2
Trabalhos Relacionados

Neste capítulo é revisado o trabalho relacionado com o objeto desta tese. Na primeira seção, são apresentados alguns conceitos do campo de Recuperação da Informação. Em seguida, na seção 2.2, é realizada uma revisão sobre a Recuperação da Informação na *Web*. Na seção 2.3, são discutidos os algoritmos de referência na literatura que são baseados apenas em hiperlink. Por último, na seção 2.4, há uma breve apresentação da utilização da estrutura de hiperlinks para melhorar o desempenho de modelos que usam texto âncora para fazer as suas classificações.

PUC-Rio - Certificação Digital Nº 0611950/CA

2.1
Recuperação da Informação

Com o crescimento da informação disponível, independente da forma que se encontra armazenada, surge a necessidade de classificá-la, indexá-la e consultá-la de forma eficiente.

O campo da Recuperação da Informação (RI) foi criado como forma de resposta oportuna a essas necessidades e as pesquisas na área focaram, ao longo dos anos, no desenvolvimento de sistemas eficientes de classificação e indexação da informação. Tais sistemas precisam ser avaliados (ROBERTSON, 2008) e dentro desse contexto nasce o conceito de relevância em RI, cuja definição é extensivamente debatida e estudada.

Como exemplo, podemos citar a existência de páginas que inicialmente são consideradas irrelevantes e, após a inserção de técnicas de recuperação *ad hoc*, passam a ser consideradas relevantes para busca direta de páginas *Web*.

Vários autores classificaram o conceito, alguns em até cinco tipos diferentes, tais como, relevância tópica, algorítmica, cognitiva, situacional e sócio-cognitiva (COSIJN; INGWERSEN, 2000).

Porém, agrupando essas definições, o conceito de relevância pode ser categorizado como *orientado ao usuário* ou *orientado ao sistema*. Esse último preocupa-se com a correspondência ou associação entre o tópico ou assunto de uma consulta e o tópico de um objeto informacional e o primeiro, por sua vez, envolve contextos relacionados à necessidade de informação, o propósito da mesma, como o usuário pretende encontrar a informação, sua experiência prévia no assunto de pesquisa, o estágio em que a busca pela informação se encontra, e assim por diante.

Destarte, dentro da categoria *orientada ao sistema* foi escolhido um conceito de relevância que está vinculado ao presente estudo, qual seja, *Relevância Algorítmica*, que é a relação entre a informação retornada pelo sistema e a consulta efetuada (SARACEVIC, 2007).

Nos experimentos realizados, adotamos a metodologia de RI *ad hoc* tradicional da *TREC*, que trata cada consulta como um documento único e assume que o usuário apenas deseja páginas que contenham informação que satisfaça a sua necessidade. Tal abordagem é muito mais restrita do que a definição de relevância algorítmica, mas é largamente utilizada para avaliação de sistemas RI.

Contextualizada a relevância, passamos para a avaliação dos sistemas RI, particularmente, a aplicável a *Web*. Habitualmente, a avaliação da estrutura de hiperlinks no sistema de RI é feita através de bases de teste compostas por um conjunto estático de páginas, um conjunto de tópicos com informações que se deseja buscar nos conjuntos de páginas e um conjunto de páginas avaliadas como relevantes para cada um dos tópicos de busca. Esse tipo de abordagem se baseia nos experimentos de *Cranfield 1 e 2* (ROBERTSON, 2008; VOORHEES, 2001; CLEVERDON, 1991).

As avaliações baseadas nesse modelo pressupõem que a relevância possa ser traduzida através da similaridade entre as páginas e o tópico de busca. Tal similaridade pode ser concretizada através de julgamentos binários, em que a página apenas é julgada como relevante ou não para um determinado tópico, e julgamentos graduados, em que a página é julgada como irrelevante, pouco relevante, relevante e altamente relevante.

Outros dois pressupostos desse modelo de avaliação são a garantia de que as listas de páginas julgadas relevantes por tópico representam o julgamento dos usuários e que as mesmas estão completas, ou seja, todas as páginas relevantes foram julgadas.

PUC-Rio - Certificação Digital Nº 0611950/CA

Normalmente, esses pressupostos não são totalmente garantidos. Conforme Voorhees (2001), os assessores da *Text REtrieval Conference* (*TREC*) *ad hoc* possuem um grau de concordância em pares sobre o julgamento de relevância em torno de 0,45, em uma escala de concordância total em 1. Em coleções grandes, tal como as utilizadas na *TREC*, é praticamente impossível julgar todas as páginas em relação a cada um dos tópicos a serem investigados.

Uma forma de lidar com esse problema é criar um subconjunto da coleção que é gerado através da contribuição dos resultados mais significativos de diferentes sistemas RI. Esse subconjunto é, então, julgado e tende a ser pequeno comparado ao tamanho da coleção total (VOORHEES, 2001).

Por fim, essa simplificação permite uma avaliação comparativa dos sistemas de RI, ou seja, se um sistema possui um desempenho melhor que o outro em uma grande coleção de teste, o primeiro sistema é considerado melhor que o outro em relação a um tipo específico de tarefa de RI considerando o julgamento de relevância para essa tarefa.

PUC-Rio - Certificação Digital Nº 0611950/CA

2.2
Recuperação da Informação na Web

Os modelos mais comuns de recuperação da informação utilizam apenas o conteúdo textual das páginas para relacioná-las com as consultas ao modelo.

Entretanto, existem outras fontes de informação nas página *Web* que podem melhorar a qualidade desse casamento entre consulta e página, tais como, estrutura lógica e física da página, os seus metadados e hiperlinks.

Na medida em que o foco do presente trabalho está em compreender como as estruturas de hiperlinks podem ajudar a melhorar os modelos de RI, na subseção 2.2.1 é feito um resumo da estrutura da *Web* e, em seguida, na subseção 2.2.2, há um breve resumo das coleções de teste baseadas na *Web* que são utilizadas para avaliar os modelos de RI.

2.2.1
Estrutura da Web

A *WWW* pode ser representada por um grande grafo direcionado, conhecido como *Grafo Web*, cujos vértices correspondem às páginas *Web* e os arcos são os hiperlinks de uma página para outra. O *Grafo Web* descarta todo o

conteúdo das páginas e contém apenas a informação de hiperlink da *Web*. Por consequência, o tamanho, em bytes, do *Grafo Web* é muito inferior ao da *Web*, mas ainda é uma estrutura de dados substancial que torna as tarefas de manipulação e armazenamento desafiadoras.

A estrutura de hiperlink da *Web* se tornou um importante objeto de estudo e muitos autores se dedicaram a essa estrutura de grafo, tais como: (KLEINBERG et al., 1999), (KLEINBERG, 1999), (PAGE et al., 1999) e (BRODER et al., 2000).

Kleinberg (1999) e Broder et al. (2000) concluíram que o grau de entrada e de saída das páginas *Web* atendiam à distribuição da Lei da Potência (*Power Law*). Broder et al. (2000) também estudaram a conectividade do *Grafo Web* e concluíram que existe um grande conjunto de páginas que podem ser alcançadas umas às outras apenas através das estruturas de hiperlink.

Essa Componente Fortemente Conectada (*SCC - Strongly Connected Component*) correspondia a 28% de um conjunto de 200 milhões de páginas. *Broder* também definiu alguns outros tipos de conjuntos de páginas (figura 2.1), quais sejam:

- *IN*: são formados por páginas que alcançam a *SCC* através dos hiperlinks, mas não são alcançadas pela SCC;
- *OUT*: são formados por páginas alcançáveis a partir do *SCC* através dos hiperlinks, mas não conseguem alcançar o *SCC*;
- *TUBES*: são formados por páginas que conectam os conjuntos *IN* e *OUT* sem passar pela *SCC*;
- *TENDRILS*: são formados por páginas que são alcançadas a partir do *IN*, ou alcançam o *OUT* e não passam pela *SCC*;
- e, *DISC*: todos as outras componentes desconectadas.

Kleinberg (1999) atribuiu a noção de importância e relevância ao hiperlink vinculando o mesmo à ideia de julgamento subjetivo do autor da página sobre um determinado assunto.

Essa noção de importância que foi atribuída ao hiperlink foi explorada para melhorar os resultados das buscas. No PageRank (PAGE et al., 1999) utiliza-se a estrutura global do *Grafo Web* enquanto o *HITS* (KLEINBERG, 1999), utiliza uma visão local do grafo.

PUC-Rio - Certificação Digital Nº 0611950/CA

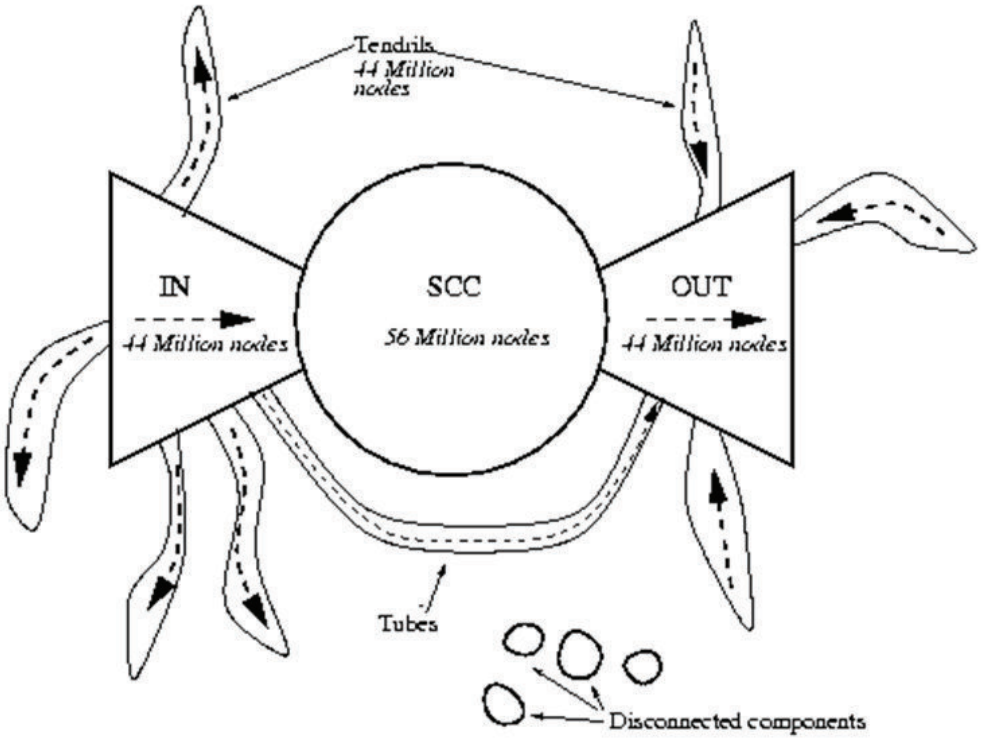

Figura 2.1: Representa a conectividade da *Web* segundo *Broder*. Extraída da *URL http://www9.org/w9cdrom/160/160.html*

Por fim, Chakrabarti et al. (2002) investigaram o grau de distribuição dos conjuntos de páginas com foco em um único tópico de busca. Como resultado, descobriram que a distribuição dos graus dos hiperlinks remete ao da *Web* como um todo.

2.2.2
TREC Web Tracks

Em 1995, com o objetivo de acompanhar a evolução das coleções de documentos, particularmente os da *Web*, a *TREC* criou a *Very Large Collection Track* (*VLC*) para estudar diversos aspectos da RI, tais como, escalabilidade, eficiência e aplicações de paralelismo (CRASWELL; HAWKING, 2004).

Assim, duas coleções foram lançadas, quais sejam:

– VLC de 20gB : lançada em 1997 e composta por uma grande quantidade de dados de jornal e do governo, muitos gigabytes da USENET e uma pequena quantidade de dados *Web*;

PUC-Rio - Certificação Digital Nº 0611950/CA

– VLC2 de 100gB: lançada em 1998 e composta por arquivos truncados da *Web* baixados em Fevereiro de 1997.

Posteriormente, em 1999, as atenções se voltaram para as avaliações que representavam melhor as características da *Web*. Neste mesmo ano, a *Small Web Task* questionou se a informação advinda do hiperlink melhoraria a qualidade das buscas *ad hoc*.

No ano seguinte, um subconjunto da *VLC2* devidamente selecionado para proporcionar maior quantidade de hiperlinks inter-domínios, gerou a *WT10g*, que posteriormente substituiu definitivamente a *VLC2* em 2001. As mesmas técnicas de busca que vinham sendo aplicadas foram repetidas usando a *WT10g* e uma nova modalidade foi inserida, chamada de *Homepage Finding*. O critério de julgamento para esse novo tipo de busca foi bem simples e consistiu em verificar se a página encontrada era da entidade descrita na consulta, a página de entrada do *website*. Como envolveu diretamente a estrutura de hiperlink, nessa nova modalidade os hiperlinks foram considerados importantes.

Novos tipos de busca foram inseridos em 2002, *Named Page Finding* e *Topic Distillation*, e uma nova coleção foi lançada, a *.GOV*. Diferentemente da *WT10g*, que possuía seleções artificiais, a *.GOV* era uma representação real e inalterada de uma parte do domínio da *Web*. A *.GOV* é formada por aproximadamente 1.25 milhões de páginas que foram baixadas do domínio *.gov* em 2002.

A busca *Named Page Finding* é uma simples variação da *Home Page Finding*, em que o resultado desejado é uma única e importante página que não era necessariamente a página de entrada do *website*.

A busca *Topic Distillation* intentava identificar os elementos chave de um tópico. Em essência, uma busca *Topic Distillation* ideal deveria retornar uma lista pequena de elementos chave que se aproximam da ideia do ser humano sobre o tópico. Assim, à luz do conceito de relevância, a definição ideal está no julgamento subjetivo do ser humano.

No ano seguinte, a busca *Topic Distillation* foi simplificada e os elementos chaves foram restringidos aos *websites* representados pelas suas páginas de entrada.

Durante o processo evolutivo da *TREC*, alguns questionamentos surgiram sobre a representatividade dos experimentos em relação à busca na *Web*.

Geralmente, os usuários que buscam informações na *Web* o fazem a

PUC-Rio - Certificação Digital Nº 0611950/CA

partir de termos curtos e buscam páginas específicas e poucos passam da primeira página com os resultados da busca (JANSEN; SPINK, 2006), e tais características não eram observadas nos tipos de busca da *Web Track* em seus primeiros anos.

Em resposta, as formas de busca foram evoluindo e, finalmente, a partir de 2001, os primeiros resultados significativos que levaram em consideração a estrutura de hiperlinks apareceram.

Segundo os resultados divulgados em 2001, os vinte e três melhores resultados na tarefa de *Home Page Finding* em um total de quarenta e três submissões oficiais, ou fez uso do texto da *URL* ou dos hiperlinks ou dos dois (CRASWELL et al., 2001).

A partir desse ponto, diversas propostas com resultados significativos de se explorar a estrutura de hiperlink e seus textos âncoras conjugado ou não com o conteúdo da página surgiram.

Dentre as mais importantes, podemos citar:

- *Craswell* utilizou o texto âncora para melhorar os resultados em buscas do tipo *Homepage Finding* (CRASWELL; HAWKING; ROBERTSON, 2001);
- *Kraaij* utilizou a profundidade da *URL* (*URL Depth*) e o grau de hiperlinks de entrada para definir probabilidades *a priori* para melhorar os resultados em buscas do tipo *Homepage Finding* (KRAAIJ; WESTERVELD; HIEMSTRA, 2002);
- *Craswell* criou um conjunto misto de tópicos de busca provindos da *Homepage Finding*, *Topic Distillation* e *Named Page Finding*, e estudou a evidência da independência desses tópicos através do *Pagerank*, do algoritmo baseado no grau dos hiperlinks de entrada, profundidade da *URL* e a distância em cliques (CRASWELL et al., 2005).

Mesmo com a evolução das coleções até 2004, ainda restava em aberto o questionamento em relação à contribuição dos hiperlinks na melhora dos resultados das buscas *ad hoc*.

Alguns autores justificaram o baixo desempenho nas buscas *ad hoc* com base na baixa densidade dos hiperlinks inter-domínio da *WT10g*. Isso restou provado quando Gurrin e Smeaton (2004) extraíram um subconjunto denso em hiperlinks de inter-domínio da *WT10g* e nesse conjunto reduzido a estrutura de hiperlinks melhorou a precisão nas buscas *ad hoc*.

PUC-Rio - Certificação Digital Nº 0611950/CA

Coleção	**Dados**	**Documentos**	**Tamanho**	**Hiperlinks**	**Ano**
VLC	Misturado	7.492.048	20GB	–	1997
VLC2 (*WT100g*)	*Web* (*.com*)	18.571.671	100GB	–	1997
WT2g	*Web* (*.com*)	247.491	2.1GB	1.166.702	1997
WT10g	*Web* (*.com*)	1.692.096	10GB	8.062.918	1997
.GOV	*Web* (*.gov*)	1.247.753	18GB	11.110.985	2002
.GOV2	*Web* (*.gov*)	25.205.179	400GB	82.711.345	2004
ClueWeb09	*Web*	1.040.809.705	25TB	≈ 12B	2009

Tabela 2.1: Resumo das informações das coleções da TREC

Assim, uma boa coleção *Web* precisa ser grande o suficiente e possuir uma alta densidade de hiperlinks inter-domínios e intra-domínios.

Entre 2004 e 2006, a *TREC* tentou satisfazer os requisitos de tamanho e densidade intra-domínio através da coleção *.GOV2*. Porém, essa coleção é muito diferente do domínio *.com* e possui muito poucos hiperlinks de entrada por página.

Finalmente, em 2010, a *TREC* liberou nova coleção, a chamada *Clueweb09* que consiste de 1.040.809.705 páginas *Web*, em dez línguas diferentes com 25TB de tamanho comprimido, baixado da *Web* entre janeiro e fevereiro de 2009 (LEMUR, 2010). Esta coleção é muito maior do que a *WT10g* e a *.GOV* e possui uma alta densidade de hiperlinks inter-domínio e intra-domínio.

Um resumo de todas as coleções pode ser visto na tabela 2.1.

2.3 Algoritmos baseados em hiperlinks

Algoritmos baseados em hiperlink utilizam apenas a estrutura dos hiperlinks para classificar as páginas *Web*. Nessa seção, são abordados os principais algoritmos baseados em hiperlinks através de uma análise dos seus funcionamentos e apresentados alguns trabalhos relacionados.

PUC-Rio - Certificação Digital Nº 0611950/CA

2.3.1
Hyperlink Induced Topic Search

Jon Kleinberg propôs um modelo baseado em hiperlinks que permite a inferência de autoridade e um conjunto de algoritmos que identifica páginas relevantes para tópicos de busca de caráter geral (KLEINBERG, 1999; KLEINBERG et al., 1999).

Esse modelo é baseado na relação entre páginas que são autoridades sobre um tópico e páginas que interligam essas autoridades (*hubs*). Jon Kleinberg observou um equilíbrio natural entre autoridades e *hubs* num grafo definido pela estrutura de hiperlinks e desenvolveu um algoritmo, conhecido como *HITS*, (*Hyperlink Induced Topic Search*), para identificar, simultaneamente, esses tipos de páginas. O algoritmo opera em um sub-grafo focado da *Web*, construído a partir do resultado de uma máquina de busca baseada somente em texto.

A partir de uma consulta de tópico geral especificada pela cadeia σ, é necessário, para analisar a estrutura de hiperlinks, extrair as páginas que são autoridades e definir qual sub-grafo da *Web* o algoritmo vai utilizar.

Tal sub-grafo (S_σ) deve conjugar três características, quais sejam:

1. Relativamente pequeno;
2. Rico em páginas relevantes;
3. Possuir a maior parte das grandes autoridades.

Para satisfazer às características 1 e 2 supracitadas, basta coletar as t primeiras páginas classificadas por uma máquina de busca baseada somente em texto - empiricamente, são escolhidas em torno de duzentas páginas. Esse grafo inicial (R_σ) geralmente está longe de satisfazer a terceira condição, pois é notório que páginas que são autoridades normalmente não possuem em seu texto muitas repetições da palavra da consulta.

Um exemplo simples consiste na procura de páginas que tratam sobre "*Harvard*". A página *www.harvard.edu*, considerada a maior autoridade sobre o assunto, não possui em seu texto repetidas vezes a palavra "*Harvard*".

Entretanto, R_σ pode ser usado como base para construir o grafo S_σ desejado. Como antes aludido, as páginas autoridades podem não pertencer a R_σ.

PUC-Rio - Certificação Digital Nº 0611950/CA

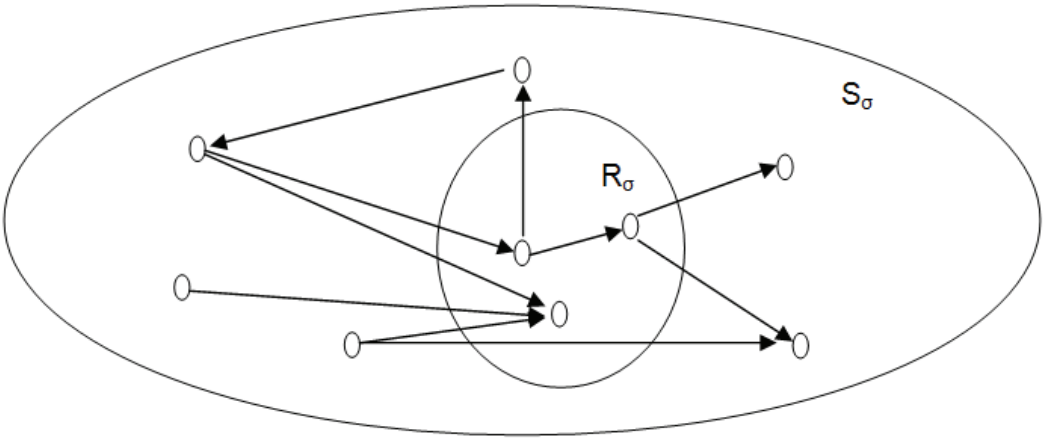

Figura 2.2: Expansão de R_σ

Considerando que essas podem ser apontadas por pelo menos uma página de R_σ, é razoável inferir que expandindo o conjunto R_σ através dos hiperlinks das páginas pertencentes a R_σ, o grafo resultante passe a satisfazer a condição 3 e, por consequência, gere S_σ (figura 2.2).

A expansão de R_σ consiste em:

- acrescer as páginas que são apontadas por cada página pertencente a R_σ;
- e, para cada página pertencente a R_σ, acrescer um número máximo d de páginas as apontam.

Os hiperlinks de navegação são desconsiderados, formando um grafo dirigido apenas com arestas inter-domínios.

A partir da definição do sub-grafo como $G = (V, E)$ temos a constituição de uma coleção V de páginas conectadas por seus hiperlinks, em que os vértices são as páginas e as arestas dirigidas $(i, j) \in E$ são hiperlinks de i para j.

O pseudocódigo da rotina de expansão do sub-grafo pode ser visto no algoritmo 2.1.

Após a construção do sub-grafo, o problema converge para a extração e classificação das autoridades existentes, considerando apenas a estrutura de hiperlinks presente. Analisando a estrutura, as autoridades relevantes presentes possuem não só um alto grau de hiperlinks que as apontam, como também possuem grupos de páginas que as apontam em comum.

Essas páginas são denominadas *hubs* e são responsáveis por vincular as autoridades comuns excluindo as páginas que possuem um alto grau de hiperlinks de chegada e não são relevantes para o assunto (figura 2.3).

PUC-Rio - Certificação Digital Nº 0611950/CA

Algoritmo 2.1: Algoritmo de construção do subgrafo expandido

Entrada:
σ : palavra da consulta
ϵ: máquina de busca baseada em texto
t: tamanho do núcleo inicial
d: número máximo de páginas que apontam para cada páginas do núcleo
R_σ: o conjunto das t mais significantes páginas retornadas por ϵ com σ
Saída: S_{sigma}: conjunto expandido

```
1  início
2      S_σ := R_σ;
3      para cada página i ∈ R_σ faça
4          Criar o conjunto F_i com todas as páginas apontadas pela i;
5          Criar o conjunto F_i* com todas as páginas que apontam para i;
6          S_σ ← S_σ ∪ F_i;
7          se ||F_i*|| < d então
8              S_σ ← S_σ ∪ F_i*;
9          senão
10             count:=0;
11             enquanto count < d faça
12                 Escolher uma pagina p aleatória em F_i* ;
13                 S_σ ← S_σ ∪ {p};
14                 count:=count+1;
15             fim
16         fim
17     fim
18 fim
```

Sendo assim, autoridades e *hubs* exibem uma relação de interdependência: uma boa autoridade será uma página apontada por bons *hubs* e um bom *hub* será uma página que aponta para boas autoridades.

Com base nessa relação, um algoritmo iterativo foi desenvolvido permitindo encontrar os *hubs* e as autoridades. Cada página i possui dois pesos não negativos associados, um para autoridade ($a^{<i>}$) e um para *hub* ($h^{<i>}$).

Desta forma, é natural reescrever a relação de interdependência como: se uma página p aponta para páginas com altos valores de a, então ela deve receber um valor alto para h, e se p é apontada por páginas com valores altos de h, essa deve receber um valor alto para a.

Em face do exposto, dois operadores foram definidos para atualizar os pesos de autoridade (I) e os de *hubs* (O), conforme as equações 2-1 e 2-2, respectivamente.

PUC-Rio - Certificação Digital Nº 0611950/CA

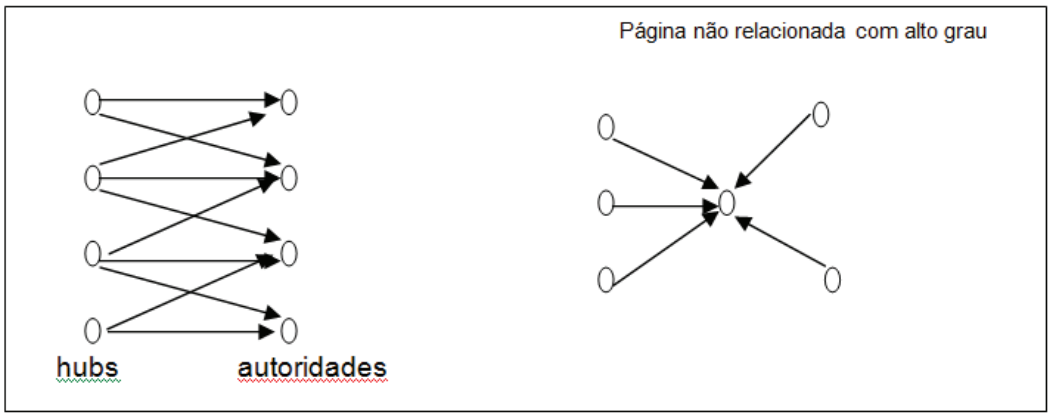

Figura 2.3: *Hubs* e autoridades

$$I : a^{<p>} \leftarrow \sum_{q:(q,p)\in E} h^{<q>} \tag{2-1}$$

$$O : h^{<p>} \leftarrow \sum_{q:(p,q)\in E} a^{<q>} \tag{2-2}$$

Tais operadores representam claramente a relação de interdependência entre *hubs* e autoridades.

Para encontrar o valor de equilíbrio entre os pesos, são aplicados, alternadamente, os operadores I e O até que a estabilidade seja alcançada, ou seja, até que os valores de a e h das páginas se tornem inalterados com a iteração do algoritmo, conforme pseudocódigo em 2.2.

Por fim, basta ordenar as coordenadas do vetor a, retornadas pelo algoritmo 2.2, para expor as páginas de maior autoridade. O mesmo se aplica ao vetor h para as páginas de maior *hub*.

2.3.2 PageRank

O algoritmo *PageRank* foi inventado pelos fundadores do *Google* Larry Page e Sergey Brin em 1998 (PAGE et al., 1999). Em seu núcleo, o *PageRank* é simplesmente uma distribuição estacionária de um caminhar aleatório em um grafo direcionado. Em outras palavras, o valor atribuído pelo *PageRank* a uma página é a probabilidade de, a qualquer momento, um navegante da *Web* qualquer visitar a página. De acordo com Brin e Page, na maior parte das vezes o navegante seguirá através de hiperlinks da página, ou seja, a partir de uma certa página i, o navegante visitará uma página fora da vizinhança de

PUC-Rio - Certificação Digital Nº 0611950/CA

Algoritmo 2.2: Algoritmo de construção do subgrafo expandido

Entrada:
S_{sigma}: conjunto expandido representando uma coleção de n páginas com hiperlinks
k: número de iterações do algoritmo
Saída:
a_k: vetor com os valores das autoridades
h_k: vetor com os valores dos *hubs*

```
1  início
2      Seja z o vetor (1,1,1...1) ∈ R^n;
3      a_0:=z;
4      h_0:=z;
5      para i := 1,2..k faça
6          Aplicar I em (a_{i-1}, h_{i-1}), obtendo a*_i;
7          Aplicar O em (a*_i, h_{i-1}), obtendo h*_i;
8          Normalizar a*_i, obtendo a_i;
9          Normalizar h*_i, obtendo h_i;
10     fim
11 fim
```

PUC-Rio - Certificação Digital Nº 0611950/CA

i.

De forma a compreender melhor o algoritmo *PageRank*, representamos a *Web* como um grafo direcionado $G = (V, E)$, em que V é o conjunto de vértices representando as páginas *Web* e E é o conjunto de arestas direcionadas (i, j) que representam a existência de um hiperlink da página i para j.

O algoritmo atribui um valor de classificação r_i para a página i em função dos valores das classificações das k páginas que apontam para i, ou seja

$$r_i = c \sum_{(k \to i) \in E} \frac{r_k}{o_i} \tag{2-3}$$

onde o_i é o número de hiperlinks que se originam em i e c é um fator de normalização.

A partir dessa definição recursiva, cada página recebe uma fração do valor das classificações das páginas que apontam para a mesma, ponderada pelo número de hiperlinks de saída. A figura 2.4 mostra um exemplo de cálculo dos pesos da página, porém sem a normalização para melhor visualização.

A forma simplificada representada pela equação 2-3 não resolve um pequeno problema que pode acontecer no grafo *Web*. Suponhamos que existam duas páginas que apontam uma para outra e não apontam para mais nenhuma outra e que alguma página aponta para uma delas.

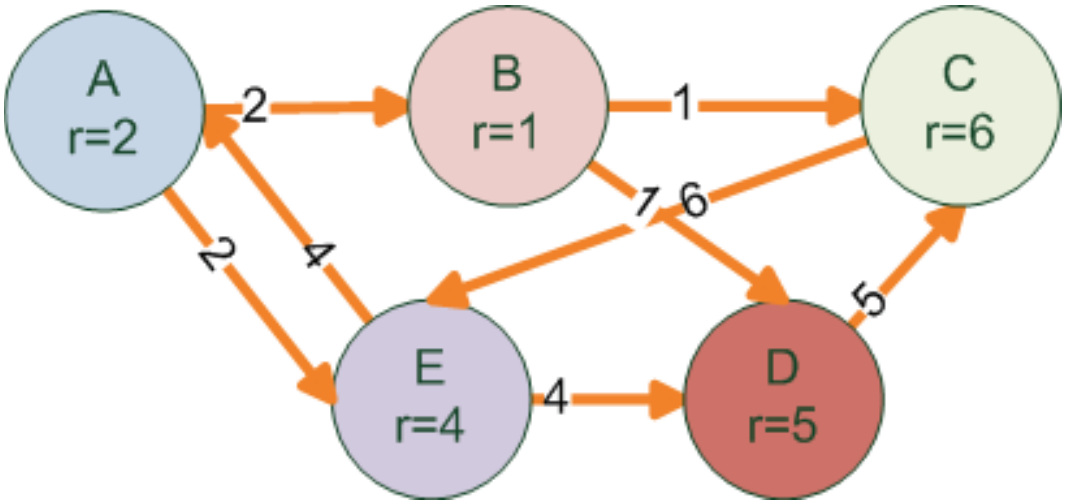

Figura 2.4: Exemplo do cálculo dos pesos do *PageRank* sem a normalização

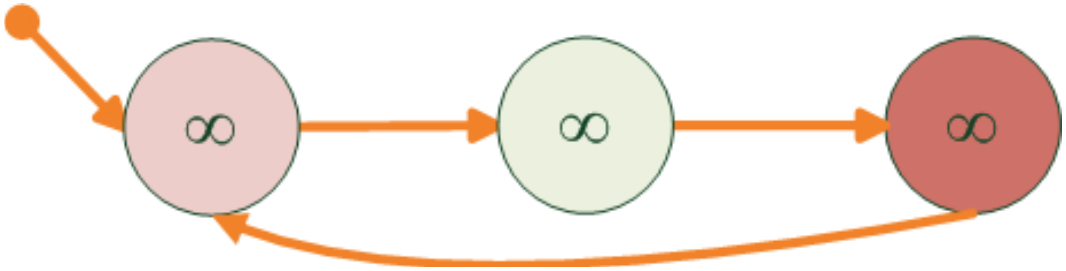

Figura 2.5: Exemplo do problema de *rank sink* no cálculo dos pesos do *PageRank*

Durante a iteração, o ciclo gerado por essas páginas vai acumular valores de classificação e não vai distribuir para as demais páginas do grafo. Esse problema foi chamado de *rank sink* (figura 2.5). Para resolvê-lo, foi inserido no cálculo recursivo um valor S_i denotado *rank source* da página i (equação 2-4).

$$r_i = cS_i + c \sum_{(k \to i) \in E} \frac{r_k}{o_i} \tag{2-4}$$

Reescrevendo em forma matricial, seja A a matriz de adjacência das páginas *Web* e seus valores compostos por $A_{i,j} = 1/o_i$, se existe um hiperlink entre i e j, e zero caso contrário. Se r é o vetor de todas as páginas *Web*, então $r = cAr$ e r é o autovetor de A com autovalor c no caso da equação 2-3.

No caso da equação 2-4, na forma matricial tem-se $r = c(Ar + S)$. Uma vez que $\|r\| = 1$, então $r = c(A + S \times 1)r$ e r é o autovetor de $(A + S \times 1)$ com autovalor igual a c.

Avaliando essa nova abordagem, verificamos que S resolve o problema das páginas *sink*, porém pode ser usado para ajustar a classificação de

PUC-Rio - Certificação Digital Nº 0611950/CA

qualquer página no grafo. Para isso, basta atribuir valores diferentes para as componentes de S.

Outro ponto importante é o fato de as páginas possuírem um valor de classificação independentemente do tipo de consulta que é feita no grafo. Quando a consulta é efetuada, as páginas relacionadas com o tópico são selecionadas e escalonadas de acordo com os seus valores r_i, ou seja, a classificação das páginas leva em consideração o grafo *Web* todo.

Finalmente, a convergência do algoritmo *PageRank* é garantida pois o grafo da *Web* é de expansão conforme descrito em (PAGE et al., 1999).

2.3.3
Stochastic Approach for Link-Structure Analysis

O algoritmo *Stochastic Approach for Link-Structure Analysis* (*SALSA*) combina ideias do algoritmo *HITS* e do *PageRank* e foi proposto por Lempel e Moran (2001). Assim como no caso do *HITS*, visualiza o grafo representativo da *Web* como um grafo bipartido em que *hubs* apontam para autoridades.

O algoritmo *SALSA* efetua um caminhar aleatório no grafo bipartido, alternando entre os lados dos *hubs* e das autoridades. O caminhar se inicia em um nó autoridade escolhido aleatoriamente, com probabilidade uniforme, prossegue alternando um passo a frente e um passo atrás, passando de uma parte do grafo para outra. Quando está no lado da autoridade, o algoritmo seleciona um dos hiperlinks de chegada aleatoriamente e caminha para o lado dos *hubs* e vice-versa. O valor das autoridades e dos *hubs* é calculado iterativamente até a convergência.

Na definição formal do *SALSA* temos um grafo bipartido $B = (V_h, V_a, E)$ gerado a partir do grafo expandido $G = (V, E)$ de R_σ da seção 2.3.1. Então, o grafo B é gerado a partir das seguintes definições:

- O lado dos *hubs* por $V_h = \{s_h | s \in G \; e \; grau - saida(s) > 0\}$;
- O lado das autoridades por $V_a = \{s_a | s \in G \; e \; grau - entrada(s) > 0\}$;
- e, $E = (s_h, r_a) | s \to r \in G$.

onde *grau-saída(s)* retorna o número de hiperlinks que se originam em s, *grau-entrada(s)* retorna o número de hiperlinks que se apontam em s e cada hiperlink $s \to r$ é representado por uma aresta não direcionada que conecta s_h e r_h.

PUC-Rio - Certificação Digital Nº 0611950/CA

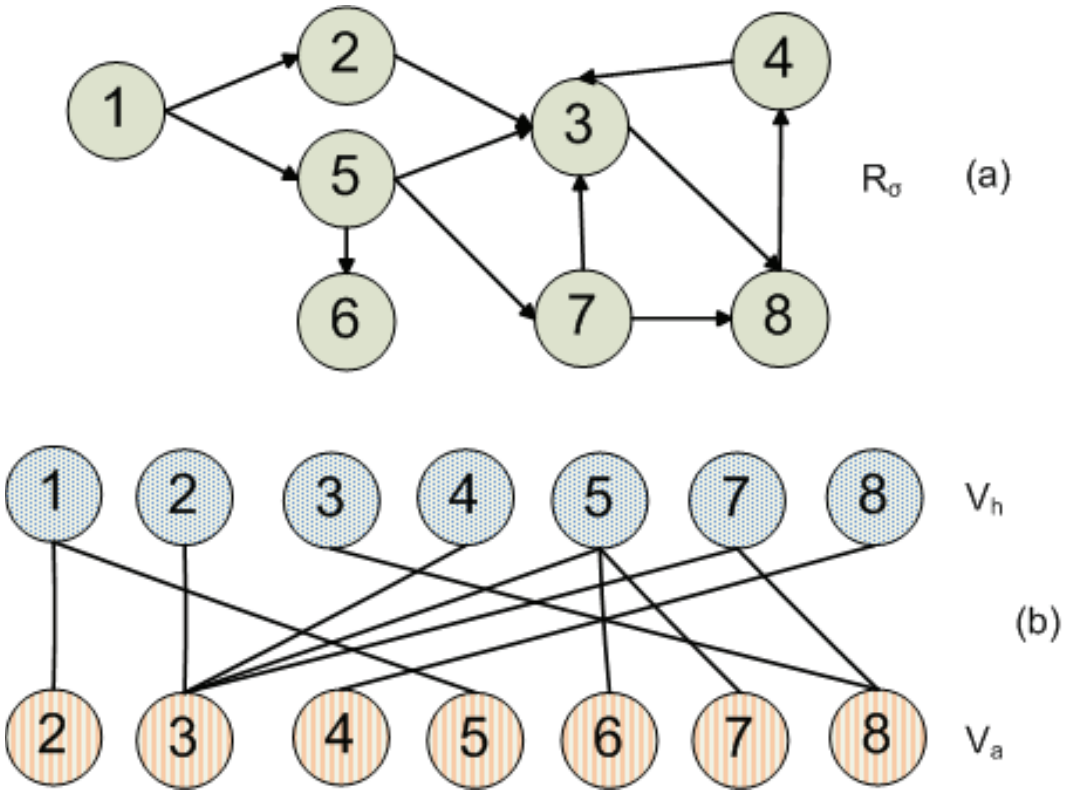

Figura 2.6: Em (a) temos o grafo estendido para a consulta σ e em (b) o grafo bipartido.

De forma a elucidar a montagem do grafo, a figura 2.6 mostra um exemplo simples e claro do processo.

Nesse grafo bipartido é feita a caminhada aleatória descrita anteriormente.

Desta forma, a caminhada é modelada por duas cadeias de *Markov*, a cadeia que visita o lado da autoridade de B e a cadeia que visita o lado dos *hubs*.

Podemos definir duas matrizes estocásticas, que são as matrizes de transição das duas cadeias de *Markov*, quais sejam:

– Matriz *Hub* H^- definida como

$$h_{ij}^- = \sum_{k|(i_h,k_a),(j_h,k_a)\in B} \frac{1}{deg(i_h)deg(k_a)} \tag{2-5}$$

– Matriz de Autoridade A^- definida como

$$a_{ij}^- = \sum_{k|(k_h,i_a),(k_h,j_a)\in B} \frac{1}{deg(i_a)deg(k_h)} \tag{2-6}$$

Finalmente, como a probabilidade de transição ocorre em dois passos, a constituição da matriz estocástica A^-, por exemplo, quando um dos seus

PUC-Rio - Certificação Digital Nº 0611950/CA

elementos for positivo, nos informa que uma autoridade alcança outra através de pelo menos um *hub*.

Então, se $a_{ij}^{-} > 0$ nos diz que existe uma página k que aponta para i e j e que para alcançar j a partir de i, no primeiro passo devemos retrair para k e depois ir para j.

2.4
Texto âncora e o Grafo Web

O texto âncora é uma parte do texto da página *Web* que está vinculado ao hiperlink e que geralmente fornece uma descrição sobre a página por ele indicada.

Normalmente, os textos âncoras são utilizados para melhorar a representação textual da página e impactam nos modelos de RI baseados em texto. Isso porque as representações das páginas amplificadas pelos textos âncora agregam as informações externas sobre o conteúdo da página.

Outro ponto importante consiste em utilizar o texto âncora para filtrar os hiperlinks vinculados à consulta. Mesmo que os hiperlinks sejam gerados independente da consulta, uma possibilidade é basear a pontuação dada à página apontada pelo hiperlink pela frequência de repetição dos termos da consulta na mesma. Desta forma, podemos separar os hiperlinks que se relacionam com o conteúdo da consulta e por fim utilizar os hiperlinks selecionados.

Uma abordagem interessante foi a realizada por Metzler et al. (2009), que propuseram uma metodologia para agregar o texto âncora e propagá-lo através do grafo *Web* e, assim, amplificar a qualidade da classificação das páginas. Na medida em que os textos âncora são passados para as páginas que os hiperlinks apontam e que resta demonstrado em pesquisas anteriores que a distribuição de hiperlinks de entrada de uma página segue a Lei da Potência (BRODER et al., 2000), apenas um pequeno grupo de páginas possui uma grande quantidade de texto âncora vinculado (problema do texto âncora esparso). Assim, a proposta consiste em propagar o texto âncora através do grafo *Web*, de forma que as páginas com pouco ou nenhum texto âncora agregado possam receber dos seus ancestrais mais distantes tal informação.

Finalmente, Eiron e McCurley (2003) mostraram que os textos âncora se comportam de forma parecida com a forma que os usuários buscam informação

PUC-Rio - Certificação Digital Nº 0611950/CA

na *Web*, pois normalmente os rótulos utilizados para descrever os hiperlinks são bem parecidos com os textos utilizados para as pesquisas na *Web*.

PUC-Rio - Certificação Digital Nº 0611950/CA