Sistemas de Recuperação de Informações e Mecanismos de Busca na *web*: panorama atual e tendências

Renato Rocha Souza

Engenheiro de Sistemas, Especialista em Tecnologia na Educação, Mestre em Engenharia de Produção e Doutor em Ciência da Informação

Busca-se traçar um panorama das características básicas dos sistemas de recuperação de informações, apresentando seus modelos de recuperação mais comuns. São apresentados com ênfase especial os mecanismos de busca na web, seu estado atual de desenvolvimento e algumas tendências para o futuro. Dentre estas tendências, destacam-se notadamente aquelas introduzidas pelo Google, como o algoritmo *pagerank*, dentre outras inovações.

Palavras-chave: Sistemas de recuperação de informações; mecanismos de busca; web; internet.

Recebido em 09.04.2006 Aceito em 10.07.2006

Introdução

Os sistemas de informação e de comunicação hoje permeiam e mesmo viabilizam virtualmente todas as atividades humanas, e não mais se pode conceber a sociedade sem a acentuada imbricação das tecnologias de informação que nela surgem e que a modificam. Acompanhando o desenvolvimento dessas tecnologias, os repositórios de informações que são produzidos no desempenho das inúmeras atividades humanas vêm migrando para o ambiente *on-line*, cada vez mais em formatos digitais, acessíveis através de redes e sistemas de computadores. Pode-se assistir em paralelo à criação destes espaços (a Internet e a *web*, as intranets empresariais, os portais corporativos, as bibliotecas digitais, etc.) o surgimento de ferramentas e sistemas de recuperação de informações, para suprir a necessidade de recuperar as informações criadas continuamente em ritmos vertiginosos.

Sistemas de recuperação de informações

A dificuldade de conceituação do que seja um sistema de recuperação de informações advem, a princípio, da ambigüidade dos conceitos de sistema e de informação em si (ARAÚJO, 1995). No âmbito destes sistemas, costuma-se evidenciar o conceito de *informação como coisa*, ou seja, registros de conhecimentos em documentos (BUCKLAND, 1991), em detrimento de outras definições e contextos. Sem embargo, há, no contexto específico supracitado, extensa literatura especializada das áreas de ciência da informação e ciência da computação, na qual podemos encontrar uma dezena de definições razoavelmente consensuais, das quais foram pinçadas as apresentadas a seguir.

KORFHAGE (1997) ressalta o caráter pessoal da informação, e aponta o fato de que sistemas de recuperação de informações armazenam dados, distinguindo as informações que foram armazenadas por um usuário das que serão apropriadas por outro. Os SRIs seriam os intermediários nesse processo mediado de troca de informações. Para LANCASTER & WARNER (1993 p. 4-5), os SRIs são a interface entre uma coleção de recursos de informação, em meio impresso ou não, e uma população de usuários; e desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários. Essa visão é abrangente, e inclui tarefas que são desempenhadas em conjunto com atores humanos. LANCASTER (1968) já havia anteriormente apontado o fato de que os SRIs não informam o usuário – no sentido de mudar seu conhecimento sobre objeto de sua questão –, mas apenas o informam sobre a possível existência de documentos atinentes à questão, além de características desses documentos; e procura, em outro trabalho, analisar os SRIs subdividindo-os em seis subsistemas: de documentos, de indexação, de vocabulário, de busca, de interface com o usuário e de *matching*[1] (LANCASTER, 1979). CHOWDHURY entende que o conceito de recuperação de informações – e como conseqüência, o conceito de sistemas de recuperação de informações – é auto-explanatório, e divide os SRIs em subsistemas de documentos, de usuários, e de busca/recuperação; detalhando cada um desses subsistemas (1999, p. 1-11). Para CHOWDHURY (Ibidem), os SRIs servem de ponte entre o mundo dos criadores de informações e os usuários dessas, e para isso,

[1] Matching pode ser definido nesse contexto como o casamento das necessidades de informação com os itens que fazem parte do acervo do sistema e que podem satisfazer esta necessidade.

colecionam-nas e as organizam. SALTON & MCGILL (1983, p. 1), e mais tarde BAEZA-YATES & RIBEIRO-NETO (1999, p. 1), definem SRIs como sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação.

Há que se notar que as definições procuram apreender um fenômeno atemporal – as necessidades de informação – e as várias metodologias e tecnologias que, através dos tempos, foram engendradas para atender a essas necessidades, desde as atividades de organização de coleções de documentos em acervos bibliográficos, até os modernos sistemas informatizados que lidam com documentos em formato digital. Partindo das definições citadas, assume-se que SRIs *organizam e viabilizam o acesso* aos itens de informação, desempenhando as atividades de:

- Representação das informações contidas nos documentos, usualmente através dos processos de *indexação e descrição* dos documentos;
- Armazenamento e gestão física e/ou lógica desses documentos e de suas representações;
- Recuperação das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as *necessidades de informação* dos usuários. Para isso é necessário que haja uma *interface* na qual os usuários possam descrever suas necessidades e questões, e através da qual possam também examinar os documentos atinentes recuperados e/ou suas representações.

Sem que seja necessário o aprofundamento da discussão conceitual sobre as diferenças entre *dado e informação*, há que se distinguirem os *sistemas de recuperação de informações* (SRI) *dos sistemas de gestão de bancos de dados* (SGBD). Dados podem ser definidos como seqüências de símbolos para os quais são atribuídos significados; símbolos estes que podem ser codificados, interpretados e manipulados por programas de computador, e enviados através de redes e dispositivos de comunicação. O conceito de informação já carrega um grau maior de abstração. A informação não prescinde do sujeito que a depreenda a partir dos dados, no ato conhecido como interpretação. No sentido estrito do conceito, nenhum programa de computador lida, sob o ponto de vista da máquina, com informações, a não ser que possua alguma capacidade de arrazoamento, e, assim mesmo, a utilização do termo dá margem a discussões. No uso corrente, porém, ambos os termos são utilizados para sistemas, apesar das diferenças entre os sistemas de recuperação de informações e sistemas de recuperação de dados, como os SGBDs.

Em sistemas gerenciadores de bancos de dados, os símbolos são armazenados em uma estrutura matricial em campos determinados, com metadados que lhes conferem certo sentido ontológico. Para recuperar dados específicos, basta especificar as restrições necessárias aos campos de pesquisa e codificá-las numa questão ou *query* (argumento de entrada no sistema) para que se tenha a resposta exata, fruto de busca completa e exaustiva.

A recuperação de informações traz dificuldades intrínsecas ao conceito de "informação", como a dificuldade da determinação da real necessidade do usuário e do seu melhor atendimento com os documentos que fazem parte do acervo do sistema (FOSKETT, 1997, p. 5). A associação entre os registros e seus conteúdos informativos é vaga, e isso pode acarretar problemas nas respostas a questões específicas, como baixas taxas *de revocação*[2] *e precisão*[3]. Um sistema de recuperação de informações deve buscar boa relação entre os

[2] A Revocação, ou "recall" ou mesmo "abrangência", é a razão do número de documentos atinentes recuperados sobre o total de documentos atinentes disponíveis na base de dados. A revocação mede o sucesso do SRI em recuperar documentos pertinentes

[3] Razão do número de documentos atinentes recuperados sobre o total de documentos recuperados. A precisão mede o sucesso do SRI em não recuperar documentos que não sejam relevantes de acordo com a necessidade de informação.

índices de revocação e precisão, para oferecer, em resposta a determinada consulta, referências ao maior número possível de documentos relevantes, ordenados por critérios que meçam esta relevância, e o menor número possível de documentos pouco ou não relevantes, de acordo com as necessidades de informação dos usuários.Dentre os diversos diagramas que descrevem o processo de recuperação de informações em sistemas, foi escolhido o proposto por BAEZA-YATES & RIBEIRO-NETO (1999), apresentado na FIG. 1, que enfatiza o processo da forma em que é realizado nos sistemas automatizados:

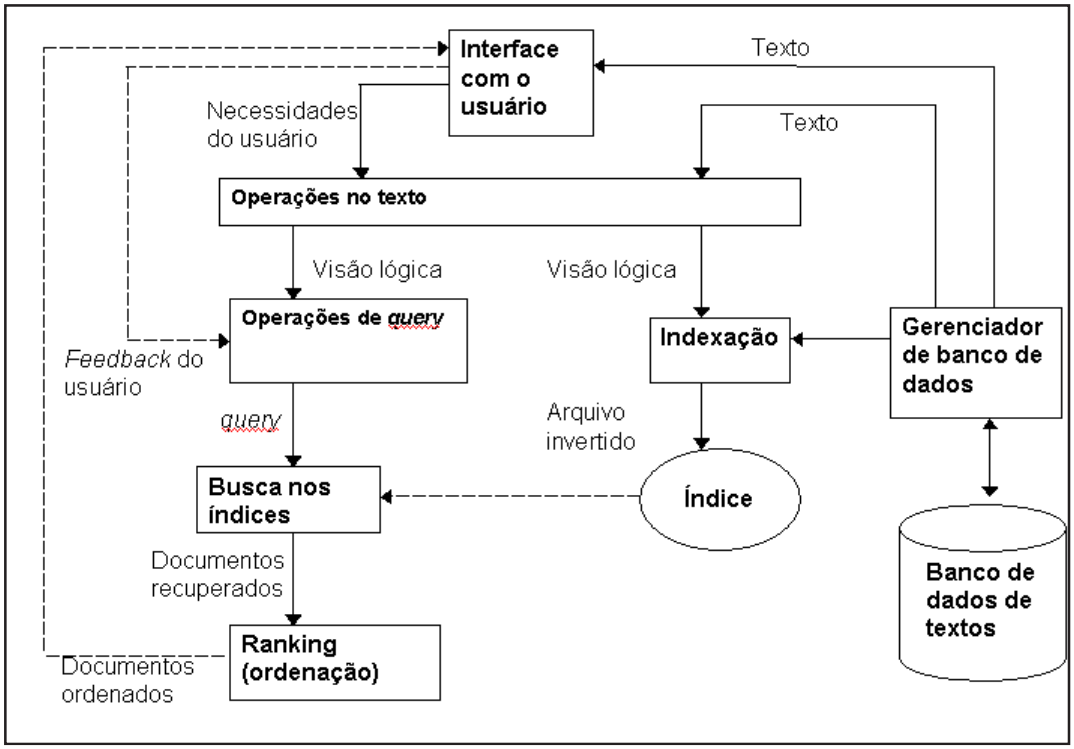

Figura 1 – O processo de recuperação de informações (adaptado de BAEZA-YATES & RIBEIRO-NETO, 1999, p. 10).

A FIG. 1 explicita as atividades de *representação* (operações no texto, indexação e criação do índice); *armazenamento e gestão* (dos documentos presentes no acervo do banco de dados de textos e do índice), e ***a*** *recuperação*, que se inicia através da análise da necessidade do usuário e redunda na apresentação de um conjunto ordenado de documentos, possivelmente permitindo ao usuário *feedback* sobre os documentos apresentados. A representação dos documentos é necessária, pois a busca na totalidade dos termos presentes geraria alta revocação e baixa precisão. E a qualidade desta representação – problema central na recuperação - é que garante boas taxas de precisão e revocação.

Recuperação de documentos em SRIs

Um dos problemas centrais da recuperação de informações em SRIs é a predição de quais são os documentos relevantes e quais devem ser descartados, e essa tarefa de "escolha", em sistemas automatizados, é executada por algum tipo de algoritmo que, baseado em heurísticas

previamente definidas, decide quais são os documentos relevantes a serem recuperados e os ordena a partir dos critérios estabelecidos (BAEZA-YATES & RIBEIRO-NETO,1999, p. 19). Quando a indexação é realizada manualmente – ou melhor expressando, intelectualmente – por seres humanos, cabe a estes descobrir conceitos que sirvam de termos-índices para serem vasculhados durante as consultas (*queries*) de usuários. Na indexação automática, existem dezenas de estratégias para a correta ponderação do valor dos documentos, de acordo com uma explicitação de necessidade de informação.

Alguns dos modelos propostos e atualmente utilizados em algoritmos de recuperação de informações são exemplificados na FIG. 2:

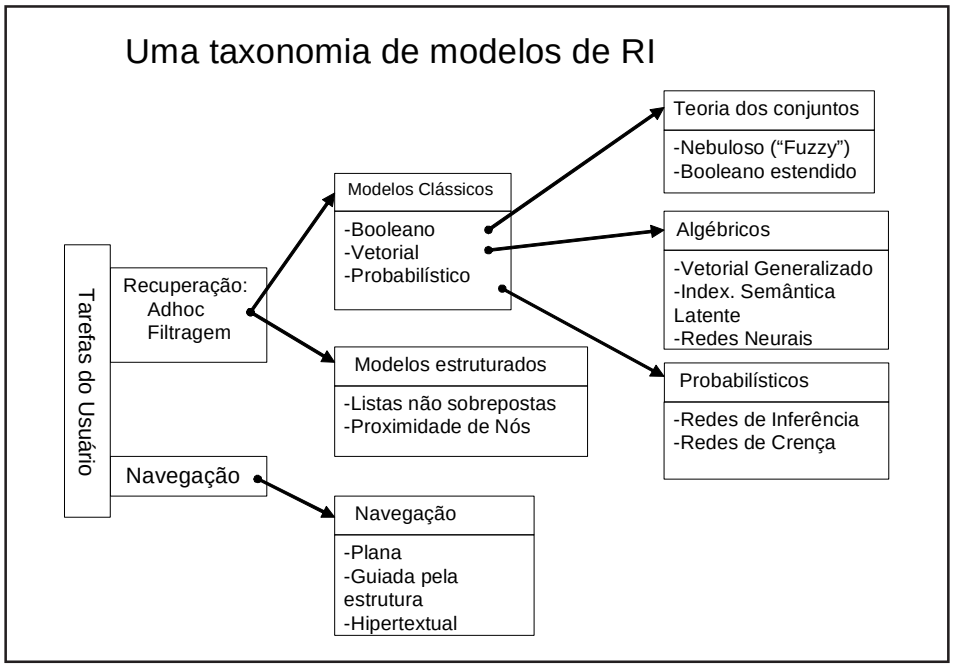

Figura 2 – Uma taxonomia de modelos de RI (adaptado de BAEZA-YATES & RIBEIRO-NETO, 1999, p. 21).

Esta figura ilustra 15 modelos de recuperação de informações, que serão detalhados nos parágrafos a seguir baseados na forma como são apresentados pelos autores, BAEZA-YATES & RIBEIRO-NETO (1999).

Nos sistemas de recuperação de informações, há usualmente interface através da qual o usuário traduz sua necessidade de informações em forma de questões ou palavras-chave, ou mesmo examina os documentos na busca de informações pertinentes. Essas ações são consideradas como papel do usuário. Os dois modos de buscar informações são classificados em modelos de recuperação e os modelos de navegação. Nestes últimos, o usuário não propõe uma questão (*query*) ou necessidade de informação ao sistema. Em vez disso, navega através dos documentos – que não foram necessariamente indexados previamente – buscando informações de interesse. A busca em estruturas de arquivos ligados em rede é usualmente executada através de navegação do tipo hipertextual.

Quando o acervo de documentos sofre poucas alterações enquanto novas *queries* são submetidas ao sistema, chama-se o modo de operação de *"recuperação adhoc*". Quando as *queries* se mantêm relativamente estáticas

enquanto novos documentos são adicionados, chama-se a esse modo de operação de filtragem (*filtering*). A filtragem acontece usualmente em processos de monitoração de fontes de informação, enquanto a recuperação *adhoc* representa as buscas usuais em SRIs.

Os modelos de recuperação se dividem em modelos *clássicos* e modelos *estruturados*. Nos modelos clássicos, cada documento é descrito por um conjunto de palavras-chave representativas – também chamadas de termos de indexação – que busca representar o assunto do documento e sumarizar seu conteúdo de forma significativa. Nos modelos estruturados, podem-se especificar, além das palavras-chave, algumas informações acerca da estrutura do texto (como seções a serem pesquisadas, fontes de letras, proximidade das palavras, entre outras informações.).

Os modelos clássicos de recuperação são três: o modelo **booleano**, o modelo **vetorial** e o modelo **probabilístico**. Para cada um deles, há modelos alternativos que visam estendê-los em funcionalidade e o desempenho. Vamos examinar brevemente esses três modelos adiante:

Modelo booleano: esse modelo, baseado na teoria dos conjuntos, é simples e elegante, embora não seja dos mais eficazes. Para cada *query*, são recuperados todos os documentos que possuem os termos nas condições especificadas pelo usuário, que ainda pode utilizar os operadores booleanos *or*, *and* e *not* para estabelecer relações específicas de ocorrência com as palavras-chave, de forma a especificar os documentos a serem recuperados. Sua maior desvantagem é o fato de trabalhar de forma binária, ou seja, os documentos são analisados sob o critério dualista relevante / não relevante, e não é criada nenhuma espécie de ordenação dos resultados que atendam às condições de consulta. Existem alguns modelos alternativos ao booleano, apresentados a seguir:

- **Lógica difusa ou nebulosa (fuzzy)**: nesse modelo, busca-se estender o conceito da representação dos documentos por palavras-chave, assumindo que cada *query* determina um conjunto difuso e que cada documento possui um grau de pertencimento a esse conjunto, usualmente menor do que 1. O grau de pertencimento pode ser determinado pela ocorrência de palavras expressas na *query*, tal como no modelo booleano, mas pode também utilizar um instrumento – como um tesauro – para determinar que termos relacionados semanticamente aos termos índice também confiram algum grau de pertencimento ao conjunto difuso determinado pela *query*.
- **Booleano estendido**: neste modelo, busca-se a superação do problema das decisões binárias do modelo clássico por meio da aferição de pesos aos termos, aproximando o modelo original do modelo vetorial, a seguir.

Modelo vetorial: nesse modelo, os documentos são modelados como "sacos de palavras" (*bags of words*), e são representados como vetores no espaço n-dimensional, onde n é o total de termos índices (palavras) de todos os documentos no sistema. No modelo, que é não binário, pode-se calcular um grau de similaridade a ser satisfeito pelos documentos para serem considerados relevantes (ex: que as palavras apareçam ao menos duas vezes, etc.) e determinar o grau

de similaridade, com vistas a construir um *ranking*. O modelo vetorial é a base da grande maioria de sistemas de recuperação de informações, mais notadamente os que têm como objeto a Internet, embora estes utilizem também outras técnicas[4] para determinar o *ranking* de documentos como resposta a uma consulta. Em seguida, apresentamos alguns modelos que se propõem a estender a funcionalidade do modelo vetorial:

- **Vetorial generalizado**: nesse modelo, questiona-se a independência dos termos índices, assumida nos modelos clássicos, e abre-se a possibilidade de considerar que certos conceitos – representados por estes termos – sejam relacionados. Uma das formas de determinar relações entre termos é examinar a co-ocorrência desses no texto de cada documento, além do exame das relações semânticas estabelecidas por um tesauro, como foi comentado.
- **Indexação semântica latente**: nesse modelo, questiona-se a significância das palavras-chave como candidatos a descritores, e busca-se estabelecer o casamento *conceitual* entre documentos e *queries*. Se nos modelos anteriores buscava-se estabelecer um mapeamento em um espaço booleano ou vetorial de palavras, no modelo em questão busca-se mapear cada documento e cada *query* em um espaço menor, construído a partir dos conceitos relevantes que possuem os documentos no acervo.
- **Redes neurais**: nesse modelo, utiliza-se o poder das redes neurais para realizar o casamento de padrões entre as *queries* e os documentos do acervo do sistema. Cada *query* "dispara" um sinal que ativa os termos índice, que por sua vez propagam os sinais aos documentos relacionados. Estes, por sua vez, retornam os sinais a novos termos índices, em interações sucessivas. O conjunto resposta é definido através desse processo, e pode conter documentos que não compartilhem nenhum termo-índice com a *query*, mas que tenham sido ativados durante o processo.

Modelo probabilístico: nesse modelo, supõe-se que exista um conjunto ideal de documentos que satisfaz a cada uma das consultas ao sistema, e que este conjunto pode ser recuperado. Através de tentativa inicial com um conjunto de documentos (para a qual se podem utilizar técnicas de outros modelos, como o vetorial) e do *feedback* do usuário em sucessivas interações, busca-se aproximar cada vez mais deste conjunto ideal, por meio de análise dos documentos considerados pertinentes pelo usuário. O valor desse modelo está em considerar a interação contínua com o usuário como um caminho para refinar o resultado continuamente. Os modelos que procuram ampliar o escopo do modelo probabilístico são os seguintes:

- **Redes de inferência**: nesse modelo, associam-se variáveis aleatórias ao evento do atendimento de uma *query* específica por um documento específico. Essas variáveis podem ser alteradas de acordo com os eventos futuros, de forma a estabelecer relacionamentos baseados nos eventos observados.
- **Redes de crença (belief networks)**: nesse modelo, similares às redes de inferência, documentos e *queries* são

[4] Nos mecanismos de busca da Internet de terceira geração, além do modelo vetorial, utilizam-se, para determinar a ordenação dos documentos, técnicas como a análise de links, que contabiliza a quantidade de documentos que apontam para um documento específico através de links hipertextuais; a análise de autoridade, que investiga a idoneidade e importância da instituição que hospeda o documento em seus servidores; e outras técnicas, como as utilizadas nas redes de inferência e redes de crença.

modelados como subconjuntos de um espaço de conceitos. A cada documento, associa-se a probabilidade de que o mesmo cubra os conceitos presentes no espaço de conceitos. Cada *query* é mapeada no espaço de conceitos, que por sua vez, está conectado ao espaço de documentos.

Os modelos apresentados são apenas amostras do que vêm sendo pesquisado, em um campo que contém muitas frentes de pesquisa, que não poderiam ser enumeradas neste artigo. As novas estratégias que surgem vêm tentar superar um aparente esgotamento das estratégias tradicionais, baseadas em modelagens estatísticas internas do *corpus*, e há consenso hoje de que a melhoria da eficácia do serviço aos usuários dos sistemas depende de esforços em diversas linhas de pesquisa, em todo o espectro da cadeia de processos de organização da informação. Algumas das novas alternativas são as seguintes:

1) a intensificação da exploração das informações semânticas intrínsecas aos documentos, de forma a expandir a compreensão das unidades e padrões de significado em textos, imagens e outras mídias;

2) o desenvolvimento de novas possibilidades de marcação semântica dos dados utilizando-se metalinguagens, criando registros de metadados acoplados aos próprios documentos com termos amplamente consensuais e não ambíguos, para que esses possam ser mais facilmente manipulados e identificados por computadores e outros dispositivos e, como conseqüência, pelos usuários;

3) o desenvolvimento de estratégias de apresentação da informação recuperada nas buscas sob formatos altamente significativos e contextuais, de forma que as relações entre os conceitos, e em conseqüência, os contextos lingüísticos subjacentes, sejam evidentes, o que permitiria aos usuários refinarem os resultados através da definição das conexões pertinentes e a exclusão das conexões geradas pelo ruído informacional;

4) A construção e a manutenção de perfis personalizados de utilização, de forma que o SRI "aprenda" com a forma de trabalho e interesses do usuário e possa utilizar essas informações específicas para melhorar a estratégia de busca do SRI;

5) A utilização das topologias hipertextuais, como no caso da *web*, e dos mapas demográficos de relacionamentos pessoais, traçados através dos usuários dos *social software*[5], para agregar elementos estatísticos de cunho bibliométrico e temático nas decisões sobre a relevância dos documentos para uma dada consulta.

Ao invés de perseguir a evolução de sistemas baseando-se puramente na imitação das heurísticas intelectuais humanas – sonho antigo das linhas de pesquisa em Inteligência Artificial – buscam-se hoje estratégias diferenciadas, que só poderiam ser implementadas em ambientes com poder computacional extremo. Algumas destas estratégias são apresentadas no mapa conceitual da FIG. 3, e em seguida delineia-se uma pequena explanação sobre os novos caminhos esboçados no mapa:

[5] "Social software" permite que indivíduos se conectem e colaborem através de comunicação mediada por computador, de forma a criar comunidades virtuais (WIKIPEDIA). Disponível em http://en.wikipedia.org/wiki/Social_software

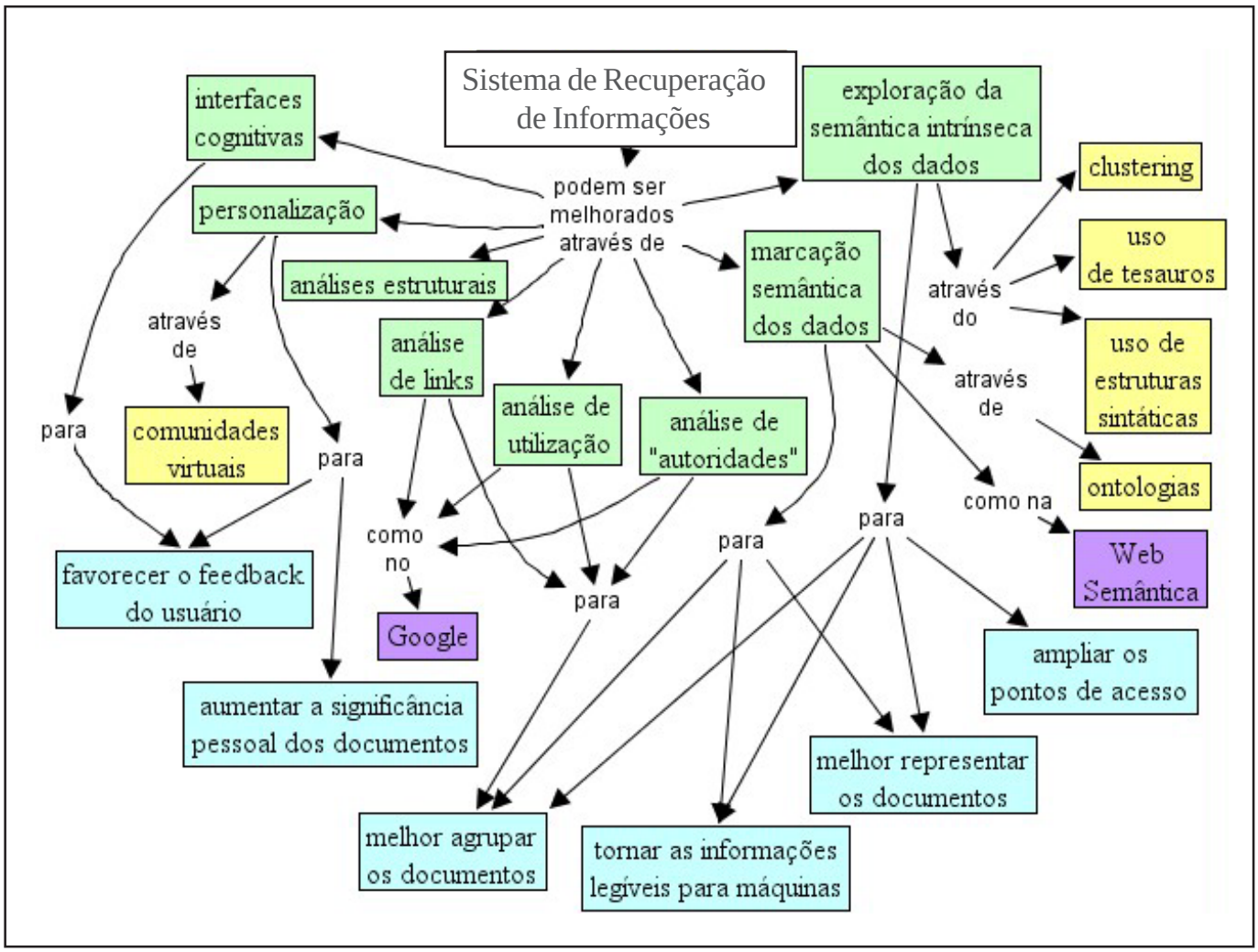

FIGURA 3 – Estratégias alternativas para melhoria dos SRIs.
Fonte: o próprio autor.

a) Interfaces cognitivas

As pesquisas para a criação de interfaces cognitivas são talvez as mais estabelecidas. A maioria procura criar uma plataforma de utilização mais intuitiva por parte dos usuários de SRIs, utilizando, como exemplo, geometrias hiperbólicas ou mapas conceituais (LAMPING et al, 1995; CAÑAS et al, 1999). São necessárias também interfaces especiais para, por exemplo, a recuperação de informações em ambientes de muitas mídias, como áudio e vídeo, pois estes sistemas necessitam apresentar interfaces multimídia. Temos como exemplo de mecanismos de busca de vídeo o YouTube[6] e de áudio o Find Sounds[7]. Temos também como exemplos de mecanismos baseados em interfaces cognitivas para recuperação o KARTOO[8], como exemplo de geometrias hiperbólicas o *Hiperbolic Browser*[9] e como exemplo de sistemas de mapas conceituais o *site* do *Institute for the Human and Machine Cognition*[10] da *umnHmaHuman University of West Florida*.

b) Personalização

A personalização das buscas de acordo com o perfil e interesses do usuário é uma grande vertente, na medida em que permite também aumentar

[6]http://www.youtube.com/

[7]http://www.findsounds.com/

[8]http://www.kartoo.com/

[9]http://www.parc.xerox.com/istl/projects/uir/projects/InformationVisualization.html

[10]http://cmap.ihmc.us/

o conhecimento e a demografia de necessidades de informação destes usuários, e estas informações podem ser usadas para database marketing e vendas personalizadas. Exemplos destas estratégias são os conjuntos de ferramentas gratuitas que compartilham *cookies*[11] no computador, como os mecanismos de busca Yahoo[12] com seu pacote de mensagens instantâneas *Yahoo Messenger*, o *webmail* gratuito *Yahoo mail*, e a ferramenta de grupos *Yahoogroups*. Ainda podemos citar o conjunto de ferramentas da Microsoft Network[13], como o *webmail* gratuito *Hotmail*, o *MSN Messenger*, o *MSN Spaces*, dentre outros. Devemos apontar, entretanto, que a empresa que está mais avançada nesta vertente é certamente a Google[14], com sua quintessência da personalização através de perfis. Estes são desenhados através dos usuários do *Orkut*, do *Google Earth*, do *Gmail*, do *Google Talk*, do *Google Desktop Search*, entre outras ferramentas gratuitas. Todos os grandes mecanismos de busca têm desenvolvido estratégias neste sentido.

c) Análises estruturais

A análise estrutural, na prática, é a possibilidade da especificação de restrições quanto ao formato dos documentos no momento de pesquisa. Esta possibilidade já é implementada na prática nas opções avançadas do Google, Yahoo, entre outros, onde se podem especificar formatos como PDF, PPT, DOC, entre outros, além do tamanho, data ou domínio de origem dos documentos procurados.

d) Análise de links

O mecanismo de busca Google foi o primeiro a desenvolver um algoritmo – chamado de "*Pagerank*" que calcula o "valor" de um *site* levando-se em conta àquelas páginas que para ele apontem. Esta estratégia já é implementada na prática em mecanismos de busca chamados de terceira geração como Google, Yahoo, etc. A análise de links consiste ainda em analisar, para fins de determinação do assunto do documento, os conceitos que descrevem o documento em páginas que apontam para este, o que poderia ser descrito pelo mote "diga-me o que dizem de você, que te direi quem és". Ultimamente, têm sido buscadas formas de extrapolar e melhorar a idéia inicial, através da ponderação dos apontadores de acordo com o assunto principal determinado na página que aponta, o que permitiria que o "voto" representado pelo *link* pudesse ter peso diferente se apontasse para páginas com a mesma temática ou com temáticas diferentes. Ainda há pesquisas para implantar o pós-processamento do ranking de relevância, levando em conta somente os apontadores das páginas que fazem parte do resultado da pesquisa. É um campo em franca evolução.

e) Análise de utilização

A análise de utilização considera as escolhas realizadas pelos usuários, dentre aquelas páginas apresentadas como resposta às consultas anteriores, de forma a valorizar aquelas mais acessadas, que seriam teoricamente mais importantes, devendo assim serem melhor ranqueadas. Alternativamente, se

[11] Um cookie é uma informação que um site web grava no disco rígido do usuário de forma a poder guardar informações sobre este usuário para uso posterior.

[12] http://www.yahoo.com

[13] http://www.msn.com

[14] http://www.google.com

o usuário volta à página de pesquisa em um tempo curto, após o exame de uma dada página, esta receberia um voto negativo. Foi implantada pelo Google com o nome de "*click-through rate*", e vem sendo adotada por outros mecanismos de busca.

f) Análise de autoridades

A análise de autoridades estabelece que alguns *sites* são mais importantes para determinados termos de busca, não importando a freqüência destes termos em outras páginas indexadas na base. Uma das formas de definir que um site seja autoridade em um assunto é o fato do termo que representa este conceito estar presente no nome do domínio.

g) Marcação semântica dos dados na origem

Os melhores exemplos desta vertente são as tecnologias exploradas no contexto da *web* semântica[15], com vistas ao projeto e à implementação de padrões de metadados, que adicionem aos dados informações significativas sobre seus contextos, marcando-os semanticamente; e mecanismos de busca que levem em conta estes dados marcados. Ainda no âmbito da *web* semântica, há pesquisas e desenvolvimento de programas de computador comumente chamados *agentes inteligentes*, que têm a possibilidade de fazer a colheita (*harvesting*) de informações em outros computadores, agentes e dispositivos eletrônicos, para então tomar decisões baseadas em heurísticas embutidas. Um exemplo de mecanismo de busca baseado nestas tecnologias é o Ontoweb[16].

h) Exploração da semântica intrínseca dos dados

Um pujante campo, que ainda aparece pouco explorado é a utilização da semântica embutida nos próprios documentos, ou seja, as potencialidades intratextuais da linguagem natural, para automatizar e melhorar as tarefas de indexação, organização e recuperação de informações. Os SRIs usualmente utilizam como descritores e unidades de recuperação as palavras isoladas que, embora sirvam de forma bastante razoável aos propósitos de recuperação de informações, falham em grande parte justamente por não considerarem o contexto informacional implícito em toda a consulta porque não estão preparados para lidar com a forma com que estas palavras ou conceitos estão relacionados. Esses relacionamentos, na prática, determinam as minúcias e especificidades dos assuntos pesquisados. Dessa forma, perdem-se informações fundamentais sobre o escopo em que as palavras estejam sendo utilizadas e, em conseqüência, a pertinência da pesquisa diminui. Pesquisas nessa área incluem o uso de estruturas profundas da linguagem natural, como os sintagmas verbais e nominais, para indexação e recuperação (KURAMOTO, 1996; MOREIRO et al, 2003; SOUZA, 2005); e de ferramentas de representação de relacionamentos semânticos e conceituais, como os tesauros, para ampliar a gama de informações recuperadas e aferição de contextos, além de outras estratégias derivadas da lingüística e da ciência da informação. Outras metodologias similares implantadas em SRIs permitem a busca de expressões

[15] "Web semântica" é o nome genérico do projeto capitaneado pelo World Wide Web Consortium que pretende embutir inteligência e contexto nos códigos XML utilizados para confecção de páginas web, de modo a melhorar a forma com que programas podem interagir com estas páginas e também possibilitar o uso mais intuitivo pelos usuários.

[16] http://www.ontoweb.org/

regulares, ou mesmo analisam a proximidade da ocorrência de alguns termos, expandindo o conceito de palavra-chave para frases ou outras hierarquias lexicais. ZIVIANI (in BAEZA-YATES & RIBEIRO-NETO, 1999, p. 169-170) aponta SRIs que utilizam a técnica de identificação de grupamentos de substantivos (*noun groups*), ao invés de palavras-chave, como estratégia para seleção de termos de indexação, assumindo que os substantivos costumam carregar a maior parte da semântica de um documento, o que não ocorre com artigos, verbos, adjetivos, advérbios e conectivos. Os grupamentos de substantivos, no escopo dessas propostas, são conjuntos de nomes para os quais a "distância sintática" (medida pelo número de palavras entre dois substantivos) não excede um limite predefinido. O objetivo destas estratégias, de maneira geral, é avançar dos níveis ortográficos ou sintáticos de representação do discurso para os níveis semânticos e pragmáticos, aproximando-se da funcionalidade esperada de algoritmos de Inteligência Artificial.

Conclusão: as novas tendências

Além destas vertentes apontadas, podemos esperar novidades nesta área em diversas outras linhas, como a recuperação multilíngüe, a recuperação por conteúdo em mídias diversas, como imagens, áudio e vídeo e as buscas em linguagem natural. Existem hoje diversas frentes de pesquisa, mais ou menos coordenadas, para tratar do problema da recuperação de informações. Uma real integração demandaria estudos concomitantes em diferentes áreas do conhecimento e campos de pesquisa, como a ciência da informação, a lingüística, a ciência da computação, com a inteligência artificial; a psicologia cognitiva, a comunicação, a sociologia, a antropologia, entre outras. Há seguramente muitas surpresas no horizonte.

Information Retrieval Systems and Search Engines on the web: present and forecast

The intention of the article is to present an overview of the basic characteristics of the information retrieval systems, with a special slant to the search engines on the web. We do present the state-of-art of the information retrieval systems and some of the tendencies for the nearby future. Among these, are the link analysis algorithms and the personalization of the queries, as introduced by Google.

Keywords: *information retrieval systems; search engines; web; internet.*

Referências

ARAÚJO, Vânia M.R.H. *Sistemas de recuperação da informação*: nova abordagem teórico conceitual. 1994. Tese (Doutorado em Ciência da Informação). Universidade Federal do Rio de Janeiro, Rio de Janeiro.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. New York: ACM Press, 1999. 511p.

BUCKLAND, Michel. Information as thing. *Journal of American Society of Information Science*. v.42, n.5, 1991. p. 351-360.

CAÑAS, A. J., LEAKE, D. B., WILSON, D. C.; Managing, Mapping, and Manipulating Conceptual Knowledge. *AAAI Workshop Technical Report WS-99-10: Exploring the Synergies of Knowledge Management & Case-Based Reasoning*, AAAI Press, Menlo Calif. Jul. 1999.

CHOWDHURY, G. *Introduction to modern information retrieval*. London: Library Association Publishing, 1999. 452 p.

FOSKETT, A. C. *The Subject Approach to Information*. 5. ed. Londres: Library Association Publishing, 1997. 119p.

KORFHAGE, Robert *Information Storage and retrieval*. New York: John Wiley & Sons, 1997. 349 p.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. *Ciência da Informação*, Brasília, v. 25, n. 2, 1996. Disponível em: <http://www.ibict.br/cionline/250296/25029605.pdf>. Acesso em: jul. 2004.

LAMPING, J, RAO, R. PIROLLI, P. *A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies*. 1995. Disponível em: <http://www.acm.org/sigchi/chi95/proceedings/papers/jl_bdy.htm>. Acesso em: jul. 2004.

LANCASTER, F. W. Information Retrieval Systems. New York: John Wiley, 1968.
_____. Information Retrieval Systems: characteristics, testing and evaluation. 2nd ed. New York: John Wiley, 1979.

LANCASTER, F. W. e WARNER, A. J. *Information Retrieval Today*. Information Resources Press, 1993.

MOREIRO, José; MARZAL, Miguel Ángel; BELTRÁN, Pilar. *Desarrollo de un Método para la Creación de Mapas Conceptuales*. Anais do ENANCIB, Belo Horizonte, 2003.

SALTON, Gerard e MCGILL, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983. 448 p.

SOUZA, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. 2005. 197 f. Tese (Doutorado em Ciência da Informação) — Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte.