

MECANISMOS DE BUSCA NA WEB: PASSADO, PRESENTE E FUTURO

Resumo: Este artigo apresenta um panorama sobre o desenvolvimento e o papel dos mecanismos de busca na recuperação da informação na *web*, destacando pormenores correlacionados à estrutura funcional dos primeiros buscadores e às especificidades da metodologia com que despontou o *Google (Google PageRank)*. A fim de sublinhar a perspectiva tecnológica vigente e as implicações decorrentes das singularidades de cada buscador, outras metodologias são examinadas. As inovações recentes do *Google Knowledge Graph*, as quais parecem levar o modelo de acesso dos buscadores a um passo a mais em direção à *web* semântica e à possibilidade de se obter maior consistência nos resultados de busca também são analisadas.

Ivan Cláudio Pereira Siqueira
Professor Doutor
Dispositivos & Recursos Informacionais
CBD - Departamento de
Biblioteconomia e Documentação
Universidade de São Paulo

naviclauper@usp.br

Palavras-chave: Buscadores. Web Semântica. *Google Knowledge Graph*. Recuperação da Informação.

WEB SEARCH ENGINES: PAST, PRESENT AND FUTURE

ABSTRACT - This paper provides an overview to the researches that have been carried out to develop of web search engine and underlies its key points in the information retrieval. It compares some general features of the first search engines with Google PageRank methodology. It also shows the current market of search engines focusing the news of Google Knowledge Graph and its promises of a new step towards the pleasing semantic web.

Keywords: Web search engine. Semantic web. Google knowledge graph. Information retrieval.

1. INTRODUÇÃO

Os modelos de recuperação da informação anteriores aos fenômenos da *internet* e da *web* tinham em comum o fato de que a informação a ser recuperada frequentemente provinha de coleções homogêneas e mais facilmente localizáveis. No caso dos documentos científicos e afins, cujo objetivo era facilitar o avanço da ciência, seus alocamentos em revistas, livros, periódicos ou dispositivos eletrônicos permitiam maior eficácia dos sistemas de recuperação da informação, tendo em vista a organização propiciada pelos padrões de Representação Descritiva (Código de Catalogação Anglo-Americano) e Representação Temática (Classificação Decimal de Dewey; Classificação Decimal Universal).

No ambiente informacional digital que eclodiu com a *web*, impuseram-se, dentre outros, o desafio da multiplicidade de documentos, os diferentes formatos e objetos *Extensible Markup Language* (XML), *Portable Document Format* (PDF), imagem, áudio, as linguagens computacionais e a interoperabilidade dos aplicativos e arquivos, e talvez o maior de todos eles – o desafio da indeterminação de controle: de autoridade, de qualidade, de emissão, de origem etc.

Isso se deve ao fato inexorável de que a *web* se constituiu como tal enquanto um universo de liberdade de criação, apresentação, acesso e apropriação da informação; mas também como um mundo de oportunidades para falsificações e do qual emergem dificuldades de se encontrar respostas à crescente diversidade de necessidades informacionais. O adensamento da rede ainda propiciou inovações e outros substratos para o incremento da pesquisa para além dos instrumentos das fontes tradicionais de informação – bibliotecas, museus, arquivos e centros de pesquisa.

É que a própria *web* tornou-se ela mesma a principal fonte das fontes de informação, não apenas para a pesquisa, estudo, criação e inovação, mas também como recurso essencial para as atividades humanas na sociedade pós-moderna – fonte de informação educacional: *Blogs*, *sites* institucionais, Educação a Distância (EAD); de lazer: *game*, música, vídeo; de atividade econômica: mercado financeiro, *net banking*; de

aplicativos: *smartphones* e *tablets*; de instrumento político: A Primavera Árabe; e de atividade social: *Content Management System* (CMS) e redes sociais.

Com o crescimento exponencial da *web*, surgiu também o imperativo de se implantar serviços de recuperação da informação compatíveis com a sua dimensão de massa. Verificou-se que modelos computacionais e algoritmos funcionais deveriam cumprir essa missão, ante ao custo e a quase inexecutabilidade de se outorgar a tarefa a procedimentos manuais.

Com base nesse contexto, este artigo faz uma revisão dos fundamentos técnicos e da estrutura metodológica dos buscadores, bem como do seu papel central de instrumento de acesso na vigente ecologia informacional da *web*. Com esse intuito, ilustram-se pormenores ligados à funcionalidade dos primeiros buscadores e às especificidades do *approach* com que despontou o *Google*.

Conclui-se com o apanhado das tecnologias de tratamento semântico da informação apresentados em ferramentas como o *Google Knowledge Graph* (Mapa do Conhecimento) e *Wolfram* (2012), que prometem levar a tecnologia dos buscadores a um passo à frente, rumo à *web* semântica; e ainda uma breve apresentação das tendências futuras dos mecanismos de busca, acesso e recuperação da informação na *web*. Financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), este trabalho está inserido num projeto maior de estudo, que integra competência, apropriação e cidadania informacional.

2. INTERNET, WEB E RECUPERAÇÃO DA INFORMAÇÃO DIGITAL

Na década de 1980, a internet já era uma realidade, notadamente em centros acadêmicos na Europa e nos Estados Unidos, mas o modelo de acesso era principalmente via diretórios de *File Transfer Protocol* (FTP) servidor e FTP cliente. O acesso aos sites de FTP permitia “navegar” sobre estruturas hierárquicas de pastas em diretórios, mas era um processo muito aquém da facilidade de visualização e manipulação que a *World Wide Web* (WWW) de Tim Berners-Lee propiciou. Na era da internet, acessavam-se diretamente

os arquivos nos servidores. Com a *web*, toda a informação disponível em servidores passou a ser acessível e mediada pela presença das páginas eletrônicas, potencializando, dentre outros, a aplicação de recursos gráficos e a interatividade (HAFNER; LYON, 1996).

O mundo da *web* emergiu a partir da concatenação de três tecnologias: o protocolo de comunicação de redes de computadores *Transmission Control Protocol/Internet Protocol (TCP/IP)*, o endereçamento e registro do domínio, a partir do qual a página da *web* é identificada, localizada e acessada *Domain Name System (DNS)*, e o protocolo de transferência de hipertextos *Hypertext Transfer Protocol (HTTP)*, cujo acesso se dá por meio de navegadores (*Chrome, Firefox, Internet Explorer, Safari* e outros). Obviamente, há que se considerar o avanço das tecnologias de telefonia e das linguagens de computação aí implicados, especialmente o *HyperText Markup Language (HTML)*. Nas palavras do criador: "I just had to take the hypertext idea and connect it to the TCP and DNS ideas and — ta-da! — the World Wide Web"¹. Em 06 de agosto de 1991, surgia a primeira página da *web*, na realidade uma espécie de metapágina, uma vez que seu objetivo era informar as suas peculiaridades técnicas: <http://info.cern.ch>.

A origem da tecnologia dos buscadores é anterior à rede de computadores (*internet*), remetendo ao sonho de organização e acesso ao conhecimento universal – o *Mundaneum* de Paul Otlet, no começo do século XX; à idealização de dispositivos como o *Memex*, de Vannevar Bush (1945), no contexto do pós II Guerra Mundial; e ao hipertexto de Ted Nelson, no bojo das tentativas de criação da rede de computadores na década de 1960 (PASSARELI, 2009).

Além disso, o desenvolvimento da tecnologia foi precedido pela formulação matemática do conceito de recuperação da informação (MOOERS, 1950) e pela codificação em linguagem computacional (algoritmos), a partir dos conceitos já há muito utilizados pela biblioteconomia – indexação, resumo e recuperação. Dentre as muitas contribuições para esse incremento, referências frequentemente citadas são o projeto *Salton's Magic Automatic Retriever of Text (SMART)* (SALTON, 1971), a aplicação de

¹ “Eu apenas juntei as funcionalidades do hipertexto ao TCP e DNS – e a Web estava criada (BERNERS-LEE, 2008, p.3, tradução nossa)

métodos quantitativos de indexação e busca de textos (CLEVERDON, 1972) e as técnicas de indexação automática do texto (indexadores de busca, tabelas de indexação, mineração de textos, processamento de linguagem natural), que já expressavam um viés epistemológico interdisciplinar: linguística, computação e matemática (DOYLE, 1961).

Nos seus primórdios, os mecanismos de busca automatizados baseavam-se em modelos de recuperação da informação em fontes de dados bibliográficos (palavras-chave e autor). Como essas fontes disponibilizavam referências informacionais e não textos completos, o desafio inicial dos buscadores na *web* era correlacionar e interpretar, a partir de buscas com expressões linguísticas, um determinado documento que satisfizesse as necessidades subjetivas enunciadas (SPARK; WILLET, 1997, p. 273).

Em parte, a problemática se devia ao fato de que, diferentemente das operações de recuperação de registros de dados em Sistemas de Gerenciamento de Base de Dados (SGBD), nos quais a taxa de precisão é de 100%, por tratar-se de valores exatos codificados em tabelas, a recuperação de documentos textuais (as primeiras páginas da *internet* assim eram concebidas) tinha como premissa incontornável a ambiguidade dos termos da linguagem natural, na busca e nos documentos indexados.

Foram elementos seminais de estruturação dos primeiros buscadores – aplicação de pesos estatísticos nos termos, relevância dos algoritmos e *feedback* das buscas, utilização de métodos quantitativos como *recall* e *precision*, pelos quais o critério de relevância era o produto da razão entre documentos relevantes encontrados e o total de documentos da coleção. Também a otimização da indexação do vocabulário e o modelo de vetor espacial. Parte desse saber é proveniente da indexação de resumos e dos experimentos na indexação de vocabulários em coleções (SALTON; WONG; YANG, 1975). Eis o esquema geral dos motores de busca na coleta, indexação, organização e armazenamento da informação:

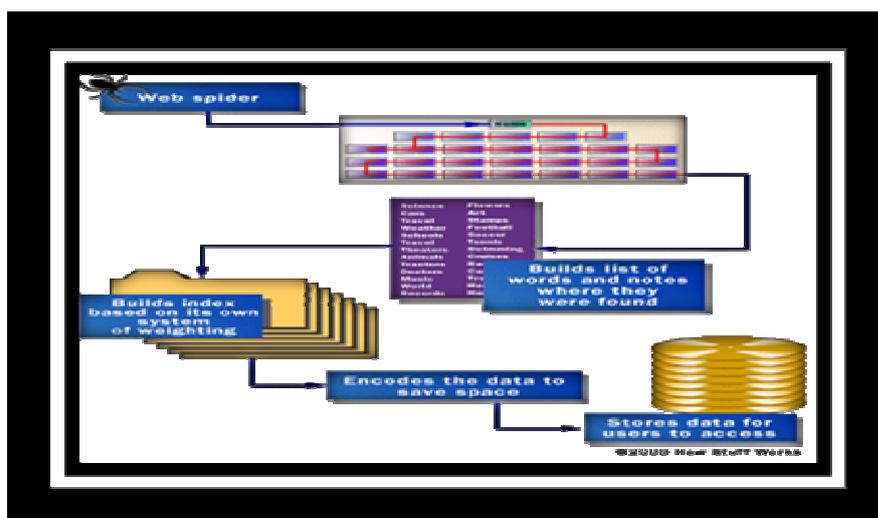


Figura 1: Funcionamento de mecanismo de busca na *web*
 Fonte: (HOW STUFF WORKS, 2000)

Dentre os dispositivos surgidos no alvorecer de 1990, o uso da metodologia de vetores espaciais veio a se concretizar no buscador criado por Brian Pinkerton (Universidade de Washington) – *WebCrawler*. Esse mecanismo aplicava a metodologia dos vetores no primeiro buscador da *web* que indexava todo o conteúdo das páginas. A metodologia de busca do *WebCrawler* visava a equacionar a busca de palavras-chave (coleta, indexação, armazenamento e recuperação da informação) e o problema da baixa qualidade das respostas oferecidas, em parte já resultante do aumento do número de páginas e da falta de padronização da *web*, mas também oriundo da não observância de um critério eficiente de “qualificação” ou “relevância” dos links indexados (BERRY; BROWNE, 2005).

O processo de coleta e indexação dos *links* partia do pressuposto de que critérios de qualidade da informação retornada na busca eram tributários do seguinte modelo – as coordenadas (x, y) de um determinado ponto espacial representariam vetores que indicariam a frequência das palavras-chave num documento. O resultado da busca deveria ser o cruzamento entre o eixo representativo da frequência das palavras-chave dos documentos indexados e os termos expressos na busca. Quanto maior a convergência entre a porcentagem de palavras-chave de um dado documento (valor) e os termos da

busca, tanto mais relevante seria aquele documento para aquela busca. Era esse o princípio:

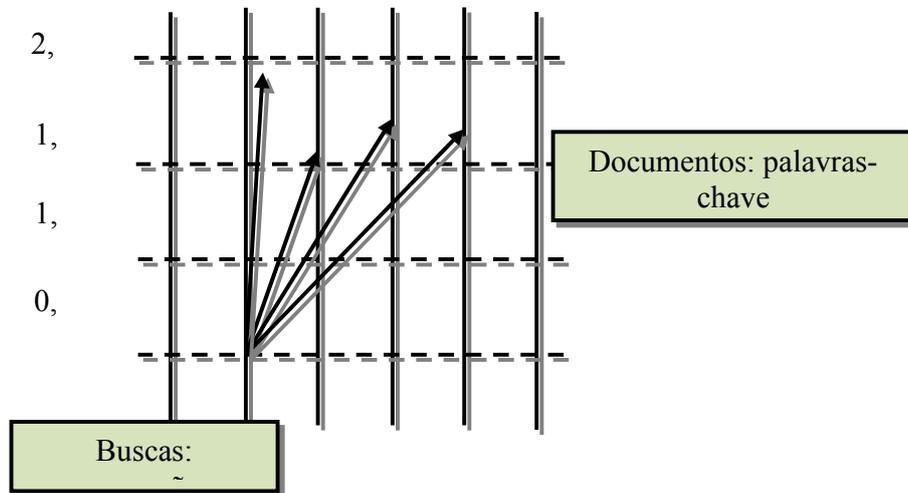


Figura 2: Relação – palavras-chave (documento) & termos de busca
Fonte: Elaboração própria, 2012

Paralelamente a essa tecnologia, havia modelos de buscadores como o *Yahoo!*, cuja metodologia de trabalho inicialmente consistia em diretórios e listas de índices com temas populares indexados manualmente. A qualidade era inegável, mas o custo, a lentidão do processo e a decorrente limitação de abrangência temática favoreceram o surgimento do *Google*, isto é, *googol*, o mesmo que 10^{100} (BRIN; PAGE, 1998).

3. GOOGLE E A MODERNIZAÇÃO DOS BUSCADORES

No limiar da década de 1990, com a multiplicação astronômica das páginas da *web*, tornou-se imprescindível o desenvolvimento de mecanismos de busca mais arrojados. Correlacionado ao aumento do número de páginas, cresciam também os percentuais de internautas e o tempo de navegação. É nesse momento, próximo à virada do milênio, que surge o buscador *Google*. Curiosamente, no seu documento de origem, os seus

fundadores enfatizavam que entre 1993 e 1997 o aumento de domínio ponto com (“.com” – comercial) passara de apenas 1,5% para 60% dos domínios; e que os buscadores em voga estavam distante dos pressupostos acadêmicos (BRIN; PAGE, 1998). Outro problema é que, por volta do ano 2000, estimava-se um número aproximado de 3.500 buscadores (LANGVILLE; MEYER, 2004). Mais ainda, quase uma década após o seu surgimento, estudo sobre a estrutura da *web* revelou que apenas 25% dos 200 milhões de páginas analisadas eram acessados (BRODER et al. 2000).

O *Google* nasceu sob o signo de uma promessa extremamente difícil – indexar a quase totalidade das páginas da *web* sem negligenciar metodologias de seleção e armazenamento de termos e *links* que possibilitassem melhorias constantes nos resultados de busca. Era a tentativa de juntar a qualidade dos diretórios organizados manualmente no primeiro *Yahoo!* e a agilidade de buscadores automatizados como o *Webcrawler*. A resposta era um método de indexação das páginas da *web* segundo um critério de leitura mais eficiente da estrutura de *hyperlinks* do HTML.

O *modus operandi* atendia pelo nome de *Google PageRank*. A sua descrição indicava uma metodologia que incorporava o modelo de citação acadêmico, com a pressuposição de que qualidade poderia ser expressa em termos de uma escala progressiva. Quanto mais “citada” (*linkada*) uma página, mais qualidade e mais pertinente à busca efetuada. Nesse sentido, o *PageRank* exibia similitudes com outro modelo contemporâneo a ele, baseado em algoritmo de indexação estruturada de *links* – o *Hyperlink Induced Topic Search* (HITS), (KLEINBERG, 1999). Mas, diferentemente do HITS, cuja procura pela “autoridade” dos *links* se processava após os termos de busca, com o seu consequente retardamento, o algoritmo do *Google* analisava os *links* previamente às buscas (BERRY; BROWNE, 2005, p.78).

Num cenário de franca expansão de buscadores, um dos objetivos do *Google* era aproximar a metodologia acadêmica ao modelo preponderantemente comercial: “with

Google, we have a strong goal to push more development and understanding into the academic realm”² (BRIN; PAGE, 1998).

Pode-se compreender o *PageRank* como um artefato conceitual de ranqueamento e de qualificação das páginas da *web*, com base numa equação que almejava verificar a probabilidade de satisfação de uma dada busca pela verificação dos constituintes (*links*) mais pertinentes de uma página, conforme critérios de citação externa de outras páginas. Tendo em vista as escolhas subjetivas dos termos de busca do usuário e o desencontro com os termos indexados pelos algoritmos, a equação procurava sistematizar um procedimento objetivo de averiguação da “qualidade da informação”. Os elementos operacionais do *PageRank* eram:

1. Probabilidade de um *link*/página ser clicado;
2. *Damping factor* (fator de redução) – probabilidade de o visitante abandonar a navegação em decorrência de frustração com a informação exibida;
3. *PageRank* alto – muitas páginas “linkadas” na página de origem; e
4. Páginas com *PageRank* alto “linkadas” na página de origem.

Inicialmente, um dos diferenciais do *PageRank* em relação aos métodos de outros buscadores de larga escala era atribuir um critério de valor aos *links* das páginas. Eis a equação:

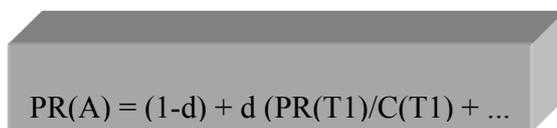
A equação do PageRank é apresentada dentro de um retângulo cinza tridimensional. O texto da equação é:
$$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots$$

Figura 3: Equação do *PageRank*
Fonte: Adaptado de (BRIN; PAGE, 1998)

² “Com o Google, temos o firme propósito de aliar o conhecimento acadêmico ao desenvolvimento dos buscadores” (BRIN; PAGE, 1998, p.12, tradução nossa).

A equação assinala que o *PageRank* de uma dada *homepage* da *web* (A) é obtido conforme a ligação com outras páginas (T1...Tn) e *links* externos direcionados. O uso do parâmetro “d” (*damping factor* – fator de redução) expressava valores entre “0” e “1” (frequentemente 0,85). O seu propósito era normalizar a probabilidade de abandono da navegação. “C” equivalia ao conjunto dos *links* internos da página, as quais direcionam para outra(s) página(s). Em síntese, o *PageRank* calculava o número de *links* apontados para uma determinada página e atribuía um valor conforme a “qualidade” desses *links* – os 04 pontos acima enumerados (BRIN; PAGE, 1998).

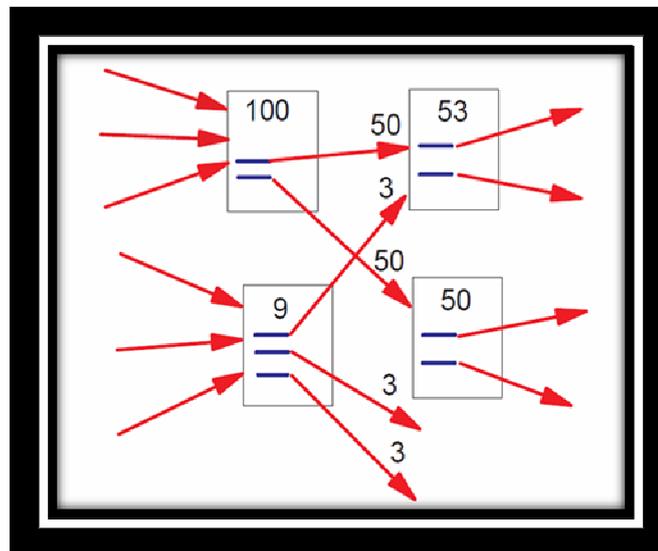


Figura 4: Modelo gráfico de mensuração do *PageRank*
 Fonte: (PAGE et al., 1998)

De um ponto de vista mais amplo, o *PageRank* era simultaneamente um algoritmo e um *score* que atribuía valoração às páginas, a princípio buscando oferecer informação de melhor qualidade aos seus usuários. Um dos problemas com a metodologia de ranqueamento é que a análise heurística dos *links* variava de 1 a 10, mas era estruturada numa escala de base 16. Dessa forma, o *PageRank* 2 expressava um valor 16 vezes maior que o *PageRank* 1, e assim sucessivamente. Não é difícil especular os problemas que o avanço na escala trazia para a correta interpretação dos dados (3 = 256; 4 = 4.096; 5 = 65.536; 6 = 1.048.576...). A mensuração entre um *PageRank* 5 e 4 (61.440) é imensamente

superior à comparação entre um 4 e 3 (3840). E isso sem considerar a implicação de valores fracionados, por exemplo, 5,1 e 5,2 (...).

Em uma página da *web*, as operações de *crawling*, assim como os critérios de indexação dos motores de busca, conforme a metodologia do *PageRank*, consistiam sobretudo na leitura das seguintes *tags* do HTML: *title* (título), *head* (cabeçalho), e *metags*. E também na avaliação da longevidade dos *sites* e dos direcionamentos internos e externos de seus *links*:



```
1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
2 <html xmlns="http://www.w3.org/1999/xhtml" lang="pt-br" xml:lang="pt-br">
3
4 <head>
5 <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
6 <title>ECA - Escola de Comunicações e Artes</title>
7 <link href="http://www3.eca.usp.br/sites/default/themes/eca_082011/home.css" media="all" rel="stylesheet" typ
8 <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
9 <link rel="EditURI" type="application/rsd+xml" title="RSD" href="http://www3.eca.usp.br/blogapi/rsd" />
10 <script type="text/javascript" src="http://widgets.twimg.com/j/2/widget.js"></script>
11 <link rel="alternate" type="application/rss+xml" title="ECA - Escola de Comunicações e Artes RSS" href="http://
12 <link rel="shortcut icon" href="/sites/default/files/picture_favicon.ico" type="image/x-icon" />
13 <link type="text/css" rel="stylesheet" media="all" href="/sites/default/themes/eca_082011/custom/modules/ddb1
upright10.css?R" />
14 <link type="text/css" rel="stylesheet" media="all" href="/sites/default/files/css/css_7b98b41d6c44605e46ddcba76
15 <link type="text/css" rel="stylesheet" media="all" href="/sites/default/themes/eca_082011/style.css?R" />
16 <script type="text/javascript" src="/sites/default/modules/igquery_update/replace/igquery.min.js?R"></script>
17 <script type="text/javascript" src="/misc/drupal.js?R"></script>
```

Figura 5: HTML de página da *web*
Fonte: (UNIVERSIDADE DE SÃO PAULO. Escola de Comunicações e Artes, 2012)

Entretanto, as boas intenções do *PageRank* foram combatidas pelos interesses de marketing, pela manipulação de *links* e de páginas (texto e *links* ocultos, falsas páginas de entrada (*doorway pages*), palavras sem sentido (*gibberish words*), *JavaScript* de redirecionamentos, *Spans*, criação de páginas adicionais, em suma, os conhecidos *Googlebombing* (BERRY; BROWNE, 2005, p. 78). Esse foi um dos motivos que acabou por inviabilizar a exteriorização do *PageRank* pelo *Google*, embora outras ferramentas prometam cumprir o mesmo propósito, a exemplo do *Google Toolbar PageRank Checker* – <http://migre.me/8BGx0>, aparentemente sem vínculos com o *Google*.

Concomitantemente a esses pormenores, ampliavam-se as exigências dos sistemas de recuperação da informação. O processo de representação, armazenamento, organização, busca e acesso passou a contemplar outras etapas, tais como: modelagem da classificação, múltipla categorização da informação, modalidades de arquitetura da informação e interfaces de usuários, dentre outros. Uma miríade de perspectivas vem alargando os horizontes dos sistemas de recuperação da informação, que podem ser paralelos, distribuídos, probabilístico, multilíngue ou específico para multimídias (BAEZA-YATES; RIBEIRO NETO, 1999).

4. BUSCADORES & NOVAS TECNOLOGIAS

Apesar dos avanços que a tecnologia de busca do *Google* trouxe, a sua massificação e o seu espraiamento em diversos aplicativos e múltiplas tarefas pareciam tê-lo afastado das “promessas” de qualidade inscritas no seu documento de origem. Abaixo, uma compilação de ícones de alguns dos seus aplicativos:

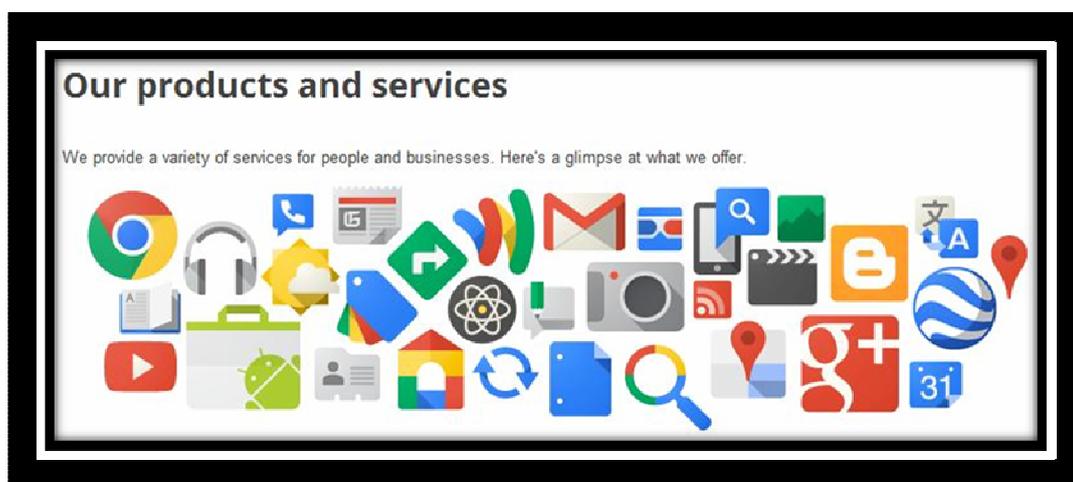


Figura 6: Ferramentas e aplicativos do *Google*
Fonte: (GOOGLE, 2012)

Apesar da incontestável hegemonia mundial do *Google* no segmento de buscadores automatizados, ao largo da sua supremacia, muito se tem discutido acerca de inovações e possibilidades de ferramentas de cunho semântico nos motores de busca e nos sistemas de recuperação da informação em geral. Dos processos iniciais que emergiram a partir da indexação de palavras-chave, já se vislumbram consideráveis avanços conceituais nessa área. Exemplos são o aperfeiçoamento das metodologias baseadas na álgebra booleana e do sistema *SMART* de leitura de textos que passaram a se aprofundar cada vez mais nas possibilidades de incorporação de elementos da lógica, da teoria dos conjuntos, da estatística, do modelo *fuzzy*, das redes neurais e dos algoritmos de aprendizagem (FERNEDA, 2003, p. 20-54).

Como resultado, houve melhorias consideráveis e uma profusão de novas terminologias para os motores de busca – metabuscadores (*Dogpile.com*); ontobuscadores (*Ontoweb.com*), surgido no Brasil e já desaparecido; buscadores especializados em temáticas como ciência (*Scirus.com*), busca de pessoas (*Radaris.com*), serviços (*City.ask.com*), busca de ocupação laboral (*Jablagoo.com*); buscadores focados em taxonomias e diretórios (*Vivissimo.com.br*); buscadores regionais (*Southernus.teoma.com*); buscadores de mídias sonoras (*FindSound.com*); buscadores de imagens (*Tiltomo.com*); e buscadores personalizados (*Eureskster.com*).

Pelo exposto, já é possível estabelecer uma considerável taxonomia dos buscadores. Dentre eles, destacam-se mais recentemente aqueles focados em processos semânticos, dentre os quais: *Sophia Search* (*SophiaSearch.com*), *Yummly* (*Yummly.com*), *GoPubmed* (*GoPubmed.com*) e o *Google Knowledge Graph* (Mapa do conhecimento), lançado recentemente nos Estados Unidos.

Diferentemente da busca de palavras e expressões em páginas da *web* e resultados de amostragens em listagens de *links*, o Mapa do Conhecimento do *Google* se baseia em componentes de Inteligência Artificial. A partir dos termos de busca, a metodologia irá subtrair “entidades”, compará-las às armazenadas nas suas bases de dados, e disponibilizar como resultado as informações relevantes. O diferencial é que as palavras não são tomadas isoladamente em fragmentos de textos e *links*, mas correlacionadas à semântica dos significados que perfazem a entidade. Isso significa que o conceito de entidade implica a sua percepção existencial enquanto um objeto ligado aos seus atributos de espaço, data, local, cultura, bem como seus relacionamentos com outras entidades.

Como transitar de um modelo de palavras-chave para relações semânticas e analogias? Parece ser esse o desafio a ser vencido não só pelo *Google*. Pode-se argumentar que a ideia de conhecimento aí implícita difere daquela de conhecimento enquanto processo e, portanto, sempre em constante volição e movimento. Por outro lado, os novos algoritmos prometem desenvolver a “capacidade” de juntar em um todo de sentido informações esparsas na rede, além de “aprender” continuamente com as

informações armazenadas. Talvez fosse mais apropriado nomeá-los como “um novo caminho de subsídios para a aprendizagem e o conhecimento”, mas certamente o efeito de marketing seria potencialmente menor do que Mapa do Conhecimento.

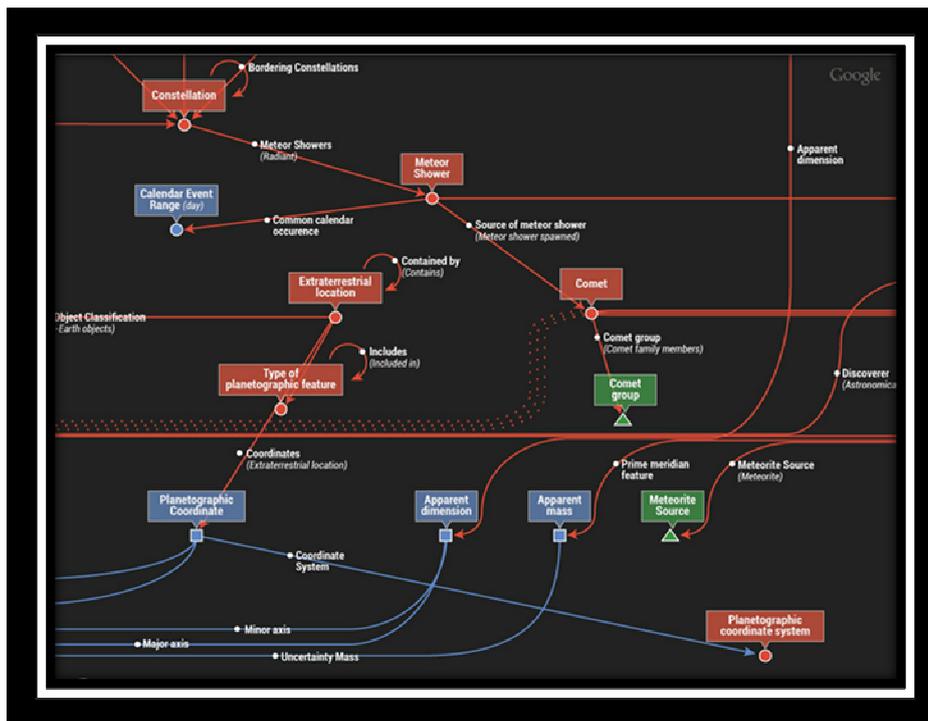


Figura 7: Relacionamento entre entidades no *Google Knowledge Graph*
 Fonte: (Google, 2012)

A tecnologia aí subjacente abre novas perspectivas. A indexação de palavras-chave, a análise dos *links* das páginas e o *PageRank* ingressam no passado. E só crescem pesquisas baseadas em conceitos de ontologia e processamento semântico da informação, com potencialidades para buscas e integração de dados em sistemas de recuperação de informação, desenvolvimento de padrões de análise, modelos de pergunta e resposta, e esquemas (GRUBER, 1993).

Há quem diga que buscadores são coisa do passado, e que as ferramentas promissoras devem ir além da amostragem de *links* (SHETH, 2011). É exemplo o *Wolfram* (WolframAlpha.com), autodenominado *Computation Knowledge Engine*, algo como “ferramenta computacional para o conhecimento”. Nele, além do teclado, há *inputs* por

imagens e dados (*upload*). É possível fazer diversas operações matemáticas (cálculo, funções, trigonometria); fórmulas da física, química e biologia; elucidar dúvidas sobre música, astronomia, estatística etc. Uma busca como “President of Brazil” resulta Dilma, informações sobre o cargo (data da posse, duração) e os seguintes relacionamentos semânticos – índices econômicos do país (*Gini*, dívida em dólar), dados sobre os antecessores na presidência, informações geográficas sobre o Brasil e dos demais países da região:

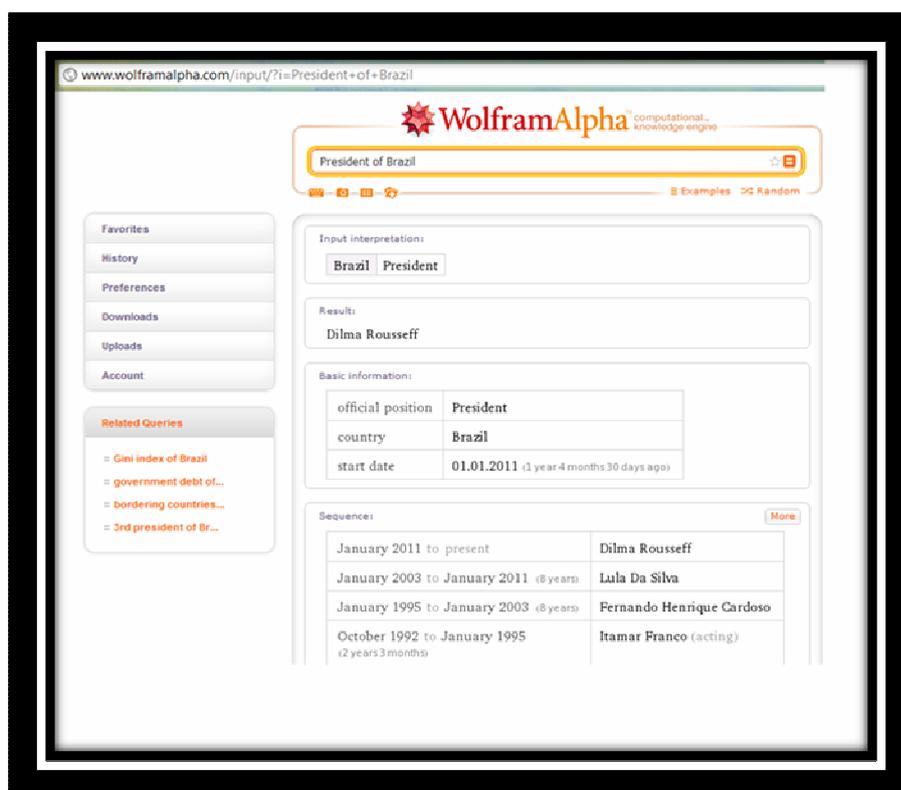


Figura 8: Exemplo de busca no aplicativo *Wolfram*
Fonte: (WOLFRAM, 2012)

Modelos de organização e recuperação da informação baseados em desenvolvimento de algoritmos semânticos e ontologias não são invenção do *Google Knowledge Graph*, como se pode verificar numa simples navegação em ferramentas como *Wolfram* e outras já citadas neste trabalho. Todavia, o fato de ser o principal instrumento

de busca na *web* pode favorecer a apropriação dessas tecnologias por parte dos usuários e quiça lhes suscitar o desejo de pesquisa em outros dispositivos.

Concomitantemente à promessa do Mapa do Conhecimento, o futuro dos mecanismos de pesquisa, busca e recuperação da informação na *web* provavelmente estará no aprofundamento das possibilidades funcionais de relacionamentos, da utilização de modelos de inteligência artificial e dos pressupostos da *web* semântica.

Exemplos dessa trajetória são os aplicativos *Gatfol* (Gatfol.com) – projetado para ser uma *interface* entre as informações existentes na *web* e as perguntas dos usuários, a partir de uma ontologia livre e um conjunto de algoritmos para processamento semântico da linguagem natural; e o *Never-Ending Language Learning* (NELL), (rtw.ml.cmu.edu/rtw), idealizado para ser um agente computacional inteligente. Trata-se de um mecanismo que pretende vasculhar a *web* diuturnamente, armazenamento e fazendo associações semânticas com os dados extraídos. O objetivo é que a ferramenta possa responder a qualquer indagação cujos objetos possam ser encontrados na rede, isto é, virtualmente, quase tudo (CARLSON, et al. 2010).

5. CONCLUSÃO

Este artigo objetivou traçar um panorama das funcionalidades e das possibilidades técnicas dos buscadores enquanto mediadores de acesso à informação na *web*. Tendo se desenvolvido concomitantemente com a *web*, os buscadores se tornaram ferramentas indispensáveis na obtenção de informações na rede. Entretanto, o uso desse recurso pode ser potencializado pelo conhecimento da sua estrutura e funcionamento, dos seus recursos, limites e implicações subjacentes às escolhas técnicas dos seus gestores. Em um momento de expansão da informação e de aumento da potencialidade do seu uso na solução de problemas, sobretudo pelas facilidades que as tecnologias oferecem, os buscadores ainda têm um papel de relevância na qualidade do acesso e do tempo envolvido na busca informacional.

Quando se considera a emergência dos *smartphones*, *tablets* e da disseminação da tecnologia *wireless* e banda larga, mais ainda crescem as utilidades e os usos dos buscadores, já então congregados numa miríade em franca expansão. Após aproximadamente uma década (1990-2000) de uso majoritário de metodologias de extração de termos e palavras-chave em *links*, os buscadores já operam com técnicas e metadados em imagens e áudio. Ainda que muito utilizado, o modelo quantitativo (*precision & recall*) já vê a companhia de tecnologias para o tratamento e recuperação de outras mídias além da textual.

Modelos de buscadores que tem por base similaridades entre imagens (*Similarity Based Image Search Engines*) já são realidade: (*Tiltomo.com*), (*TinEye.com*), (*Terragalleria.com*), (*FacesAerch.com*) e outros; assim como os que atuam com áudio (*Retrieving Musical Information*). Mais do que um conjunto de *links* como resultado da busca, esses buscadores e tecnologias permitem o processamento do áudio (tonalidade, altura, duração) e da imagem (busca reversa, alterações). Com os dispositivos portáteis e a melhora da qualidade da conexão, as possibilidades de busca, acesso e uso se agigantam, sobretudo para as atividades ligadas à educação.

Assim como a diversidade e o coletivo são apontados como valores sociais do século XXI, as fontes de informação e os seus mediadores não podem ser tributários exclusivamente de instrumentos ou canais unívocos. A ecologia informacional, os desafios globais e a cidadania que queremos construir não cabem mais no modelo unidirecional – navegar é preciso, mas em direção ao conhecimento!

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO NETO, B. **Modern information retrieval**. Boston, EUA: Addison Wesley Longman, 1999.

BERNERS-LEE, T. **Answers for young people**. [S.l]: World Wide Web Consortium, 2008. Disponível em: <<http://migre.me/9fvAQ>>. Acesso em: 01 set. 2012.

BERRY, M.; BROWNE, M. **Understanding search engines: mathematical modeling and text retrieval**. 2. ed. Philadelphia: Society for Industrial and Applied Mathematics, 2005.

BRIN, S.; PAGE, L. **The anatomy of a large-scale hypertextual web search engine**. Stanford: Stanford University, 1998. Disponível em: <<http://migre.me/9hwaw>>. Acesso em: 01 set. 2012.

BRODER, A. et al. Graph structure in the web. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 9., 2000, New York. **Proceedings...** New York: Association for Computing Machinery, 2000, p. 309-320.

BUSH, V. **As we may think**. Boston: Atlantic Magazine, 1945. Disponível em: <<http://migre.me/9fsEZ>>. Acesso em: 01 set. 2012.

CARLSON, A. et al. Toward an architecture for never-ending language learning. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE, 30., 2010, Georgia. **Proceedings...**, Georgia: Association for the Advancement of Artificial Intelligence, 2010. Disponível em: <<http://migre.me/9jc00>>. Acesso em: 01 set. 2012.

CLEVERDON, C. On the inverse relationship of recall and precision. **Journal of Documentation**, London, n. 23, p. 195-201, 1972.

DOYLE, L. B. Semantic road map for literature searchers. **Journal of the Association for Computing Machinery**, New York, n. 8, p. 553-578, 1961.

FERNEDA, E. **Recuperação de Informação**: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. 137 f. Tese (Doutorado em Ciência da Comunicação) - ECA-USP, São Paulo, 2003.

GOOGLE. Disponível em: <www.google.com.br>. Acesso em: 01 set.2012.

GRUBER, T. A Translation approach to portable ontology specifications. **Knowledge Acquisition**, Stanford, v. 5, n. 2, p.199-220, 1993. Disponível em: <<http://migre.me/9j7R7>>. Acesso em: 01 set. 2012.

HAFNER, K.; LYON, M. **Where wizards stay up late**: the origins of the internet. New York: Simon & Schuster Paperbacks, 1996.

HOW stuff works. 2000. Disponível em: <<http://migre.me/9ibRj>>. Acesso em: 01 set. 2012.

KLEINBERG, J. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, California, n. 46, 1999.

LANGVILLE, A.; MEYER, C. The use of linear algebra by Web search engines. **IMAGE Newsletter**, Maine, n. 33, p.2-6, 2004.

MOOERS, C. Information retrieval viewed as temporal signaling. In: INTERNATIONAL CONFERENCE OF MATHEMATICIANS, 6., 1950, Cambridge. **Proceedings...**, Cambridge, 1950, p. 572-573.

PAGE, L. et al. The PageRank citation ranking: bringing order to the Web. **Technical report**. Stanford InfoLab, 1998. Disponível em: <<http://migre.me/9hBM0>>. Acesso em: 01 set. 2012.

PASSARELLI, B. O Bibliotecário 2.0 e a emergência de novos perfis profissionais. **DataGramZero** – Revista de Ciência da Informação, Rio de Janeiro, v. 10, n. 6, dez. 2009.

SALTON, G. (Ed.) **The SMART retrieval system**. New Jersey: Englewood Cliffs; Prentice Hall, 1971.

SALTON, G.; WONG, A.; YANG, C. A vector space model for automatic indexing. **Communications of the Association of Computing Machinery**, New York, n. 18, p. 613-620, 1975.

SHETH, A. Semantics Scales Up: beyond search in web 3.0. **IEEE Internet Computing**, Washington, v. 15, n. 6, p. 3-6, Nov/Dec, 2011. Disponível em: <<http://migre.me/9j6Bq>>. Acesso em: 01 set. 2012.

SPARK, Karen; WILLET, Peter (Ed.). **Readings in information retrieval**. San Francisco: Elsevier Science, 1997.

UNIVERSIDADE DE SÃO PAULO. Escola de Comunicações e Artes, 2012. Disponível em: <www.eca.usp.br>. Acesso em: 01 set. 2012.

WOLFRAM TECHNOLOGY CONFERENCE, 2012. Illinois, EUA. **Proceedings...** Illinois: WOLFRAM, 2012.

AGRADECIMENTO