de banco de dados centralizados ou serviços de terceiros;
> *c)* RDF em atributos (RDFa): [...] é uma forma de codificar dados dentro de páginas web em HTML e XHTML, permitindo assim que as pessoas e máquinas forneçam itens de dados estruturados, embutidos diretamente dentro de páginas da Web;
> *d)* *Web Ontology Language* (OWL): [...] é uma extensão do modelo de dados RDF para fornecer um conjunto muito rico de semântica para a construção de modelos de dados complexos, vocabulários, lógicas de software. (POLLOCK, 2010, p. 77, 79, 81, 84).

[16] A arquitetura RDF é um modelo que permite a representação de dados com um vocabulário distinto para a modelagem da informação. O XML é uma linguagem que fornece um conjunto extensível de marcações que podem ser utilizadas para capturar a estrutura semântica do documento. (LIMA, 2006, p. 112).

As linguagens RDF e OWL como especificações da tecnologia da Web Semântica, construídas para o uso como linguagens de metadados, são utilizadas para originarem outras linguagens, tanto de programação quanto de dados (SPARQL, SWRL, SAML, UML2 ODM, SAWSDL, GRDDL, ISO 15926 – Parte 7, etc). Em resumo, o objetivo do RDF é tornar a semântica de recursos Web acessível a máquinas, porque embora a informação seja lida automaticamente, sua semântica não é definida. A OWL tem o objetivo de atender as necessidades das aplicações para a Web Semântica no que se refere a construção de ontologias, explicitar fatos sobre um determinado domínio e racionalizar sobre ontologias e fatos (BREITMAN, 2010).

Com a utilização das tecnologias (RDF, RDFS, Oil, OWL) e lógicas de descrição, a Web Semântica proporcionará um salto na evolução da Web porque o:

> - Conhecimento poderá ser organizado em espaços conceituais, de acordo com seu significado. Essa organização será assistida por máquinas que serão capazes de fazer a seleção e a filtragem da informação. Ontologias serão cruciais para essa tarefa.
> - Ferramentas automatizadas vão ser responsáveis pela verificação de consistência e mineração de novas informações.
> - Mecanismos de busca baseados em palavras-chave serão substituídos por *queries* sofisticadas. A informação requisitada poderá ser recuperada, extraída e apresentada de maneira amigável. (BREITMAN, 2010, p. 12, grifo da autora).

No entanto, acredita-se que a Web Semântica mesmo desenvolvendo tecnologias sofisticadas para permitir que as máquinas façam o processamento que atualmente é realizado por humanos, não conseguirá realizar a representação da informação e do conhecimento de todo conteúdo da Web, como não resolverá todos os problemas relacionados à busca e recuperação de informação, porque sempre se esbarrará na questão da linguagem natural.

4.2 WEB PRAGMÁTICA

O interesse na Web Pragmática tem sido crescente nos últimos anos, como uma extensão da Web Semântica; entretanto, embora existam diversas pesquisas científicas e aplicações para essa Web, os pesquisadores ainda não chegaram a um consenso sobre a caracterização e definição dos aspectos pragmáticos dessa Web. (AABERGE; AKERKAR; BOLEY, 2011). Acredita-se que a

Web Pragmática está mais associada a Web 2.0 ou Web Social, do que como uma extensão da Web Semântica.

Em termos de evolução da Web, observa-se que a Web Pragmática tem procurado melhorar significativamente a proposta inicial, já que o contexto tem se tornado a cada dia mais importante para o leitor. Na Web Pragmática, conforme Liang, Rong e Liu (2007), o contexto pragmático é a capacidade de combinação das intenções, das comunicações de contexto e de negociação entre os agentes e o leitor. Ela usa o conjunto do contexto pragmático para manipular os recursos semânticos e oferecer aos leitores serviços mais personalizados.

Para Gracioso (2010, p. 287), o conceito de Web Pragmática está relacionado ao "[...] conjunto de processos e produtos gerados a partir do uso social da internet - que refletem em modelagens e configurações da rede [...]". Nessa mesma linha de pensamento, a Wikipédia refere-se à Web Pragmática como um

> [...] conjunto de ferramentas, práticas e teorias que descrevem como e por que as pessoas usam informação. Em contraste com a Web Sintática e a Web Semântica, a Web Pragmática não está focada só no significado da informação, mas também na interação social no significado da mesma como por exemplo: consensos e entendimentos.

Conforme afirmam Repenning e Sullivan (2003, p. 213, tradução livre), "Em contraste com a Web Sintática e a Semântica, a Web Pragmática não é sobre ou para o significado da informação mas sobre **como a informação é usada**." (Grifo dos autores).

Assim sendo, a Web Pragmática transforma a informação existente em informação relevante para um leitor ou grupo de leitores, considerando o modo como os leitores usam, localizam, filtram, acessam, processam, sintetizam e partilham a informação. Em outras palavras, é a Web prática, que se preocupa com a pragmática da interação entre as pessoas, os agentes virtuais que executam *scripts*, buscas e de processamento e os criadores de conteúdo, *websites*. (AABERGE; AKERKAR; BOLEY, 2011).

Desse modo, embora o propósito da Web Semântica seja ter um impacto no mundo real, com suas fontes de significados múltiplos, cambiantes e imperfeitos, a modelagem adequada ao contexto é essencial e, por essa perspectiva, o contexto de uso é o foco da Web Pragmática, porque é fundamental

para lidar com questões como a sobrecarga e a relevância das informações. (MOOR, 2005). Na abordagem pragmática, o controle sobre a representação deve mudar do produtor para o consumidor da informação. No que se refere à busca no ciberespaço, Battelle (2006, p. 226) chama “[...] isso de buscar sua Web pessoal – a busca ampliada por tudo o que você viu, toda consulta que digitou e toda página que salvou para uso futuro ou com a qual interagiu de qualquer maneira.”, uma busca que requer do leitor compreensão, identificação, seleção, decisão e avaliação da informação recuperada devido aos múltiplos sentidos e os diversos contextos em que a informação se apresenta. Desse modo, o leitor deve verificar se as informações recuperadas estão de acordo com o sentido buscado e dentro contexto necessário a produção de conhecimento.

4.3 INDEXAÇÃO NA WEB

Verificamos ao longo da evolução da Web que o trabalho de indexá-la é gigantesco e interminável tendo em vista que milhares de páginas são publicadas diariamente. Desse modo, o acesso rápido e preciso à informação torna-se cada vez mais difícil mesmo com a evolução da tecnologia da informação. Sem dúvida a indexação na Web não é uma tarefa fácil, porque a Web cresce mais rápido do que as tecnologias disponíveis para indexar os *sites* e o seu conteúdo é atualizado continuamente.

A indexação na Web ainda está em desenvolvimento, tanto em termos de definição de conceitos quanto em relação à prática, pois a representação e a organização do conhecimento estão se configurando de maneira diferenciada do ambiente tradicional da linguagem verbal escrita (impressa). Nesses ambientes cognitivos – no ciberespaço e na Web, seu principal constructo – os signos, as linguagens híbridas e os processos informacionais estão contribuindo intensamente para essa nova configuração. A indexação operada tanto na linguagem natural quanto por máquinas semióticas tornou-se o modelo possível de organização do conhecimento, mas “Como conseqüência não há uma sintaxe geral e, por isso mesmo, o fechamento semântico não parece possível, por ser o sentido um reflexo dessa nova linguagem hipertextual.” (MONTEIRO; GIRALDES, 2008, p. 24).

Monteiro e Giraldes (2008, p. 24) afirmam que

> No ciberespaço as máquinas indexam os textos, não mais verbais escritos, mas híbridos, não mais fixos, antes, dinâmicos e desterritorializados, operando essa indexação na equivocidade e na polissemia da Linguagem Natural. Percebe-se a passagem do significado, ou conceito adotado para os múltiplos sentidos.

Essa passagem do significado para os múltiplos sentidos ocorre porque a linguagem natural é facilmente reconhecida por humanos, mas não por máquinas, porque ainda não tornaram possível identificar as relações como estabelecidas na mente quando da recuperação da informação na Web. Dreyfus (2001) ilustra de forma objetiva a diferença entre a cultura do impresso e a cultura do digital, no que se refere à recuperação da informação (Quadro 2), bem como destaca a dificuldade de indexar conteúdos na Web, comparada à indexação praticada na cultura bibliotecária.

Quadro 2 - Oposição entre sistemas novos e antigos de recuperação da informação.

Old Library Culture	**Hyperlinked Culture**
Classificação	**Diversificação**
a. estável	a. flexível
b. organizado hierarquicamente	b. nível único
c. definido por interesses específicos	c. permite todas as associações possíveis
Seleção cuidadosa	**Acesso a tudo**
a. qualidade das edições	a. inclusão de edições
b. autenticidade do texto	b. disponibilidade de textos
c. eliminação de material antigo	c. salva tudo
Coleções permanentes	**Coleções dinâmicas**
a. preservação de um texto fixo	a. evolução intertextual
b. navegação interessada (*browsing*)	b. navegação lúdica (*playful surfing*)

Fonte: Dreyfus (2009, p. 13, grifos do autor, tradução livre).

A Web, como espaço social, permite a livre expressão e, nesse sentido, observa-se que a indexação está sendo realizada por humanos (*folkosonomia*) e por máquinas (*search engine indexing*). Embora existam outras categorias de *Web indexing* ou indexação na Web, como o *back-of-the-book style indexing* e o *subject trees indexing* (TAYLOR; JOUDREY, 2009), nos deteremos às formas de indexação mencionadas, porque retratam essa nova configuração da organização, da representação e da recuperação da informação e do conhecimento no ambiente digital. Entretanto, nessa pesquisa o nosso interesse e análise está relacionado à indexação realizada pelos mecanismos de busca (*search engine indexing*).

4.3.1 Folkosonomia

A Web 2.0 ou Web Social possibilitou que os criadores de conteúdos produzissem seus próprios descritores por meio da aplicação da *folksonomia* e da etiquetagem, utilizando a linguagem natural da comunidade. A *folksonomia* é utilizada para organizar vários tipos de recursos como artigos científicos, referências, *bookmarks*, fotos, vídeos, arquivos de áudio, postagens dos blogues, discussões, eventos, lugares, pessoas, etc. (WELLER, 2010).

De acordo com Catarino e Baptista (2007), "*Folksonomia* é o resultado da etiquetagem dos recursos da *Web* num ambiente social (compartilhado e aberto a outros) pelos próprios usuários da informação visando a sua recuperação." Para as autoras a *folksonomia*, essencialmente, está relacionada a três fatores:

1) é resultado de uma indexação livre do próprio usuário do recurso;
2) objetiva a recuperação *a posteriori* da informação e
3) é desenvolvida num ambiente aberto que possibilita o compartilhamento e, até, em alguns casos, a sua construção conjunta.

Essa forma de organização dos conteúdos produzidos pelos leitores e disponibilizados na Web faz com que haja "[...] pouca precisão na recuperação da informação pois um mesmo termo pode ter significados diversos para os vários usuários que atribuíram as etiquetas." (CATARINO; BAPTISTA, 2007). Assim, nesse tipo de atribuição de etiqueta, "Sempre haverá ambigüidade, porque as marcações são criadas por pessoas comuns que usam palavras que têm significado para elas." (WEINBERGER, 2007, p. 95). Nesse caso,

> A falta de controle de vocabulário, ou seja, o não uso de instrumentos de terminologia tais como listas de cabeçalhos de assunto ou *tesauros,* e de regras gerais para a aplicação das palavras-chave, singular ou plural, termos simples ou compostos causam vários problemas que poderão afetar a recuperação da informação. (CATARINO; BAPTISTA, 2007).

No entanto, como afirma Lancaster (1993, p. 208), "[...] o uso do vocabulário controlado costuma ser preferido pelo especialista em informação, que domina inteiramente as diretrizes e regras que o respaldam, enquanto a linguagem natural conta com a preferência do usuário especialista num assunto." Por isso o uso

da linguagem natural nesse tipo de indexação, não só provoca a dispersão semântica, mas também a equivocidade ou ambiguidade na recuperação da informação.

Conforme Brascher (2002, p. 3), “A ambiguidade pode ser ocasionada por diversos fatores [1]: polissemia, homografia, policategorização, relação contextual e estrutura sintática das frases. Segundo o fator que a ocasiona, a ambigüidade pode ser classificada em diferentes tipos.” De acordo com a classificação de Fuchs (1996), a ambiguidade pode ser morfológica, lexical (homografia, polissemia), sintática, predicativa, semântica e pragmática.

Na visão de Ferrarezi Jr. (2010, p. 242), “[...] certas peculiaridades do contexto e do cenário de enunciação podem gerar uma possibilidade de dupla interpretação pelo interlocutor.” E, por isso, a ambiguidade é “[...] uma operação em que os interlecutores operam com sentidos dos sinais ou com cenários diferentes para um dado contexto.” Entretanto, “Existem senhas sob as palavras de ordem. Palavras que seriam como que passagens, componentes de passagem, enquanto as palavras de ordem marcam paradas, composições estratificadas, organizadas.” Assim, é possível a desambiguação transformando “[...] as composições de ordem em componentes de passagens.” (DELEUZE; GUATTARI, 1997, v. 2, p. 59).

Contudo, além da ambiguidade, as *folksonomias* apresentam outros problemas, porém possuem também benefícios relevantes (Quadro 3).

Quadro 3 - Benefícios e problemas das *folksonomias*.

Benefícios das *Folksonomias*	**Problemas das *Folksonomias***
Folksonomias • representa o uso autêntico da língua, • permite interpretações múltiplas, • reconhece neologismos, • são métodos baratos de indexação, • são a única forma de indexar a informação em massa na Web, • deixa o “controle” de qualidade para as massas, • permite a busca (*searching*) e – talvez melhor ainda – a navegação (*browsing*), • ajuda a identificar comunidades, • são fontes para sistemas de recomendação colaborativa, • são fontes para o desenvolvimento de ontologias, tesauros ou sistemas de classificação, • torna as pessoas sensíveis às questões de indexação de informação.	*Folksonomias* • não possuem controle de vocabulário e não reconhecem sinônimos e homônimos, • não usam relações semânticas entre as etiquetas, • misturam graus diferentes de especificidade, • misturam linguagens diferentes, • não distingue o conteúdo formal das etiquetas descritivas, • Incluem etiquetas *spam*, etiquetas específicas do usuário (que não podem ser interpretadas por outros usuários do que o autor), e outras palavras-chave enganosas.

Fonte: Weller (2010, p. 73, tradução livre), modificado de Peters e Stock (2007b).

Ao analisar o Quadro 3, nota-se que os problemas das *folksonomias* são poucos se comparados aos benefícios que elas geram para o ambiente digital, principalmente por servirem de fontes para o desenvolvimento de ontologias, tesauros e sistemas de classificação, bem como de padrões para os mecanismos de busca.

4.3.2 Search Engine Indexing

A *search engine indexing* é uma categoria de indexação na Web ou *Web indexing*, dentre outras apresentadas por Taylor e Joudrey (2009). Os autores dividem a indexação na Web em três categorias:

> a) *Back-of-the-book style indexing*: as vezes chamada de indexação A-Z, utiliza um índice codificado de *links* dentro do *Web site*;
> b) *Subject trees indexing*: essencialmente classificação, categorias de *Web site* que fornecem palavras-chave para busca (por ex., Yahoo); e
> c) *Search engine indexing*: mais exatamente descrita como a indexação automática de *Web sites* em que: (1) a procura de *Web sites* baseia-se em termos da pergunta do usuário, (2) o índice das palavras encontradas e onde foram encontradas é mantido, e (3) a busca futura por meio das mesmas perguntas usa os índices salvos. (TAYLOR; JOUDREY, 2009, p. 22-3, tradução livre).

A categoria *search engine indexing* é uma forma de indexação automática, na qual o mecanismo de busca utiliza um programa chamado robô, também conhecido por *spider, crawler* ou agente, para coletar automaticamente recursos da Web e registros das bases de dados ou indexar texto completo. O índice criado para a base de dados do mecanismo que é pesquisado quando os leitores entram com o termo na caixa de busca.

Para a criação dos índices existem duas alternativas básicas:

> - O índice pode ser construído manualmente por indexadores profissionais. A vantagem óbvia está na utilização da insubstituível capacidade humana em julgar relevância e categorizar documentos, refletindo diretamente na qualidade do índice gerado e, consequentemente, na precisão da recuperação, desde que exista algum tipo de controle de vocabulário.
> - O índice pode ser gerado automaticamente, permitindo uma cobertura mais ampla e rápida das páginas web. (FERNEDA, 2012, p. 122).

Ferneda (2012, p. 123), afirma que essa indexação automática é realizada em duas etapas:

1. Seleção de endereços (URLs) de páginas;
2. Indexação das páginas, gerando para cada uma um conjunto de termos de indexação.

O uso de algoritmos de *software* para extrair os termos de indexação é o método predominante no processamento da grande massa de conteúdos da Web. Nesse tipo de indexação não há intervenção humana, são os programas que analisam, extraem e atribuem aos conteúdos os termos de indexação. (GIL LEIVA, 2008).

Os mecanismos de busca utilizam técnicas de indexação desenvolvidas nos anos 60, a grande parte, sendo que alguns utilizam listas de *stop words*[17] para eliminar as palavras pouco significativas, outros utilizam técnicas estatísticas ou processamento da linguagem natural para atribuir pesos às palavras e técnicas de extração de radicais (*stemming*) para normalizar os termos de indexação. (FERNEDA, 2012). O processo de como os mecanismos de busca, coletam, indexam, e apresentam os resultados para o leitor pode ser visualizado na Figura 2.

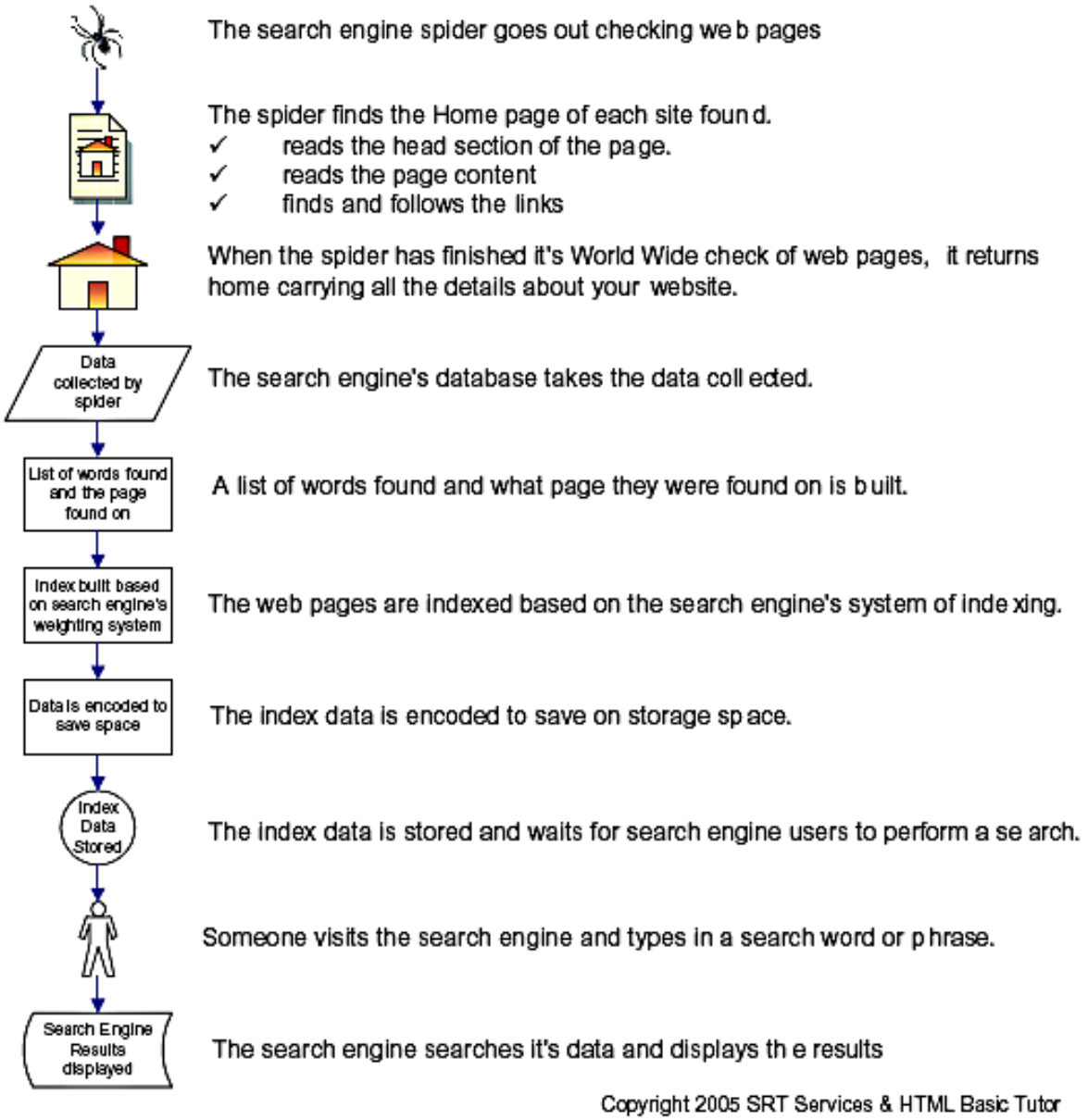
Kuramoto (2006, p. 126) entende que o processo de indexação automática "[...] deveria extrair dos documentos informações que fizessem referência a algum objeto ou fato do mundo real e que pudessem facilitar a sua recuperação, e não extrair símbolos sem referência como o são as palavras." Todavia, Kuramoto (2006, p. 125) refere-se à "[...] recuperação da informação em base de dados que contêm documentos textuais, escritos pelo homem, e que utilizam a sua linguagem, a linguagem natural." Entretanto, os mecanismos de busca também utilizam palavras na indexação, por isso são comuns alguns problemas na recuperação da informação:

a) alto retorno com baixa precisão;
b) baixo ou nenhum retorno;
c) resultados altamente sensíveis ao vocabulário;

[17] Lista com palavras de pouco valor semântico, tais como preposições, artigos, conjunções, etc. (FERNEDA, 2012, p. 125). Para Gil Leiva (2008, p. 329) são "[...] palavras que não fornecem informação temática que possibilitem o processamento automático de texto para fins de indexação, resumo, extração ou classificação de documentos ou recuperação da informação." (Tradução livre).

d) os resultados são apenas páginas da Web;
e) a intervenção humana é necessária para interpretar os resultados;
f) o resultado obtido não é diretamente acessível por outras ferramentas de *software*, exceto por mecanismos de busca por palavras-chave como, por exemplo, Google, Yahoo etc.

Figura 2 - Processo de indexação realizado pelo mecanismo de busca.



Fonte: http://www.htmlbasictutor.ca/search-engine-indexing.htm

No entanto, se considerarmos que os mecanismos de busca, com todas as suas sintaxes e heterogênese, são máquinas capazes de buscar quaisquer elementos, estejam eles em uma biblioteca digital, em um meio de comunicação, no mercado eletrônico, em um espaço para criação de mundos digitais ou em qualquer lugar do ciberespaço (MONTEIRO, 2007), acredita-se que existam outros problemas que ocasionam a obtenção de resultados poucos significativos, além da recuperação por palavras.

Nesse sentido, alguns mecanismos de busca têm adotado novas técnicas de indexação para melhorar os resultados da busca, como é o caso do Google, que "[...] além de utilizar as evidências indicadas por métodos de *ranking*, utiliza outras técnicas, como a análise de vínculos, que proporciona melhoria considerável na precisão dos resultados." (KURAMOTO, 2006, p. 118).

Outros mecanismos de busca utilizam as tecnologias da Web Semântica para melhorar os resultados. Segundo Favaretto (2006),

> Para que o resultado das buscas na Web por meio de *search engines*, receba cada vez mais um toque humano, **diversas novas idéias e teorias** estão em discussão. Além de métodos específicos para **acesso segmentado *da deep Web***, mecanismos de buscas especialistas em determinados assuntos, a teoria da ***Web Semantic*** e sites de **busca em comunidades** ou aqueles que recebem o trabalho participativo ou colaborativo dos internautas. (Grifos do autor).

Nessa direção, Catarino e Baptista (2008) relatam que a Web Semântica adicionará um nível de significado compreensível por máquinas à Web de conteúdos que todos conhecem, pois, atualmente, o significado do conteúdo das páginas não é diretamente acessível por máquinas pela ausência de semântica. De acordo com as autoras,

> A adição de 'semântica' aos conteúdos da Web permitirá que o intercâmbio e reuso das informações tenham maior qualidade, estejam elas disponíveis em quaisquer tipos de fontes de informações. Neste contexto, as recomendações da W3C para a WS são fundamentais para que a Web continue evoluindo, não apenas em termos de volume de sites e, consequentemente, quantidade de informações, mas em termos de qualidade das informações e serviços disponíveis. (CATARINO; BAPTISTA, 2008, p. 48).

No entanto, embora a Web Semântica possibilite a construção de hipertextos mais estruturados semanticamente, é patente a necessidade de uma

modelagem conceitual no momento da construção do hiperdocumento (KURAMOTO, 2006). Mas, por outro lado, a semântica embutida nos documentos permitirá aos dispositivos de recuperação evitar os problemas comuns de polissemia e sinonímia.

5 OS MECANISMOS DE BUSCA

O primeiro robô de busca, o “World Wide Web Wanderer”, foi criado em 1993 por Matthew Gray, do Massachsetts Institute of Tecnhology (MIT), com o objetivo inicial de medir o tamanho da Web. Ele incluiu o recurso de capturar URLs, dando origem a primeira base de dados de *sites*: o Wandex.

Em 1994, Brian Pinkerton (da empresa Next, de Steve Jobs) desenvolveu o Webcrawler, primeira ferramenta a indexar todo o texto dos documentos da Web e, foi essa tecnologia que permitiu o desenvolvimento dos algoritmos de vários mecanismos de busca, como Yahoo e o Google, entre outros (BUSBY, 2004).

Acredita-se que os mecanismos de busca evoluíram em relação ao seu conceito e arquitetura na mesma proporção em que a Web, e investem em tecnologias para otimização dos resultados da busca.

5.1 Conceitos, Definições e Evolução

Os mecanismos de busca são partes integrantes do ciberespaço e de acordo com Monteiro (2008), “[...] recebem várias nomenclaturas na literatura científica, como buscadores, ferramentas de busca, serviços de busca, motores de busca, entre outros [...].”

Conforme Cohen (1998, tradução nossa) o “[...] mecanismo de busca é uma base de dados de arquivos da Internet coletados por um programa de computador (conhecidos como *wanderer*, *crawler*, *robot*, *worm*, *spider*).”

Para Bueno e Vidotti (2000, p. 7) as ferramentas de busca, máquinas de busca ou *search engines*

> [...] são programas computacionais desenvolvidos com o objetivo de indexar informações descritivas e temáticas das páginas e/ou sites da Internet em bases de dados, com a finalidade de possibilitar a recuperação de documentos solicitados, pelos usuários da Internet, segundo as estratégias de busca e os critérios adotados.

A anatomia dos mecanismos de busca é composta por:

- ***Crawler/Spider***: programa que vasculha a Web, link por link, identificando e lendo as páginas. Procura outras páginas relevantes para alimentar e atualizar as páginas do mecanismo de busca.
- ***Index***: base de dados contendo uma cópia de cada página obtida pelo *crawler/spider*.
- **Search engine mechanism**: *software* que possibilita aos usuários consultarem o índice e o qual devolve resultados da busca pela relação numa ordem de relevância. (COHEN, 1998, BERRY; BROWNE, 2005, WEN-CHEN et al., 2005, v. 5).

Nahuz (1999), Berry e Browne (2005) e Wen-Chen et al. (2005, v. 5) apresentam o mesmo tipo de estrutura composta por três elementos básicos descrita por Cohen (1998), contudo, Nahuz classifica os mecanismos em: a) mecanismos de busca (busca direta e indireta), b) aplicação e c) conteúdo. Já Cendón (2001) adota o termo ferramentas de busca, porém não apresenta qualquer definição. De acordo com Müller (2003, p. 31, tradução nossa),

> [...] um mecanismo de busca é uma base de dados pesquisável de arquivos da Internet coletados por um programa de computador. Um mecanismo de busca consiste em três partes: uma aranha [*spider*], um índice e um mecanismo de busca.
> A aranha (às vezes chamada de *crawler*, *robot*, *worm* ou *bot*) é um programa que percorre a Web link por link, identificando e lendo páginas e as armazena no índice.
> O índice é um banco de dados que contém uma cópia de cada página da Web coletada pela aranha.
> O mecanismo de busca é o *software* que o permite examinar o índice e normalmente devolve os resultados classificados em ordem de relevância. O programa recebe seu pedido de busca, o compara com as entradas no índice e retorna os resultados para você. Não há nenhum critério de seleção para a coleção de arquivos, entretanto uma avaliação pode ser aplicada à classificação dos resultados.

Zanier (2006, p. 22), em uma concepção mais evoluída dos mecanismos de busca, afirma que

> A anatomia dos serviços de mecanismo de busca é similar, formado de componentes como o Web *Crawler* (Rastejador), *Document Index* (Índice de Documentos) e *Query Processor* (Processador de Consultas), entre outros. Um mecanismo de busca consiste de seis componentes principais:
> 1. *Crawler* ou *spider* que examina os Web *sites*;

> 2. Índice de documentos com a listagem dos Web *sites*;
> 3. Cachê [sic] de documentos que armazena páginas Web;
> 4. Processador de consultas;
> 5. Sistema de *ranking* (raqueamento) de documentos;
> 6. *Software* de interface, interrogação e recuperação. (Grifos do autor).

Moura (2001), diferentemente de Cohen (1998), Müller (2003), Favaretto (2006) e Zanier (2006), prefere o termo “sistemas de busca”, que, segundo ele é

> [...] um conjunto organizado constituído de computadores, índices, bases de dados e algoritmos tudo isso reunido com a missão de:
> - analisar e indexar as páginas da Web,
> - armazenar os resultados dessa análise e indexação numa bases de dados e mais
> - quando de uma consulta de um usuário, o sistema de busca vai pesquisar a sua base de dados e
> - fornecer os resultados da pesquisa ao usuários.

Contudo, Moura (2001), assim como Cohen (1998), Nahuz (1999) e Müller (2003) nos mostra que os “sistemas de busca” apresentam três componentes principais:

> - programa de computador denominado ***robot***, ***spider***, ***crawler***, ***wanderer***, ***knowbot***, ***worm*** ou ***web-bot***. Aqui, vamos chamá-los indistintamente de **robô**. Esse programa "visita" os sites ou páginas armazenadas na Web. Ao chegar em cada site, o programa robô "pára" em cada página dele e cria uma cópia ou réplica do texto contido na página visitada e guarda essa cópia para si. Essa cópia ou réplica vai compor a sua base de dados.
> - o segundo componente é a **base de dados** constituída das cópias efetuadas pelo robô. Essa base de dados, às vezes também denominada índice ou catálogo, fica armazenada no computador, também chamado servidor do mecanismo de busca.
> - o terceiro componente é **o programa de busca propriamente dito**. Esse programa de busca é acionado cada vez que alguém realiza uma pesquisa. Nesse instante, o programa sai percorrendo a base de dados do mecanismo em busca dos endereços - os URL - das páginas que contém as palavras, expressões ou frases informadas na consulta. Em seguida, os endereços encontrados são apresentados ao usuário. (Grifo do autor).

Observa-se que em relação à anatomia dos mecanismos de busca é consenso entre os autores a composição em três componentes básicos: *crawler/spider*, *index* e *search engine mechanism;* porém seu conceito evoluiu e se ampliou, pois para Monteiro (2006, p. 34),

> [...] os indexadores (mecanismos de busca) da Internet, como modelo de organização do conhecimento, detêm os mesmos atributos do rizoma, operando a multiplicidade do sentido existente na forma hipertextual (ou rizomática) à recuperação da informação e do conhecimento, não incorrendo no bom senso (sentido único) e senso comum (identidade fixa), ambos elementos da doxa, ou seja, no fechamento semântico (do significado/conteúdo) e físico (da materialidade/forma) das obras e bibliotecas que deram origem à referência fixa do conhecimento.

Diante dos conceitos e definições ora apresentadas, reconhece-se os mecanismos de busca como ferramentas da tecnologia da informação capazes de indexar e buscar a informação disponível nesse imenso universo digital que é o ciberespaço, basicamente o conteúdo disponível na Web visível[18].

No que se refere à evolução dos mecanismos de busca, observa-se que ao longo do desenvolvimento das tecnologias da informação, e dos estudos no âmbito da Ciência da Informação, a tipologia desses mecanismos evoluiu das categorias baseadas na forma geral de organização ou indexação à categoria do paradigma semiótico, apresentada por Monteiro (2008, 2009b) a partir da tese das múltiplas sintaxes de organização e busca do conhecimento no ciberespaço.

No entanto, Monteiro (2012) apresenta outras categorias além da abordada (Diagrama 1), mas nosso objetivo ao destacá-la é para ressaltar o início da evolução dos mecanismos de busca com os diretórios ou catálogos e com os programas ou robôs de busca, por terem marcado o início de uma possível organização virtual do conhecimento e destacar o paradigma semiótico como o reflexo de um momento “[...] em que o conhecimento está rompendo com a cultura verbalista e o ciberespaço desterritorializando os signos, permitindo todas as hibridizações possíveis.” (MONTEIRO, 2009b, p. 91).

[18] Refere-se ao conteúdo disponível e acessível pelos mecanismos de busca. Na Web existem conteúdos que estão nas *intranets*, nas bases de dados proprietárias, nos bancos de dados do governo etc., que são acessíveis mediante senhas porque são informações sigilosas ou proprietárias, e não estão disponíveis para indexação e busca pelos mecanismos de busca.

Diagrama 1 - Tipologia dos mecanismos de busca de acordo com Monteiro (2012).

CATEGORIAS	1) **ANATOMIA**	*Crawling* (varrer) *Indexing* (indexar ou gerar o índice a partir da base de dados) *Searching* (buscar através da interface de busca)
	2) **FORMA GERAL DE ORGANIZAÇÃO OU INDEXAÇÃO DOS MECANISMOS DE BUSCA** (*indexing*)	Diretórios ou Catálogos Programas ou robôs de Busca Híbridos
	3) **ORDENAÇÃO DOS RESULTADOS** (*searching*)	Localização Freqüência do termo Análise de *links* Relevância Pagos, orgânicos e híbridos
	4) **APRESENTAÇÃO DOS RESULTADOS** (*searching*)	Agrupamento ou Clusterização: a) Verbais b) Visuais Especializados Personalizados Ontologias Federados Web Profunda Web Semântica
	5) **PARADIGMA SEMIÓTICO** (*indexing* e *searching*)	Sonoros Visuais e Geo-referenciais a) Espacial Verbal escrito Híbridos

Fonte: http://www.uel.br/grupo-pesquisa/ciberespaco/cubos/cubo2.html

Assim sendo, a categorização quanto ao paradigma semiótico constitui o diferencial na abordagem tipológica apresentada por Monteiro (2009b) porque utiliza a Semiótica peirciana para o estudo das linguagens nos mecanismos de busca e apóia-se nas três matrizes da linguagem-pensamento, a sonora, a visual e a verbal, apresentadas por Santaella (2009). De acordo com Monteiro (2008, p. 117), “[...] os mecanismos podem ser classificados segundo o paradigma em que operam seus algoritmos de busca.” Nesse caso, a tipologia dos mecanismos de busca baseada no paradigma semiótico seria classificada de acordo com a “[...] forma analógica de busca, ou seja, usar o som para buscar o som, a forma para imagem, diretrizes espaciais para localização geográfica e a palavra para os textos.” (MONTEIRO, 2009b, p. 92).

Os mecanismos de busca classificados dentro das categorias elaboradas por Monteiro (2009b) foram pesquisados, exemplificados e discutidos, exceto os mecanismos da Web Semântica, que aparecem na categoria de apresentação dos resultados (*searching*), os quais são objetos desta pesquisa.

Na categoria da forma geral de organização ou indexação, se enquadram os diretórios ou catálogos que precederam os programas ou robôs de busca e surgiram quando a quantidade de recursos disponíveis ainda permitia a coleta das informações manualmente, segundo Branski (2004), e constituem a primeira solução para organizar e localizar os recursos da Web. São listas de assuntos organizadas com a ajuda de editores em categorias e subcategorias (base de dados), setores de atividade econômica ou ramos do conhecimento, geralmente com uma estrutura hierárquica (árvore), classificadas por assunto com tópicos de interesse amplo (educação, esporte, entretenimento, viagens, compras, etc) para atender um público variado (CENDÓN, 2001).

De acordo com Moura (2001), os **Diretórios** possuem dois componentes:

a) uma base de dados, também chamada de índice ou catálogo e
b) um programa de computador que faz a pesquisa na base de dados.

São características dos diretórios ou catálogos:

a) localização da informação: navegação nas categorias através do mouse e/ou busca via formulário e palavras-chave;
b) editores tomam conhecimento de novos *sites* através de sugestões de usuários, pesquisas na Internet (listas de anúncios de novas páginas) ou robôs;
c) *sites* coletados passam pela seleção de editores, o que pode indicar qualidade dos dados;
d) apenas os melhores recursos informacionais são escolhidos para inclusão;
e) necessidade de um grande número de editores;
f) grandes diretórios podem conter dezenas de milhares de categorias e subcategorias;

g) a montagem ou criação da base de dados de um diretório é realizada por humanos. São os humanos, que fazem a análise e a indexação dos *sites* da Web;
h) mantêm em suas bases de dados apenas um resumo do conteúdo dos sites por ele catalogados (CENDÓN, 2001; BRANSKI, 2004; MOURA, 2001).

Já os programas ou robôs de busca surgiram quando o número de recursos na Web adquiriu proporções que impediam a sua coleta manual e também a busca através de navegação (BRANSKI, 2004), e o resultado de uma busca é classificado e apresentado por um método conhecido como “relevância”. Como características dos programas de busca, destacamos:

a) os documentos encontrados pelos robôs são encaminhados aos indexadores, que extraem a informação das páginas *html* e as armazenam em uma base de dados;
b) com relação a localização da informação: uma página Web é usada para efetuar a pesquisa na base de dados; o usuário formula a consulta por meio de combinações de palavras-chave, que é transmitida ao motor de busca propriamente dito; o programa (mb) localiza na base de dados os itens que devem constituir a resposta e ordena os resultados colocando os mais relevantes em primeiro lugar na lista de resultados (descrição dos *sites* e *links*); cada mecanismo de busca utiliza método próprio de classificação;
c) foco na abrangência das bases de dados, que podem alcançar centenas de milhões de itens, e não na seletividade;
d) o usuário pode sugerir sua URL ao em vez de esperar que o *site* seja encontrado pela varredura do robô (ou robôs trabalhando em paralelo) (CENDÓN, 2001; BRANSKI, 2004; MOURA, 2001).

Dentro da mesma categoria supracitada surgiram os *mecanismos de meta busca, metabuscadores* ou *metamotores*, que são constituídos de componentes como algoritmos de fusão, organização e filtragem dos resultados encontrados, interface de consulta e apresentação, e, conexão com os mecanismos de busca

utilizados (HUANG et al., 2001, RUNG; MING; CHUNG, 2005, v. 5, LEUF, 2006). O surgimento dos metabuscadores foi motivado pelo paradigma de que melhores resultados em uma pesquisa (busca) são obtidos com o uso de várias ferramentas diferentes. Nesse particular, Stanley (1998) considerou que os metabuscadores constituiriam o próximo estágio dentro da cadeia alimentar da informação dos mecanismos de busca convencionais, porque possibilitam uma abordagem sofisticada da busca em vários bancos de dados, com as seguintes vantagens:

a) acesso a uma única página web para formular a busca;
b) necessidade de conhecimento somente da interface de uma página para a busca;
c) a estratégia de busca é formulada só uma vez;
d) os resultados permitem redirecionar a busca a outros mecanismos;
e) obtém-se resultados integrados, a partir de vários mecanismos.

Os metabuscadores ou metamotores não possuem base de dados de páginas web próprias. Eles enviam as consultas de forma simultânea às bases de dados mantidas por outros mecanismos de busca ou diretórios, combinam e apresentam as respostas de todos eles ao mesmo tempo, mas, elimina resultados duplicados ou triplicados e geram uma lista final. Esses mecanismos fornecem uma interface que permite ao leitor formular uma busca e “clicar” em um botão para receber os resultados, no entanto, fazem um pré-processamento da consulta do leitor para prepará-la para submissão a cada ferramenta (CENDÓN, 2001).

Chaín Navarro (2004) afirma que a literatura diferencia os metabuscadores dos multibuscadores. Segundo o autor, os metabuscadores realizam a busca em diferentes buscadores, mas é o próprio programa que faz a busca e seleciona os buscadores, sem deixar a opção para o leitor, enquanto os multibuscadores fornecem informação buscador por buscador e, geralmente, o leitor assinala os buscadores que quer utilizar. Geralmente os multibuscadores oferecem estatísticas dos resultados de cada buscador, com o qual o leitor pode selecionar os que parecem mais adequados de acordo com os dados obtidos, e os metabuscadores, apesar de serem mais rápidos, geralmente não especificam as fontes (buscadores) de onde foram obtidas as informações e muitas vezes há duplicações.

Entretanto Cendón (2001, p. 47) nos alerta que “[...] algumas ferramentas que se intitulam metamotores são, na realidade pseudometamotores, pois que apenas fornecem uma interface onde vários motores são listados sem que haja um mecanismo de busca integrada.” Atualmente, nota-se que os metabuscadores evoluíram para o que chamamos de mecanismos de busca federados, ou seja, houve uma inversão na maneira de buscar o conhecimento, porque antes os metabuscadores buscavam os conteúdos em uma única coleção, e os federados realizam a busca em múltiplas coleções. A busca federada segundo Codina, Abadal e Rovira (2010), consiste em enviar a mesma pergunta (*query*) a vários mecanismos de busca.

5.2 Mecanismos de Busca Semântica

Estudiosos da área da Ciência da Computação têm empreendido esforços consideráveis em pesquisas para o desenvolvimento dos mecanismos de busca semântica e para o aperfeiçoamento dos mecanismos tradicionais para que estes possam operar com a semântica.

Os mecanismos de busca tradicionais, de acordo com Ramalho, Vidotti e Fujita (2007, p. 6), “[...] baseiam-se exclusivamente na recuperação de dados, não levando em consideração as semânticas contidas nas páginas da Web, recuperando apenas seqüências de caracteres que satisfaçam determinadas condições de busca.”

Kassim e Rahmany (2009) observaram que os mecanismos de busca tradicionais não estão sendo mais capazes de prover resultados precisos, porque além de serem baseados apenas em palavras-chave, eles não sabem lidar com aspectos de polissemia e sinônimos, e por isso os resultados da busca não satisfazem às necessidades dos usuários. Para ilustrar melhor essa questão, os autores fazem uma comparação entre os mecanismos de busca tradicionais e os semânticos (Quadro 4).

Quadro 4 - Diferenças entre os mecanismos de busca tradicionais e os semânticos.

Mecanismos de Busca Tradicionais	**Mecanismos de Busca Semânticos**
- No *prompt* do mecanismo, voce entra uma palavra-chave. - Não compreende polissemia e sinonímia. - Desconhecimento do significado dos termos. - Não consideram as *stop words*, tais como um/uma, e, é, em, de, ou, a, o, as, os, com, por, depois de. - Ao procurar por uma página web, um mecanismo de busca convencional procura pela distribuição das palavras dentro da página web para tentar encontrar quais são relevantes à questão de busca do usuário. Basicamente, isto significa que uma página web com palavras semelhantes a desses tipos de usuário dentro de um mecanismo de busca será mais relevante, e aparecerá em uma posição mais elevada na página de resultados da busca. - Inabilidade para manusear *queries* longas.	- No *prompt* do mecanismo, voce entra uma questão. - Compreende polissemia e sinonímia. - Conhecimento do significado dos termos. - Consideram as *stop words*, tais como um/uma, e, é, em, de, ou, a, o, as, os, com, por, depois de. - É projetado para tentar entender o contexto em que as palavras são usadas na página web tentando combiná-las com mais precisão à questão de busca do usuário. - Habilidade para manusear *queries* longas.

Fonte: Kassim e Rahmany (2009, p. 384, tradução livre).

Também Pollock (2010, p. 37) considera que enquanto

> [...] os mecanismos de busca tradicionais operam principalmente em índices de palavras-chave e páginas de resultados simples, os mecanismos de busca semântica tentam dar resultados mais inteligentes, procurando primeiro por conceitos e, em seguida, tornando os resultados mais navegáveis para pessoas.

Dessa forma, um mecanismo de busca semântica (SSE – *Semantic Search Engine*), conforme Renteria-Agualimpia et al. (2010, p. 613-4),

> [...] pode ser entendido como um aplicativo da Web semântica que pode responder às perguntas com base no significado da questão especificada pelo usuário, recursos nos repositórios e em muitos casos baseia-se em domínios semânticos predefinidos ou em modelo de conhecimento. SSE pode retornar resultados relevantes em seus tópicos que não necessariamente mencionam a palavra que você procurou explicitamente. (Tradução livre).

Para Berry e Browne (2005) e Levene (2010), nos mecanismos de busca tradicionais, como Google e Yahoo, não há inteligência porque simplesmente combinam texto em uma lista de palavras, ou seja, não existe lógica complexa ou

raciocínio com os dados, apenas simples algoritmos de combinação de palavras-chave. Contudo, Angotti (2012) relata que o Google está aproveitando o poder das tecnologias semânticas[19] para criar um banco de dados, que está sendo chamado de Gráfico do Conhecimento, para proporcionar a busca semântica de verdade, provendo resultados melhores, não só baseados em palavras simples.

Papoutsidakis et al. (2009) afirmam que a Web Semântica possibilitará muitas inovações para os mecanismos de busca e os leitores serão capazes de formular mais livremente suas questões/perguntas, sem necessariamente usarem palavras-chave ou operadores booleanos para que eles lhes tragam resultados satisfatórios. De acordo com Pollock (2010, p. 37),

> [...] os mecanismos de busca semântica tentam dar resultados mais inteligentes, procurando primeiro por conceitos e, em seguida, tornando os resultados mais navegáveis para pessoas que querem analisar os resultados dos dados. Em geral, pesquisas semânticas tentam aumentar e melhorar buscas tradicionais alavancando dados formatados pela Web Semântica para adicionar mais significado à consulta de pesquisa e ao texto da Web, a fim de aumentar a precisão dos resultados, bem como torná-los mais fácil de navegar para o melhor resultado.

Hendler (2010, p. 78) afirma que

> Embora os detalhes internos da maior parte desses sistemas ainda são proprietários, em geral, eles parecem combinar uma abordagem pragmática para processamento de linguagem natural com uma semântica leve que lhes permite melhor coletar e processar informações sobre áreas específicas. A pesquisa completa está além do mandato desta coluna, mas algumas aplicações serão suficientes para destacar algumas das diferenças entre esses e o sistema tradicional. (Tradução livre).

Enfim, como explica Pollock (2010, p. 38), “O objetivo de um mecanismo de busca semântica é de fornecer exatamente a informação consultada por um usuário, em vez de retornar uma lista de resultados de palavras-chave relacionadas livremente nas quais o usuário tenha que clicar.” (POLLOCK, 2010, p. 38). Essa segunda geração de mecanismos de busca utiliza as tecnologias da Web Semântica na representação do conhecimento para obterem resultados satisfatórios,

[19] Creio que por uma questão de sigilo a empresa não disponibiliza essas informações em seu site. Contudo, Santos e Nicolau (2012, p. 11, grifo nosso) afirmam que “A representação do Gráfico do Conhecimento demonstra o caminho que a seleção das informações, a partir de **ontologias**, **metadados** e **agentes**, contribuirão para os resultados serem muito mais assertivos.”

uma vez que os algoritmos de busca na Web não são públicos e visíveis, bem como a grande maioria dos leitores não estão familiarizados com eles.

6 BUSCA DE INFORMAÇÃO NA WEB

Os conteúdos no ciberespaço não são apresentados apenas para um leitor, abrem

> [...] a possibilidade de se navegar entre os diferentes *sites* em um processo que implica habilidades de decodificação das suas linguagens, habilidades estas que são aprendidas, sobretudo, por meio de visitações assíduas em que são praticadas operações mentais exploratórias ou de ensaio e erro [...]. (SANTAELLA, 2011b, p. 180).

O acesso, traço mais marcante do ciberespaço, se dá por meio de interfaces, o que obriga-nos a encontrar novas formas de orientação e busca.

Nesse contexto, o acesso aos conteúdos do ciberespaço é realizado por diferentes tipos de leitores, os quais utilizam métodos diferenciados para localização da informação desejada. Santaella (2011a, b) apresenta, com base em uma pesquisa conceitual e empírica, três tipos de perfis cognitivos ou estilos de navegação para caracterizar o leitor que navega de uma informação a outra no ciberespaço: o errante/novato, o detetive/leigo e o previdente/experto. Esses leitores manifestaram habilidades, nas quais Santaella (2011a) pode perceber conexões muito evidentes com os três tipos de raciocínio que Peirce estudou detalhadamente: o abdutivo, o indutivo e o dedutivo (Quadro 5).

Em relação aos métodos adotados por esses leitores para localização da informação, recorremos a Salazar (2005) que os classificam em:

a) navegar: processo simples e intuitivo que consiste em seguir os *links* do hipertexto, criados por outros usuários;
b) buscar: processo que requer o uso de mecanismos de busca e o leitor precisa aprender a utilizá-lo, desenvolvendo, na prática, a habilidade para obter resultados satisfatórios.

Quadro 5 - Perfil cognitivo do leitor imersivo.

Leitor	Raciocínio	Habilidades
Errante/Novato	Abdutivo	- Desorientação diante da profusão de signos que se apresentam na tela, ansiedade e insegurança nas operações de navegação. - Navega utilizando seu instinto para advinhar. - Explora aleatoriamente o campo de possibilidades aberta pela trama hipermidiática, pois lhe falta a internalização de esquemas gerais e a conseqüente capacidade de recuperar esses esquemas para adaptá-los às situação em curso. - Não temem o risco de errar. - Não traz consigo o suporte da memória.
Detetive/Leigo	Indutivo	- São lentos e hesitantes, realizam repetidamente operações de busca, avançam, erram e se autocorrigem, retornam e tentam outro caminho para encontrar uma solução. - Orientado por inferências indutivas, segue, com muita disciplina, as trilhas dos índices de que os ambientes hipermidiáticos estão povoados. - Possui memória operativa aguçada, e movimenta-se no campo do contingente.
Previdente/ Experto	Dedutivo	- Hábil no desenvolvimento das inferências dedutivas. - Já passou pelo processo de aprendizagem, adquiriu tal familiaridade com os ambientes informacionais que neles se movimenta seguindo a lógica da previsibilidade. - É capaz de antecipar as conseqüências de cada uma de suas escolhas, que são mais escolhas necessárias do que contingentes. - Sua atividade mental mestra é a da elaboração.

Fonte: Elaborado pela autora com base em Santaella (2011a, p. 65-72; 2011b, p. 322-3).

Broder (2002) e Marcos e González-Caro (2010) apresentam três tipos de consulta, e as relacionam as três intenções que os leitores manifestam em suas buscas:

a) *informacional*: quando a intenção do usuário é adquirir algumas informações sobre um tópico e presume-se presente em uma ou mais páginas *web*;
b) *navegacional*: quando a intenção do usuário é encontrar um determinado *site* a partir do qual iniciará a navegação, por exemplo, *site* de uma universidade onde quer estudar, de uma empresa onde irá fazer uma entrevista de trabalho ou de um jornal que gostaria de ler;
c) *transacional*: quando a intenção do usuário é realizar uma atividade ou ação que é mediada por um *website*, por exemplo

> compras *on-line*, *download* de *software*, navegação na biblioteca *on-line*, ou algum outro serviço especializado.

A esse respeito, embora o ato de navegar seja utilizado frequentemente pelos leitores ou cibernautas, o buscar, nessa pesquisa, é o método adequado para identificarmos como são processadas a busca e como a informação é recuperada em um contexto mais semântico e pragmático, dentro de um ambiente acadêmico onde a intenção do leitor é informacional, conforme apresentada por Broder (2002). Nesse sentido, o

> [...] *buscar* depende de um programa que se encarrega de combinar as palavras-chave que o usuário especifica com os documentos mais relevantes que existem na rede. A busca eficaz requer a aprendizagem para usar as ferramentas de busca, assim como para desenvolver, com a prática, a habilidade para obter resultados satisfatórios. (SALAZAR, 2005, p. 46, tradução livre).

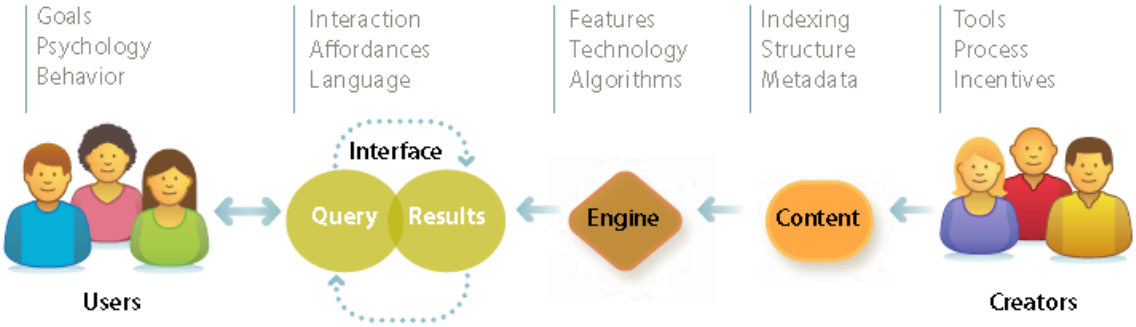
Conforme afirma Monteiro et al. (2011, p. 2550)

> A busca é uma sintaxe em devir, de recursos variados, segundo as necessidades do usuário e os limites dos índices compilados pelos mecanismos, de tal modo que os operadores pragmáticos, envolvidos no processo, permitem construir um mapa de significados vigentes e atualizados, no momento da busca, no universo simbólico, virtual e movente que é o ciberespaço.

Por essa perspectiva, os mecanismos de busca, mais que ferramentas de manipulação da informação, são, efetivamente, tecnologias da inteligência, tecnologias mediadoras e essencialmente tecnologias da linguagem com capacidade para presentificar, apresentar, indicar e representar a realidade (SANTAELLA, 2011b), por isso é importante considerar os esforços empregados pelos pesquisadores das Webs Semântica e Pragmática no aprimoramento dos sistemas e das interfaces destes mecanismos com o objetivo de atender a demanda por informação relevante ao contexto de busca do leitor.

Nesse aspecto, Morville e Callender (2010) propõem uma anatomia da busca (Figura 3), deslocando o foco de atenção do *hardware* e do *software* para os elementos da experiência dos usuários, como afirmam Monteiro et al. (2011).

Figura 3 - Anatomia da busca.



Fonte: Morville e Callender (2010, p. 25).

Na Figura 3, observa-se nitidamente que a linguagem é o elemento central determinante na elaboração da *query* e na obtenção dos resultados, por meio de um processo cíclico de retroalimentação. São os leitores que mantêm o processo em funcionamento, colaborando entre si e fazendo o *upload* da linguagem para que os mecanismos de busca possam retornar os resultados de acordo com o contexto desejado.

Observa-se também na literatura sobre busca de informação ou *information seeking* essa evidente mudança de paradigma, porque no início os estudos eram focados nos sistemas e hoje estão centrados no usuário e nos procedimentos heurísticos com que indaga e manipula os recursos de informação, conforme apontam Saracevic e Kantor (1988a,b) e Saracevic et al. (1988) e, que de acordo com Figueiredo (2011):

a) está orientada por uma finalidade que requer compreensão e mudança de um estado prévio de conhecimento;
b) as estratégias de busca seriam mais oportunistas, não-planejadas;
c) atende a procedimentos e estruturas de interação;
d) o julgamento de relevância acompanha cada passo da busca e pode efetuar-se pelo acesso direto ao texto completo, e não só a partir de suas representações.

Assim sendo, se considerarmos a atuação do profissional da informação[20] com experiência no processo de busca e recuperação da informação, embora não seja o foco principal do estudo, podemos afirmar que esses profissionais se encaixam no perfil do leitor previdente/experto, que desenvolveu durante sua formação e prática profissional a habilidade das inferências dedutivas e a familiaridade com os ambientes informacionais, tornando-se *expert* na elaboração de estratégias de busca.

Essa *expertise* possibilita, por exemplo, que um bibliotecário de referência almeje a realização da "busca perfeita", uma vez que ela depende da habilidade do leitor imersivo em estruturar sua estratégia de busca de maneira adequada ao mundo digital, onde a informação é manipulada dinamicamente, diferentemente da informação armazenada no computador, que pode ser imediatamente recuperável e livremente variável (SANTAELLA, 2011b).

A busca perfeita, segundo Battelle (2006, p. 217), não teria apenas a capacidade de trazer a resposta precisa, mas "[...] uma resposta adequada ao contexto e ao objetivo de sua pergunta, uma resposta que, com uma precisão assustadora, é formulada por quem é você e por que está perguntando." Tal resposta será "[...] capaz de incorporar todo o conhecimento investigável do mundo à tarefa em questão – seja ele capturado em formato de texto, vídeo ou áudio." (BATTELLE, 2006, p. 218), imagem ou programas informáticos, uma vez que de acordo com Santaella (2011b), o aspecto semiótico mais proeminente do ciberespaço encontra-se na convergência das mídias.

Nesse aspecto, a busca perfeita não depende somente da competência/ habilidade do leitor imersivo previdente em elaborar estratégias de busca eficientes, mas, sobretudo, da interatividade com as interfaces dos mecanismos de busca. Todavia,

[20] As inovações e os avanços tecnológicos ampliaram o leque de atuação dos profissionais da informação, principalmente na Web. Algumas profissões tiveram que ser criadas e outras de se adaptarem ao contexto digital. Dentre as novas profissões destacamos: arquiteto da informação, analista de mídia online, analista de anúncios de links patrocinados, blogger profissional, copywriter, especialista em otimização de sites (SEO), especialista em redes sociais, gestor de conteúdo, profissional de suporte, web developer, web designer, web master, web writer, etc. A profissão de bibliotecário é uma das que tiveram que se adaptar às novas formas de produção, organização, gestão e acesso à informação.

> A interatividade ciberespacial não seria possível sem a competência semiótica do usuário para lidar com as interfaces computacionais. Essa competência semiótica implica vigilância, receptividade, escolha, colaboração, controle, desvios, reenquadramentos em estados de imprevisibilidade ou de acasos, desordens, adaptabilidades, que são, entre outras, as condições exigidas para quem prevê um sistema interativo e para quem o experimenta. (SANTAELLA, 2011b, p. 80).

Desse modo, o bibliotecário de referência deve conhecer e estudar as interfaces dos mecanismos de busca uma vez que muitos conteúdos da Web não estão formalmente indexados em base de dados e por isso dependem desses mecanismos para serem recuperados. Assim, o bibliotecário de referência, experiente e conhecedor dos mecanismos de busca também saberá, após alguns momentos de conversa com o leitor, fazendo perguntas e ouvindo-o, indicar os mecanismos que poderão satisfazer suas necessidades de informação, porque a entrevista de referência serve de ponte entre a questão do usuário e a solução do problema, envolve fatores pessoais e impessoais e depende basicamente do bibliotecário, do usuário, e das fontes de informação (GROGAN, 2001; LUZ GARCÍA; PORTUGAL, 2008). Entretanto, além da indicação dos mecanismos de busca pertinentes na solução do problema, o bibliotecário de referência, com sua habilidade prática de busca da informação, deverá instruir e treinar o leitor nas técnicas e estratégias de busca.

6.1 Classificação da Busca

Acredita-se que de certa forma, a busca de informação na Web aprimorou-se conforme a Web foi evoluindo. De acordo com Battelle (2006, p. 7-8),

> Por sua natureza, a busca é um dos problemas mais desafiadores e interessantes de toda a ciência da computação e muitos especialistas afirmam que a pesquisa continuada de seus mistérios irá prover a massa crítica comercial e acadêmica, que nos permitirá criar computadores capazes de agir, sob qualquer critério, como um ser humano.

Essa “[...] idéia de que um dia a busca irá assumir uma forma humanóide permeia quase toda discussão sobre o futuro da aplicação.” “Assim, Hillis [Danny, gênio e cientista de computadores da MacArthur Foundation] afirma que o

futuro da busca terá mais a ver com compreender do que com simplesmente descobrir.” (BATTELLE, 2006, p. 13).

Para Battelle (2006, p. 241),

> A busca não é mais um aplicativo isolado, uma ferramenta útil, mas impessoal, para achar algo num novo meio chamado World Wide Web. A busca é, cada vez mais, nosso mecanismo para que conheçamos a nós mesmos, nosso mundo e nosso lugar nele. É nosso modo de navegar pelo único recurso infinito que move a cultura humana: o conhecimento.

Assim sendo, considerando a evolução da Web e, particularmente os esforços dos mecanismos de busca, como organizadores do conhecimento no ciberespaço, de não só descobrirem mas de compreenderem a *query* formulada pelo leitor, acredita-se que a busca pode ser classificada em: busca sintática, busca semântica e busca pragmática, pois conforme afirma Santaella (2011a, p. 50), “A estrutura flexível e o acesso não linear da hipermídia permitem buscas divergentes e caminhos múltiplos no interior do documento.”

6.1.1 Busca Sintática

A busca sintática é realizada basicamente por meio de palavras-chave, ou seja, uma busca tipicamente tradicional estabelecida principalmente sobre a ocorrência de palavras em documentos (sintático), por isso mais sensível a ambiguidades. Para Brascher (2002), a ambiguidade sintática “[...] ocorre na estruturação da frase em constituintes hierarquizados, quando se definem as ligações que se estabelecem entre os sintagmas.”

Nesse aspecto, Antoniou et al. (2005, v. 5, p. 2464, tradução nossa) aponta cinco problemas relacionados à busca por palavras:

a) alta revocação[21], baixa precisão: a maior parte das páginas recuperadas são irrelevantes;

[21] **Revocação** – A extensão com que todos os itens numa base de dados que são considerados relevantes ou pertinentes são recuperados durante uma busca nessa base de dados. Uma busca com ‘alta revocação’ será aquela em que a maioria dos itens relevantes (pertinentes), se não todos, forem recuperados. O coeficiente de revocação – uma medida da extensão com que ocorre a recuperação de itens relevantes (pertinentes) – é o número de itens relevantes (pertinentes) recuperados dividido pelo número total de itens relevantes (pertinentes) existentes na base de dados. (LANCASTER, 1993, p. 306, grifo do autor).

b) baixa ou nenhuma revocação: páginas chave e relevantes não são recuperadas;
c) sensibilidade ao vocabulário escolhido: pequenas modificações no vocabulário podem causar modificações significantes nos resultados;
d) resultados de páginas web únicas: informação extensa e expressa por meio de várias páginas;
e) o envolvimento humano é necessário para interpretar as páginas recuperadas, e combinar a informação.

Marcondes (2011, p. 84) ressalta que as

> Buscas por palavras-chave ligadas pelos operadores booleanos não dão conta da expressividade e precisão necessária para a recuperação de conteúdo semântico contido no crescente número de artigos científicos e outras fontes de informação agora disponíveis em toda a *Web*. (Grifo do autor).

No entanto, observa-se que na busca sintática os operadores booleanos podem ser eficientes quando utilizados nos sistemas de recuperação da informação convencionais/formais (catálogos de bibliotecas, bases de dados, repositórios etc), onde a indexação é manual ou semi-automática valendo-se da linguagem documentária, mas na Web seu uso é restrito devido a natureza ambígua da linguagem natural em que é expressa a *query*, bem como pela inabilidade em ordenar os documentos resultantes de uma busca, conforme afirma Ferneda (2012). Entretanto, Ferneda (2012, p. 29) conclui que mesmo com suas limitações, o modelo booleano

> [...] pode ser considerado o modelo mais utilizado não só nos sistemas de recuperação de informação e nos mecanismos de busca da Web, mas também nos sistemas de bancos de dados, onde o seu poder se expressa por meio da linguagem de consulta SQL (*Structured Query Language*).

Dentre as limitações do modelo booleano Ferneda (2012, p. 27-28) destaca:

- Sem um treinamento apropriado, o usuário leigo será capaz de formular somente buscas simples. Para buscas que exijam expressões mais complexas é necessário um conhecimento sólido da lógica booleana;
- Existe pouco controle sobre a quantidade de documentos resultante de uma busca. O usuário é incapaz de predizer quantos registros satisfarão a restrição lógica de uma determinada expressão booleana, sendo necessárias sucessivas reformulações antes que seja recuperado um volume aceitável de documentos;
- O resultado de uma busca booleana se caracteriza por uma simples partição do corpus em dois subconjuntos: os documentos que atendem à expressão de busca e aqueles que não atendem. Presume-se que todos os documentos recuperados são de igual utilidade para o usuário. Não há nenhum mecanismo pelo qual os documentos possam ser ordenados.
- Não existe uma forma de atribuir importância relativa aos diferentes termos da expressão booleana. Assume-se implicitamente que todos os termos têm o mesmo peso.

Contudo, na Web, embora sejam aplicadas modernas técnicas de mineração de dados e de textos à recuperação da informação, Marcondes (2011, p. 84) afirma que na busca por palavra elas

> [...] se mostram como técnicas de busca “cega”, com base somente no poder computacional, não conseguem identificar significados. São baseados somente em técnicas computacionais de correspondência entre padrões de caracteres nos termos de busca, que remontam aos primórdios da era do computador.

Segundo Breitman (2010), esta é uma situação que estamos vivenciando atualmente, porque mecanismos de busca como Google e Yahoo, os quais utilizam palavras na busca da informação, não conseguem ainda identificar os significados implícitos na *query* formulada pelo leitor, ocasionando problemas na recuperação da informação, conforme apresentados por Antoniou et al. (2005, v. 5). Por outro lado, para um leitor capacitado em lógica booleana, mesmo que não lhe pareça natural, esse tipo de *query* pode ser mais fácil e mais exato de processar, porque a importância de cada palavra pode ser medida da estrutura semântica da sentença/oração. O descarte de palavras insignificantes permite que um sistema de recuperação da informação seja capaz de determinar quais palavras são mais importantes e por isso são usadas para extrair grupos de documentos e/ou termos relacionados. (BERRY; BROWNE, 2005, p. 8). Entretanto, nas *queries* elaboradas em linguagem natural as palavras eliminadas pelo uso da lista de *stop words* ocasionam a perda do contexto em que elas são usadas.

Outra questão relevante, relacionada à busca sintática, é que ela foi introduzida na Web reproduzindo o modelo dos sistemas formais de recuperação da informação (referência fixa, transporte do físico para o digital), sem considerar a natureza intrinsecamente ambígua da linguagem natural em que a necessidade de informação é expressa.

6.1.2 Busca Semântica

A busca semântica se tornou uma alternativa para superar as deficiências dos mecanismos de busca tradicionais. Os mecanismos tentam analisar e compreender o que o leitor deseja na pesquisa em um contexto, através de "raciocínio lógico", possibilitando melhores resultados e, conforme afirma Battelle (2006, p. 231-2), "[...] a busca tornou-se bastante sofisticada com o uso de palavras-chave e análise do padrão de conexões."; porém, "[...] a tecnologia de busca ainda não tem idéia do real significado de um documento – no sentido humano."

Para Mangold (2007) e Reis (2011), a busca semântica é um processo de recuperação de documentos que explora o conhecimento no domínio, que pode ser formalizado por meio da ontologia. Na opinião de Bonino et al. (2004) o ponto chave para o processo de refinamento de uma busca semântica está na disponibilidade de uma ontologia de domínio[22], e na capacidade de compreender as relações semânticas entres os conceitos ontológicos. Esse refinamento é importante porque os significados diversos de uma mesma palavra permitem que as buscas sejam bem dependentes dos contextos em que a palavra é empregada.

Dessa forma, conforme afirmam Alesso e Smith (2009), os métodos de busca semântica podem aumentar e melhorar os resultados de busca tradicionais usando, não apenas palavras, mas conceitos e relações lógicas; por meio de duas abordagens:

a) uso direto de metadados da Web Semântica (*Semantic Web documents*);

[22] "[...] são ontologias que podem ter seus conceitos reutilizados dentro de um domínio específico (médico, farmacêutico, direito, financeiro, entre outros)." (BREITMAN, 2010, p. 40).

b) indexação semântica latente (*Latent Semantic Indexing* - LSI)[23].

Assim, a busca semântica objetiva encontrar documentos que tenham conceitos similares, e não apenas palavras semelhantes, ou seja, ela procura entender o que o leitor busca e não realizar apenas uma busca mecânica por *sites* a partir das palavras-chave. Nessa busca há a interpretação do sentido da palavra ou conceito antes de se apresentar os resultados. No entanto, verifica-se que esta análise ainda é realizada com base em sentidos já construídos, em um tipo de semântica formal dependente de contextualização; por isso Bonino et al. (2004) propuseram o uso de uma ontologia de domínio para o refinamento de uma busca semântica.

6.1.3 Busca Pragmática

No nosso entendimento, a busca pragmática engloba tanto a busca sintática quanto a busca semântica, porque se trata de uma busca que deve estar adequada ao contexto da *query* formulada pelo leitor, independente da palavra de ordem adotada, que pode ser uma palavra-chave, um descritor, uma frase, um som, uma imagem etc.. Percebe-se que nas novas configurações dos mecanismos de busca, eles estão procurando se adequar aos perfis dos leitores e, nesse particular, sugerem e/ou encontram padrões de busca tanto para estabelecer sentidos para o leitor quanto contexto.

Observa-se essas novas configurações nos recursos de *mashup* (lista de possíveis sentidos, no final da página de resultados) (MONTEIRO et al., 2011) e de *autocomplete* e *autosuggest* (MORVILLE; CALLENDER, 2010).

Os recursos autocomplete e autosugestão são distintos, porém os mecanismos de busca os utilizam juntos, quando o leitor inicia a digitação na caixa de busca as sugestões aparecem automaticamente. O autocomplete proporciona economia de tempo, agiliza a digitação e auxilia na correção de erros, enquanto que a autosugestão oferece opções melhores de busca e temas relacionados (Figura 4). O Yahoo foi o mecanismo de busca pioneiro na disponibilização desses recursos.

---

[23] Indexação semântica latente (ISL) é um método estatístico de recuperação de informação capaz de recuperar o texto com base nos conceitos que ele contém, não apenas combinando palavras-chave específicas.

Figura 4 – Recurso *autocomplete*.

Fonte: Morville e Callender (2010, p. 86).

Entretanto, além dos recursos de *mashup, autocomplete* e *autosuggest* não encontramos de maneira explícita na literatura sobre a Web Pragmática, conceitos ou outros aspectos que envolvem a busca pragmática, mas acredita-se que ela está relacionada ao contexto de uso, quando o leitor faz o *upload* da linguagem na Web; porém, assim como a busca semântica, deve ser melhor investigada.

6.2 Estratégia de Busca

Na Web, a maneira que é formulada a questão da busca influencia substancialmente na recuperação de conteúdos relevantes ao contexto do leitor, por isso uma estratégia de busca bem elaborada pode aumentar a precisão e relevância dos resultados obtidos. Entende-se que a elaboração da estratégia de busca requer do leitor o domínio da técnica e da tecnologia para a busca de informação na Web, pois a

> A técnica envolve conhecimento para a realização de determinada tarefa, como desempenhar-se de certo modo. Assim, ela se define como um saber fazer, referindo-se a habilidades, a uma bateria de procedimentos que se criam, se aprendem, se desenvolvem. (SANTAELLA, 2011b, p. 257).

Porém, “Enquanto a técnica é um saber fazer, cuja natureza intelectual se caracteriza por habilidades que são introjetadas por um indivíduo, a tecnologia, como conhecimento acerca da própria técnica, avança além desta.” (SANTAELLA, 2011b, p. 258).

No âmbito da recuperação da informação, a estratégia de busca pode ser definida como uma técnica ou conjunto de regras para tornar possível o encontro entre a pergunta formulada e a informação que pode estar armazenada em uma base de dados, ou nesse caso, no ciberespaço. Cunha e Cavalcanti (2008, p. 158) referem-se à estratégia de busca como uma “[...] pergunta ou conjunto de perguntas, formada por palavras da linguagem natural, por palavras-chave ou descritores, podendo estar unidos por operadores lógicos booleanos, que possibilitam a recuperação da informação.” Para os autores, quando se trata da recuperação da informação por meio dos mecanismos de busca, há dois tipos de estratégias:

> a) básica ou simples: realizada mediante a digitação de termos ou palavras num campo predeterminado;
> b) avançada: a que permite ao usuário especificar os campos onde deverão ser pesquisados os termos da estratégia de busca. (CUNHA; CAVALCANTI, 2008, p. 158).

Entretanto, com a evolução dos mecanismos de busca, observa-se que hoje na Web, as estratégias de busca podem ser elaboradas de diferentes formas e, nesse sentido, Canter et al. (1985) e Forrai (2003) citados por SANTAELLA, 2011b, p. 325), apresentam cinco tipos:

> 1. escaneamento (cobrir uma vasta área sem profundidade);
> 2. *browsing* (seguir um caminho até que um alvo seja encontrado;
> 3. busca (insistir na busca de um alvo explícito);
> 4. exploração (descobrir a extensão de uma dada informação);
> 5. passeio (navegar de modo desestruturado e sem propósito definido).

Como explicitado anteriormente, a “busca” conforme descrita pelos autores supracitados é a que se adéqua ao contexto desta pesquisa. No ciberespaço, segundo Levene (2010), a “busca” refere-se ao uso de um mecanismo de busca para nos auxiliar na recuperação da informação que procuramos, e o "navegar" ou "surfar" na Web refere-se ao emprego de uma estratégia *link-following* iniciada a partir de uma determinada página web, para satisfazer nossa necessidade de informação.

A estratégia de busca de informação que está se tornando mais útil e utilizada na Web é a estratégia que combina busca e navegação de uma forma natural. (LEVENE, 2010). Assim sendo, a técnica ou o “saber fazer” na busca de informação na Web deve ser desenvolvida e Santaella (2011a, b) apresenta como o

leitor imersivo elabora sua estratégia de busca, de acordo com seu perfil cognitivo ou estilo de navegação (Quadro 6).

Quadro 6 - Relação entre o leitor e a forma de elaboração da estratégia de busca.

Leitor	Estratégia de Busca
Errante/Novato	- Suas rotas são idiossincráticas, turbulentas e, no mais das vezes, dispersivas e desorientadoras.
Detetive/Leigo	- São acionadas através de avanços, erros e autocorreções. - Seu percurso se caracteriza, como um processo auto-organizativo próprio daquele que aprende com a experiência. - É capaz de usar regras situacionais para diminuir a aleatoriedade das escolhas.
Previdente/Experto	- Como adquiriu a habilidade de ligar os procedimentos particulares aos esquemas gerais internalizados, sua navegação se dá em percursos ordenados, norteados por uma memória de longo prazo que o livra dos riscos do inesperado.

Fonte: Elaborado pela autora com base em Santaella (2011a, b).

Na busca de informação por meio de um mecanismo de busca, a experiência do leitor conta muito na construção da *query* ideal. Nesse sentido, Torres Pombert (2003) alerta que para se obter resultados relevantes na Web é necessário um misto de experiência, técnicas, habilidades, criatividade e boa sorte, tudo isso combinado com a capacidade de definir uma direção de forma metódica e clara, ou seja, "navegar com uma finalidade". O autor sugere a seguinte sequência de passos para uma recuperação eficaz:

- Determinar o tipo de informação que necessita (artigos científicos, estatísticas, documentos governamentais) e, em seguida, determinar que tipo de organização pode fornecer esses documentos.
- Criar uma lista de todas as possíveis palavras-chave e seus sinônimos.
- Determinar qual o tipo de instrumento utilizado na pesquisa (diretório, motor geral ou especializado, metabuscador) dependendo do que se está procurando.
- Construir a estratégia de busca e conduzi-la (dependendo do buscador, definir as combinações de buscas, deve ser tão precisa quanto possível e explorar as opções disponíveis).

- Avaliar os resultados da busca (se os primeiros 15 registros não são considerados relevantes deve-se repensar a estratégia várias vezes ou mudar de motor de busca se o resultado persistir, se os resultados são relevantes deve-se verificar a atualidade dos registros e se o site é proveniente de uma fonte confiável). (Tradução nossa).

Torres Pombert (2003), com base em vários autores, também sugere outros elementos para se conseguir melhores resultados, e que são aplicáveis a maioria dos mecanismos de busca:

- Escrever letras minúsculas e sem acentos.
- Não utilizar uma única palavra na sua busca porque se obtém muitos resultados.
- Empregar várias palavras-chave que definam ou determinem o que está sendo buscado.
- Colocar "entre aspas" as palavras que deseja que sejam encontradas juntas (frases). [...]
- Consultar sempre a informação disponível de cada mecanismo de busca para saber quais opções você pode usar para interrogar suas bases de dados. (Tradução nossa).

Para construir uma *query* ideal, pensando em um mecanismo de busca, além de seguir as orientações básicas apresentadas por Torres Pombert (2003), que ainda se aplicam aos dias atuais, é importante conferir alguns itens por meio de um *checklist* (Quadro 7).

Quadro 7 - *Checklist* para construção da *query*.

1. Quantos conceitos-chave (ideias importantes) encontram-se na pergunta?
2. Quantos conceitos-chave eu buscarei em uma única questão?
3. Quais palavras-chave são provavelmente eficazes "como é?"
4. Para quais conceitos provavelmente serão necessárias palavras-chave mais eficazes?
5. Há hipônimos ou linguagem profissional para qualquer uma das palavras intermediárias?
6. Há palavras com múltiplos significados?
7. Usei todas as *stopwords* ou cortei algumas palavras?
8. Escrevi corretamente as palavras?
9. Inseri as palavras mais importantes em primeiro lugar?

Fonte: https://sbp-portal.wikispaces.com/file/detail/handout+-+Query+Checklist.docx[24]

[24] Confira o *checklist* completo com as respostas para cada pergunta no Anexo.

No entanto, além das dicas apresentadas, existe o *insight,* presente em todos os níveis de leitores, que “[...] significa a capacidade de mudar de estado, descoberta de uma rota eficaz no caminho para um resultado final. As mudanças se dão tanto no estado interior do usuário quanto no estado físico da tela.” (SANTAELLA, 2011a, p. 69-70).

Desse modo, na prática, conhecendo o perfil do leitor, pressupõe-se que o bibliotecário poderá capacitá-lo para que adquira ou desenvolva a habilidade técnica para elaboração de estratégias de busca que possibilitem a recuperação da informação de forma mais eficaz, porque “Quanto mais a prática é executada, mais o desempenho se aperfeiçoa.” (SANTAELLA, 2011a, p. 71).

7 METODOLOGIA

O enfoque desta pesquisa foi o teórico-informal porque não partiu de uma premissa teórica, mas sim de uma questão derivada, ou problema de estudo. O objeto de pesquisa foi investigado a partir do aporte teórico da Filosofia, da Linguística, da Ciência da Computação e da Tecnologia da Informação para construí-lo na Ciência da Informação.

A abordagem eleita para realização da investigação foi a qualitativa uma vez que não foram empregados procedimentos estatísticos, mas a análise e compreensão dos fatos e variáveis que envolveram o fenômeno estudado. A abordagem qualitativa proporcionou uma melhor visão e compreensão do contexto do problema. Conforme Gil (1999), um estudo que utiliza este tipo de abordagem visa proporcionar um maior conhecimento para o pesquisador acerca do assunto, a fim de que esse possa formular problemas mais precisos ou criar hipóteses que possam ser pesquisadas por estudos posteriores.

De acordo com Godoy (1995, p. 58), a pesquisa qualitativa

> [...] considera o ambiente como fonte direta dos dados e o pesquisador como instrumento chave; possui caráter descritivo; o processo é o foco principal da abordagem e não o resultado ou o produto; a análise dos dados é realizada de forma intuitiva e indutivamente pelo pesquisador [...].

E, nesse caso, como o tema pesquisado ainda é pouco explorado na área da Ciência da Informação, adotamos a pesquisa documental por se adequar melhor às necessidades e aos interesses da investigação, no que se refere ao alcance dos objetivos propostos. De acordo com Gil (2007), a pesquisa documental se assemelha muito à pesquisa bibliográfica, porém, a diferença essencial entre ambas está na natureza das fontes. Enquanto a pesquisa bibliográfica se utiliza fundamentalmente das contribuições de diversos autores sobre determinado assunto, a pesquisa documental vale-se de materiais que não receberam ainda um tratamento analítico.

Para Witter (1990, p. 19),

> A pesquisa documental é estritamente a que é feita tendo por base qualquer um dos suportes de informação decorrentes de momentos anteriores a pesquisa, quer em andamento, quer relatadas, ou então de informações resultantes do fazer humano ligado a outras áreas, que não a ciência.

Acrescenta que “A pesquisa documental é aquela cujos objetivos ou hipóteses podem ser verificados através da análise de documentos bibliográficos ou não-bibliográficos [...]”. (WITTER, 1990, p. 22). Nesse sentido, a pesquisa documental foi extremamente vantajosa uma vez que para solucionar o problema de pesquisa foi necessário recorrer às fontes de informação diversificadas com vistas à construção do *corpus* teórico que auxiliou posteriormente na fase de identificação, seleção e classificação dos mecanismos de busca semânticos e na exemplificação da busca contextual.

7.1 Análise Documental

Como método de investigação utilizou-se a análise documental que tem por objetivo a identificação, verificação e apreciação de documentos. De acordo com Moreira (2005, p. 272), a análise documental é, ao mesmo tempo, método e técnica: “Método porque pressupõe o ângulo escolhido como base de uma investigação. Técnica porque é um recurso que complementa outras formas de obtenção de dados [...]”.

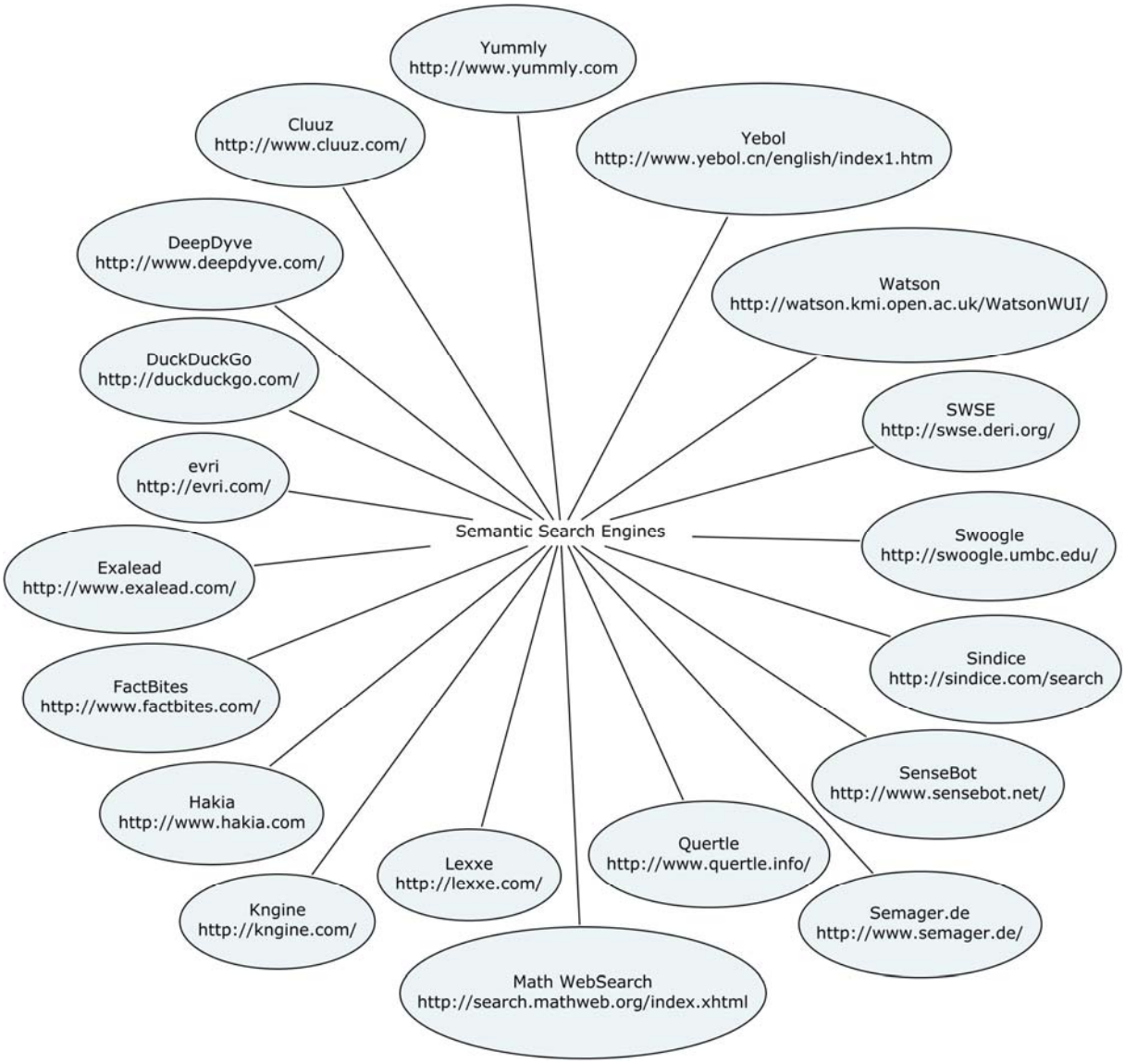
Na concepção de Lüdke e André (1986), os documentos não constituem apenas uma fonte de informação contextualizada, eles surgem em um determinado contexto e podem também fornecer informações sobre esse contexto. Nesse sentido, a adoção desse método permitiu interpretar o documento a luz do contexto do problema uma vez que a linguagem utilizada nos documentos constitui-se elemento fundamental para a investigação. Diferentemente da pesquisa social, em que se estuda o sujeito, a pesquisa teórica possibilita a investigação em documentos, que podem ser coletados nas mais variadas fontes e formatos. Dessa maneira, a análise documental, como método, na Ciência da Informação, utiliza-se de um *corpus*, que são documentos finitos e selecionados *a priori*, ou *a posteriori* para a investigação e estudo.

7.1.1 Corpus de Pesquisa

Como *corpus*, entende-se o “Conjunto de documentos, dados e informações sobre determinada matéria [...].” (FERREIRA, 2004, p. 557). Assim, o *corpus* (objeto) de pesquisa foi constituído pelos *sites* dos mecanismos de busca, os

quais foram identificados por meio de uma busca no Google, utilizando o termo “semantic search engine”. Do resultado dessa busca extraímos um *corpus* inicial com 19 mecanismos (Figura 5) e, posteriormente, selecionamos dois exemplares para análise da forma de organização (indexação) e do processo de busca e um exemplar, fora desse *corpus*, que opera com semântica e oferece os recursos de busca pragmática, conforme apresentado na revisão de literatura desse estudo (Quadro 8).

Figura 5 - “Semantic Search Engines” - *Corpus* inicial.



Quadro 8 - Mecanismos de busca selecionados e analisados.

MECANISMO	URL
Hakia	http://www.hakia.com
Lexxe	http://lexxe.com/index.html
Google	http://www.google.com.br

7.2 COLETA DE DADOS

A coleta de dados foi executada em três etapas.

Na primeira etapa foi realizada a coleta de documentos que não tinham recebido ainda um tratamento analítico ou que poderiam ser reelaborados (GIL, 2007), uma vez que se observou a necessidade de ampliação da compreensão do fenômeno em estudo.

Na segunda etapa, os mecanismos de busca foram identificados e selecionados do *corpus* inicial e as categorias de análise foram definidas para compor a lista de verificação (Figura 6) de acordo com as características observadas no desenvolvimento do *corpus* teórico. As informações foram coletadas diretamente no *site* dos mecanismos de busca e em outros sites que tinham informações sobre eles.

Figura 6 - Lista de verificação – Categorias de análise dos mecanismos de busca.

Nome: **URL**: **Proprietário**:
Descrição:
Categorias de Análise a) Anatomia do mecanismo b) Forma de organização (indexação) • *Indexing* – indexa ou gera o índice a partir da base de dados c) Processo de busca • *Searching* – busca através da interface do mecanismo

No que diz respeito à terceira etapa, foram elaboradas questões (*queries*) e executadas nos mecanismos de busca identificados e selecionados na etapa anterior. A análise dos resultados da busca foi realizada somente em relação

ao aspecto funcional, não medimos a precisão, a relevância, nem a revocação dos conteúdos recuperados.

8 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

A sistematização e análise dos dados coletados foram realizadas por meio da análise documental que constitui uma técnica importante na pesquisa qualitativa, seja complementando informações obtidas por outras técnicas, seja desvelando aspectos novos de um tema ou problema, conforme relatam Ludke e André (1986).

A análise documental da literatura científica e das informações contidas na estrutura dos *sites* dos mecanismos de busca identificados como semânticos, ou que operam com semântica, bem como em documentos como guias, tutoriais, manuais técnicos, manuais de usuários etc., nos permitiu analisar como eles organizam, processam, buscam e recuperam conteúdos no ciberespaço.

Dessa maneira, a partir da análise dos *corpora* e do estudo teórico realizado para fundamentar este trabalho, passamos a seguir à apreciação crítica dos dados e apresentação dos resultados.

8.1 Forma de Organização (Indexing)

Como abordado no referencial teórico, a *search engine indexing* é uma forma de indexação automática de *websites*. Nesse tipo de indexação, no entanto, os mecanismos de busca não conseguem sozinhos identificar conceitos. Eles apenas recuperam ocorrências de palavras no texto sem, contudo, diferencia-las. Assim sendo, os mecanismos de busca semântica que operam com a linguagem natural procuram indexar o conteúdo da Web não só com base em palavras-chave, mas em conceitos, cujo sentidos já estão construídos ou que estão sendo determinados pelo contexto que o leitor fornece.

Nessa direção, o Hakia e o Lexxe são exemplos de mecanismos de busca semântica, e o Google, um exemplo de mecanismo tradicional que tem agregado ou desenvolvido tecnologia para operar com a semântica na indexação de páginas web.

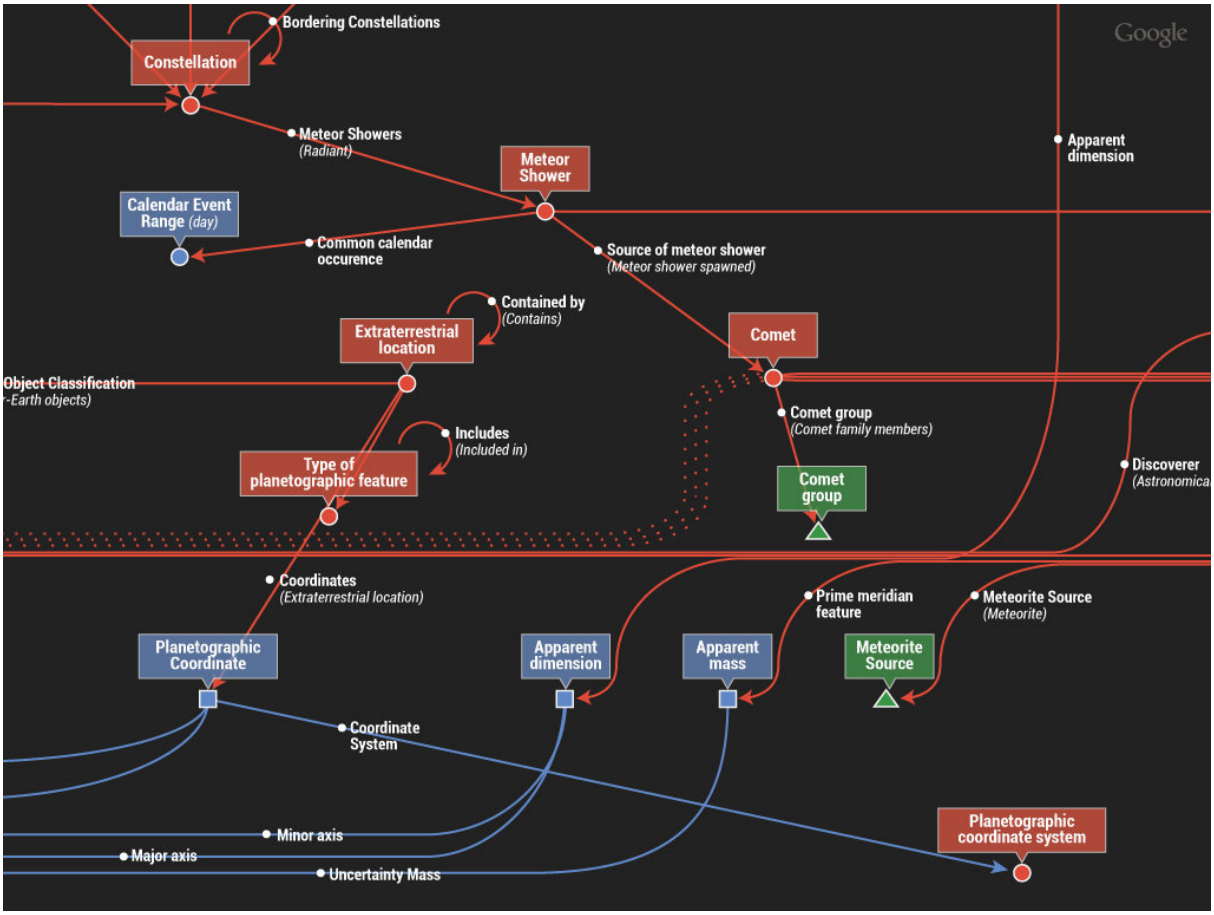
O Google faz o rastreamento para descobrir páginas novas e atualizadas para serem incluídas no índice por meio do Googlebot, seu *web crawling bot* (aranha) e, para isso, utiliza um grande número de computadores. O Googlebot utiliza um processo algorítmico, ou seja, programas de computador que determinam

quais *sites* devem ser rastreados, com que frequência e quantas páginas devem buscar a partir de cada local (GOOGLEBOT..., 2012). Dessa maneira, ele processa cada página que rastreia para compilar um índice imenso com todas as palavras que encontra e sua localização em cada página. Também processa informações incluídas nas principais *tags* de conteúdo e atributos; entretanto, não processa todos os tipos de conteúdos como, por exemplo, alguns arquivos de mídia ou páginas dinâmicas. Esse processo de rastreamento inicia-se com uma lista de URLs de páginas web, gerada a partir de processos anteriores de rastreamento e que é aumentada com dados dos *site maps*[25] fornecidos por *webmasters*. O Googlebot visita cada um desses *sites*, detecta os *links* em cada página e os adiciona à sua lista de páginas ratreadas. Assim, os *sites* novos ou alterações em *sites* existentes e *links* inativos são detectados e utilizados para atualizar o índice do Google (GOOGLEBOT..., 2012).

Em maio de 2012, o Google lançou o “Gráfico do Conhecimento” (Figura 7), um banco de dados com mais de 500 milhões de pessoas do mundo real, lugares e coisas, com 3,5 bilhões de atributos e conexões entre eles (SINGHAL, 2012). Na construção desse gráfico a empresa Google Inc está utilizando tecnologias semânticas e espera que com essa inovação possa auxiliar o leitor a encontrar o resultado correto mais rapidamente quando houver significados diferentes. Desse modo, o Google poderá dar sugestões diferentes de entidades do mundo real na caixa de busca enquanto o leitor digita. Por enquanto esse recurso foi disponibilizado para os Estados Unidos (EUA), mas gradativamente estará disponível para outros países onde o Google tem domínio.

[25] **Mapa do *site*** - Uma página no seu *site* que possui *links* para as partes importantes do *site* para facilitar a navegação do usuário. (http://www.feedthebot.com/sitemaps.html, tradução livre)

Figura 7 – Representação do Gráfico do Conhecimento do Google.



Fonte: http://agenciadeinternet.com/wp-content/uploads/2012/05/goo.jpg

O Hakia, diferentemente do Google, utiliza novas formas de indexação e algoritmos semânticos construídos, essencialmente, por três tecnologias: o QDEX^{TM} (*Query indexing techinique*), OntoSem^{TM} (*sense repository*) e o algoritmo SemanticRank^{TM}, com a finalidade de indexar e buscar conteúdo semântico. O QDEX^{TM} e o OntoSem^{TM} são as tecnologias do Hakia que estão relacionadas à indexação (*indexing*) e o SemanticRank^{TM} à busca (*searching*). O Hakia rastreia a Web e analisa apenas as páginas web que são autorizadas pelos proprietários do *site*.

A empresa Hakia, Inc inventou o sistema QDEX^{TM} para detecção e extração de perguntas (*queries*) para possibilitar a análise semântica de páginas web. Trata-se de uma nova forma de analisar e armazenar o conteúdo da página em termos de *bits*, em substituição ao método de índice invertido[26] mais comumente

[26] "**Inverted index,** índice invertido. Estrutura de dados onde cada palavra tem associada uma lista de documentos onde ocorre. Um índice invertido é análogo a um índice remissivo de um livro. Os

usado pelos mecanismos. De acordo com Baeza-Yates e Ribeiro-Neto (1999), o índice invertido é a estrutura mais comum para indexar informação de forma a permitir um melhor desempenho durante a execução de uma busca, desse modo, observa-se que o que o QDEX^{TM} gera é um índice semântico que permite a identificação de termos em contextos específicos.

O OntoSem^{TM} trata-se de um repositório de conceitos, com os seus diferentes sentidos e o SemanticRank^{TM} do Hakia é um algoritmo que tem o propósito de classificar os resultados por ordem de relevância. O OntoSem^{TM} é um banco de dados linguístico onde as palavras são categorizadas de acordo com os diversos "sentidos" que transmitem (ISKOLD, 2006). Dessa maneira, segundo seus criadores, o Hakia deve ser capaz de reconhecer as variações morfológicas (tempo verbal, gênero e número) e diferenciar sinônimos e conceitos para determinar qual informação é relevante (LOS NUEVOS..., 2012).

O Hakia, segundo Pollock (2010), está atualmente indexando o conteúdo confiável em setores verticalizados como a medicina, finanças, leis, ciência, viagens, artes, história etc., mas também oferece sua tecnologia de busca semântica para negócios. O Pubmed[27] está adotando a tecnologia do Hakia e já possui mais de 20 milhões de *abstracts* indexados na nova interface.

O mecanismo de busca Hakia oferece também a oportunidade para que as pessoas interessadas colaborem em projetos de pesquisa ou anotem semanticamente suas próprias páginas. Dentre essas pessoas, destacamos a participação dos bibliotecários que fazem a recomendação de páginas web confiáveis para o Hakia e, dessa maneira, os resultados de busca do mecanismo podem apresentar um grau de revisão por pares (*peer-review*) e controle editorial (CHO, 2009). Segundo Cho (2009), ao permitir que os bibliotecários e profissionais da informação sugiram URLs de fontes confiáveis sobre uma variedade de tópicos, que o Hakia pode verificar usando seu sistema QDEX, a empresa está solicitando a perícia de um recurso inexplorado, mas que já faz parte do espaço *online*. A empresa Hakia, Inc possui em seu *site* o "Canto dos Bibliotecários", em que bibliotecários e profissionais da informação podem apresentar recursos que atendam a critérios Hakia, incluindo informações revisada por pares, livre de comércio, moeda

índices invertidos são utilizados pelos motores de busca de modo a retornarem rapidamente as páginas onde uma determinada palavra ocorre." (GLOSSÁRIO..., 2012, grifo do autor).

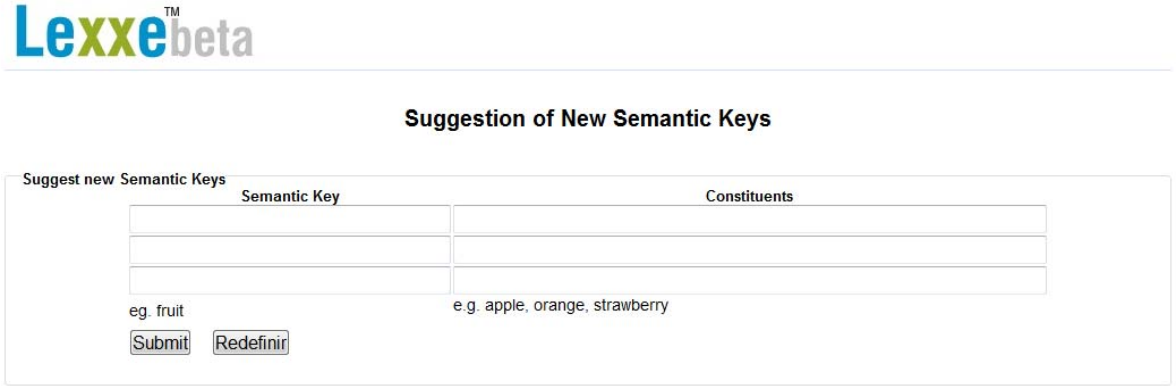
[27] http://newpubmed.com

de conteúdo (*currency of content*) e autenticidade de origem (originalidade do material).

O Lexxe, da empresa Lexxe Pty Ltd, assim como o Hakia, é considerado um mecanismo de busca de terceira geração que tem desenvolvido tecnologia de processamento de linguagem natural avançada. Como o nome do mecanismo sugere, é derivado do termo linguístico "léxico", que significa "relacionado com palavras". O Lexxe enfatiza o processamento da linguagem a partir do nível de palavras e os significados associados a elas, o que é uma questão central para ele.

Em 2005, Lexxe lançou sua versão Alpha, objetivando responder questões (*queries*) concretas em linguagem natural e, sua versão Beta lançada em 2011, agregou tecnologia de busca pela chave semântica (*semantic key*); entretanto, continua se esforçando para melhorar sua tecnologia de resposta às questões (*queries*), já que a maior parte das respostas são retiradas de outros *sites* da Web e de textos não estruturados. Esse mecanismo tem explorado formas mais inteligentes de encontrar informações para os leitores e acredita que este método irá, eventualmente, trazer resultados de busca muito mais precisos e relevantes do que a tecnologia de busca tradicional, pois eles são apresentados com base no significado (LEXXE..., 2012). O Lexxe também possibilita que os leitores contribuam, sugerindo novas chaves semânticas de forma a auxiliar na cobertura de áreas de conhecimento específicas, muito amplas para que o mecanismo e sua equipe trabalhem sozinhos para defini-las (Figura 8, LEE, 2011).

Figura 8 – Caixa do Lexxe para sugestão de chave semântica.



Fonte: http://www.lexxe.com/suggest.html

Observamos que embora as informações obtidas nos *sites* dos mecanismos sejam insuficientes para uma análise mais profunda da sua indexação, identificamos que o Google não trabalha com métodos de indexação ou com sistemas de organização conhecimento (SOC)[28], de acordo com a concepção de Weller (2010). O Google trabalha com algoritmos, pois estes podem ser processados sem qualquer conhecimento sobre seus significados (SANTAELLA, 2009), ou seja, não analisam o conteúdo do texto em um sentido linguístico. Entretanto, ele utiliza a linguagem natural e indexa os conteúdos da Web de acordo com os critérios do algoritmo, aproveitando as palavras utilizadas pelos leitores nas buscas para construir seu índice e retorna-as em buscas futuras adequando-as ao contexto de uso do leitor.

Diferentemente do Google, o Hakia desenvolveu um método de indexação, o QDEX™ (*Query indexing techinique*) e o OntoSem™ (*sense repository*), um sistema de organização do conhecimento que categoriza os diversos sentidos das palavras, que o posiciona, de fato, como um mecanismo de busca semântica, tanto do ponto de vista desta pesquisa quanto do da representação do conhecimento.

O Lexxe, ao contrário do Google, utiliza inovadores algoritmos linguísticos e bases de dados de conhecimento para identificar os tópicos realmente verdadeiros de cada documento (LEXXE..., 2012) e retornar resultados mais relevantes de acordo com a *query* formulada na busca. Quando ocorre o empate em relevância entre páginas da web o Lexxe analisa a popularidade.

Dessa maneira, como a língua somente adquire sentido em uso e os sentidos não são fixos e pré-determinados, não existe razão para haver uma Pragmática *stricto sensu*, bem como uma Semântica *stricto sensu*, e por isso o estudo do sentido (Semântica) torna-se um estudo semântico-pragmático *lato sensu* (FERRAREZI JR., 2010). Assim sendo, a semântica e a pragmática são interdependentes quando se trata do estudo dos mecanismos de busca, porque não dá para desvincular a indexação e a busca, uma vez que entendemos que o mecanismo é o interpretante (imediato) da enunciação da busca e o leitor, o interpretante (dinâmico) dos resultados.

---

[28] Tradução para o português do original inglês "*Knowledge Organization System*" (KOS), proposto pelo *Networked Knowledge Organization Systems Working Group* na 1ª Conferência da ACM *Digital Libraries* em 1998, Pittsburgh, Pennsylvania.

Por fim, verificamos ao identificarmos e selecionarmos os mecanismos de busca que operam com semântica, que eles podem ser classificados de duas maneiras:

a) mecanismos que organizam ou indexam o conteúdo semanticamente, de uso geral; e
b) mecanismos voltados para as tecnologias e linguagens da Web Semântica: ontologias, RDF, XML, etc., de uso dos especialistas.

Desse modo, sugerimos que os mecanismos das ontologias sejam migrados como uma subclasse dos mecanismos da Web Semântica (WS), assim, a categoria “apresentação dos resultados (*searching*)” seria reapresentada conforme o Diagrama 2.

Diagrama 2 – Rediagramação da categoria “Apresentação dos Resultados”.

4) **APRESENTAÇÃO DOS RESULTADOS** (*searching*)	Agrupamento ou Clusterização: a) Verbais b) Visuais Especializados Personalizados Federados Web Profunda Web Semântica a) Semantização da Web por mecanismos gerais – ex. Google (Knowledge Graphic). b) Gerais – ex. Cluuz, Google, Hakia, Lexxe, OntoWeb (inativo). c) Ontologias, RDF etc. – ex. Swoogle, Watson.

8.2 PROCESSO DE BUSCA (SEARCHING)

A semântica e a pragmática são interdependentes, porém, para exemplificarmos como os mecanismos estão processando a busca focamos a análise na pragmática, ou seja, no contexto de uso da linguagem. Entendemos que a *query* é esclarecida por seu contexto e como afirma Lévy (1997) a interface condiciona a dimensão pragmática, porque ela é a ferramenta que garante a ligação da “parte dos fundos” com a “parte da frente” dos mecanismos de busca, segundo Batelle (2006).

Como abordado, consideramos que acompanhando a evolução da Web os mecanismos de busca estão investindo em tecnologias que possibilitem não apenas descobrir, mas compreender a *query* formulada pelo leitor em um determinado contexto. Nessa direção, a busca pragmática que, no nosso entendimento, engloba tanto a busca sintática quanto a busca semântica, é uma busca para a qual os mecanismos estão definindo, sugerindo ou encontrando padrões de busca para apresentarem resultados que façam sentido para o leitor. Dentre as novas configurações utilizadas pelos mecanismos de busca para atribuir sentido e contexto a *query*, destacamos os recursos de *mashup* (lista de possíveis sentidos), o autocomplete e a autosugestão.

No caso do Google, identificamos que além do algoritmo *PageRank*, ele incorporou recentemente uma tecnologia semântica ao seu sistema de busca, o Gráfico do Conhecimento, para oferecer resultados aos leitores que estejam de acordo com o contexto.

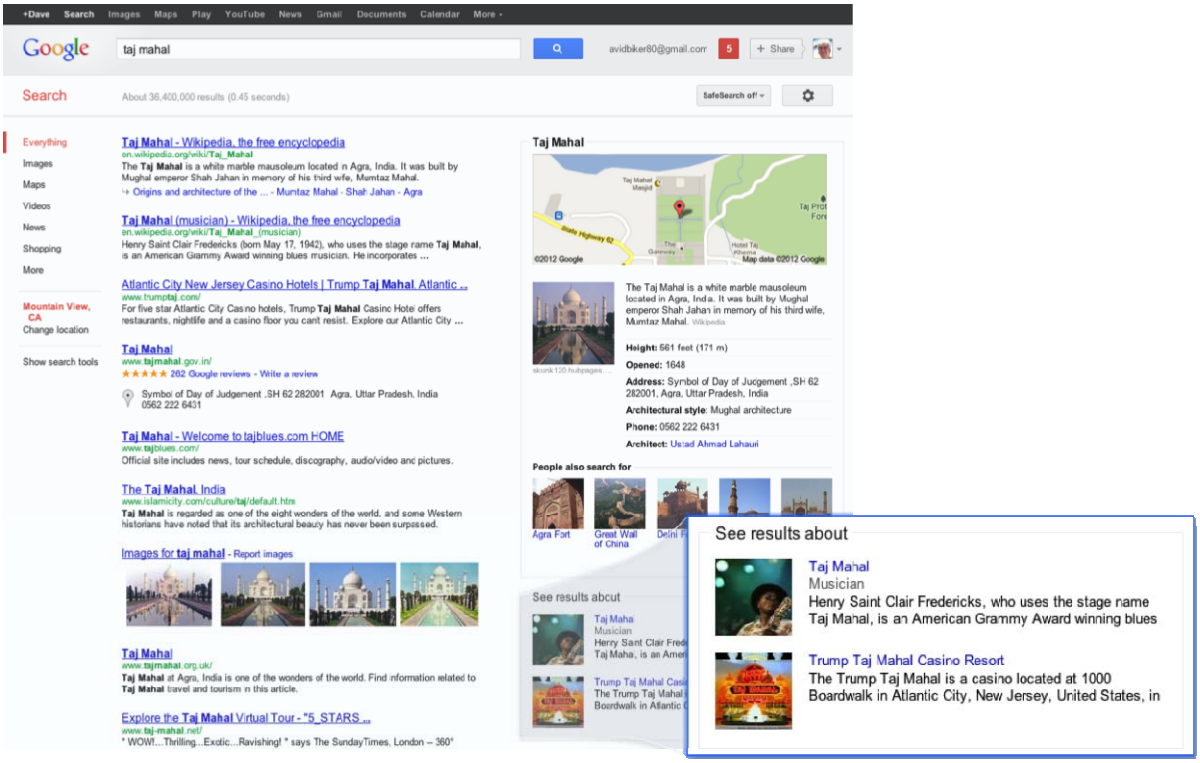
O algoritmo *PageRank* do Google organiza em formato de ranqueamento (*ranking*) os resultados obtidos de seu índice de documentos. Os resultados com maior relevância em relação à consulta do leitor são dispostos por ordem de importância. Trata-se de um sistema para dar notas para páginas web, desenvolvido pelos fundadores do *Google Incorporated*. O *PageRank* usa a estrutura gráfica de *links* da Web como uma ferramenta organizacional e como um indicador do valor de uma página individual. O *PageRank* interpreta um *link* da página A para a página B como um voto da página A para a página B. Mas, analisa também o valor da página que dá o voto. Os votos dados por páginas importantes pesam mais e ajudam a tornar outras páginas importantes (MARCHIORI, 2007).

Desse modo, com o uso do *PageRank*, o Google apresentou uma melhora significativa na classificação dos resultados da busca, mas com o Gráfico

do Conhecimento pretende ser capaz de não só combinar palavras-chave para compreender seu significado, mas entender o contexto e, para isso, foi programado para utilizar cerca de 3,5 bilhões de atributos diferentes para organizar os resultados das buscas por pessoas do mundo real, coisas e lugares. A informação recuperada é proveniente de várias fontes, incluindo a CIA World Factbook, Freebase e Wikipedia. Como o Gráfico do Conhecimento foi disponibilizado apenas para os leitores norte-americanos, utilizamos os exemplos de Singhal (2012) para apresentar como são realizadas as buscas nos três principais formatos: “encontrar a coisa certa” (*find the right thing*), “obter o melhor resumo” (*get the best summary*) e “ir mais profundo e mais amplo” (*go deeper and broader*).

Se o leitor deseja “encontrar a coisa certa”, ao buscar por Taj Mahal, que tanto pode ser o monumento da Índia quanto o músico americano de *blues*, o Google entende agora essa diferença, ou seja, a ambiguidade da linguagem, e restringe os resultados de acordo com o que o leitor quer dizer realmente, e para o leitor ver os resultados específicos é só clicar em um dos *links*. Conforme Singhal (2012), essa é uma forma de busca mais inteligente proporcionada pelo Gráfico do Conhecimento, seus resultados são mais relevantes porque entende estas entidades e as nuances de seu significado como o leitor faz (Figura 9).

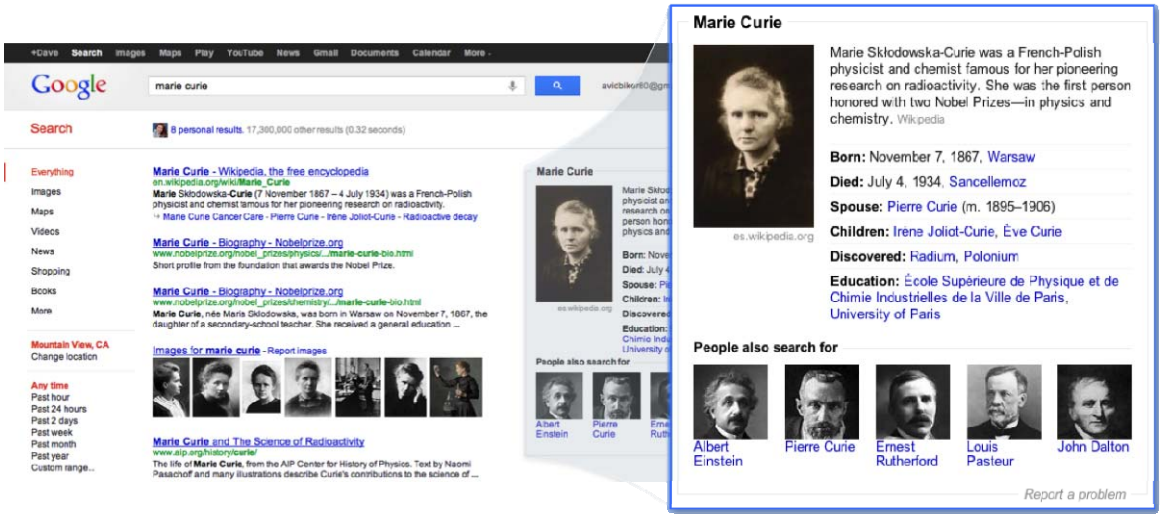
Figura 9 – “Encontrar a coisa certa” - Gráfico do Conhecimento do Google.



Fonte: http://googleblog.blogspot.com.br/2012/05/introducing-knowledge-graph-things-not.html

O Google também pode entender melhor a busca e com a opção “obter o melhor resumo”, resumir o conteúdo relevante sobre um tema, incluindo fatos chave em particular que o leitor provavelmente precisará sobre ele. Por exemplo, se o leitor buscar por Marie Curie, ele saberá quando ela nasceu e morreu, mas também poderá obter mais detalhes sobre a sua educação e suas descobertas científicas, bem como que teve dois filhos, um dos quais também ganhou um Prêmio Nobel, assim como o marido, Pierre Curie, que conseguiu um terceiro Prêmio Nobel para a família (Figura 10). Todas essas informações estão ligadas no gráfico. Aqui observamos que o Google sabe quais fatos são mais importantes para cada item, porque está aproveitando as contribuições dos leitores quando fazem o *upload* da linguagem no momento da busca. Segundo Singhal (2012), o Gráfico do Conhecimento não é apenas um catálogo de objetos, é também todos os modelos dessas inter-relações e é essa inteligência *entre* estas entidades diferentes, que é a chave.

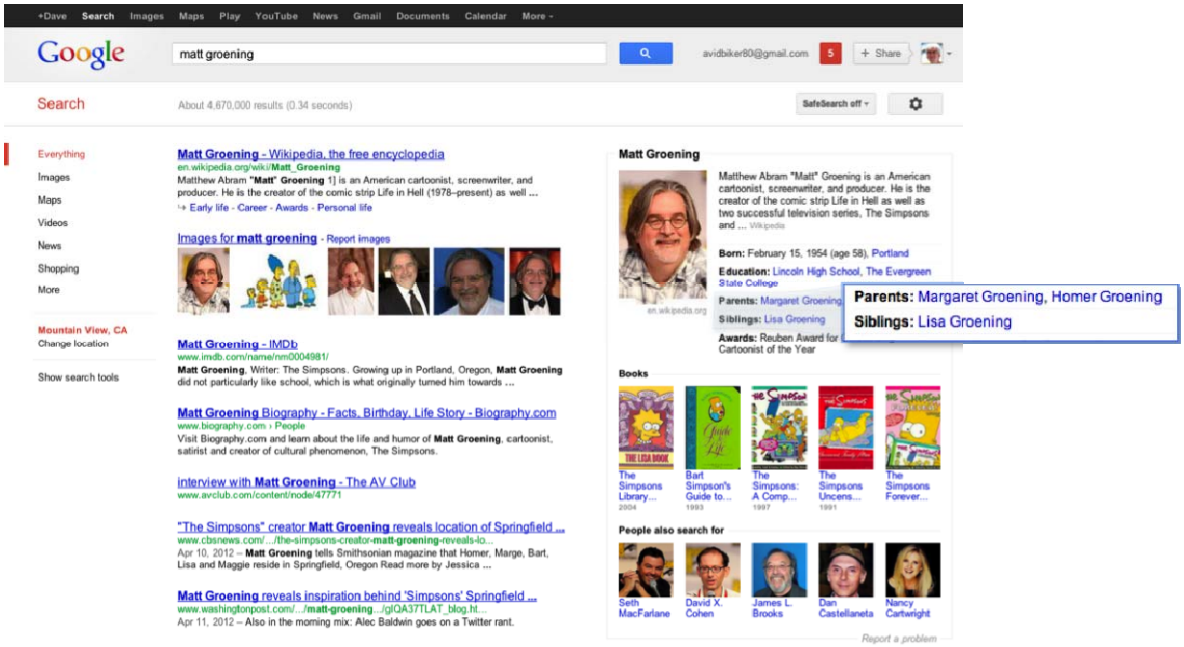
Figura 10 – “Obter o melhor resumo” – Gráfico do Conhecimento do Google.



Fonte: http://googleblog.blogspot.com.br/2012/05/introducing-knowledge-graph-things-not.html

Com o Gráfico do Conhecimento o Google também possibilita que o leitor faça uma busca mais profunda e mais ampla sobre um tema e se surpreenda com descobertas inesperadas e interessantes. De acordo com Singhal (2012), o leitor sempre almejou que um mecanismo de busca perfeito pudesse entender exatamente o que ele quer dizer e retornar exatamente o que ele quer encontrar e o Google agora pode, algumas vezes, responder a próxima pergunta, mesmo antes de ser solicitada, pois os fatos mostrados informam o que as outras pessoas têm buscado. Singhal (2012) afirma que podemos aprender sobre um novo fato ou uma nova conexão que dependa de uma nova linha de investigação. Ele cita, por exemplo, se todos sabem de onde Matt Groening, o criador dos Simpsons, tirou a idéia dos nomes para seus personagens, Homer, Marge e Lisa (Figura 11), para demonstrar que quando o leitor buscar por pessoas que estão em destaque, automaticamente o Google puxará uma caixa de resumo com as principais informações sobre elas.

Figura 11 – “Ir mais profundo e mais amplo” – Gráfico do Conhecimento do Google.



Fonte: http://googleblog.blogspot.com.br/2012/05/introducing-knowledge-graph-things-not.html

O próximo passo do Google será responder às questões mais complexas, como por exemplo, "Quais são os 10 lagos mais profundos dos Estados Unidos?", de acordo com Singhal (2012). Nesse tipo de busca, o mecanismo será capaz de não só entender a *query*, mas informar a quantidade de água, a profundidade, a área de superfície, a temperatura e até mesmo a salinidade de cada lago.

Essa é a nova configuração do Google para busca de pessoas do mundo real, lugares e coisas, e apresentação dos resultados, entretanto, a análise foi prejudicada porque está disponível somente na versão norte-americana e ao acessar a URL do Google internacional somos direcionados automaticamente à página brasileira. Dessa maneira, não foi possível fazer buscas para analisar como estão funcionando os recursos autocomplete e autosugestão nesse tipo de configuração de busca. No entanto, na versão brasileira do Google, observamos que esses recursos funcionam em conjunto (Figura 12).

Figura 12 - *Autocomplete* e *autosuggest* do Google.

Fonte: http://www.google.com.br

Com o *autocomplete*, quando o leitor digita, o algoritmo do Google prevê e exibe buscas baseadas em seu histórico e em atividades de outros leitores. O *autocomplete* ajuda o leitor a evitar erros de digitação e escolher entre os termos apresentados aquele que lhe faça mais sentido ou pode continuar inserindo sua *query* na caixa de busca.

O *autosuggest*, por outro lado, recomenda *queries* relacionadas que incluem os termos de busca inicial. O objetivo desse recurso é apresentar alternativas de busca ao leitor antes mesmo de a busca ser realizada. O leitor poderá selecionar uma das sugestões ou reformular sua busca, pois ao identificar conceitos relacionados, o *autosuggest* o auxilia no refinamento ou expansão de sua *query*.

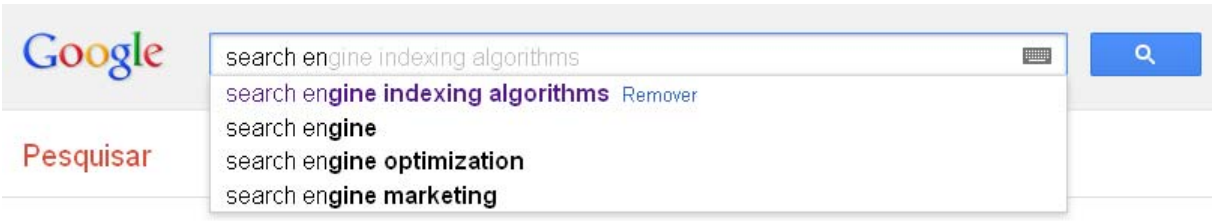
O Google também apresenta no final da página de resultados sugestões de “pesquisas” relacionadas a *query* que o leitor buscou, que também é uma forma de reformular, refinar ou expandir a busca (Figura 13). Observamos que o mecanismo está aproveitando os percursos linguísticos deixados pelos leitores nas buscas, e está construindo uma semântica para retornar a eles como sugestões.

Figura 13 – Buscas relacionadas sugeridas pelo Google.

Fonte: http://www.google.com.br

Com relação aos resultados, observamos que à medida que o leitor digita sua *query*, o Google apresenta automaticamente os resultados, alterando-os de acordo com as palavras que estão sendo inseridas na caixa de busca. Esse recurso chama-se *Google Instant*, é um avanço na tecnologia e um investimento em infraestrutura para auxiliar o leitor a obter resultados de busca melhores e mais rapidamente (GOOGLE, 2012). Todavia, a ordem em que os resultados aparecem depende da popularidade dos termos, os mais procurados ficam melhores posicionados. Contudo, se o leitor não quiser que o Google utilize os termos buscados anteriormente, ele pode removê-los do histórico de busca (Figura 14).

Figura 14 - Remoção do histórico de busca.

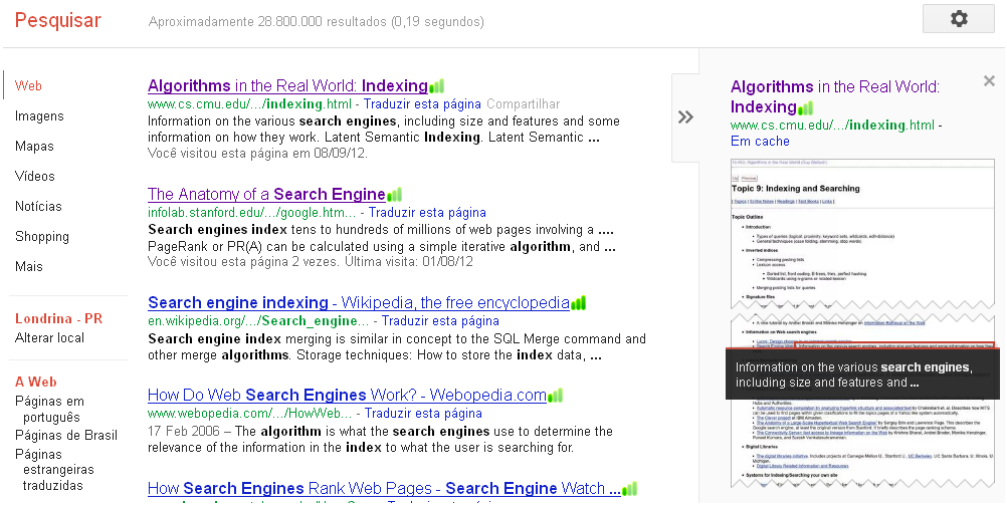


Fonte: http://www.google.com.br

O *Google Instant* também disponibiliza para o leitor o acesso a prévia da página antes de clicar em um resultado da busca, é só passar o cursor sobre um resultado e em seguida sobre as setas exibidas ao lado para visualizá-lo (Figura 15). Em algumas visualizações, partes relevantes da página são exibidas em caixas de chamada de texto sobre a imagem de visualização para ajudar o leitor ver

onde sua consulta aparece na página. Todavia, não são todos os resultados que possuem visualizações disponíveis.

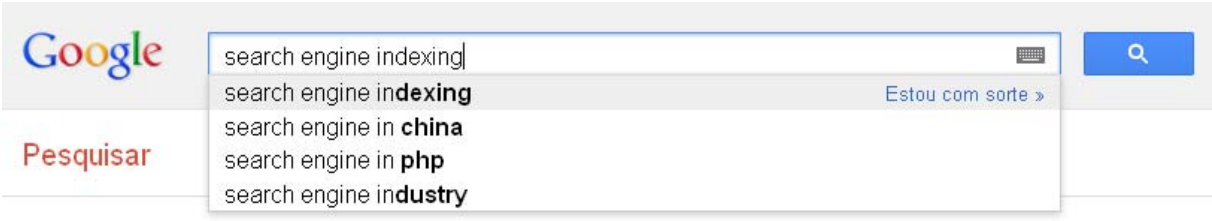
Figura 15 – Pré-visualização do resultado com o *Google Instant*.



Fonte: http://www.google.com.br

Outro recurso disponibilizado pelo Google é o “Estou com sorte”, que traz resultados rápidos para o que o leitor busca (Figura 16). Para utilizá-lo o leitor deve digitar um termo para buscar e clicar neste botão na página inicial do Google para pular a lista e chegar diretamente na primeira página exibida nos resultados.

Figura 16 – Recurso “Estou com sorte” do Google.



Fonte: http://www.google.com.br

O mecanismo de busca Lexxe também disponibiliza os recursos *autocomplete* e *autosuggest* (Figura 17), porém diferentemente do Google, utiliza chaves semânticas para sugerir para os leitores na busca. Entretanto, o Lexxe também aproveita as palavras de busca utilizadas pelos leitores para recuperar conteúdos, mas solicita que eles sugiram chaves semânticas para compor sua base, conforme descrito na análise da forma de organização (*indexing*). O Lexxe possibilita também que o leitor limpe seu histórico de busca (Figura 18).

Figura 17 – *Autocomplete* e *autosuggest* do Lexxe.

Fonte: http://www.lexxe.com/index.html

Figura 18 – Remoção do histórico de busca do Lexxe.

Fonte: http://www.lexxe.com/index.html

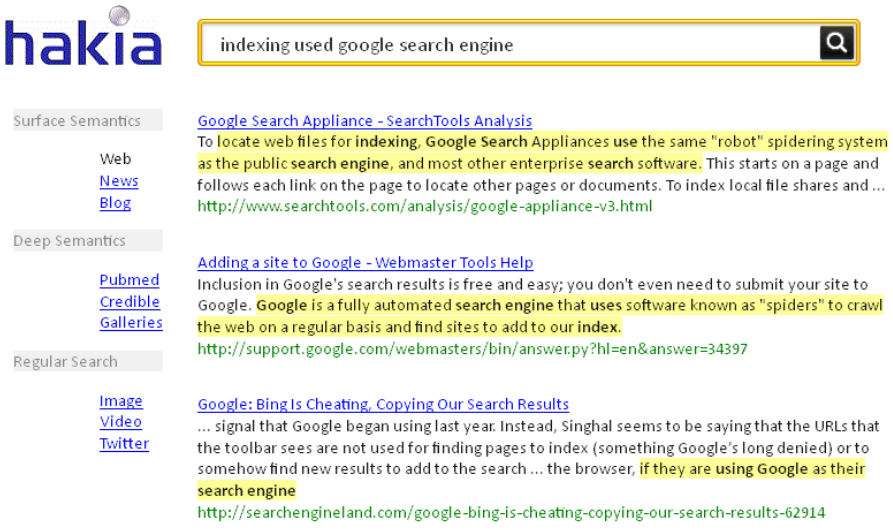
O Hakia, por sua vez, não disponibiliza os recursos *autocomplete* e *autosuggest*, bem como nenhum outro recurso que auxilie o leitor na busca, do ponto de vista pragmático da interface.

Com relação ao objetivo de exemplificar como os mecanismos estão processando a busca (pragmática), selecionamos a *query* “indexing used google search engine” e realizamos a busca no Google, no Hakia e no Lexxe (Figuras 19, 20 e 21).

Figura 19 – Busca no Google por “indexing used google search engine”.

No Google, observamos que embora a *query* não seja tão ampla, pela quantidade dos resultados, aproximadamente 3.500.000.000, não é muito buscada pelos leitores, porque o mecanismo não apresenta ao final da página a lista das “pesquisas” relacionadas, e sugere somente uma outra alternativa de busca (*type indexing used google*). Assim sendo, o leitor terá que ir interagindo com o mecanismo, acrescentando ou substituindo termos, reformulando a busca de acordo com seu contexto. O leitor pode utilizar a “pesquisa avançada” do Google, onde poderá buscar a expressão ou frase exata ou colocar os termos da *query* entre “aspas” na caixa de busca que exerce a mesma função. Nesse caso, não houve resultados na busca pela *query* “indexing used google search engine”, pela opção da “pesquisa avançada”. Entretanto, tanto no Google (Figura 19) quanto no Hakia (Figura 20) a busca foi realizada sem o uso das aspas.

Figura 20 – Busca no Hakia por “indexing used google search engine”.



O Hakia é um mecanismo pioneiro na tecnologia de busca semântica e diferentemente do Google faz a busca por meio da semântica usada nas frases e como não disponibiliza os recursos de *autocomplete* e *autosuggest*, o leitor tem que digitar a *query* diretamente na caixa de entrada. Todavia, por meio de um *add-on*[29] para navegadores os leitores podem encontrar no Hakia documentos exatos que contém a resposta para os termos de busca. Dessa forma, observamos que esse mecanismo está preocupado com a qualidade, não com a popularidade, como o Google, que usa métodos estatísticos de classificação (*ranking*) para definir a popularidade de um site.

No caso específico da busca, observamos que o Hakia destacou nos resultados o parágrafo onde os termos apareceram com maior relevância, embora não esteja entre os objetivos dessa pesquisa medir esse coeficiente. O Hakia utiliza critérios de relevância e data na apresentação dos resultados, mas isso só está explícito nas outras subcategorias, dentro da categoria *Surface Semantics*, onde está inserida a Web. Diferentemente do Google, o Hakia não apresenta o tempo de busca e a quantidade de resultados obtidos na busca.

No Lexxe, fizemos inicialmente a busca livre com a mesma *query* executada nos outros mecanismos, e obtivemos quase 40.000 resultados (Figura 21); um número bem inferior a do Google, porém demonstrou que a busca estava ainda muito ampla. Como o Lexxe possibilita a seleção de uma chave semântica, por meio do recurso *autocomplete* e *autosuggest*, selecionamos chave "*search engine:*" e acrescentamos a ela os termos "*indexing* Google" na caixa de busca, e observamos que houve uma pequena diminuição na quantidade dos resultados (Figura 22).

[29] Também conhecido por *plug-in* e *add-in*, o *add-on* "[...] é um programa de computador usado para adicionar funções a outros programas maiores, provendo alguma funcionalidade especial ou muito específica." (http://pt.wikipedia.org/wiki/Plugin)

Figura 21 – Busca no Lexxe por “indexing used google search engine”.

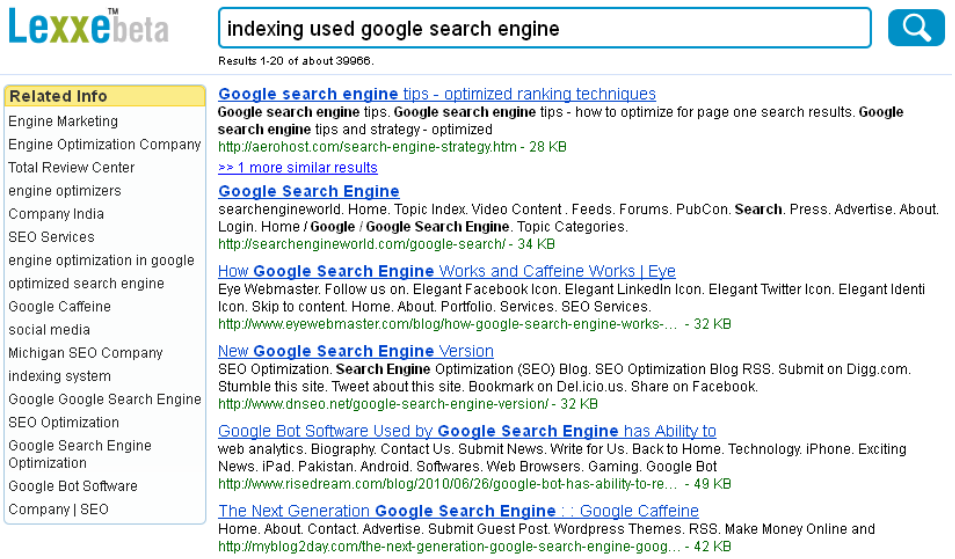
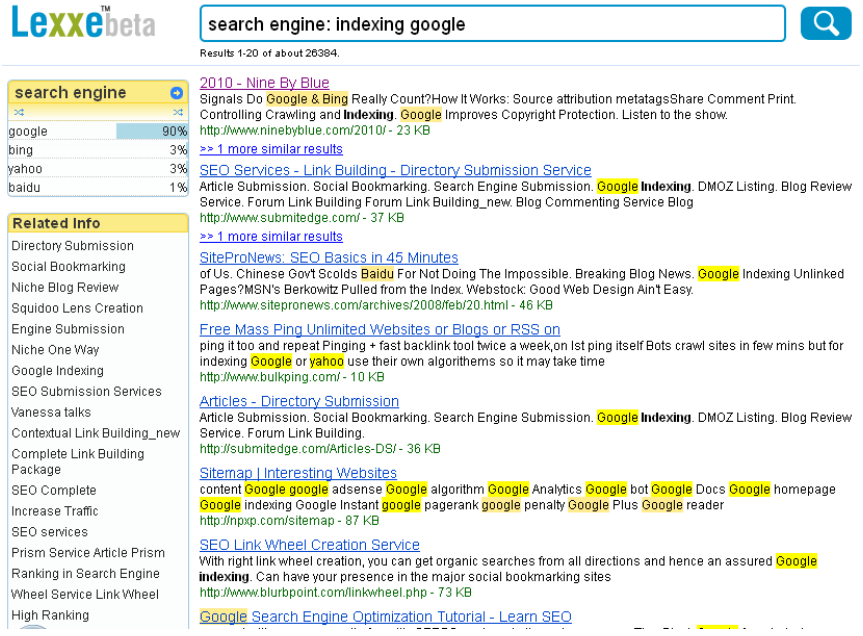


Figura 22 – Busca no Lexxe com a chave semântica “*search engine*:”.



Entretanto, como o Lexxe apresenta do lado direito da sua interface, uma lista de informações relacionadas (*Related info*), que o leitor pode utilizar para melhorar os resultados da busca, substituímos os termos “*indexing* Google” por “Google *indexing*”, sugerido nesta lista; porém não houve uma diminuição significativa na quantidade dos resultados apresentados (Figura 23). Entretanto, percebemos que a ordem em que os termos são posicionados na *query*, interfere nos resultados. Dessa forma, voltamos a utilizar os termos da *query* inicial

associados à chave semântica e observamos que com a inserção da palavra *used* a quantidade dos resultados diminuíram de forma significativa (Figura 24).

Figura 23 – Busca no Lexxe utilizando sugestão da lista "*Related info*".

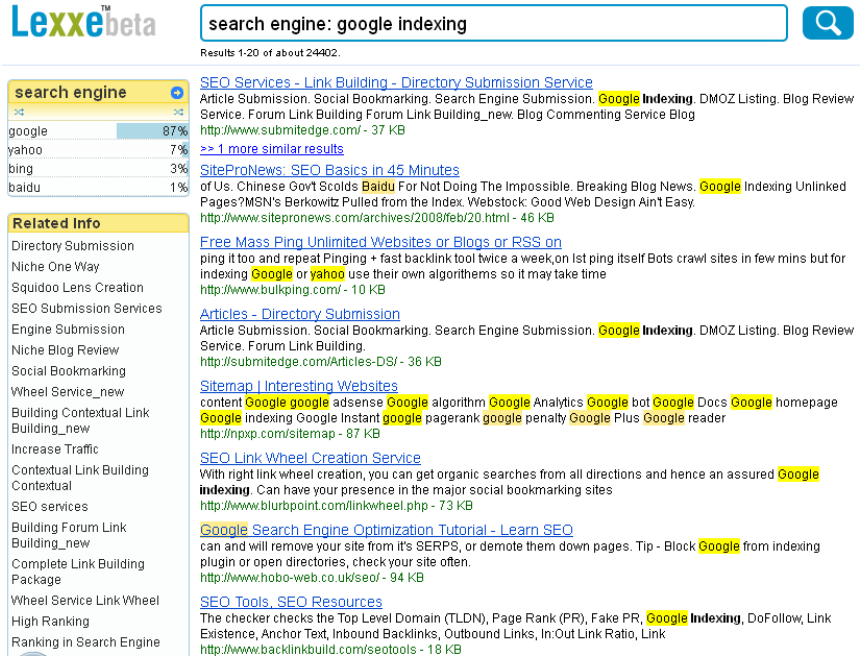
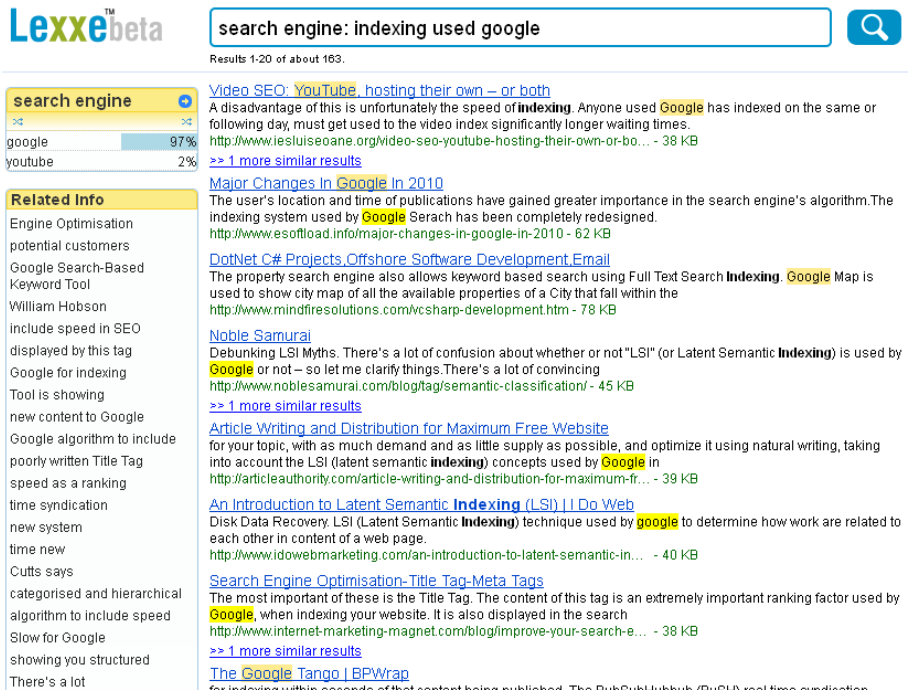


Figura 24 – Busca no Lexxe com a adição da palavra *used* à chave semântica "*search engine*".



Por fim, deduzimos que a adição da palavra “*used*” possibilitou a diminuição significativa na quantidade dos resultados porque atribuiu mais contexto a *query*. Contudo, sabemos que os resultados da busca dependem também do tamanho da base de dados (índice) do mecanismo de busca, entretanto, no caso do Lexxe, a maioria das respostas são extraídas tanto de outros sites da Web quanto de textos não estruturados.

9 CONSIDERAÇÕES FINAIS

A Web e os mecanismos de busca, como tecnologias da informação, são objetos que estão em constante modificação, tanto do ponto de vista técnico quanto conceitual. Assim, nessa pesquisa buscamos os aportes da Linguística e da Filosofia para desenvolver os conceitos referentes à dimensão semântica e pragmática da Web e dos mecanismos de busca objetivando estudar os mecanismos de busca que operam com semântica e a busca contextual.

Os mecanismos de busca, como objetos de mediação tecnológica, fazem a organização virtual do conhecimento por meio da indexação e por sua interface de busca possibilitam o acesso aos signos e as linguagens do ciberespaço. Entretanto, como objetos de estudo, os mecanismos de busca ainda são poucos pesquisados na área da Ciência da Informação, por isso a necessidade de buscar em outras áreas a fundamentação teórica para analisá-los porque alguns conceitos não estão muito evidentes na literatura.

Dessa forma, com base no aporte teórico da Linguística e da Filosofia foi possível relacionarmos o conceito de semântica ao sentido e o conceito de pragmática ao contexto de uso da linguagem. Assim sendo, compreendemos que a semântica no ciberespaço pode ser analisada melhor sob uma perspectiva sígnica uma vez que tanto os mecanismos de busca quanto os leitores são interpretantes do fundamento, sendo o primeiro, o interpretante da enunciação, e o segundo, o interpretante dos resultados da busca. Também foi possível compreender que a pragmática está relacionada ao contexto de uso da linguagem e ao final da pesquisa, entendemos que o contexto é no fundamento do signo, o objeto dinâmico, é o objeto em si mesmo, objeto real ou abstrato, verdadeiro ou falso, em suma, é a dimensão da realidade ou daquilo que achamos realidade. Conforme Santaella (2009), o objeto é algo diferente do signo e que está fora dele, um ausente, mas que pode se tornar mediatamente presente a um intérprete graças à mediação do signo.

Observamos que essa relação triádica entre estes três elementos, o fundamento, o objeto e o interpretante, que estão íntima e inseparavelmente interconectados (SANTAELLA, 2009), se reflete na forma de organização (*indexing*) e no processo de busca (*searching*); por isso, chegamos à conclusão de que a semântica e a pragmática são interdependentes quando se trata do estudo dos mecanismos de busca.

Nesse sentido, a dificuldade de desvincular a indexação e a busca tem impulsionado os mecanismos de busca a aperfeiçoarem seus índices e suas interfaces para adequarem-se ao contexto dos leitores que estão cada vez mais exigentes no quesito busca de informação na Web.

Desse modo, verificamos que dentre os mecanismos de busca analisados, os que operam com a semântica, de acordo com os pressupostos delineados para o estudo, são o Google e o Lexxe, porque têm utilizado da colaboração dos leitores tanto para melhorar na geração do índice quanto na definição de padrões de busca para que os resultados sejam obtidos em um contexto pragmático.

O Google, por exemplo, tem aproveitado todo volume de *queries* que são inseridas pelos leitores para gerar padrões de busca, ou seja, ele aproveita a linguagem humana para desenvolver esses padrões, porque sua tecnologia possibilita a colaboração de forma explícita, tanto que quando o leitor aceita colaborar, o mecanismo faz o registro do computador para capturar as informações das estratégias de busca elaboradas e manipular os sentidos do conteúdo das buscas e capturar os contextos de uso. Dessa forma é que o Google tem trabalhado a dimensão pragmática na *query*, na busca e no comportamento de busca. Ele aproveita os traços linguísticos deixados pelos leitores para construir uma semântica e retornar aos leitores como sugestão. Observamos, então, que a dimensão pragmática está mais presente na Web 2.0 ou Social do que na Web Semântica e que os sentidos são atualizações do contexto.

Com o Hakia identificamos que os mecanismos de busca estão ampliando o campo de atuação dos profissionais da informação e que, no caso do bibliotecário, ele pode tanto recomendar páginas web confiáveis para o mecanismo indexar quanto trabalhar em conjunto com outros profissionais na área de otimização de *websites* (*Search Engine Optimization* - SEO), pois com o conhecimento que possui poderá contribuir significativamente nos aspectos relacionados ao tratamento, organização da informação e do conhecimento, para representação de conteúdos eletrônicos na *Web* (OLIVEIRA et al., 2011).

Acredita-se que o estudo também servirá para que os bibliotecários adquiram mais conhecimento das funcionalidades dos mecanismos de busca, uma vez que a tendência atual das interfaces de busca dos catálogos online (OPACs) é reunir em um índice único, todas as informações, metadados, texto completo,

arquivos multimídia etc, para facilitar o acesso em um ambiente similar ao dos *sites* de busca da Web, aos quais os leitores estão familiarizados, e a recuperação da informação por meio da busca federada, simultânea e em tempo real, em todas as coleções de responsabilidade da biblioteca (biblioteca digital, repositórios etc.).

Durante a revisão da literatura e depois na coleta e análise dos dados constatamos também que a pesquisa pode contribuir para a categorização ou tipologia dos mecanismos de busca no ciberespaço, por isso sugerimos a rediagramação da categoria “apresentação dos resultados”, conforme apresentada e discutida nos resultados.

Nesse sentido, sugerimos também que o estudo desses mecanismos de busca no âmbito do grupo de pesquisa “Informação e Conhecimento no Ciberespaço” tenha continuidade uma vez que a *search engine indexing* é uma área que está em desenvolvimento e ainda possui diversos aspectos pouco explorados no meio acadêmico e na Ciência da Informação.

Por fim, acreditamos que será interessante também, para um futuro próximo, o desenvolvimento de estudos mais complexos que contemplem o aprofundamento dos conceitos teóricos e práticos da busca semântica e da busca pragmática, que não foi possível desenvolver nessa pesquisa, bem como sobre a reconfiguração da profissão do bibliotecário para atuar no ambiente da Web.

REFERÊNCIAS

AABERGE, Terje; AKERKAR, Rajendra; BOLEY, Harold. An intensional perspective on the semantic and pragmatic web. **International Journal of Metadata, Semantics and Ontologies**, v. 6, n. 1, p. 74-80, 2011.

ABBAGNANO, Nicola. **Dicionário de filosofia**. 5. ed. São Paulo: Martins Fontes, 2007.

ALESSO, H. Peter; SMITH, Craig F. **Thinking of the Web**: Berners-Lee, Gödelm and Turing. New Jersey: John Wiley & Sons, 2009.

ALMEIDA Maurício Barcellos, SOUZA, Renato Rocha. Avaliação do espectro semântico de instrumentos para organização da informação. **Encontros Bibli**, Florianópolis, v. 16, n. 31, p.25-50, 2011.

ANGOTTI, David. Algorithm Update: Google Focusing on Semantic Search Technology. **SEJ Search Engine Journal**, 16 March 2012. Disponível em: <http://www.searchenginejournal.com/google-algorithm-semantic-search/41477/>. Acesso em: 8 ago. 2012.

ANTONIOU, Grigoris et al. Semantic Web fundamentals. In: KHOSROW-POUR, Mehdi. **Encyclopedia of Information Science and Tecnholgy**. Hershey: Idea Group Reference, 2005. v. 5, p. 2464-8.

ARMENGAUD, Françoise. **A pragmática**. 2. ed. São Paulo: Parábola Editorial, 2008.

BADR, Youakim et al. (Ed.). **Emergent web intelligence**: advanced semantic technologies. London: Springer, 2010.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern information retrieval.** New York: ACM Press, Addison-Wesley, 1999.

BATTELLE, John. **A busca**. Campinas: Campus; Rio de Janeiro: Elsevier, 2006.

BERNERS-LE, Tim; FISCHETTI, Mark. **Weaving the Web**: the original design and ultimate destiny of the World Wide Web. New York: HarperCollins, 2000. 246p.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, may 2001. Disponível em: <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: 19 fev. 2011.

BERRY, Michael W.; BROWNE, Murray. **Understanding search engines**: mathematical modeling and text Retrieval. 2nd ed. Philadelphia: Society for Industrial and Applied Mathematics, 2005.

BLIKSTEIN, Izidoro. **Kaspar Hauser ou a fabricação da realidade**. São Paulo: Cultrix, 1995.

BONINO, Dario. et al. Ontology Driven Semantic Search. **WSEAS Transaction on Information Science and Application**, v. 1, n. 6, p. 1597-1605, 2004.

BORBA, Francisco S. **Dicionário de usos do português do Brasil**. São Paulo: Ática, 2002.

BOUYER, Gilbert Cardoso. Pragmatismo e cognição: *self*, mente, mundo e verdade na teoria pragmática do conhecimento. **Ciências & Cognição,** v. 15, n. 3, p. 164-179, 2010. Disponível em: <http://www.cienciasecognicao.org>. Acesso em: 2 out. 2011.

BRANSKI, Regina Meyer. Recuperação de informações na web. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 9, n. 1, p. 70-87, jan./jun. 2004.

BRASCHER, Marisa. A ambigüidade na recuperação da informação. **DataGramaZero** [Online], v. 3, n. 1, fev. 2002. Disponível em: <http://www.datagramazero.org.br/fev02/Art_05.htm>. Acesso em: 25 maio 2011.

BREITMAN, Karin. **Web semântica**: a internet do futuro. Rio de Janeiro: LTC, 2010.

BRODER, Andrei. **A taxonomy of web search.** 2002. Disponível em: <http://www.sigir.org/forum/F2002/ broder.pdf>. Acesso em: 23 abr. 2012.

BUENO, Márcia Correa; VIDOTTI, Silvana Aparecida Borsetti Gregório. Ferramentas de busca na Internet: para quê, por quê e como utilizá-las? In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 11., 2000, Florianópolis. **A biblioteca universitária no século XXI**. Florianópolis: FEBAB, 2000. Disponível em: <http://snbu.bvs.br/snbu2000/docs/pt/doc/t100.doc>. Acesso em: 18 jul. 2011.

BUSBY, Michael. All about search engines. In: _____. **Learn Google™**. Plano, Texas: Wordware Publishing, 2004. Chapter 1, p. 1-34.

CATARINO, Maria Elisabete; BAPTISTA, Ana Alice. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **DataGramaZero** [Online], v. 8, n. 3, jun. 2007. Disponível em: <http://www.datagramazero.org.br/jun07/Art_04.htm>. Acesso em: 19 out. 2011.

_____; _____. Web Semântica e a qualidade no intercâmbio da informação. In: TOMAÉL, Maria Inês (Org.). **Fontes de informação na Internet**. Londrina: EDUEL, 2008. p. 31-51.

CENDÓN, Beatriz Valadares. Ferramentas de busca na web. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001.

CHAÍN NAVARRO, Célia. **Técnicas y métodos de recuperación de información**. Murcia: DM, 2004.

CHO, Allan. Hakia, search engine, and librarians: how expert searchers are building the next generation Web. **Internet@suite101®**, jan 10, 2009. Disponível em:

<http://suite101.com/article/hakia-search-engines-and-librarians-a89266>. Acesso em: 16 jul. 2012.

CODINA, Lluís; ABADAL, Ernest, ROVIRA, Cristòfol. Búsqueda federada en el ecosistema de la e-ciencia: el caso Science Research. **El Profesional de la información**, v. 19, n. 1, p.77-85, 2010. Disponível em: <http://www.lluiscodina.com/scienceResearch.pdf>. Acesso em: 16 jul. 2012.

CODINA, Lluís; ROVIRA, Cristòfol. La Web semántica. In: TRAMULLAS, Jesús (coord.). **Tendencias en documentación digital**. Guijón: Trea, 2006. Cap. 1, p. 9-54.

COHEN, Laura. **Checklist of internet research tips**. 1998. Disponível em: <http://www.hsc.wvu.edu/aap/education/Faculty_Development/internet/checklist.html>. Acesso em: 16 mar. 2011.

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008.

DELEUZE, Gilles. **Lógica do sentido**. 5. ed. São Paulo: Perspectiva, 2009.

DELEUZE, Gilles; GUATTARI, Félix. **Mil platôs**: capitalismo e esquizofrenia. São Paulo: Ed. 34, 1997. v. 2.

DREYFUS, Hubert L. **On the internet**: thinking in action. 2nd ed. New York: Routledge, 2009.

FAVARETTO, Eduardo. **Uma nova Torre de Babel**. 2006. Disponível em: <http://www.ibuscas.com.br/site/artigos/uma_nova_torre_de_babel.html>. Acesso em: 18 jul. 2011.

FERNEDA, Edberto. **Recuperação da informação**: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. 137f. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo.

_____. **Introdução aos modelos computacionais de recuperação da informação**. Rio de Janeiro: Editora Ciência Moderna, 2012.

FERRAREZI JR., Celso. **Introdução à semântica de contextos e cenários**: de langue à la vie. Campinas: Mercado de Letras, 2010.

FERREIRA, Aurélio Buarque de Holanda. **Novo dicionário Aurélio da língua portuguesa**. 3. ed. rev. e atual. Curitiba: Positivo, 2004.

FIGUEIREDO, Márcia Feijão de. **Busca e validação da informação imagética na *web***. 2011. 108f. Dissertação (Mestrado em Ciência da Informação) - Convênio Instituto Brasileiro de Informação em Ciência e Tecnologia e Universidade Federal do Rio de Janeiro/ Faculdade de Administração e Ciências Contábeis, Rio de Janeiro.

FRANCO, Maria Laura Puglisi Barbosa. **Análise de conteúdo**. Brasília: Plano Editora, 2003.

FUCHS, C. **Les ambiguïtés du français**. Paris : Orphys, 1996.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

_____. **Métodos e técnicas de pesquisa social.** São Paulo: Atlas, 1999.

GIL LEIVA, Isidoro. **Manual de indización**: teoria y práctica. Gijón: Ediciones Trea, 2008.

GOOGLE Inc. **Sobre o Google Instant.** Disponível em: <http://www.google.com/insidesearch/features/instant/about.html>. Acesso em: 4 jul. 2012.

GOOGLEBOT – Webmaster Tools Help. Disponível em: <http://support.google.com/webmasters/bin/answer.py?hl=en&answer=182072>. Acesso em: 4 jul. 2012.

GLOSSÁRIO de informática e tradução de termos informáticos. **Visibilidade.net**. Disponível em: <http://visibilidade.net/tutorial/glossario-informatica.html#I>. Acesso em: 16 jul. 2012.

GODOY, Arilda Schmidt. Introdução à pesquisa qualitativa e suas possibilidades. **Revista de Administração de Empresas**, São Paulo, v. 35, n. 2, p. 57-63, abr. 1995.

GRACIOSO, Luciana de Souza. Justificação e a ação de informação no contexto da pragmática virtual. **Liinc em Revista**, Rio de Janeiro, v. 6, n. 2, p. 286-300, set. 2010. Disponível em: <http://www.ibict.br/liinc>. Acesso em: 11 ago. 2011.

GRESHAM, Keith. Surfing with a Purpose: Process and strategy put to the test on the Internet. **Educom Review** [online], v. 33, n. 5, 1998. Disponível em: <http://www.educause.edu/ir/library/html/erm9851.html>. Acesso em: 16 mar. 2011.

GROGAN, D. **A prática do serviço de referência**. Brasília: Briquet de Lemos Livros, 2001.

HENDLER, James. Web 3.0: the dawn of semantic search. **Computer**, p. 77-80, jan. 2010.

HOUAISS, Antônio; VILLAR, Mauro de Salles; FRANCO, Francisco Manoel de Mello. **Dicionário da língua portuguesa**. Rio de Janeiro: Objetiva, 2004.

HUANG, Lieming et al. Adaptively constructing the query interface for meta-search engines. In: INTERNATIONAL CONFERENCE ON INTELLIGENT USER INTERFACES (IUI '01), 6., 2001, New York. **Proceedings…** New York: ACM, 2001.

ILARI, Rodolfo; GERALDI, João Wanderley. **Semântica**. 11. ed. São Paulo: Ática, 2006.

ISKOLD, Alex. Hakia – first meaning-based search engine. **ReadWriteWeb**, Dec 2006. Disponível em: <http://www.readwriteweb.com/archives/hakia_meaning-based_search.php>. Acesso em: 3 ago. 2012.

JOHNSON, J. David. On contexts of information seeking. **Information Processing and Management,** v. 39, p. 735-60, 2003.

KASSIM, Junaidah Mohamed; RAHMANY, Mahathir. Introduction to Semantic Search Engine. 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia. p. 380-6. p. 384.

KELION, Leo. The future of Google search: thinking outside the box. **BBC News. Technology**, June 2012. Disponível em: <http://www.bbc.co.uk/news/technology-18327263>. Acesso em:

KOBASHI, Nair Yumiko; FERNANDES, Joliza Chagas. Pragmática linguística e organização da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 10., 2009, João Pessoa. **A responsablidade social da Ciência da Informação**. João Pessoa: ANCIB, 2009. v. 1.

KOEPSELL, David R. **A ontologia do ciberespaço**: a Filosofia, a lei e o futuro da propriedade intelectual. São Paulo: Madras, 2004.

KOO, Lawrence. O papel da Web 3.0 no consumo contemporâneo. **Pensamento & Realidade**, v. 24, n. 2, p. 109-24, 2009. Disponível em: <http://revistas.pucsp.br/index.php/pensamentorealidade/article/view/7086/5127>. Acesso em: 18 jun. 2012.

KURAMOTO, Hélio. Sintagmas nominais: uma nova abordagem no processo de indexação. In: NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Org.). **Organização da informação**: princípios e tendências. Brasília, DF: Briquet de Lemos, 2006. Cap. 8, p. 117-137.

LANCASTER, F. Wilfrid. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2003.

_____. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos/Livros, 1993.

LEÃO, Lucia. **O labirinto da hipermídia**: arquitetura e navegação no ciberespaço. 3. ed. São Paulo: Iluminuras, 2005.

LEE, Michael. Aussie Lexxe challenges search engines. **ZDNet**, nov. 2011. Disponível em: <http://www.zdnet.com/aussie-lexxe-challenges-search-engines-1339325469/>. Acesso em: 30 ago. 2012.

LEUF, Bo. **The semantic web**: crafting infrastructure for agency. Chichester: John Wiley & Sons, 2006.

LEVENE, Mark. **An introduction to search engines and web navigation**. New Jersey: John Wiley & Sons, 2010.

LÉVY, Pierre. O ciberespaço ou a virtualização da comunicação. In: _____. **Cibercultura**. 2. ed. São Paulo: Ed. 34, 2000. p. 85-107.

_____. **As tecnologias da inteligência.** Rio de Janeiro: Ed. 34, 1997.

LEXXE search engine. Disponível em: <http://www.lexxe.com>. Acesso em: 30 ago. 2012.

LIANG, Lin; RONG, Wenge; LIU, Kecheng. Intelligent Agents for Pragmatic Web Services. In: INTERNATIONAL CONFERENCE ON ADVANCED LANGUAGE PROCESSING AND WEB INFORMATION TECHNOLOGY (ALPIT 2007*)* (ALPIT '07), 6., 2007, Washington. **Proceedings…** Washington, DC: IEEE Computer Society, 2007. p. 530-6.

LIMA, Gercina Ângela Borém. Organização da informação para sistemas de hipertextos. In: NAVES, Madalena Martins Lopes; KURAMOTO, Hélio (Org.). **Organização da informação**: princípios e tendências. Brasília, DF: Briquet de Lemos, 2006. Cap. 7, p. 99-116.

LOS NUEVOS buscadores inteligentes que prometem revolucionar la Web. **NuevoDiarioWeb**, Santiago del Estero, 22 feb. 2012. Disponível em: <http://www.nuevodiarioweb.com.ar/notas/2012/2/22/nuevos-buscadores-inteligentes-prometen-revolucionar-387011.asp>. Acesso em: 30 jul. 2012.

LÜDKE, Menga; ANDRÉ, Marli E.D.A. **Pesquisa em educação**: abordagens qualitativas. São Paulo: EPU, 1986.

LUZ GARCÍA, Irma; PORTUGAL, Mercedes. **Servicio de referencia**: uma propuesta integradora. Buenos Aires: Alfagrama, 2008.

MANGOLD, Christoph. A survey and classification of semantic search approaches. **International Journal Metadata, Semantics and Ontology**, v. 2, n. 1, p. 23–34, 2007.

MARCHIORI, Patrícia Zeni. Google we trust? Redesenhando o acesso a recursos de informação. In: GIANNASI-KAIMEN, Maria Júlia; CARELLI, Ana Esmeralda (Org.). **Recursos informacionais para compartilhamento da informação**. Rio de Janeiro: E-Papers, 2007. Cap. 4, p. 99-123.

MARCONDES, Carlos Henrique. Um modelo semântico de publicações eletrônicas. **Liinc em Revista**, Rio de Janeiro, v. 7, n. 1, p. 82-103, mar. 2011. Disponível em: <http://www.ibict.br/liinc>. Acesso em: 11 jul. 2011.

MARCOS, Mari-Carmen; GONZÁLEZ-CARO, Cristina. Comportamiento de los usuarios en la página de resultados de los buscadores. Un estudio basado en *eye tracking*. **El profesional de la información**, v. 19, n. 4, p. 348-358, julio-agosto 2010. Disponível em: <http://grupoweb.upf.es/WRG/dctos/marcos__gonzalez_2010.pdf>. Acesso em: 30 jul. 2012.

MIRANDA, Marcos Luiz Cavalcanti de. **Organização e representação do conhecimento**: fundamentos teórico-metodológicos na busca e recuperação da informação em ambientes virtuais. 2005. 351f. Tese (Doutorado em Ciência da Informação) - Convênio CNPq/IBICT – UFRJ/ECO, Rio de Janeiro.

MONTEIRO, Silvana Drumond. A organização virtual do conhecimento no ciberespaço. **DataGramaZero [online]***,* v. 4, n. 6, dez. 2003.

_____. O ciberespaço e os mecanismos de busca: novas máquinas semióticas. **Ciência da Informação**, Brasília, v. 35, n. 1, p. 31-8, jan./abr. 2006.

_____. O ciberespaço: o termo, a definição e o conceito. **DataGramaZero [online]**, v. 8, n. 3, jun. 2007.

_____. Os mecanismos de busca: à guisa de uma tipologia das múltiplas sintaxes. In: TOMAÉL, Maria Inês (Org.). **Fontes de informação na Internet**. Londrina: EDUEL, 2008. p. 97-122.

_____. A organização do conhecimento no ciberespaço [Editorial]. **Informação & Informação**, Londrina, v. 14, n. esp., 2009a.

_____. As múltiplas sintaxes dos mecanismos de busca no ciberespaço. **Informação & Informação**, Londrina, v. 14, n. esp., p. 68-102, 2009b. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/2027/3223>. Acesso em: 20 fev. 2012.

_____. **Os mecanismos de busca: investigação das múltiplas sintaxes de organização e busca de informação e conhecimento no ciberespaço:** representação visual do projeto. Disponível em: <http://www.uel.br/grupo-pesquisa/ciberespaco/cubos/cubo2.html>. Acesso em: 3 set. 2012.

MONTEIRO, Silvana Drumond; GIRALDES, Maria Júlia Carneiro. Aspectos lógico-filosóficos da organização do conhecimento na esfera da ciência da informação. **Informação & Sociedade**, João Pessoa, v. 18, n. 3, p. 13-27, set./dez. 2008.

MONTEIRO, Silvana Drumond et al. Em busca da compreensão da “busca” n o ciberespaço. In: ENANCIB ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília. [**Anais...**]. Brasília: ANCIB, 2011. p. 2536-2551.

MOOR, Aldo de. Patterns for the Pragmatic Web. In: INTERNATIONAL CONFERENCE ON CONCEPTUAL STRUCTURES (ICCS 2005), 13., 2005, Kassel, Germany. **Proceedings…** Berlin: Springer Verlag, 2005.

MOOR, Aldo de; KEELER, Mary; RICHMOND, Gary. Towards a Pragmatic Web. In*:* PRISS, Uta; CORBETT, Dan; ANGELOVA, Galia (Eds.). **Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces (ICCS '02)***.* London: Springer-Verlag, 2002. Disponível em: <http://www.cspeirce.com/menu/library/aboutcsp/richmond/web.pdf>. Acesso em: 23 nov. 2011.

MOREIRA, Sonia Virgínia. Análise documental como método e como técnica. In: DUARTE, Jorge; BARROS, Antônio (Org.). **Métodos e técnicas de pesquisa em comunicação.** São Paulo: Atlas, 2005. p. 269-279.

MORVILLE, Peter; CALLENDER, Jeffery. **Search patterns**. Beijing: O'Reilly, 2010.

MOURA, Gevilacio Aguiar Coêlho de. **Sistemas de busca da web**: diretórios e mecanismos de busca. 2001. Disponível em: <http://www.quatrocantos.com/tec_web/sist_busca/index.htm>. Acesso em: 6 abr. 2011.

MÜLLER, Jeane Froidevaux. Search engines, directories and gateways. In: _____. **A librarian's guide to the internet**: searching and evaluating information. Oxford: Chandos Publishing, 2003. p. 31-85.

NAHUZ, Fernanda. World Wide Web: aspectos teóricos dos mecanismos de busca. **Informação & Sociedade**: Estudos, João Pessoa, v.9, n. 2, p. 1-7, 1999.

NIECHWIEJ, Bartlomiej. More prediction in autocomplete. **Google™. Official Blog**, April 2012.

OLIVEIRA, Adriano Mendes de et al. Search engine optimization – SEO: a contribuição do bibliotecário na otimização de websites para os mecanismos de busca. **Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 1, Número Especial, p. 137-159, out. 2011. Disponível em: <http://periodicos.ufpb.br/ojs2/index.php/pg c/article/view/10792/6087>. Acesso em: 13 set. 2012.

PAPOUTSIDAKIS, Markos et al. Guidelines for web search engines: from searching and filtering to interface. In: SICILIA, Miguel-Angel; LYTRAS, M. Miltiadis D. **Metadata and semantics**. New York: Springer, 2009. p. 383-392.

PEIRCE, Charles Sanders. **Semiótica**. 4. ed. São Paulo: Perspectiva, 2010.

PETERSON, Donald. Context and the e-condition. Ind: NYÍRI, Kristóf (Ed.). **Mobile learning. Essays on philosophy, psychology and education**. Viena: Passagen Verlag, 2003. p. 117-125.

PIETARINEN, Ahti-Veikko. The semantic + pragmatic web = the semiotic web. In: IAÍAS, Pedro; KARMAKAR, Nitya (Ed.). **Proceedings of the IADIS International Conference on WWW/Internet.** Algarve: IADIS, 2003. p. 981-4. Diponível em:

<http://www.helsinki.fi/~pietarin/publications/The%20Semiotic%20Web-Pietarinen.pdf>. Acesso em: 23 abr. 2012.

POLLOCK, Jeffrey T.. **Semantic web for dummies**. Indianapolis: Wiley Publishing, 2009.

_____. **Web Semântica para leigos**. Rio de Janeiro: Alta Books, 2010.

PRAGMATIC Web. In: WIKIPEDIA: the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Pragmatic_web>. Acesso em: 18 mar. 2012.

QUERY checklist. Disponível em: <https://sbp-portal.wikispaces.com/.../handout+-+Query>. Acesso em: 18 jul. 2012.

RABAÇA, Carlos; BARBOSA, Gustavo G. **Dicionário de comunicação**. 2. ed. Rio de Janeiro: Campus, 2001.

RAMAL, Andrea Cecília. **Educação na cibercultura**: hipertextualidade, leitura, escrita e aprendizagem. Porto Alegre: Artmed, 2002.

RAMALHO, Rogério Aparecido; VIDOTTI, Silvana Aparecida Borsetti; FUJITA, Mariângela Spotti Lopes. Web semântica: uma investigação sob o olhar da Ciência da Informação. **DataGramaZero [online]***,* v. 8, n. 6, dez. 2007.

REIS, Julio Cesar dos. **Busca informada por abordagem semiótica em redes sociais inclusivas online**. 2011. 116f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Campinas.

RENTERIA-AGUALIMPIA, Walter et al. Exploring the Advances in Semantic Search engines. **Advances in Intelligent and Soft-Computing**, v. 79, p. 613-620, Springer, 2010.

REPENNING, Alexander; SULLIVAN, James. The Pragmatic Web: Agent-Based Multimodal Web Interaction with no Browser in Sight. In: PROCEEDINGS of IFIP INTERACT03: Human-Computer Interaction 2003. Zurich, Switzerland: [s. n.], 2003. p. 212-219. Disponível em: <http://www.cs.colorado.edu/~ralex/papers/PDF/I2003%20Pragmatic%20Web.pdf>. Acesso em: 14 jan. 2012.

RIEH, Soo Young. On the Web at Home: Information Seeking and Web Searching in the Home Environment. **Journal of the American Society for Information Science and Technology**, v. 55, n. 8, p. 743–753, 2004.

RODRÍGUEZ PEROJO, Keilyn; RONDA LEÓN, Rodrigo. Web semántica: un nuevo enfoque para la organización y la recuperación de información en el Web. **Acimed**, v. 13, n. 5, 2005. Disponível em: <http://bvs.sld.cu/revistas/aci/vol13_6_05/aci03605.htm>. Acesso em: 10 jul. 2011.

RUNG, Ching Chen; MING, Yung Tsai; CHUNG, Hsun Hsieh. Similarity web pages retrieval technologies on the Internet. In: KHOSROW-POUR, Mehdi. **Encyclopedia**

of Information Science and Tecnholgy. Hershey: Idea Group Reference, 2005. v. 5, p. 2486-91.

SALAZAR, Idoia. **Las profundidades de Internet**: accede a la información que los buscadores no encuentran y descubre el futuro inteligente de la Red. Somonte: Edicciones Trea, 2005.

SANTAELLA, Lucia. **Matrizes da linguagem e pensamento**: sonora visual verbal: aplicações na hipermídia. São Paulo: FAPESP: Iluminuras, 2009.

_____. **Navegar no ciberespaço**: o perfil do leitor cognitivo. 4. ed. São Paulo: Paulus, 2011a.

_____. **Linguagens líquidas na era da mobilidade**. 2. ed. São Paulo: Paulus, 2011b.

_____. A tecnocultura atual e suas tendências futuras. **Signo y Pensamiento 60 – Eje Temático**, v. 30, p. 30-43, enero/junio 2012.

SANTOS, Emanuelle; NICOLAU, Marcos. Web do Futuro: a Cibercultura e os Caminhos Trilhados Rumo a uma Web Semântica ou Web 3.0. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 35., Fortaleza, 2012. [**Anais...**]. Fortaleza: INTERCOM, 2012. Disponível em: <http://www.intercom.org.br/sis/2012/resumos/R7-1985-1.pdf>. Acesso em: 4 set. 2012.

SARACEVIC, Tefko et al. A study of information seeking and retrieving. I. Background and methodology. **Journal of American Society for Information Science***,* v. 39, n. 3, p. 161-176, 1988.

SARACEVIC, Tefko; KANTOR, Paul. A study of information seeking and retrieving. II. Users, questions, and effectiveness. **Journal of American Society for Information Science**, v. 39, n. 3, p. 177-196, 1988a.

_____; _____. A study of information seeking and retrieving. III. Searchers, searches, and overlap. **Journal of American Society for Information Science**, v. 39, n. 3, p. 197-216, 1988b.

SCHOOP, Mareike; MOOR, Aldo de; DIETZ, Jan L. G. The Pragmatic Web: a manifesto. **Communications of the ACM**, v. 49, n. 5, p. 75-6, May 2006.

SHADBOLT, Nigel; HALL, Wendy; BERNERS-LEE, Tim. **The Semantic Web revisited**. 2006. Disponível em: <eprints.ecs.soton.ac.uk/12614/01/Semantic_Web_Revisted.pdf>. Acesso em: 13 fev. 2011.

SHERMAN, Chris; PRICE, Gary. **The invisible web**: uncovering information sources search engines can't see. New Jersey: Information Today, 2001.

SILVA, Carlos Alberto F. da; TANCMAN, Michéle. A dimensão socioespacial do ciberespaço: uma nota. **GEOgrafia**, v. 1, n. 2, p. 55-66, 1999. Disponível em: <http://www.uff.br/geographia/ojs/index.php/geographia/article/view/18/16>. Acesso em: 20 out. 2011.

SILVA, Tércio de Morais Sampaio. **Extração de informação para busca semântica na Web baseada em ontologias**. 2003. 79f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Santa Catarina, Florianópolis.

SINGHAL, Amit. Introducing the Knowledge Graph: things, not strings. **Google™. Official Blog**, May 2012. Disponível em: <http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html>. Acesso em: 25 ago. 2012.

STANLEY, Tracey. Search engines corner: meta-search engines. **ARIADNE**: Web Magazine for Information Professionlas, n. 14, March 1998. Disponível em: <http://www.ariadne.ac.uk/issue14/search-engines>. Acesso em: 10 ago. 2011.

TAMBA, Irène. **A semântica**. 2. ed. São Paulo: Parábola, 2009.

TAYLOR, Arlene G.; JOUDREY, Daniel N. **The organization of information**. 3rd ed. Westport: Libraries Unlimited, 2009.

TORRES POMBERT, Ania. El uso de los buscadores en Internet. **ACIMED**, Habana, v. 11, n. 3, mayo/jun. 2003.

TREDINNICK, Luke. Web 2.0 and business: a pointer to the intranets of the future? **Business Information Review**, v. 23, n. 4, p. 228-34, 2006.

USCHOLD, Michael. Where are the semantics in the Semantic Web? **AI Magazine**, v. 24, n. 3, p. 25-36, Fall 2003.

VELTMAN, Kim H.. Towards a Semantic Web for Culture. **Journal of Digital Information**, v. 4, n. 4, 2004. Disponível em: <http://journals.tdl.org/jodi/article/viewArticle/113>. Acesso em: 27 out. 2011.

VIANELLO OSTI, Marina. Representación y creación de conocimiento en la WWW. In: CARIDAD SEBASTIÁN, Mercedes; NOGALES FLORES, J. Tomás (Coord.). **La información en la posmodernidad**: la sociedad del conocimiento en España e Iberoamérica. Madrid: Editorial Centro de Estudios Ramón Areces, 2004. Cap. 2, p. 15-26.

VISE, David A.; MALSEED, Mark. **Google**: a história do negócio de mídia e tecnologia de maior sucesso dos nossos tempos. Rio de Janeiro: Rocco, 2007.

WEINBERGER, David. **A nova desordem digital**: os novos princípios que estão reinventando os negócios, a educação, a política, a ciência e a cultura. Rio de Janeiro: Elsevier, 2007.

WELLER, Katrin. **Knowlwdge representation in the Social Semantic Web**. Berlin: De Gruyter Saur, 2010.

WEN-CHEN, Hu et al. World Wide Web Search Technologies. In: KHOSROW-POUR, Mehdi. **Encyclopedia of Information Science and Tecnholgy**. Hershey: Idea Group Reference, 2005. v. 5, p. 3111-7.

WITTER, Geraldina Porto. Pesquisa bibliográfica, pesquisa documental e busca de informação. **Estudos de Psicologia**, v. 7, n. 1, p. 5-30, jan./jul. 1990.

W3C. **World Wide Web Consortium**. 2011. Disponível em: <http:www.w3.org>. Acesso em: 27 maio 2011.

W3C Brasil. **Consórcio World Wide Web**. 2011. Disponível em: <http://www.W3C.BR/Home/WebHome>. Acesso em: 27 maio 2011.

XAVIER, Antonio Carlos dos Santos. **O hipertexto na sociedade da informação**: a construção do modo de enunciação digital. 2002. Tese (Linguística) – Instituto de Estudos da Linguagem, Universidade Estadual de Campinas.

ZANIER, Andrien Marcus de Almeida. **A evolução dos mecanismos de busca on-line**: a melhoria nos resultados obtidos. 2006. 162f. Dissertação (Mestrado Profissional em Administração) – Faculdade de Economia e Finanças IBEMC, Rio de Janeiro, 2006.

ANEXO

QUERY CHECKLIST

Pense como um mecanismo de busca. Utilize esta lista para construir uma *query* ideal.

1. Quantos conceitos-chave (ideias importantes) encontram-se na questão?
2. Quantos conceitos-chave eu buscarei em uma única questão?
3. Quais palavras-chave são provavelmente eficazes "como é?"
4. Para quais conceitos provavelmente serão necessárias palavras-chave mais eficazes?
5. Há hipônimos ou linguagem profissional para qualquer uma das palavras intermediárias?
6. Há palavras com múltiplos significados?
7. Usei todas as *stopwords* ou cortei algumas palavras?
8. Escrevi corretamente as palavras?
9. Inseri as palavras mais importantes em primeiro lugar?

1. Quantos conceitos-chave (ideias importantes) encontram-se na questão?

Após explorar o tópico, defina as palavras conceitos que se relacionam com sua necessidade de informação. Inclua terminologia específica, nomes, sinônimos, e palavras importantes relacionadas ao tópico.

Por exemplo: se voce está interessado em **tocadores mp3,** as duas palavras envolvem apenas um conceito. Por outro lado, se você quer saber **'Quantos búfalos existem hoje na América do Norte?'**, então você tem quatro conceitos chave com que confrontar:

o que - quantos (número);

búfalo;

onde - América do Norte;

quando - hoje.

2. Quantos conceitos-chave eu buscarei em uma única questão?

Geralmente, quanto mais definido o objetivo, mais conceitos existem. Buscando por apenas um conceito chave ou por mais de três conceitos em uma pesquisa causa problemas. Tentando encontrar as mesmas palavras utilizadas por um autor torna-se mais difícil quanto mais palavras você utiliza. Você terá mais sucesso, provavelmente, buscando por **dois ou três conceitos de uma vez,** mesmo que

existam outros conceitos importantes em mente. É melhor **manter a consulta simples** a menos que você tenha uma boa ideia das palavras exatas que um especialista tenha utilizado. Isto requer que se mantenham outros conceitos importantes em mente conforme você faz uma varredura nos resultados da pesquisa.

3. Quais palavras-chave são provavelmente eficazes "como é?

Palavras que são comumente boas "como é" são substantivos próprios e números. Quando se transforma uma pergunta em uma consulta, pense se não é um nome próprio, que pode ser utilizado no lugar de um dos conceitos. Palavras que tendem a ser ineficazes são verbos, adjetivos e advérbios: partes de discurso para as quais há muitas opções. Pronomes e preposições tendem a ser "*stopwords*" e são ignoradas pelos mecanismos de busca.

4. Para quais conceitos provavelmente serão necessárias palavras-chave mais eficazes?

Pesquisas efetivas exigem palavras de busca eficazes. Não faça pesquisas utilizando apenas as primeiras palavras que vem a mente ou aquelas utilizadas para explicar a atribuição. Pesquisas efetivas tipicamente envolvem a busca pela palavra chave "correta", ou seja, a **palavra utilizada pelos especialistas.** Você pode precisar se familiarizar com o tópico para ter certeza que utilizará as palavras de busca mais específicas.

5. Há hipônimos ou linguagem profissional para qualquer uma das palavras intermediárias?

Possíveis palavras chave caem, ao longo de um continuum (uma sequência contínua, em que os elementos adjacentes não são perceptivelmente diferentes uns dos outros, embora os extremos são bastantes distintos), do muito específico para o muito geral. Os termos técnicos utilizados para eles são **hipônimos** (muito específico) e **hiperônimos** (muito gerais). Por exemplo, palavras profissionais devem ser utilizadas como hipônimos para restringir sua pesquisa. Muitas pesquisas podem ser melhoradas utilizando palavras mais específicas. Troque termos gerais por mais específicos, embora ocasionalmente seja uma palavra-chave tão específica que produza poucos resultados ou irrelevantes.

6. Há palavras com múltiplos significados?

Pesquisas com uma palavra são ineficazes se a palavra tem mais de um significado. No entanto, se sua questão inclui **um número adequado de palavras-chave** - e não mais que o necessário - uma palavra com múltiplos significados pouco influencia num impedimento de você encontrar o que procura. Atualmente os mecanismos de busca utilizam palavras pares de acordo com o seu contexto dentro do resto da cadeia de pesquisa e exclui outros usos do termo. Enquanto o termo acompanhante é suficientemente único, utilizar uma palavra de múltiplos significados não é um problema.

7. Usei todas as *stopwords* ou cortei algumas palavras?

Stopwords são termos não indexados nos mecanismos de busca porque são partes comuns da língua que não adicionam significado para a pesquisa, como pronomes, preposições e conjunções. Elas são usualmente ignoradas por mecanismos de busca, então você precisa fazer algo diferente para dar atenção a eles se são importantes em alguma pesquisa em particular. Palavras que são utilizadas comumente como "a", "um", "para", são usualmente ignoradas, mas algumas vezes o mecanismo de busca fará suposições e incluir uma palavra de parada na busca se ela parece fazer diferença.

Por exemplo, para a pesquisa [the who] os resultados do Google irão relacionar com a banda enquanto que a pesquisa [who] será interpretada como referencia a **W**orld **H**ealth **O**rganization. Como o Google "sabe" que não precisa ignorar a palavra de parada "the" na primeira pesquisa? A análise de linguagem do Google é configurada para distinguir entre a banda e a Wolrd Health Organization. Ela não ignora a palavra de parada, porque reconhece que é parte do título da banda. O Google interpreta a pesquisa [who] como a World Health Organization, mesmo quando você não soletra ela. Quando em dúvida de como sua consulta será interpretada, utilize o operador AND (E) entre as palavras pesquisadas ou coloque a frase entre aspas.

Palavras desordenadas (***Clutter words***) são menos comuns que as *stopwords* mas não adicionam valor à consulta, também. Elas podem até forçar o mecanismo de

busca a procurar as palavras que você pensa ser importantes, mas não acrescenta nada especial para a pesquisa e pode limitá-la desnecessariamente.

Por exemplo, a consulta [terremoto e danos] são desnecessariamente redundantes. A palavra "dano" será provavelmente inclusa em qualquer coisa escrita sobre terremotos, então você não precisa bagunçar a pesquisa com ele. Além disso, verbos, adjetivos e advérbios são termos desordenados frequentes. Uma regra de ouro para manter em mente é "se voce não pode ver claramente, não utilize a palavra". Atenha-se a objetos, substantivos e números.

8. Escrevi corretamente as palavras?

Alguns mecanismos de busca possuem verificadores ortográficos, então você pode ou não conseguir os resultados que deseja quando erra a palavra-chave, dependendo do mecanismo de busca que utilizar. Quando as palavras têm mais de uma escrita ou acontece de uma palavra diferente coincidir com a errada, o mecanismo de busca pode buscar por algo que não tem relação com o que você quer.

9. Inseri as palavras mais importantes em primeiro lugar?

A ordem das palavras na busca de termos podem ou não fazer diferença para os mecanismos de busca. Em testes, Google retornou resultados similares, mas outros mecanismos de busca podem variar em seus resultados, dependendo da ordem que você colocou as palavras chave na caixa de busca. Para se assegurar, tente colocar os termos de busca mais importantes no início da *query*.

Adaptado de uma Query Checklist online apresentada pela *Twenty-first Century Information Fluency*.