ENSAIO

Recebido em: 15/09/2016

Aceito em: 29/06/2017

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 22, n.50, p. 161-175, set./dez., 2017. ISSN 1518-2924. DOI: 10.5007/1518-2924.2017v22n50p161

Sistemas de recuperação da informação e o conceito de relevância nos mecanismos de busca: semântica e significação

Information retrieval systems and the relevance concept in search engines: semantics and significance

Silvana Drumond MONTEIRO (silvanadrumond@gmail.com)*
Rogério Paulo Muller FERNANDES (rogerio_muller@uol.com.br)*
Gian Carlo DECARLI (giandecarli@hotmail.com)**
Gustavo Lunardelli TREVISAN (gutrevisan@gmail.com)**
* Professor(a) do Departamento de Ciência da Informação UEL.
** Mestrando do Programa de Pós-Graduação em Ciência da Informação - UEL.

Resumo
Com o objetivo de estudar o conceito de relevância, algumas concepções de Sistemas de Recuperação da Informação são retomadas, para contextualizar os atuais mecanismos de busca considerados Sistemas Digitais ou de Significação. A partir de uma pesquisa documental, de uma revisão de literatura e da compreensão dos resultados de uma interface de busca (SERP), discute-se, a partir do algoritmo de busca e do *Knowledge Graph* do Google, a correlação entre otimização semântica e a relevância. Como resultado, verificou-se a forte ligação entre a noção de (*Web)* Semântica com o conceito de relevância, nos mecanismos de busca. Também pode ser estabelecida a relação do significado (sentido e referência) e a informação e, dessa maneira, conclui-se que a relevância é uma importante propriedade da informação. Percebeu-se que a relevância, nesta última década, no contexto dos mecanismos de busca, engloba, especialmente na relação sistema-usuário, a coleção de entidades descritas, a personalização e a contextualização.
Palavras-chave: Sistema de Recuperação da Informação; Recuperação da Informação; Ciência da Informação; Mecanismos de busca; Semântica; Relevância.

Abstract
In order to study the concept of relevance, some conceptions of Information Retrieval Systems are resumed to contextualize the current search engines, considered Digital or Significance Systems. Based on documentary research, literature review and understanding of the results of a search interface (SERP), it is argued, based on the search algorithm and the Google Knowledge Graph, the correlation between (web) semantic optimization and relevance. As a result, it was verified the strong association between the notion of semantics with the concept of relevance in search engines. It can also be established the relation of meaning (sense and reference) and information and thus come to the conclusion that relevance is an important property of information. It was noticed that the relevance, in the last decade, in the context of search engines, encompasses, especially in the system-user relationship, the collection of described entities, personalization and contextualization.
Keywords: Information Retrieval Systems; Information Retrieval; Information Science; Semantic; Search Engines; Relevance.

v. 22, n. 50, 2017.
p. 161-175
ISSN 1518-2924

1 INTRODUÇÃO

A Recuperação da Informação (RI) é uma área originária da Ciência da Computação (CC) e a expressão foi atribuída ao engenheiro estadunidense Calvin Mooers, em 1951 (p. 51), que a definiu, à época, como processo que "[...] engloba os aspectos intelectuais de descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação."

Como a Ciência da Informação (CI), em seu nascedouro, na década de 1960, tinha o problema de organizar Sistemas de Informação, mesmo que manuais, e permitir a recuperação da informação em uma época já consagrada pela "explosão da Informação", logo incorporou o termo como uma solução bem-sucedida para a questão, consolidando-se como campo de estudo da área.

Nesse sentido, Saracevic (1996, p.2) constatou, com muita propriedade, que um campo é definido pelos problemas que lhe são propostos e o "[...] problema era (e basicamente ainda é) a tarefa massiva de tornar acessível um acervo crescente do conhecimento.", e Mooers (1951) prossegue afirmando que a RI é crucial para a documentação e organização do conhecimento.

Desde a instituição desses campos, o cenário da produção científica e de signos, de uma forma geral, modificou-se com o advento do ciberespaço e do surgimento dos mecanismos de busca. Assim, na esteira de Saracevic (1996), problematizando os atuais Sistemas Digitais de Recuperação da Informação, por meio dos mecanismos de busca, pode-se questionar: quais os aspectos intelectuais da descrição de informações? Quais as suas especificidades para a busca? Quais as técnicas, máquinas ou sistemas empregados para o desempenho da operação?

Responder todas essas questões (embora não seja o objetivo deste artigo) requer, cada vez mais, a interdisciplinaridade com a Ciência da Computação, dada a complexidade dos atuais Sistemas Digitais de Recuperação da Informação, entretanto outra questão se impõe: qual seria o conceito de relevância utilizado pelos mecanismos de busca?

Para tanto, o artigo fará uma revisão de literatura sobre o conceito de "relevância" no contexto da CI, depois a sua apropriação em estudos de Sistemas da Recuperação da Informação (SRI) e atualmente em Sistemas Digitais, especialmente no mecanismo do Google por meio de seus algoritmos.

Mooers (1951), como pioneiro na RI, foi um dos primeiros a utilizar conceito de "relevância" no contexto dos SRI, e embora o termo já tivesse sido empregado por outros autores, em outros contextos da CI (conforme abordado na Seção 3 deste artigo), o conceito sofreu influências dos sistemas, técnicas e máquinas utilizados no processo de Recuperação da Informação, ao longo das décadas.

Percebe-se, hoje, que a relevância está associada ao significado, isto é, com a Semântica, conforme descrição dos atributos dos algoritmos do Google e de uma forma geral com os aspectos da *Web* Semântica (WS). Tratar desta temática implica, mesmo que de forma breve, esbarrar nos questionamentos supramencionados, isto é, abordar a busca nos atuais "Sistemas de Significação" e máquinas de busca no ciberespaço, rumo aos dados estruturados semanticamente.

2 SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO

Em artigo que discute a origem e a evolução da CI, Saracevic (2006) aponta a influência da RI no surgimento da primeira, na década de 1960, em razão da emergente necessidade de armazenamento e recuperação em uma sociedade cada vez mais especializada. O contexto americano propiciou o ambiente intelectual e industrial para o surgimento da "Sociedade da Informação".

Embora haja na literatura muitas definições sobre SRI tanto na CC quanto na CI, em grande medida, elas convergem semanticamente, de forma que algumas

interessantes são citadas para o propósito deste artigo, por outro lado, vários são os conceitos ou medidas de relevância ao longo do desenvolvimento dos SRI.

De acordo com Araújo (1995, p. 15), Sistemas de Informação são sistemas de comunicação e podem ser denominados "Sistemas humanos de processamento de informação, sistemas eletrônicos de processamento de dados ou sistemas de recuperação da informação [...]" e têm como função dar acesso aos conteúdos.[1] Dessa forma, vários autores da CI (ARAÚJO, 1995; ARAÚJO 2012; SARACEVIC, 1975) consideram os SRI como sistemas de comunicação ou parte deles e a relevância como uma das propriedades desse processo, conforme abordado na Seção 3.

Para Souza (2006, p. 163), um SRI deve desempenhar as seguintes atividades:

a) dar informações contidas nos documentos, usualmente através dos processos de indexação e descrição dos documentos;
b) armazenamento e gestão física e/ou lógica desses documentos e de suas representações;
c) recuperação das informações representadas e dos próprios documentos armazenados, representação de forma a satisfazer as necessidades de informação dos usuários.

Os Sistemas de Recuperação da Informação, de acordo com Araújo (2012), devem representar, armazenar, organizar e localizar os itens de informação e o referido autor aponta que a indexação é a principal função de um SRI, e seus componentes devem incluir: documentos; necessidades do usuário; consulta formulada e o processo de recuperação propriamente dito.

No exame aos modelos de SRI, tanto Cardoso (2000), Souza (2006) quanto Silva, Santos e Ferneda (2013) constatam - com pequenas variações - que existem os seguintes modelos:

a) modelos clássicos: booleano, vetorial e probabilístico;
b) modelos mais avançados: lógica fuzzy, redes neurais (baseados em bases de conhecimento) e algoritmos genéticos.

Já para Baeza-Yates e Ribeiro Neto (2013), os modelos de SRI podem ser divididos em três diferentes categorias: baseados em texto, baseados em *links* e baseados em objetos multimídia, conforme a Figura 1.

[1] Percebe-se que a literatura da CC utilizada neste artigo denomina o campo de Recuperação da Informação, embora abordem Sistemas e modelos (SRI), ao passo que a CI utiliza os dois termos, ou seja, RI e SRI.

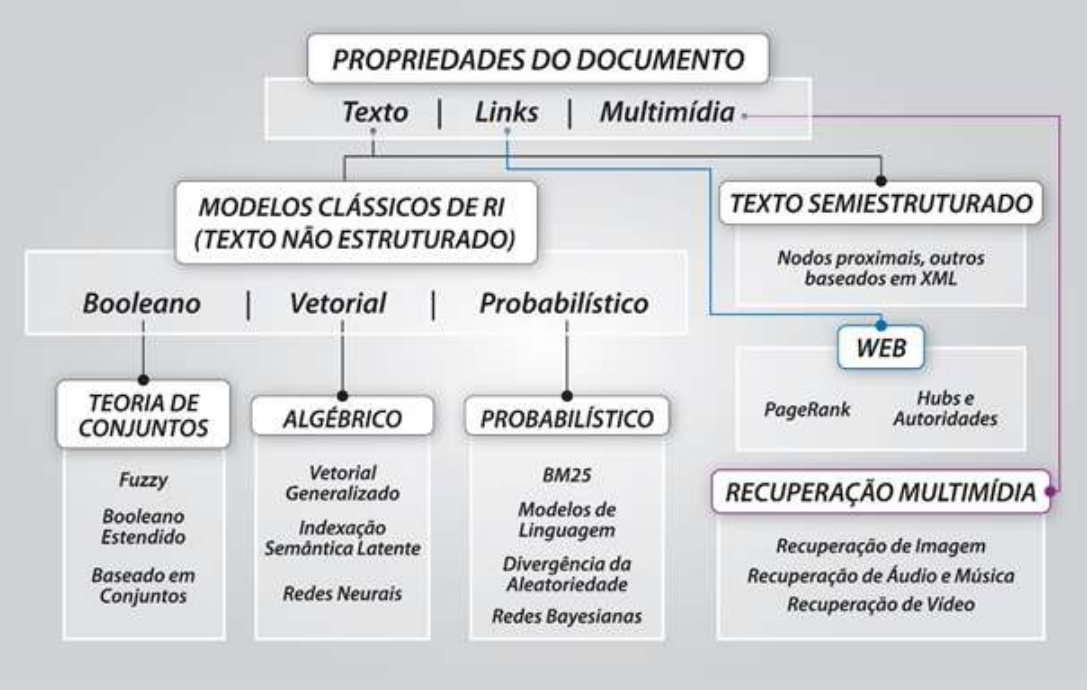

Figura 1: **Taxonomia de modelos de RI.**
Fonte: adaptado de Baeza-Yates e Ribeiro Neto (2013, p. 24).

Na primeira categoria, subdividida entre modelos para texto não estruturado e modelos que levam em conta a sua estrutura, o texto dos documentos é usado para ranqueá-los em relação à consulta. Nos modelos clássicos de RI, booleano, vetorial e probabilístico, o texto é modelado puramente como uma sequência de palavras, enquanto no texto semiestruturado os componentes como títulos, seções e parágrafos são parte integral do modelo.

A partir dos modelos clássicos, outros foram desenvolvidos visando à ampliação da performance e operacionalidade: baseados na teoria dos conjuntos, a lógica difusa ou nebulosa (*fuzzy*), o booleano estendido e o baseado em conjuntos; baseados nos modelos algébricos, vetorial generalizado, indexação semântica latente e redes neurais; e baseados nos modelos probabilísticos, BM25, modelos de linguagem, divergência da aleatoriedade e redes bayesianas.

Nos modelos que tratam com a estrutura fornecida pelo texto, os autores destacam técnicas de indexação como os nodos proximais e métodos de indexação baseados em XML.

A categorização dos modelos baseados em *links* utiliza a informação sobre a sua estrutura para alcançar um bom ranqueamento na *Web*, uma vez que devido ao grande número de documentos (ou páginas *Web*) o ranqueamento baseado somente em texto não é suficiente. Por isso, os *links* entre as páginas *Web* também devem ser considerados no modelo, como acontece no *PageRank* e no *Hubs* & Autoridades.

Já em relação aos objetos multimídia, por causa das particularidades em suas formas de representação que possibilitam, por exemplo, a recuperação de uma imagem apenas pela comparação entre imagens relacionadas, sem a necessidade de escrever uma consulta, o ranqueamento é feito de forma diferente ou mesmo é inexistente em uma recuperação. Dentre as estratégias de recuperação multimídia, a mais simples é a recuperação de imagem, por ser estática. Na recuperação de áudio, música e vídeo, o problema se torna mais difícil, pois é necessária a inclusão de uma dimensão temporal que torna os arquivos maiores (BAEZA-YATES; RIBEIRO NETO, 2013).

Após algumas definições e apresentação, mesmo que breve, dos modelos de SRI (uma vez que não é o objetivo principal deste artigo), vale observar que os atuais

mecanismos de busca são considerados SRI, conforme Cendón (2005), Souza (2006), Silva, Santos e Ferneda (2013) já constataram.

Os mecanismos de buscas geram os índices contemporâneos, refletem os agenciamentos híbridos e podem ser definidos como uma enorme base de dados de informações importantes a respeito de *sites* na *Web* (MONTEIRO, 2015). Para Battelle (2006), os índices são povoados com etiquetas, outro tipo de metadados. Sangirardi (2014, p. 228) afirma que "Os mecanismos de busca são tecnologias cognitivas [...]."

Dessa forma, podemos considerá-los como um novo paradigma na Recuperação da Informação, seja pela escalabilidade de processamento e indexação, seja pelos índices de múltiplas semióticas, seja pelos algoritmos e aprimoramento na interface de busca. Aliás, é o ramo do ciberespaço e da computação (juntamente com as mídias sociais e comércio eletrônico) que mais cresce e lucra na atualidade.

Sua anatomia (ou arquitetura) pode variar, mas deve apresentar: a) processos de coleta e indexação; b) geração de índices; c) processos de busca. Utilizando a taxonomia ou tipologia de Monteiro (2009), um SRI possui em sua anatomia os seguintes componentes: a) *crawling*; b) *indexing*; c) *searching,* conforme a Figura 2*.*

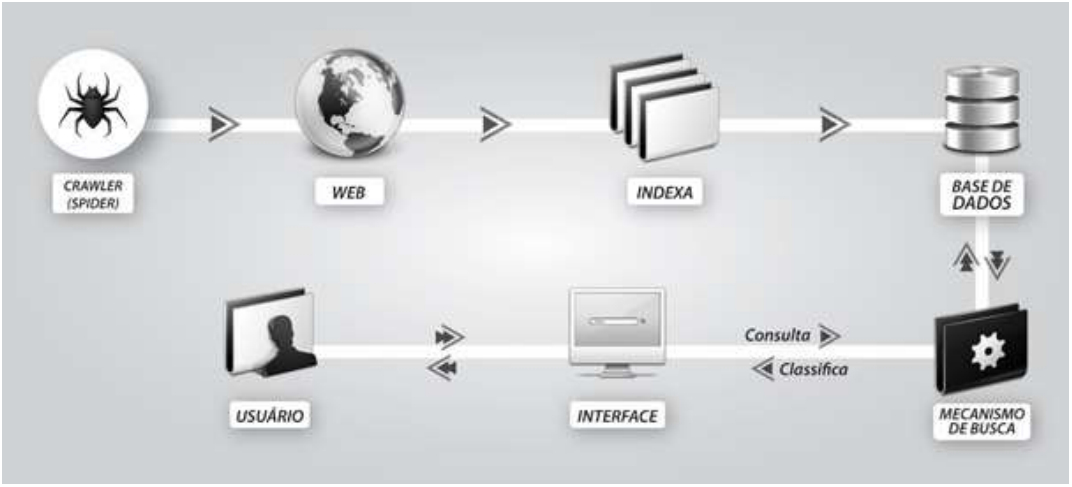

Figura 2: **Anatomia dos mecanismos de busca.**
Fonte: elaborado pelos autores

Os atuais SRI ou máquinas de busca possuem sofisticados processos de coleta, indexação, análise e interface de busca em que: linguagens e programas são utilizados para representar e descrever semanticamente a informação; ontologias para relações de domínio; bases de conhecimento (as chamadas bases secundárias com entidades descritas semanticamente) que fornecem dados estruturados para os agentes inteligentes; programas e tecnologias de visualização e apresentação dessa informação em contextos (*Knowledge Graph* ou mapa do conhecimento) nas interfaces de busca.

Dessa maneira, um motor de busca faz parte de um sistema maior, que utiliza aplicativos (*crawler, spider* ou robô) para capturar palavras-chave e frases que identifiquem o conteúdo de uma página *Web* para indexá-la e armazená-la em sua base de dados. Do outro lado, conta com uma interface destinada ao usuário para consultar essas informações mediante a inserção de termos de pesquisa que irão acionar o algoritmo de busca para exibir os resultados recuperados da base de dados (LEDFORD, 2009). As ferramentas para recuperação de informações da rede ou em plataformas específicas da *Web* são entendidas nesta pesquisa como

> [...] mecanismos de busca, sistemas de busca, motores de busca,

> buscadores, *search engine*, são os termos utilizados para indicar um conjunto organizado de computadores, algoritmos, bases de dados e índices reunidos com a função de analisar e indexar as páginas da *Web* e apresentá-las de forma organizada aos usuários. (OLIVEIRA; ARAÚJO, 2012, p. 64-65).

Essas máquinas de busca permitem a organização do conhecimento no ciberespaço por meio da indexação que realizam (MONTEIRO, 2006).

A partir dessas definições, a próxima seção investiga o conceito de relevância na Ciência da Informação e sua incorporação nos SRI que utilizam dados estruturados semanticamente nos processos de Recuperação da Informação em uma interface de busca.

3 O CONCEITO DE RELEVÂNCIA

Na CI o conceito de relevância, ao que tudo indica, foi empregado pela primeira vez por Bradford em 1934, como conceito de medida no contexto das fontes de informação, estudos esses que mais tarde foram rotulados de bibliométricos e que objetivaram medir a produtividade das fontes, a partir do enfoque probabilístico. Bradford constatou que havia periódicos (e autores) mais produtivos e, portanto, mais relevantes para determinadas áreas. Nessa direção, Saracevic, em estudos sobre relevância, alertou "[...] para o fato de que distribuições bibliométricas são, de fato, 'distribuições associadas à relevância'." (SARACEVIC, 1975, p. 330).

O uso do termo foi incorporado ao campo da RI, conforme mencionado, na década de 1950, por Mooers, mas, segundo Figueiredo (1977, p. 76), "[...] a primeira exploração extensa do conceito de relevância apareceu em 1958, durante a *'International Conference for Scientific Information'* (ICSI) com Fairthorne, Vickery, Bar-Hillel e outros.", em que a CI concordou com os seguintes aspectos:

a) relevância não é, exclusivamente, uma propriedade do documento;
b) relevância não é uma propriedade dicotômica [isto é, entre documentos satisfatórios e insatisfatórios ao pedido do usuário];
c) existe uma 'relevância para o usuário' que deve ser julgada [configurando-se, portanto, em um julgamento da relevância].

De acordo com Figueiredo (1977, p. 76), a primeira Teoria da Relevância foi formulada na década de 1960 por Maron e Kuhns e tinha uma noção quantitativa, mas ligada à satisfação, uma vez que foi definida como "[...] a derivação de uma medida indicando a probabilidade de um documento satisfazer a um determinado pedido [...]."

Após a formulação matemática do fenômeno da relevância por Goffman (a*pud* FIGUEIREDO, 1977), o enfoque voltou-se para a relação pergunta formulada-documento. Mais tarde reconheceu que tributar à medida somente a relação pergunta-documento não contempla o conceito, e que a "relevância" é uma relação comparativa, isto é, deve-se considerar também a relação entre documentos.

Dessa forma, os SRI tornaram-se arena de várias disputas de testes de relevância, na década de 1960, especialmente em relação à revocação e precisão dos resultados, para aferir o sistema perfeito. Mas, de acordo com Figueiredo (1977), os questionamentos cresceram na mesma proporção da aplicação dos experimentos.

Apesar de serem usados como termos equivalentes, relevância e pertinência são diferentes. O primeiro busca contemplar o universo que envolve os termos de uma pesquisa em um sistema e apresentá-los ao requisitante enquanto o segundo depende da interpretação e apropriação do leitor/usuário sobre o que é

útil/pertinente para ele em meio ao universo de resultados relevantes (FOSKETT, 1972).

À medida que cada modelo de SRI é aplicado, a relevância é alterada e ordena de diferentes maneiras a apresentação dos resultados. No entanto, independentemente do modelo de SRI empregado, é possível avaliar os resultados por intermédio de índices de precisão e revocação. Assim, a "Precisão é a fração dos documentos já examinados que são relevantes e revocação é a fração dos documentos relevantes observada dentre os documentos examinados." (CARDOSO, 2000, p. 2). Nesse sentido, Lancaster (1969) alerta que ao realizar uma indexação exaustiva em um sistema, a revocação será alta, mas com precisão baixa, porém, ao considerar apenas conceitos importantes na indexação, haverá uma alta precisão com baixa revocação.

Apesar da profusão semântica do conceito de relevância, ele tem a vocação e o viés avaliativo associado, em um primeiro momento, à medida probabilística de produção na avaliação de fontes de informação, à satisfação do usuário, à relação pergunta-documento e entre documento e depois ao significado, mas sempre vinculado a uma medida, envolvendo uma relação comparativa entre partes no processo de comunicação.

Nessa direção, Saracevic (1975) alerta que o conceito apresenta vários pontos de vista, a saber:

a) o conhecimento sobre o assunto;
b) o aspecto lógico;
c) o aspecto do sistema;
d) o lado do destinatário;
e) o lado do conhecimento do destinatário – pertinência;
f) o aspecto pragmático.

Saracevic (1971) constata, ainda, que as dimensões dos estudos de relevância podem incluir:

a) documentos e representações;
b) questões (queries);
c) situações e condições de julgamento;
d) formas de expressão;
e) características humanas (pessoas);
f) usuários.

Hoje, nas várias modelagens de SRI, a relevância é propriedade essencial para o desenvolvimento desses sistemas, porém apresenta-se de diferentes maneiras para cada um dos modelos de RI. Em um modelo booleano, os resultados recuperados são equivalentes para o sistema, pois ao utilizar a teoria dos conjuntos restringe-se apenas aos resultados que pertencem à lógica de busca, sem a possibilidade de ordenar/classificá-los do mais para o menos relevante. No modelo vetorial, palavras e expressões possuem valores de representação, assim os resultados podem ser ordenados do mais relevante (maior valor) para o menos relevante (menor valor). O probabilístico baseia-se na teoria matemática das probabilidades para apresentar a primeira leva de resultados que o sistema considera relevante, em que relevância é determinada a partir da interação do usuário com o sistema (SILVA; SANTOS; FERNEDA, 2013).

O modelo *fuzzy* atribui valores presentes no intervalo 0 e 1 para elencar os mais relevantes na apresentação dos resultados. Nas redes neurais artificiais simula-se computacionalmente o sistema nervoso humano de recuperação de informações. Nesse modelo, os resultados relacionados a termos equivalentes podem ser recuperados como itens relevantes. Também pautado em aspectos humanos, o modelo de algoritmos genéticos exibe resultados prévios para que o usuário possa

aplicar mais filtros e torná-los mais relevantes, posteriormente registra essas interações para otimizar pesquisas futuras (SILVA; SANTOS; FERNEDA, 2013).

Nessa perspectiva, a CC e a CI têm sido a base para os estudos de RI. Na CC a preocupação está em avaliar e evoluir tecnicamente a RI, enquanto na CI o enfoque permeia o usuário e a avaliação deste sobre o sistema. No entanto, os motores de busca para *Web* desafiam essas áreas a trabalhar a interdisciplinaridade, pois seu uso na rede desperta também o interesse de outras disciplinas, em virtude do impacto social desses mecanismos na rede (LEWANDOWSKI, 2012).

Na CI, a relevância tornou-se central quando as preocupações práticas e teóricas se voltaram para a RI e não apenas na organização de informações. A Biblioteconomia envolve-se com a temática, com a organização das informações, enquanto a CI preocupa-se com a recuperação, trabalhando com a relevância (SARACEVIC, 2012). Os índices dos motores de busca *Web* são semelhantes aos modelos anteriores, porém três características os diferenciam: a primeira refere-se ao barateamento do acesso a diversas fontes de informação, a segunda diz respeito aos avanços das TIC que promoveram o acesso à rede, e a terceira à liberdade de uso e publicação de conteúdo na *Web* (BAEZA-YATES; RIBEIRO-NETO, 1999).

Ao longo dos anos, os SRI evoluíram significativamente, porém o desafio inicial de indicar ao usuário o que é relevante em seus resultados ainda é um desafio. Basicamente, o problema reside em informar à máquina o que é relevante para um ser humano, pois até mesmo entre as pessoas há divergências em determinar com unanimidade o que é relevante para elas, talvez por se tratar de uma questão subjetiva, pessoal, intrínseca e imparcial. No entanto, a CI tem a relevância na pauta de seus estudos e alimenta expressivamente a produção e discussão de cunho científico sobre o tema.

Apesar de os mecanismos de busca considerarem questões culturais, sociais e contextuais na apresentação de seus resultados, ainda há a fragilidade de que a noção de relevância é humana e não técnica. Esse fato a torna de difícil formalização por possuir muitas variáveis. Em concordância, Silva, Santos e Ferneda (2013, p. 37) consideram que a relevância consiste em:

> [...] mostrar os resultados possivelmente mais relevantes em forma de ranque (*ranking*), do mais relevante ao menos relevante. Entretanto, o conceito de relevância é subjetivo e inexato, não podendo ser definido por fórmulas matemáticas e implementadas em sistemas computacionais.

Desse modo, a construção de SRI eficazes somente será possível quando se traduzir a noção de relevância intuitiva em formalismo escrito (LAVRENKO, 2009). Seja como for, os humanos usam o contexto para a resolução de problemas e os mecanismos usam algoritmos e processos tecnológicos aliados à probabilidade na apresentação de seus resultados (SARACEVIC, 2012).

3.1 O Conceito de relevância nos mecanismos de busca

Nos mecanismos de busca, a relevância tem como correlato o "significado", em especial no Google, mas também na própria *Web* Semântica, conforme será abordado nesta seção, tanto que surge uma inclinação para chamar esses Sistemas Digitais de Recuperação da Informação de "Sistemas de Significação", em razão das camadas Semântica e Semiótica em seus processos de indexação, geração de índices e interpretação das *queries* e dos resultados.

Desde os primeiros SRI automatizados até o aparecimento dos mecanismos de busca, observa-se hoje que o usuário visa recuperar a informação "[...] mas a recuperação de informações relevantes para seus propósitos de pesquisa [...]." (ARAÚJO, 2012, p. 142) e não exatamente o documento, mas, na busca no

ciberespaço, tanto faz se ela inclui a recuperação ou simplesmente proporciona a descoberta.

Nessa mudança de paradigma (que foi possível a partir da digitalização dos textos e da virtualização da informação), os algoritmos utilizados pelos serviços de busca têm procurado o aperfeiçoamento nos processos de indexação e busca, mas desde o início dos mecanismos, final da década de 1990, os engenheiros debruçaram-se sobre o aspecto da "relevância".

Assim, as palavras semanticamente significativas para a busca são analisadas de acordo com a frequência e localização no documento, análise de *links,* além mais outros 200 sinais e 500 aprimoramentos, por ano, utilizados pelo Google (GOOGLE, 2016).

Já em 2009, Fioravante chama a atenção para relevância afirmando que ela surge como principal vetor de novas tendências dos mecanismos de busca, "Assim, a relevância de uma busca só existirá se a ferramenta de busca conseguir, de forma eficaz, identificar o contexto da pesquisa e seus elementos." (2009, p. 1).

Nessa direção, em 2013 no Brasil e 2012 em língua inglesa, foi possível perceber traços semânticos dos resultados de busca nos mecanismos, com o aparecimento do "*Knowledge Graph*[2]" do Google (KG ou Mapa do Conhecimento e hoje por outros mecanismos de busca, como o *Bing*), e foi descrito como um enorme mapa de elementos do mundo real e as conexões entre eles, de modo a oferecer resultados mais relevantes (GOOGLE, 2012).

Percebe-se nas informações contidas no *infobox*, no lado direito de um *Search Engine Results Page* (SERP - Página de Resultados de um Mecanismo de Busca), que os resultados apontam para um referente existente no mundo real e com ligações de contexto, em redes de significações no ciberespaço.

Nesse momento, emergia para o usuário a *Web* Semântica, isto é, os dados descritos semanticamente (metadados) e ligados, noções essas estabelecidas por Tim Berners-Lee, no começo da década de 2000 (BERNERS-LEE *et al.*, 2001).

Mas como ligar a noção de Semântica com o conceito de relevância? Figueiredo (1977) estabelece a relação entre significado e relevância ao abordar que conteúdo significativo é relevante, e dessa maneira a relevância é uma propriedade da informação.

A Semântica (assim como o conceito de relevância) apresenta várias nuances e matrizes teóricas. De maneira geral, pode-se afirmar que a Semântica é a ciência da significação, mas sua apropriação na *Web* é da Semântica Formal (baseada na Semântica verifuncional), em que o significado é entendido como a relação entre as palavras e o mundo, entre o **sentido** e a **referência**, pois "[...] a referencialidade é fundamental para entendermos o fenômeno do significado. A intuição básica dessa semântica é que, ao usarmos a linguagem, falamos sobre um mundo." (OLIVEIRA, 2012, p. 89).

A ontologia da Semântica Formal admite que as entidades que compõem o mundo contemplam apenas os objetos reais, dentre eles os valores de verdade (verdade ou falso). Não por acaso, uma das denominações da WS é "*Unambiguous Web*" (SIEGEL, 2010).

[2] Segundo Meloni (2015, *on-line*, grifos do autor), "O Grafo do Conhecimento é um banco de dados dos fatos sobre '**o que é pedido**' no mundo e como elas se relacionam entre si." Para Monteiro (2015, p. 7) "Trata-se o KG de uma engenharia de Recuperação da Informação, [...] que tem como objetivo a aprendizagem da máquina e a semantização dos resultados de busca.", e isso engloba algoritmos, agentes computacionais e banco de conhecimento associado.

As propriedades de apresentação de uma entidade no KG, conforme sua patente (MONTEIRO, 2015), deve conter: título; imagem; descrição; fato associado, conforme ilustrado na Figura 3.

Figura 3: **Exemplo do *Knowledge Graph* em um resultado de busca.**
Fonte: elaborado pelos autores com base no Google (2016).

É importante observar que o *Frame* ou espaço que contém o KG (também denominado *infobox*) é um documento gerado a partir de uma busca com a descrição e interligação do objeto buscado com suas respectivas fontes, afirmando assim em seu significado o **sentido e sua referência**, ou seja, dois aspectos fundamentais da significação sob a ótica da Semântica Formal.

No contexto dos SRI, Araújo (2012, p. 141) apresentou um novo modelo incorporando uma teoria semântica (aparelho formal de enunciação, de Benveniste), em que "A noção de referência aparece como definidora do sentido no nível semântico, sendo a referência ao mundo parte integrante do ato de enunciação: aqui entra a noção de contexto."

No algoritmo de ranqueamento do Google, o *PageRank*, desenvolvido há 18 anos (PAGE *et al*., 1998), baseado em métrica de citações, foi atualizado por entidades (relacionamentos) e se chama "*hummingbird*", segundo os desenvolvedores, o foco principal desse algoritmo é aumentar a **compreensão semântica** e a rapidez. Um dos principais sinais desse algoritmo é o *rankbrain,* que segundo Meloni (2015) visa ajudar na interpretação das pesquisas, não se atendo a mesma sintaxe da *query* utilizada na busca.

Também se deduz, até o momento, que o conhecimento semanticamente modelado e executável por máquina permite conectar informações sobre pessoas, eventos, locais, horários, entre outros, em fontes de conteúdo diferentes e vários processos de aplicação da WS.

Assim, de acordo com Simoni (2013), o Google, e, portanto o KG, realiza sua indexação e busca a partir dos atributos de personalização, otimização semântica, contexto e localização. A partir da Figura 4 é possível verificar os atributos tanto do KG quanto do *hummingbirds.*

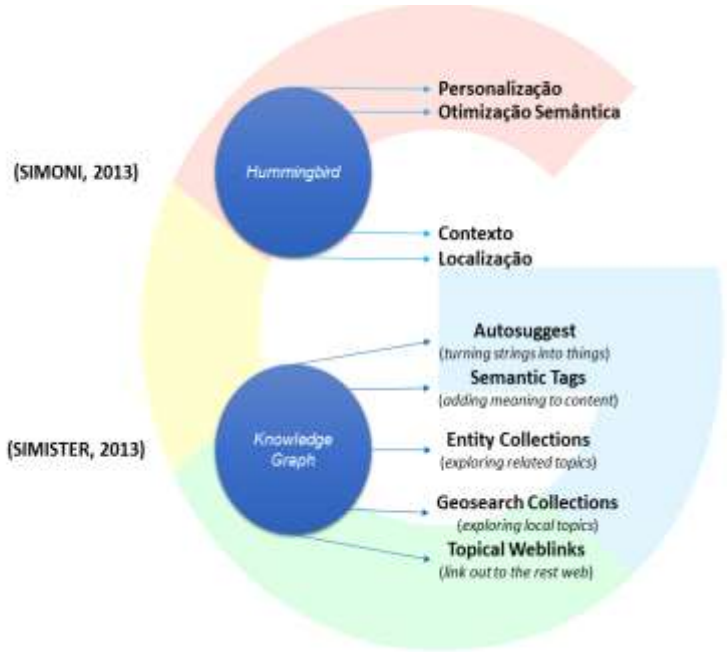

Figura 4: **Atributos de busca no Google: KG e *Hummingbirds*.**
Fonte: elaborado pelos autores.

A personalização se faz a partir da intenção de busca, que é inferida por meio das pragmáticas ou trilhas sígnicas semânticas deixadas pelo sujeito navegador, tanto no buscador quanto no ciberespaço. Hoje, é difícil fazer uma "busca pura", isto é, sem algum tipo de personalização. O fato de estar "logado", a cidade em que o usuário se localiza, a plataforma utilizada, o sistema operacional, a navegação, enfim, tudo é capturado por sistemas de busca. Isso quer dizer que as máquinas estão ouvindo o que milhões de pessoas falam, por meio de ferramentas linguísticas e estatísticas.

A "otimização" semântica, nesse ambiente, **é sinônima de "relevância**". A marcação semântica na estrutura dos dados potencializa o sentido, afirma a referência e constrói uma rede de significação, em que o conteúdo significativo é relevante para o usuário, conforme Figueiredo (1977), a relevância torna-se atributo da informação, como mencionado.

O contexto pode ser interpretado tanto do ponto de vista linguístico, e isso teria relação com a otimização semântica, quanto do ponto de vista pragmático, aproximando-se da personalização e localização. No primeiro caso, considerando a Semântica como a relação do signo com a realidade, pode-se tê-la como contexto. No segundo caso, a personalização e a localização trazem trilhas sígnicas que são pragmáticas de interação e intenção dos sujeitos no ciberespaço.

Já o KG torna possível a visualização dos dados descritos semanticamente (dados estruturados) por meio do formato e padrões da WS e possui as seguintes propriedades para a busca: *autosuggest* (torna trilhas ou linhas de computação em coisas); *semantic tags* (adiciona significado ao conteúdo); *entity collections* (explora tópicos ou itens relacionados); *geosearch collections* (explora tópicos locais); *topical weblinks* (*links* para o resto da *Web*) (MONTEIRO, 2015; SIMISTER, 2013).

As *tags* (*queries*) são palavras e como tais são unidades semânticas que guardam em si mesmas as referências e significações próprias. Já a autossugestão são frases encapsuladas, a partir do histórico de busca de outros usuários, e expressões semânticas, pois "[...] quando há o agenciamento de palavras para a construção da frase, a referência passa a ser o próprio ato de enunciação; dele decorre o sentido da frase." (ARAÚJO, 2012, p. 140).

A coleção de entidades advém dos bancos de dados de conhecimento, como, por exemplo, no caso do Google, o Wikidata (da *Wikimedia Foundation*), que é um desdobramento da *Web* Semântica, e reforça, como mencionado, o sentido e a referência (do estado de coisas no mundo), uma vez que itens são descritos, propriedades e valores são atribuídos a eles.

A geolocalização diz respeito, além da referência, ao contexto, pois, de acordo com Araújo (2012), o contexto faz parte da referência ao mundo, no ato de enunciação.

Por fim, os *weblinks* para outros tópicos vêm reforçar as teias de significações dos índices contemporâneos, no ciberespaço, gerados por esses Sistemas de Significação que são os mecanismos de buscas.

Nessa direção, Corcoglioniti *et al.* (2015) afirmam que o objetivo de um Sistema de Recuperação da Informação é determinar, para uma consulta, os documentos relevantes em uma coleção de textos, classificando-os de acordo com o seu grau de relevância. Afirmam, ainda, que os Sistemas Semânticos de Recuperação da Informação aumentam em 3.5 pontos percentuais sua relevância quando combinam a análise semântica à textual (de acordo com a *Mean Average Precision* – MAP, métrica utilizada em experimento realizado pelos autores).

A relevância nos mecanismos de busca (especialmente no Google) significa menos esforço de processamento de informação e mais efeitos cognitivos para a satisfação do usuário. Apoiada na Teoria da Relevância de Yus, Sperber e Wilson, Caldeira (2015, p.92) explica que os efeitos cognitivos são:

a) implicação textual: ocorre quando suposições resultam da síntese de uma informação antiga com uma informação nova;
b) fortalecimento ou enfraquecimento de suposições, que ocorre quando a informação nova fornece evidências que fortalecem ou enfraquecem uma informação antiga;
c) eliminação de suposições contraditórias, que ocorrem quando a informação nova fornece evidências que levam ao abandono de suposições antigas.

A partir desses pressupostos, a autora considera que o Google, ao oferecer a personalização e a contextualização do usuário, além de tecnologias da *Web* Semântica,

> [...] maximiza a relevância *input* (a consulta, a busca) para disponibilizar *outputs* (resultados do mecanismo de busca) que atendam à relevância de cada indivíduo, aumentando os efeitos cognitivos e diminuindo o esforço de processamento do usuário para que este alcance a meta que o levou à consulta. (CALDEIRA, 2015, p. 91, grifos da autora).

Dessa forma, atualizando os estudos sobre a relevância nos SRI, de acordo com o propósito deste artigo e em especial nos mecanismos de busca, pode-se inferir que ela envolve, especialmente:

a) a comunicação, processamento e a cognição, tanto de sistemas quanto de usuários;
b) relação sistema-sistema;
c) relação sistema-usuário;
d) coleção de entidades e seus aspectos semânticos: sentido e referência;
e) personalização e contextualização.

A partir do exposto, e concordando com Caldeira (2015), os mecanismos de busca são um *locus* profícuo de estudo sobre a relevância e sobre o comportamento de busca da informação e de conhecimento da sociedade, na atualidade.

4 CONSIDERAÇÕES FINAIS

A RI apresenta-se no cenário da CI e da CC como parte consagrada de estudos e aplicações. Sua extensão alcança, nesta última década, o desenvolvimento dos mecanismos de busca no ciberespaço e suas nuances em relação à recuperação do conteúdo de forma relevante ao usuário.

Apresentamos as visões de diversos autores estudiosos da CI que entendem os SRI como Sistemas de Comunicação, ou parte dele, dentre os quais também corroboram com a visão de que os mecanismos de busca também são considerados SRI.

Intentando responder as questões levantadas, na introdução deste artigo, sobre os mecanismos de busca, ou sistemas de significação, destacamos algumas observações. Os aspectos intelectuais da descrição da informação envolvem, na camada de estruturação (da *Web* Semântica), aspectos semânticos de descrição de dados, que são ligados e contextualizados com outras fontes e informações textuais e imagéticas. A Semântica Formal entende que as relações de significação podem ser descritas formalmente.

As especificidades para a busca englobam interfaces, pragmáticas de interação, contextos, negociações do sentido e a referencialidade que permite alcançar os objetos no mundo (OLIVEIRA, 2001).

Os sistemas, técnicas ou máquinas envolvem, entre outros, etiquetas, dados estruturados, variáveis existentes nos algoritmos, agentes computacionais, arquiteturas de coleta e indexação, mapa do conhecimento, enfim, sistemas híbridos de organização e significação no ciberespaço.

Nesse sentido, o conceito de relevância confunde-se com a otimização semântica, isto é, com os significados e também com o menor esforço e maior efeito cognitivo do usuário, mas baseando-se nas várias apropriações do termo, ao longo do desenvolvimento dos SRI e estudos da RI, a relevância é um conceito em devir, à espera de novas atualizações conceituais, cognitivas e técnicas. Em especial, nos mecanismos de busca, o conceito está fortemente relacionado às tecnologias semânticas do ciberespaço, à personalização e à contextualização da busca.

REFERÊNCIAS

ARAÚJO, V. M. R. H de. Sistemas de informação: nova abordagem teórico-conceitual. **Ciência da Informação**, v.24, n.1, p. 1-39, 1995. Disponível em: <http://revista.ibict.br/ciinf/article/view/610/612>. Acesso em: 22 de jul. 2016.

ARAÚJO, V. M. A. P. de. Sistemas de recuperação da informação: uma discussão a partir de parâmetros enunciativos. **TransInformação,** v. 24, n. 2, p. 137-143, maio/ago. 2012. Disponível em: <http://www.scielo.br/pdf/tinf/v24n2/a06v24n2.pdf>. Acesso em: 22 de jul. 2016.

BAEZA-YATES; R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM Press, 1999.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação de informação:** conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

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

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **The Scientific american.com,** 17, May 2001. Disponível em: <file:///C:/Users/User/Desktop/10.1.1.115.9584.pdf>. Acesso em: 18 ago. 2016.

CALDEIRA, F. H. O mecanismo de busca do Google e a relevância na relação sistema-usuário. ***Letrônica:*** Revista digital do Programa de Pós-Graduação em Letras PUCRS. Disponível em: <http://revistaseletronicas.pucrs.br/ojs/index.php/letronica/article/view/19616>. Acesso em: 12 de jun. 2016.

CARDOSO, O.N.P. Recuperação da Informação. **INFOCOMP:** Journal of Computer Science, v. 2, n. 1, p. 1-6, 2000. Disponível em: <http://www.dcc.ufla.br/infocomp/index.php/INFOCOMP/article/view/46/31>. Acesso em: 08 jun. 2016.

CENDÓN, B. V. Sistemas e redes de informação. In: OLIVEIRA, M. (Coord.). **Ciência da Informação e Biblioteconomia:** novos conteúdos e espaços de atuação. Belo Horizonte: UFMG, 2005. p. 45-75.

CORCOGLIONITI, F. *et al.* Knowledge extraction for information retrieval. In: SACK, *et al.* (Ed.). **The Semantic Web. Latest Advances and New Domains.** Berlin: Springer, 2016. p. 317-333.

FIGUEIREDO, L. M. de. O conceito de relevância e suas implicações. **Ciência da Informação**, v. 6, n. 2, p. 75- 78, 1977.

FIORAVANTE, Felipe. **Tendências emergentes em mecanismos de busca**. Disponível em: < www.terraforum.com.br/>. Acesso: 20 ago. 2009.

FOSKETT, D. J. A note on the concept of "relevance". **Information Storage and Retrieval**, Great Britain, v. 8, n. 1, p. 77-78, 1972.

GOOGLE. **The *knowledge graph*.** 2012. Disponível em: <http://www.google.com.br/intl/pt-PT/insidesearch/features/search/knowledge.html>. Acesso em: 20 jun. 2016.

GOOGLE. Rastreamento e indexação: como funciona a pesquisa. In: _____. **Por dentro da pesquisa. <**https://static.googleusercontent.com/media/www.google.com/pt-BR//intl/pt-Br/insidesearch/howsearchworks/assets/searchInfographic.pdf>. Acesso em: 18 ago. 2016.

LANCASTER, F. W. MEDLARS: report on the evaluation of its operating efficiency, **American Documentation**, p. 119-142, 1969.

LAVRENKO, V. **A generative theory of relevance**. Berlin: Springer, 2009.

LEDFORD, J. L. **SEO**: Search Engine Optimization – Bible. Indianapolis: Wiley Publishing Inc, 2009. 389 p.

LEWANDOWSKI, D. New perspectives on web search engine research. In: _____. **Web search engine research**. 4. ed. Bingley: Emerald Group Publishing Limited, 2012. p. 1-16.

MELONI, L. F**. Tudo sobre *RankBrain*, o novo algoritmo do Google**. 2015. Disponível em: <http://www.profissionaldeecommerce.com.br/faq-tudo-sobre-rankbrain-o-novo-algoritmo-google/>. Acesso em: 08 jul. 2016.

MONTEIRO, S. D. As múltiplas sintaxes dos mecanismos de busca no ciberespaço. **Informação & Informação,** v. 14, n.1, p. 68-102, 2009. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/2027>. Acesso em: 28 abr. 2016.

MONTEIRO, S. D. *Knowledge Graph* e a significação: novos agenciamentos semióticos dos índices contemporâneos. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 8, n. 2, p. 203-218, jul./dez. 2015.

MONTEIRO, S. D. O ciberespaço e os mecanismos de busca: novas máquinas semióticas. **Ciência da Informação**, v. 35, n. 1, p. 31-38, jan./abr. 2006.

MOOERS, C. N. Zatocoding applied to mechanical Organization of Knowledge. **American Documentation,** v. 2, n. 1, p. 20-32, 1951.

OLIVEIRA, R. P. **Semântica formal.** Campinas: Mercado de Letras, 2001.

OLIVEIRA, G.; ARAÚJO, W. Usar ou não usar – qual a relevância das metatags na recuperação da informação pelos mecanismos de busca? **Biblionline**, v. 8, n. 1, p. 60-77, 2012.

PAGE L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking: bringing order to the Web**. Disponível em: <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>. Acesso em: 18 ago. 2016.

SANGIRARDI, Pedro. Tecnologias cognitivas, comunicação e a crise de representação política. ***Revista Com Política,*** v. 1, n. 4, p. 218-234, jan./jul. 2014. Disponível em: <http://compolitica.org/revista/index.php/revista/article/view/145>. Acesso em: 08 out. 2015.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v. 1, n. 1, p. 41-62, jan./jun. 1996.

_____. Why is relevance stull the basic notion in Information Science? In: PEHAR, F.; SCHÖGL, C.; WOLFF, C. (Ed.). Re:inventing Information Science in the Netwoked Society. INTERNATIONAL SYMPOSIUM ON INFORMATION SCIENCE, 14, 2015, Zadar. **Proceedings...** Zadar: ISI, 2015.

_____. Relevance: a review of and a framework for the thinking on the notion in Information Science. **Journal of the American Society for Information Science**, v. 58, n. 13, p. 321-343, nov./dez. 1975.

_____. Research on relevance in Information Science: a historical perspective. In: CARBO, T.; HAHN, T. B (Ed.). International Perspectives on the History of Information Science and Technology. ASIS&T – PRE-CONFERENCE ON THE HISTORY OF ASIS&T AND INFORMATION SCIENCE AND TECHNOLOGY, 75, 2012, Medford. **Proceedings ...** Medford: ASIS&T, out. 2012.

SIEGEL, D. *Pull* – **The power of Semantic *Web* to transform your business**. Portfolio Hardcover, 2010.

SIMISTER, S. The Freebase API: tapping into Google's Knowledge Graph. In: GOOGLE DEVELOPERS. ***Freebase API***, 2013. Disponível em: <https://developers.google.com/freebase/>. Acesso em: 15 nov. 2014.

SIMONI, R. **SEO e semântica *web*: uma nova revolução com a busca de entidades**. 2013. Disponível em: < http://www.conversion.com.br/blog/seo-e-semantica-*web*-uma-nova-revolucao-com-a-busca-de-entidades/>. Acesso em: 02 out. 2015.

SILVA, R. E. da; SANTOS, P. L. V. A. da C.; FERNEDA, E. Modelos de recuperação de informação e *web* semântica: a questão da relevância. **Informação & Informação**, v. 18, n. 3, p. 27 – 44, set./dez. 2013.

SOUZA, R.R. Sistemas de recuperação de informações e mecanismos de busca na *web*: panorama atual e tendências. **Perspectivas em Ciência da Informação**, v. 11, n. 2, p. 161-173, maio/ago. 2006.