Incluindo Abordagens de Recuperação de Informação em Serviços de Criação de Hiperligações

Alessandra Alaniz Macedo, José Antonio Camacho-Guerrero, Maria da Graça Pimentel

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - São Carlos/SP - Brazil
{ale,jcamacho,mgp}@icmc.usp.br

Abstract

Especially due to the increasing popularity of the Web, more and more hyperdocuments have been created. A major task in creating hypertext document is link generation. Therefore, at least some support for link generation is highly desirable. We have created linking services that create semantic hyperlinks automatically among Web information contextualized. Using those services, users are able to exploit relationships among Web documents without having to suffer all cognitive overhead implied by manual authorship of documents. However, those services did not make any preconceived assumptions about the information needs of their users. This paper presents an infrastructure for generating links by considering relevance feedback given by users. Relevance feedback information is an information retrieval method required by the proposed infrastructure during the presentation of semantic links identified. At that time, users indicate which of returned semantic links are useful. The original context is automatically reformulated upon those relevance judgments and the new "feedback context" is then compared to the collection of documents, returning an improved set of documents to the user. That process can continue until the users considered relevant all presented links. To demonstrate the utility of our infrastructure, we have run an experiment in which latent semantic links extracted from the Sport section of two online versions of Brazilian newspapers were evaluated via relevance feedback information by users.

Resumo

Especialmente devido ao crescimento da popularidade da Web, mais e mais hiperdocumentos são criados. Um grande desafio na criação desses documentos é a criação de ligações. Portanto, ferramentas de suporte à criação de ligações são altamente desejadas. Os autores deste artigo têm trabalho no desenvolvimento de serviços de criação automática de ligações semânticas entre informações Web contextualizadas. Utilizando esses serviços, usuários podem explorar relacionamentos entre informações sem ter que sofrer sobrecarga cognitiva ao definí-los manualmente. Contudo, esses serviços não consideram as necessidades de seus usuários. Este artigo propõe uma infra-estrutura para geração de ligações considerando *relevance feedback* fornecido por usuários. *Relevance feedback* é um método de Recuperação de Informação solicitado pela infra-estrutura proposta durante a apresentação das ligações semânticas identificadas. Nesse momento, usuários devem indicar quais das ligações retornadas são relevantes. O contexto original é automaticamente reformulado de acordo com as ligações julgadas como relevantes, e o novo *feedback context* é comparado aos documentos, retornando um conjunto aprimorado de ligações. Esse processo pode ser repetido até os usuários considerarem relevantes todas as ligações apresentadas. Para demonstrar a utilidade da infra-estrutura proposta, foi realizado um experimento com ligações semânticas extraídas de dois jornais *online* brasileiros, particularmente das seções de esporte. As ligações definidas foram avaliadas através de interações com usuários.

Keywords: Web, Recuperação de Informação, Hipermidia, Método de *Relevance Feedback*

1 Introdução

Com o crescimento da disponibilidade de informações na Web, torna-se interessante identificar relacionamentos entre documentos Web disponíveis em repositórios distintos e, através desses relacionamentos, gerar hiperligações que facilitam seu acesso. Pesquisadores de diferentes áreas investigam métodos para apoiar a identificação de relacionamentos entre informações de coleções de documentos. Uma dessas áreas é a de Recuperação de Informação (*Information Retrieval* - IR) que investiga formas de representação, armazenamento, interrelacionamento e acesso de informação na qual o usuário está interessado [2].

Um dos primeiros sistemas a combinar abordagens de Recuperação de Informação na manipulação de hiperdocumentos foi o Intermedia, o qual definiu mecanismos de recuperação de informação sobre um modelo hipertexto implícito [18]. Usuários podiam localizar seus documentos utilizando uma interface de consulta a qual fornecia uma lista dos documentos encontrados. Qualquer documento nessa lista poderia ser navegado como em um hipertexto. A inovação desse sistema adveio do fato que o Intermedia forneceu aos usuários o rompimento das fronteiras estruturais e de localização de um documento.

Em 1993, utilizando fundamentos da área de Recuperação de Informação e de Hipertexto, Agosti e Crestani apresentaram uma metodologia para a criação automática de hiperdocumentos [1]. Além de Agosti e Crestani, outros pesquisadores têm realizado estudos e experimentos na tentativa de otimizar a recuperação de informação e automatizá-las para a criação de ligações hipertextuais [6] [8] [9] [19]. Características dos próprios hiperdocumentos também têm sido utilizadas para análise [21] e classificação de documentos recuperados a partir de sistemas clássicos de IR ou da Web [12] [22]. No contexto da Web, artifícios de busca [15], de indexação [14], de navegação [3] e de classificação [5] têm sido criados e estendidos.

Em trabalhos prévios, os autores deste artigo criaram serviços que, explorando abordagens de IR, identificam e criam automaticamente ligações entre repositórios Web com informações homogêneas.[1]

A técnica IR de Análise da Semântica Latente (*Latent Semantic Analysis* – LSA) pode ser utilizada para a extração das estruturas semânticas de coleções de documentos, contornando assim, problemas relacionados a polissemia[2] e sinonímia[3] [7].

Em uma abordagem de criação automática de ligações semânticas entre repositórios Web foi explorada a técnica de LSA [16]. Nesse trabalho, foi definida uma infra-estrutura para a geração automática de ligações baseadas em estruturas semânticas extraídas de repositórios homogêneos cujas informações eram provenientes de dois repositórios complementares de informações associadas a um mesmo curso de graduação.

Na tentativa de estender as contribuições de LSA com as de Hipermídia Aberta sobre repositórios Web, foi proposta uma infra-estrutura aberta para a manipulação das ligações semânticas via interface Web independente e para o seu armazenamento em bases de dados independentes, denominadas bibliotecas de ligações (*linkbases*) [17]. Hipermídia Aberta, especificamente a definição de bibliotecas de ligações e de APIs (*Application Programming Interfaces*), permitiu a utilização de uma abordagem geral para criar ligações sobre qualquer repositório Web pois, em sistemas hipermídia abertos, as ligações hipertextuais são manipuladas em bibliotecas de ligações que são repositórios independentes. Esse tipo de abordagem flexibiliza a criação de hiperdocumentos, sendo que funcionalidades hipermídia podem ser incorporadas a qualquer documento sem a necessidade de mudança de formato do documento ou inclusão de especificações de ligações [10]. A infra-estrutura aberta proposta foi experimentada sobre informações homogêneas de jornais *online* de um mesmo país, em um mesmo dia e em uma única seção.

Através da definição dos serviços de criação automática de ligações semânticas suportados pelas infra-estruturas citadas, usuários são capazes de obter páginas Web com âncoras (títulos das páginas Web relacionadas) para ligações criadas entre repositórios Web especificados pelos usuários. Essas páginas-índice de ligações são criadas sem sobrecarregar cognitivamente os autores Web, que para criá-las, teriam que navegar através de todo o espaço de informação a ser interrelacionado e definir nós e ligações que comporiam essas páginas. Porém, esses serviços de criação automática de ligações têm oferecido grandes variações de precisão na definição dos relacionamentos.

Relevance feedback é um método de IR utilizado para aprimorar o desempenho de sistemas de IR [2]. Explorando esse método, consultas a sistemas de IR podem ser seletivamente modificadas na tentativa de recuperar documentos mais relevantes da coleção. A consulta pode ser modificada através de ajustes de pesos de termos da consulta, da inclusão de novos termos ou da combinação das duas abordagens. Todo o processo de ajustes ou inclusão de termos pode ser repetido, através de múltiplas iterações, até o usuário considerar satisfatório os resultados retornados. As iterações do método de *relevance feedback* podem ser efetuadas através de interação com usuários ou através da utilização automática de propriedades dos documentos recuperados.

Neste artigo é proposta uma infra-estrutura para serviços de criação automática de ligações semânticas que permite a participação de usuários, após a geração das ligações, através da inclusão de conjuntos de ligações consideradas relevantes pelos usuários. A participação de usuários é apoiada pelo método de *Relevance Feedback*. Para demonstrar a utilidade da infra-estrutura proposta, é apresentado um serviço Web de criação de ligações semânticas que permite a interação com usuários. Utilizando esse serviço, foi realizado um experimento com usuários que fornecem *feedback* para a geração de ligações referentes à seção de esportes de dois jornais *online* brasileiros.

[1] Por informações homogêneas entende-se informações baseadas em um único contexto como, por exemplo, informações relativas a páginas Web com o mesmo material de curso didático ministrado no mesmo período de tempo de maneiras complementares

[2] Palavras com a mesma grafia mas com significados distintos.

[3] Emprego freqüente de sinônimos ou de vocábulos que, embora não sejam sinônimos, têm significados muito próximos.

As outras seções desse artigo são assim organizadas: na Seção 2 são descritos (a) os conceitos da técnica de LSA, a qual é utilizada para modelagem dos repositórios a serem relacionados e para extração da semântica implícita entre documentos, e (b) o método *relevance feedback*, que foi utilizado para permitir a participação de usuários durante a definição de hiperligações. Na Seção 3 são apresentadas infra-estruturas e serviços relacionados à criação automática de ligações semânticas entre repositórios Web com base na técnica de LSA. Na Seção 4 são apresentados a infra-estrutura proposta, o serviço Web e o algoritmo que implementa mecanismos para permitir a utilização do método de *relevance feedback* durante a definição de ligações entre os repositórios Web com informações textuais. Na Seção 5 são apresentadas as conclusões sobre o trabalho apresentado e discutidos trabalhos futuros.

2 Abordagens de Recuperação de Informação

Algumas abordagens da área de Recuperação de Informação têm sido utilizadas nas infra-estruturas e serviços de criação de hiperligações propostos pelos autores deste artigo. Um dos mais populares modelos de recuperação de informação é o modelo vetorial e a técnica de LSA é uma extensão proposta para esse modelo com o intuito de aumentar a sua efetividade. Outra abordagem proposta para aumentar o desempenho de modelos de recuperação de informação, inclusive do modelo vetorial, é o método de *relevance feedback*.

A técnica de LSA tem sido utilizada para a modelagem da informação a ser semanticamente interrelacionada e o método de *relevance feedback* é utilizado na infra-estrutura proposta para permitir a interação de usuários no processo de geração de hiperligações semânticas.

2.1 A Técnica de LSA

A técnica de Análise da Semântica Latente (*Latent Semantic Analysis* – LSA) oferece mecanismos para a representação de informações textuais em estruturas semânticas, as quais podem ser utilizadas para a recuperação de informações e para a navegação [7]. A técnica de LSA tem como objetivo reduzir os problemas de comparações léxicas de termos ao considerar uma estrutura semântica latente implícita pela variabilidade das palavras. As descrições dos documentos, dos termos e das consultas são representadas no modelo vetorial estendido, subjacente à técnica. Técnicas algébricas e estatísticas são utilizadas para manipular a estrutura semântica latente implícita aos documentos. O modelo algébrico de Decomposição de Valores Singulares (*Singular Value Decomposition* – SVD) é utilizado para a representação da estrutura semântica manipulada pela técnica de LSA.

2.1.1 Decomposição de Valores Singulares

O modelo algébrico de SVD representa termos, documentos e consultas como vetores de dimensão reduzida. O co-seno dos pontos desse espaço vetorial indica a similaridade dos objetos representados. Segundo Furnas, o SVD é um modelo com representação explícita, ajustável e passível de processamento computacional [7]. No contexto de recuperação de informação, para utilizar o SVD deve-se, primeiramente, definir e construir uma matriz X cujas células representam pesos dos termos nos documentos da coleção de documentos a ser manipulada. De acordo com o SVD, essa matriz X é retangular e expressa pela Equação 1.

$$X = T_m S_m D'_m \tag{1}$$

A Equação 1 é denominada de Equação de Decomposição de Valores Singulares de X, e as matrizes T, D e S são as matrizes componentes com dimensão *m=min(número de termos, número de documentos)*. A matriz T_m é composta pelos vetores singulares (autovetores) esquerdos de X; a matriz D_m é composta pelos vetores singulares direitos de X e a matriz S_m é a matriz-diagonal de valores singulares (autovalores). Na Equação 1, D'_m é a transposta de D_m. Normalmente, a matriz X e conseqüentemente as matrizes T_m, S_m e D_m, são de dimensões grandes e, por isso, custosas de serem manipuladas. Analisando a matriz S_m, observa-se que possui células com valores próximos a zero que, segundo a teoria de SVD, podem ser desconsideradas para conservar apenas os componentes linearmente independentes de X. Conseqüentemente, deve-se eliminar os respectivos termos em T_m e D_m. Dessa maneira, obtém-se um modelo reduzido com dimensão k denominado $\hat{X}$. A partir daí, a matriz reduzida $\hat{X}$ representa a estrutura semântica, a qual é uma aproximação da informação sobre similaridade entre os documentos.

Consultas. Na técnica de LSA, uma consulta deve ser considerada como mais um documento (pseudo-documento) da coleção de documentos indexados, isto é, mais um vetor-coluna da matriz $\hat{X}$. Assim, o vetor de consulta é considerado aproximado, representado por $\hat{V}_q$, e seus valores são calculados a partir dos valores da matriz T_k componente da matriz reduzida $\hat{X}$, com T'_k sendo a transposta de T_k, conforme a Equação 2.

$$\hat{V_q} = T_k T_k' V_q \tag{2}$$

O grau de similaridade entre $\hat{V_q}$ e os documentos pode ser calculado através do co-seno de $\hat{V_q}$ e cada vetor-coluna da matriz $\hat{X}$.

Interpretação Geométrica. Interpretar geometricamente as matrizes componentes de SVD auxilia na construção da relação do modelo de SVD para a definição de similaridades semânticas [7]. Ao considerar os vetores das matrizes T_k e D_k como coordenadas do espaço com dimensão "k" e a matriz S_k como um prolongamento dos eixos ortogonais desse espaço, pode-se indicar uma estrutura de similaridade de alguns pontos em relação a essas coordenadas. O modelo vetorial, originalmente, faz comparações léxicas para o cálculo de similaridade entre dois documentos através do co-seno dos vetores que representam os documentos, porém o resultado é simplesmente de *matching* de termos. Em LSA, usando modelo vetorial estendido, calcular essa similaridade entre todos os pares de documentos da matriz reduzida é equivalente a realizar a multiplicação da matriz ($\hat{X}$) pela sua transposta ($\hat{X}'$). De acordo com a decomposição de SVD, esse cálculo é algebricamente equivalente à Equação 3.

$$\hat{X}'\hat{X} = (D_k S_k)(D_k S_k)' \tag{3}$$

Assim, a comparação de um documento i com um documento j é realizada através do co-seno das linhas i e j da matriz $D_k S_k$. De forma análoga, tem-se a Equação 4. Então, a comparação de um termo i com um termo j é realizada através do co-seno das linhas i e j da matriz $T_k S_k$.

$$\hat{X}\hat{X}' = (T_k S_k)(T_k S_k)' \tag{4}$$

Finalmente, a comparação de um termo i e de um documento j é realizada através da Equação 5.

$$\hat{X} = (T_k S_k^{1/2})(D_k S_k^{1/2}) \tag{5}$$

Interpretação Gráfica. As operações definidas em (3), (4) e (5) podem ser apresentadas de forma gráfica. Se os eixos do espaço tiverem escalas com base nos termos da matriz-diagonal $S_k^{\frac{1}{2}}$, o co-seno entre os pontos de termos e os de documentos fornece comparações interessantes. Para completar a visualização da estrutura de LSA, o vetor de consulta deve também ser apresentado graficamente. A consulta, considerada um pseudo-documento, deve ser apresentada no modelo SVD como o co-seno de seus termos com os outros pontos dos documentos. A partir do novo documento $\hat{V_q}$ da matriz $\hat{X}$, uma nova linha D_k na matriz D pode ser calculada conforme mostra a Equação 6.

$$D_q = \hat{V_q}' T S^{-1} \tag{6}$$

Apesar de um dos grandes problemas da técnica de LSA ser o custo computacional despendido em cálculos, o resultado obtido para aproximação semântica é significativo. A maior contribuição da técnica de LSA é que consultas e documentos não precisam possuir termos em comum para serem considerados semelhantes. Caso o vetor de consulta e um documento estejam próximos no espaço geométrico semântico, esse documento é considerado similar à consulta.

2.2 *Relevance Feedback*

Relevance feedback é um processo através do qual consultas podem ser seletivamente modificadas na tentativa de recuperar documentos mais relevantes de uma coleção de documentos [2]. A consulta pode ser modificada através de ajustes de pesos de termos da consulta, da inclusão de novos termos ou da combinação das duas abordagens. Essas abordagens de modificação da consulta original podem ser efetuadas via interação com usuários ou através da utilização automática de propriedades dos documentos recuperados.

O cálculo de *relevance feedback* de Rocchio [20], apresentado na Equação 7, inclui à consulta inicial, após análise de relevância realizada por usuários, um conjunto de documentos relevantes. Os pesos dos vetores que modelam os documentos considerados relevantes são somados ($\overrightarrow{d}_j$) e os vetores são normalizados em um novo vetor ($|D_r|$) dividindo-se o vetor-soma de documentos relevantes pelo número de documentos relevantes. Um processo similar pode ser realizado para os documentos considerados não-relevantes. Os vetores resultantes podem ser modificados através da multiplicação de valores de ajuste (α, β e γ). Na fórmula original de Rocchio, o valor de α deve ser considerado "1". Finalmente, o vetor da consulta ($\overrightarrow{q}$) original deve ser modificado incluindo-se o novo vetor de documentos relevantes e subtraindo-se o novo vetor de documentos não-relevantes. Todo esse processo pode ser repetido através de múltiplas iterações até o usuário considerar satisfatório os resultados retornados.

$$\overrightarrow{q}_m = \alpha \overrightarrow{q} + \frac{\beta}{|D_r|} \sum_{\forall \overrightarrow{d}_j \in D_r} \overrightarrow{d}_j - \frac{\gamma}{|D_{n-r}|} \sum_{\forall \overrightarrow{d}_j \in D_{n-r}} \overrightarrow{d}_j \quad (7)$$

O trabalho de Ide [13], baseado no trabalho de Rocchio, propõe a eliminação dos vetores normalizados ($|D_r|$ e $|D_{n-r}|$) como apresentado na Equação 8. Ide considera $\alpha = \beta = \gamma = 1$. Ide também propõe a diminuição do valor de ajuste de γ, considerando apenas os documentos não-relevantes mais expressivos como apresentado na Equação 9. Algumas vezes, γ pode ser considerado zero ocasionando em uma estratégia de *feedback* totalmente positiva, isto é, baseado apenas no conjunto de documentos relevantes.

$$\overrightarrow{q}_m = \alpha \overrightarrow{q} + \beta \sum_{\forall \overrightarrow{d}_j \in D_r} \overrightarrow{d}_j - \gamma \sum_{\forall \overrightarrow{d}_j \in D_{n-r}} \overrightarrow{d}_j \quad (8)$$

$$\overrightarrow{q}_m = \alpha \overrightarrow{q} + \beta \sum_{\forall \overrightarrow{d}_j \in D_r} \overrightarrow{d}_j - \gamma \, max_{non-relevant}(\overrightarrow{d}_j) \quad (9)$$

Segundo Baeza-Yates e Ribeiro-Neto, as principais vantagens das técnicas de *relevance feedback*, apresentadas nas Equações 7, 8 e 9, são a simplicidade e os bons resultados [2]. A simplicidade está relacionada ao fato de que os pesos dos termos modificados são calculados diretamente a partir do conjunto de documentos recuperados. Já os bons resultados estão relacionados ao fato de que o vetor modificado de consulta ($\overrightarrow{q}_m$) procura refletir a consulta semântica pretendida. Porém, a grande desvantagem da utilização do método de *relevance feedback* é a dependência do julgamento de relevância por usuários. Uma solução que tem sido adotada para esse problema é a utilização automática de propriedades de documentos como informação de *feedback*.

Harman realizou experimentos comparando o desempenho do método de *relevance feedback* para informações modeladas através do uso de vetores (modelo vetorial) e para informações representadas de acordo com propriedades do conjunto de documentos manipulado (modelo probabilístico) [11]. Harman concluiu que técnicas de *relevance feedback* são mais efetivas para informações representadas através do modelo vetorial como, por exemplo, da técnica de LSA.

3 Serviços de Criação Automática de Ligações Semânticas

Com o intuito de explorar a homogeneidade natural entre repositórios Web e de evitar problemas de dependência de vocabulário associados a comparações léxicas quando se deseja relacionar documentos, foi utilizada a técnica de LSA para a geração automática de ligações semânticas entre repositórios Web [16]. Para reutilizar o processo implícito ao serviço semântico criado foi definida a infra-estrutura descrita a seguir.

Infra-estrutura. A Figura 1 apresenta a infra-estrutura proposta para o serviço de criação automática de ligações semânticas. O nível de estruturação (*Structural Level*) define os relacionamentos semânticos entre os repositórios Web utilizando a técnica de LSA [7]. Nesse nível, a infra-estrutura manipula componentes de conteúdo e de estrutura como nós e ligações semânticas, respectivamente. O processamento interno ao nível de estruturação é dividido nas seguintes fases:

1. O módulo *Indexing* indexa cada repositório Web especificado. A implementação desse módulo incluiu uma adaptação da ferramenta de licença pública mnoGoSearch[4] para a coleta e indexação de repositórios Web. Os ajustes realizados no mnoGoSearch foram, por exemplo, inclusões de palavras de contexto dos repositórios manipulados no dicionários de *stopwords* e modificações no seu código de indexação a fim de permitir a extração de palavras a partir de páginas Web ativadas via JavaScript. Os repositórios Web são indexados separadamente.

2. O módulo *Generate Terms by Documents Matrix* gera, segundo a técnica de LSA, a partir dos índices dos repositórios, a matriz X de termos por documentos. Os elementos dessa matriz são preenchidos com base na freqüência de ocorrência dos termos que representam cada documento. Nesta fase, que aplica o modelo vetorial (geração da matriz X) ao espaço de informação (repositórios indexados) a ser relacionado, perde-se a estrutura inicial do documento.

3. O módulo *Compute SVD* executa o algoritmo do modelo de SVD; dados a matriz X e o valor de "k" para redução, este módulo gera as matrizes componentes T, S e D e realiza a redução de dimensão dessas matrizes de acordo com "k".

[4] http://www.mnogosearch.ru

4. O módulo *Compute Similarity* utiliza um limiar de similaridade (99%) de modo a filtrar e assim gerar a matriz semântica com relacionamentos considerados pertinentes. Esses relacionamentos podem ser utilizados para definir as ligações entre os repositórios manipulados. Neste módulo são calculados os co-senos entre os documentos dos repositórios manipulados.

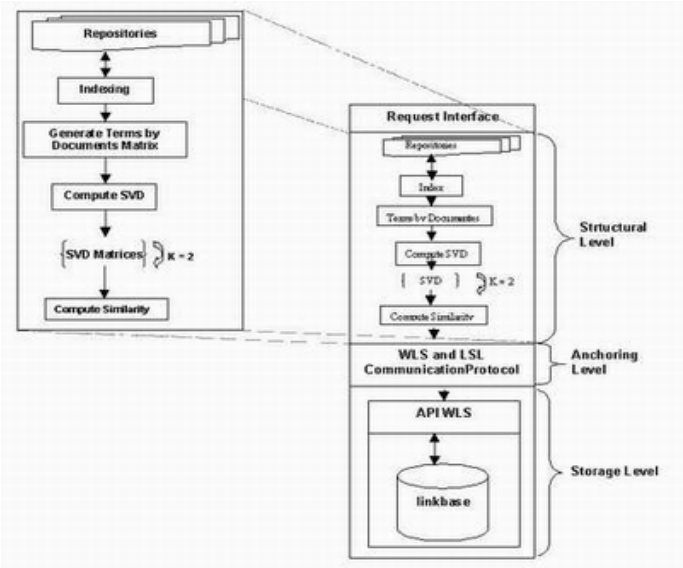

Figura 1: Infra-estrutura para representação da geração automática de ligações semânticas abertas entre repositórios Web [17].

O nível de armazenamento (*Storage Level*) da infra-estrutura proposta para o serviço de criação automática de ligações semântica é composto por um serviço aberto de ligações (*Web Linking Service* – WLS) [4]. O WLS é utilizado para armazenar as ligações criadas no nível de estruturação. A inclusão do WLS estendeu a infra-estrutura proposta anteriormente [17]. A comunicação entre os níveis de estruturação e de armazenamento é realizada por meio de chamadas a funções do nível de ancoragem (*Anchoring Level*).

Implementação. Para validar a infra-estrutura proposta, foi construída uma aplicação, denominada *LinkDigger Service*[5] que permite: (a) a indexação e a identificação de relacionamentos semânticos entre repositórios Web, (b) a criação de ligações entre documentos de um mesmo repositório, no caso em que apenas uma URL seja fornecida e (c) o acesso aos resultados através de uma interface Web que apresenta de forma hierárquica as páginas relacionadas pelo LinkDigger para os *sites* manipulados.

Dessa forma, à infra-estrutura foi acrescentado um nível de apresentação que suporta interação através da Web para que os usuários possam fazer requisições ao serviço de criação de ligações semânticas abertas, e ter acesso à visualização dos resultados correspondentes.

A partir desta implementação, para executar o serviço, é requerido que o usuário: (1) selecione o número de repositórios Web a serem relacionados semanticamente, (2) disponibilize um *e-mail* de contato para o serviço avisar quando o processamento estiver finalizado e (3) ative a execução do processamento. Na implementação, algumas decisões importantes de projeto foram a permissão de, no máximo, cinco repositórios a serem relacionados semanticamente e a execução em *background* do serviço. A primeira decisão foi tomada com o intuito de dimensionar o espaço de informação a ser manipulado, e a segunda decisão planeja suportar a indexação e criação de ligações entre repositórios de qualquer tamanho. Por esse motivo, é solicitado o *e-mail* de contato do usuário durante a interação inicial.

Resultados. Um experimento usando LinkDigger foi realizado para relacionar páginas com notícias *online* dos seguintes jornais: The New York Times (NYT) e New York Post (NYP). O experimento considerou 25 páginas referentes a notícias internacionais publicadas no dia 13 de janeiro de 2002 naqueles jornais (16 notícias do jornal NYT e 9 do jornal NYP). O serviço identificou e criou 174 ligações semânticas latentes entre as 25 páginas dos jornais, uma média de 7 ligações por página. A página com o maior número de ligações (14) foi a página inicial do NYT. Esse resultado era esperado uma vez que essa página é tipicamente índice de notícias. Além disso, das 14 ligações criadas para essa página, 6 delas eram para páginas do NYP. Outro fato que chamou a atenção foi que a página principal do NYP foi relacionada exclusivamente às páginas do próprio jornal. Tal fato talvez possa ser explicado pelo fato do NYT ter um número maior de notícias além de uma maior variedade de informações.

[5] http://mexcal.intermidia.icmc.sc.usp.br/LSL/

4 *Relevance Feedback* para Serviços de Criação de Ligações Semânticas

Os serviços automáticos de criação de ligações semânticas apresentados anteriormente não requerem a participação de usuários na definição de hiperligações como, por exemplo, fazem as *search engines* através da especificação de consultas para iniciar o processo de busca por documentos similares. Esse fator provocou adaptações na utilização das abordagens de recuperação de informação baseadas em sistemas de IR do tipo recuperação de documentos baseada em consultas formuladas por usuários. Por exemplo, para definir ligações entre documentos em repositórios de informações baseando-se na técnica de LSA, os serviços desenvolvidos utilizam todo o conjunto de palavras dos documentos de cada repositório como uma espécie de consulta sobre todos os documentos dos repositórios. Essa adaptação implica: (a) em num número muito maior de palavras a serem utilizadas como consulta por esses serviços do que em consultas tradicionais e (b) na exclusão da participação de usuários em todo o processo de geração de ligações.

Na tentativa de verificar o efeito da participação de usuários na definição de ligações semânticas entre repositórios Web, propõe-se a incorporação de outra abordagem de Recuperação de Informação aos serviços criados, o método de *relevance feedback*. Essa abordagem é proposta através da definição de uma infra-estrutura para serviços de criação automática de ligações que permite a incorporação de um conjunto de ligações consideradas relevantes por usuários.

4.1 Infra-estrutura para geração de ligações de semântica latente com suporte de *relevance feedback*

A Figura 2 apresenta a infra-estrutura proposta para serviços de criação automática de ligações semânticas que permite a participação de usuários, após a geração das ligações, através da inclusão de conjuntos de ligações consideradas relevantes por eles.

No nível de estruturação são definidas as ligações hipertexto entre os repositórios manipulados. Esse nível foi estendido a partir do nível de estruturação da infra-estrutura apresentado na Figura 1 de maneira a permitir a interação de usuários através do método IR de *relevance feedback*. Essa extensão é suportada pelo módulo "Uso de *Relevance Feedback*". Os processamentos numerados na Figura 2 como (2), (3) e (4) correspondem aos passos de execução do módulo após a interação do usuário apresentada no processamento (1).

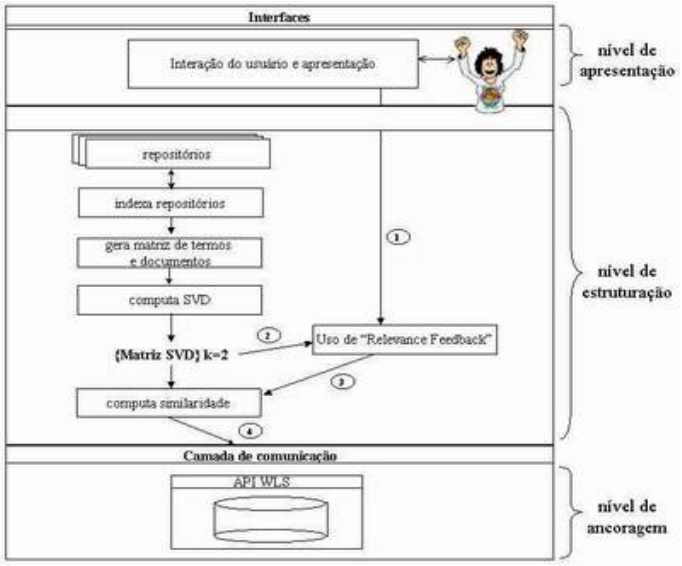

Figura 2: Infra-estrutura proposta para geração de ligações semânticas entre repositórios Web através da interação com usuário suportadas pelo método de *relevance feedback*.

No nível de apresentação, dois tipos de interfaces possibilitam a especificação de informações para a inicialização do serviço e apresentação de resultados com possibilidade de interação de usuários. A interface de ativação do serviço é responsável por habilitar o fornecimento das informações necessárias para execução dos módulos de programa do nível de estruturação. A interface de apresentação de resultados mostra as informações advindas do nível de armazenamento e permite que usuários definam, dentro do conjunto de ligações retornadas, as ligações que são relevantes.

Como na infra-estrutura da Figura 1, a comunicação entre os níveis de estruturação e de armazenamento

continua sendo suportada por chamadas de funções à API (*Application Programming Interface*) do WLS. Essas funcionalidades estão disponíveis nos módulos do nível de ancoragem.

4.2 Algoritmos e Serviço baseados na Infra-estrutura proposta

Para a implementação dos módulos da infra-estrutura proposta apresentados na Figura 2, os algoritmos codificados em *SemanticServiceLinking* e *Uso de Relevance Feedback* foram considerados. O algoritmo *SemanticServiceLinking* foi criado para a infra-estrutura da Figura 1 e foi utilizado para fornecer informações à abordagem proposta suportada pelo algoritmo *Uso de Relevance Feedback*.

4.2.1 Algoritmos do nível de estruturação e de ancoragem

O procedimento *SemanticServiceLinking* é formado por módulos de programa que compõem basicamente os níveis de estruturação e de ancoragem:

- "indexa repositórios" - indexa cada repositório Web. Esses repositórios são indexados separadamente e armazenados em banco de dados.
- "gera matriz termos documentos" - gera a matrix X de termos por documentos com base nos índices gerados a partir dos repositórios.
- "computa SVD" - dada a matrix X gerada anteriormente e o valor k de redução, executa o modelo algébrico SVD gerando as matrizes componentes T, S e D e suas respectivas matrizes reduzidas de dimensão k.
- "computa similaridade" - utiliza o grau de similaridade selecionado na configuração para gerar a matriz semântica a qual é usada para identificar o relacionamento entre os repositórios.
- "serviço aberto de ligações WLS" - módulo do nível de ancoragem que faz requisições ao serviço de armazenamento o qual gera e armazena as hiperligações na base de dados do WLS.

```
Procedure SemanticServiceLinking
Begin varchar(50)
 URLs[]; URLs[]:=URLs /* URLs a serem indexadas */
 For (i:=1 to length(URLs);i++) do {
     /* chamada a engine de indexacao*/
     tbTerms, tbURLs:= indexa_repositórios(URLs[i]);
 }
 matrix_X[N][Z]:= gera_matriz_termos_documentos(tbTerms, tbURLs);
 matrix_T[N][K],matrix_S[K][K],matrix_D[K][Z]:= computa_SVD(matrix_X,K);
 matrixReduced_X[N][Z]:= matrix_T[N][K]*matrix_S[K][K]* transpose(matrix_D[K][Z]);
 /*computa_similaridade() é equivalente a */
 For (i:=1 to Z;i++) do {
    matrixSimilarity[i]:= coseno(matrixReduced[1..N][i], matrixReduced[1..N][i+1]);
    If (matrizSimilarity[i] >= Similarity_Threshold) then
       linkbase_WLS(matrixSimilarity[1..N][i], matrixSimilarity[1..N][i+1]);
} End.
```

Para a abordagem proposta, o algoritmo *Uso de Relevance Feedback* foi definido. Após sua ativação, através da interação de usuários no nível de apresentação, as informações armazenadas na matriz semântica, são recalculadas para redefinir as ligações semânticas. Os módulos básicos do procedimento *Uso de Relevance Feedback* são os seguintes:

- "carrega matriz" - busca a matriz de similaridades semânticas gerada pelo procedimento *SemanticServiceLinking* a ser manipulada a partir dos *feedbacks* de usuários.
- "calcula somatorias" - soma todos os vetores de documentos relevantes após sugestão dos usuários de quais são ou não relevantes.

```
Procedure "Uso de Relevance Feedback"
Begin
 X_Reduced:= carrega_matriz(matrixReduced_X); /* Matriz obtida a partir do procedimento ‘‘SemanticServiceLinking’’ */
 Dr[]:= Drs; /* Conjunto de documentos relevantes a serem processados */
 Dn-r[]:= Dn-rs; /* Conjunto de documentos relevantes a serem processados */
 For (i=0 to i < length(X_Reduced);i++) do {
    calcula_somatorias(X_Reduced[i],SumDr,SumDn-r); /* Soma todos os vetores de documentos relevantes e não-relevantes*/
    X[][q_m]:= alfa*(X[][q_m])+((beta)*SumDr)-(gama*SumDn-r); /* Equacao de Ide para relevance feedback */
 }
 /* Repete os processamentos finais do algoritmo de ‘‘SemanticServiceLinking’’ */
 For (i:=1 to Z;i++) do {
    matrixSimilarity[i]:= coseno(X_Reduced[1..N][i], X_Reduced[1..N][i+1]);
    If (matrizSimilarity[i] >= Similarity_Threshold) then
       linkbase_WLS(matrixSimilarity[1..N][i], matrixSimilarity[1..N][i+1]);
} End.
```

Na codificação do módulo *Uso de Relevance Feedback* foi utilizada a Equação 8 de Ide apresentada na Seção 2 com os seguintes valores para as variáveis de ajuste $\alpha = 0,5$, $\beta = 0,7$ e $\gamma = 0,1$. O método de *relevance feedback*, suportado pela equação de Ide, considera consultas e documentos a serem manipulados interativamente por usuários. Porém, o processo de criação de hiperligações gera pares de documentos relacionados em vez de documentos retornados a partir de comparações de consultas. Por isso, na equação de Ide as consultas $\overrightarrow{q}$ foram consideradas como conjuntos de vetores de palavras de um repositório a ser comparado com outros repositórios. Uma abordagem semelhante foi utilizada na modelagem dos repositórios com o uso da técnica de LSA. A diferença está no fato de que para modelar os repositórios para gerar ligações usando LSA, todas as informações dos repositórios são consideradas, enquanto que no contexto de *relevance feedback*, apenas as informações componentes da matriz semântica, resultante de LSA, são alteradas através da equação de Ide.

Para permitir diversas interações de usuários, indicadas por m, a equação de Ide foi ajustada, como apresentado na Equação 10.

$$\overrightarrow{q}_{m+1} = \alpha\overrightarrow{q_m} + \beta \sum_{\forall \overrightarrow{d}_j \in D_r} \overrightarrow{d}_j \text{ - } \gamma \sum_{\forall \overrightarrow{d}_j \in D_{n-r}} \overrightarrow{d}_j \tag{10}$$

Durante a execução dos passos de iterações, algumas decisões de projeto foram importantes. Por exemplo, decisões em relação à maneira de combinar os resultados do processo de iteração. Existem algumas soluções para combinar conjunto de informações, mas todas elas têm algumas desvantagens: a concatenação de resultados provoca a permanência das ligações não-relevantes na lista total de resultados, a sobreposição de resultados ocasiona a perda da lista de resultados anteriores, e a combinação de resultados pode mesclar resultados relevantes com não-relevantes.

Para combinar resultados na infra-estrutura proposta para geração de ligações, os autores realizam concatenações apenas de resultados relevantes a cada iteração (ligações consideradas relevantes por usuários).

Devido ao fato do contexto manipulado tratar-se de ligações hipertexto bidirecionais, o vetor $\overrightarrow{q_m}$ da Equação 10, pode ser formado a partir de qualquer documento relacionado a cada uma das âncoras correspondentes às ligações semânticas geradas. Por esse motivo, na reconstrução da matriz semântica pelo processo *Uso de Relevance Feedback*, a equação de Ide foi aplicada duas vezes para cada ligação relevante considerada. Variou-se a âncora origem de acordo com a direção que se desejava atender.

4.2.2 Serviço LinkDigger

O LinkDigger foi remodelado de maneira a suportar a interação de usuários durante a criação e a apresentação das ligações definidas e criadas pelo serviço. Para a codificação do algoritmo *InteractiveServiceLinking*, no LinkDigger, foram utilizadas a linguagem de programação PHP[6] para a definição das interfaces do nível de apresentação e as linguagens de programação OX[7] e C++ para a definição dos processos do nível de estruturação.

Um exemplo da interface do LinkDigger que permite aos usuários interagirem com o serviço é apresentada na Figura 3. Essa interface apresenta todas as ligações geradas pelo serviço e, ao lado de cada ligação semântica, existem campos que permitem aos usuários indicar se a ligação relativa àquele campo é relevante (*relevance: 1*) ou não-relevante (*relevance: 0*). Os documentos relacionados em cada ligação podem ser acessados através de âncoras compostas pelos títulos dos documentos Web. Desse modo, usuários podem acessar facilmente as informações relativas às ligações definidas.

Através do botão *Re-create links*, usuários podem iniciar o processo de *relevance feedback* sobre as ligações apresentadas. Na Figura 3, o usuário já realizou interações com o LinkDigger e definiu algumas ligações como relevantes (*relevance: 1*). Para realizar uma nova interação, o usuário deve manipular novamente as variáveis de marcação de relevância e reiniciar o processo através do botão *Re-create links*.

4.3 Experimentos e Resultados

Nesta seção é apresentado o ambiente experimental dos testes realizados com a infra-estrutura proposta e implementada pelo serviço LinkDigger apresentado. A seguir, são analisados os resultados obtidos.

4.3.1 Coleção de documentos

Para verificar a utilidade da infra-estrutura proposta, foi realizado um experimento com dois repositórios Web referentes a dois jornais *online* brasileiros para a criação automática de ligações semânticas mediante a participação de usuários nesse processo.

[6] http://www.php.net

[7] Ferramenta orientada a matrizes desenvolvida em Oxford University. Disponível em http://www.timberlake.co.uk

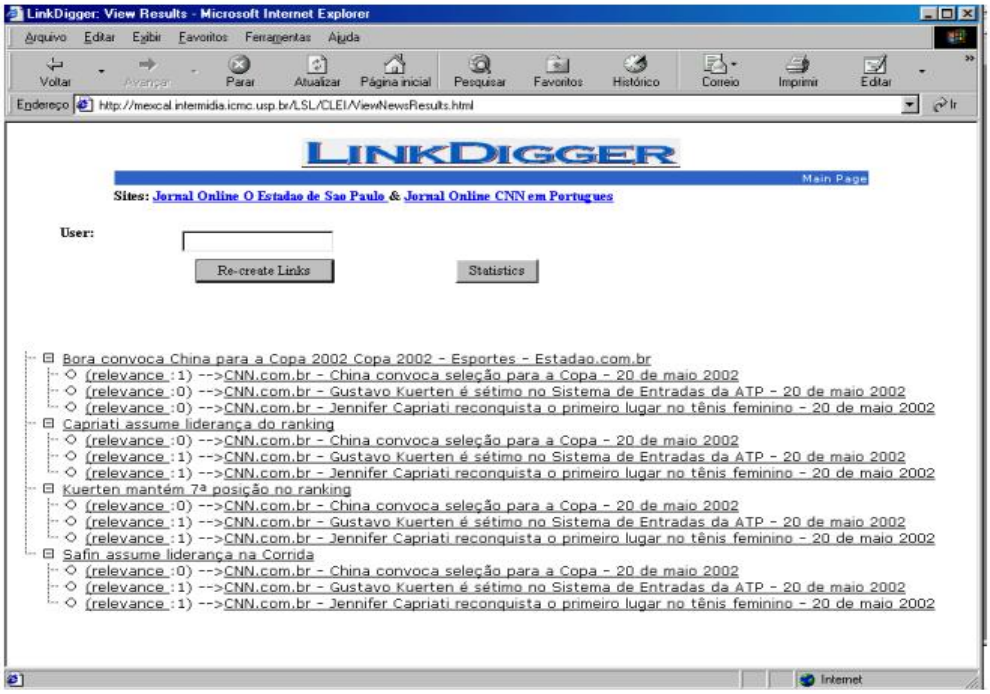

Figura 3: Interface do LinkDigger para apresentação das ligações semânticas criadas e para interações de usuários

Os jornais manipulados foram a versão brasileira do jornal CNN (*http://www.cnn.com.br*) e o jornal *O Estado de São Paulo* (*http://www.estadao.com.br* - OESP). Os *sites* dos jornais OESP e CNN foram analisados na seção de *Esportes*. No contexto deste trabalho, a informação contida nesses repositórios, por serem de uma mesma categoria, é considerada homogênea.

Os repositórios homogêneos foram especificados como fontes de informação para o serviço LinkDigger no dia 20 de maio de 2002. O jornal OESP contava com 59 documentos na seção de *Esportes*. O jornal CNN possuía 15 documentos na mesma seção.

Com o objetivo de indexar esses dois jornais, o LinkDigger considerou um dicionário de *stopwords* em português com 169 palavras e diferentes atribuições de pesos às palavras localizadas no título (peso 2) e no corpo do documento (peso 1). Após indexar os dois repositórios, gerar as ligações e apresentá-las, foram requeridas as participações de cinco usuários para a realização de *relevance feedback* sobre as informações geradas. Através da interface de apresentação de resultado do LinkDigger, os usuários indicaram se a ligação acompanhada pelo campo *relevance* era relevante ou não. Para realizar essa avaliação, foi solicitado aos usuários a leitura de cada par de documentos Web relacionado semanticamente pelo LinkDigger.

4.3.2 Resultados obtidos

O experimento realizado, considerando a infra-estrutura proposta e o contexto apresentado, resultou nos dados resumidos na Tabela 1. A coluna "estado inicial" fornece a quantidade de ligações geradas pelo LinkDigger antes das interações de usuários, isto é, o contexto original. As demais colunas da tabela correspondem a informações geradas pelo LinkDigger após interações de cada um dos cinco participantes do experimento (usuário 1 a usuário 5).

As colunas correspondentes aos usuários na linha (1) indicam as ligações marcadas como relevantes por cada usuário das 44 ligações geradas pelo LinkDigger após a primeira interação. As linhas (2) e (3) indicam, respectivamente, a quantidade de ligações totais e as consideradas relevantes após a interação do usuário. Na coluna "estado inicial", os valores 44 e 8 nas linhas (2) e (3) correspondem às quantidades de ligações totais e relevantes geradas inicialmente pelo LinkDigger.

Para permitir a avaliação dos resultados obtidos em termos do clássico par de medidas de avaliação de sistemas de recuperação de informações, índice de revocação (*recall*) e índice de precisão (*precision*), os dois repositórios manipulados foram analisados por dois usuários conhecedores do conteúdo manipulado com o intuito de criar uma coleção de referência. Nessa análise, os especialistas indicaram que, de todo o conjunto de informação a ser relacionado, apenas 27 relacionamentos semânticos poderiam ser considerados relevantes.

Tabela 1: Informações obtidas a partir do experimento com LinkDigger considerando *relevance feedback*

Variáveis de avaliação	est. inicial	usuário 1	usuário 2	usuário 3	usuário 4	usuário 5
(1) Qtde de ligações marcadas como relevantes na interação do usuário	0	5	12	7	8	4
(2) Qtde de ligações relevantes e não relevantes geradas pelo LinkDigger(A)	44	37	12	16	12	9
(3) Qtde de ligações relevantes geradas pelo LinkDigger (Ra)	8	5	7	6	7	2
(4) Qtde de ligações relevantes (R)	27	27	27	27	27	27
(5) Índice de precisão - Ra/A (%)	18,18	13,51	58,33	37,5	58,33	22,22
(6) Índice de revocação - Ra/R (%)	29,63	18,51	25,92	22,22	25,92	7,40

Esse valor é apresentado na linha (4) da Tabela 1 como (R). Entende-se por índice de revocação, a fração de documentos que foram recuperados pelo sistema de IR (no caso do LinkDigger, ligações relevantes), e por índice de precisão, a fração de ligações recuperadas que são relevantes para o usuário [2]. Os índices de revocação e precisão são calculados como $Rev = \frac{Ra}{R}$ e $Prec = \frac{Ra}{A}$ e são apresentados, respectivamente, nas linhas (5) e (6) da tabela.

Analisando os resultados desse experimento, pode-se observar que, após a interação dos usuários através do método IR de *relevance feedback*, os índices de revocação e de precisão foram positivos para todos os usuários. O índice de precisão demonstra um ganho significativo após as interações de usuários, enquanto que o índice de revocação apresenta uma seqüencia de pequenas quedas de valores se comparados ao contexto inicial de 29,63%. Esses números demonstram a eficiência da abordagem proposta, uma vez que com um bom ganho de precisão, obteve-se pouca geração de informação não-relevante.

Insatisfeito com os resultados da primeira interação, o quinto usuário realizou uma nova interação com o serviço LinkDigger e obteve 77,7% de índice de precisão e 25,9% de índice de revocação. Esses valores demonstraram o aprimoramento de resultados a cada interação.

5 Conclusão

A pesquisa aqui reportada é motivada pela necessidade de investigar mecanismos que forneçam apoio à atividade de identificação de ligações por parte de autores e usuários entre repositórios homogêneos de informações.

Os resultados reportados indicam a viabilidade da utilização de abordagens de Recuperação de Informação que exploram a indexação semântica latente combinada com o método de *relevance feedback* no ambiente da Web.

A pesquisa apresentada deverá ser continuada no sentido de incorporar mecanismos que permitam a usuários a inclusão de termos léxicos ou termos estatisticamente relacionados às ligações que se deseja gerar. Pretende-se também analisar as vantagens de realizar *relevance feedback* de modo automático através de análise de propriedades e características das ligações retornadas.

Agradecimentos

Alessandra Alaniz Macedo e José Antonio Camacho-Guerrero recebem apoio da FAPESP (99/115270 – 00/141036). Maria da Graça Pimentel coordena o projeto de parceria internacional InCA-SERVE que recebe apoio financeiro do CNPq no Brasil e do NSF nos USA. Os autores também gostariam de agradecer a participação de alguns pesquisadores do ICMC-USP/São Carlos na realização dos experimentos apresentados.

Referências

[1] M. Agosti and F. Crestani. A methodology for the automatic constructions of a hypertext for information retrieval. In *Proceedings of Hypertext'93*, pages 745–753, 1993.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1ª edition, Janeiro 1999.

[3] R. Bodner and M. Chignell. Dynamic hypertext: Querying and linking. *ACM Computing Surveys*, 31(4), December 2000.

[4] R. F. Bulcão Neto. An open linking service supporting the authoring of web documents. In *Proceedings of the ACM Document Engineering Conference*, Virginia, USA, 2002. ACM Press.

[5] C. Dwork, R. Kumar, M. Naor, and D. Sivakimar. Rank aggregation methods. In *Proceedings of WWW'10*, pages 613–622, May 2001.

[6] S. R. El-Beltagy, W. Hall, D. DeRoure, and L. Carr. Linking in context. In *Proceedings of the Hypertext'01*, pages 151–160, August 2001.

[7] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th International Conference on Research & Development in Information Retrieval*, pages 465–480, 1988.

[8] G. Golovchinsky. What the query told the link: The integrations of hypertext and information retrieval. In *Proceedings of the ACM Hypertext'97*, pages 30–39, 1997.

[9] S.J. Green. Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering*, 11(5):713–730, September/October 1999.

[10] K. Grønbæk and R.H. Trigg. Design issues for a Dexter-based hypermedia system. *Communications of the ACM*, 37(2):41–49, 1994.

[11] D. Harman. Relevance feedback revisited. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, June 21-24 1992.

[12] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring quality using walks on the Web. In *Proceedings of the 8th International World Wide Web Conference*, pages 213–225, Toronto/Canada, May 1999.

[13] E. Ide. New experiments in relevance feedback. In G. Salton, editor, *The Smart Retrieval System*, pages 337–354. Prentice Hall, 1971.

[14] M. Kobayashi and K. Takeda. Information retrieval on the Web. In *Proceedings of the ACM Computing Survey 32(2)*, pages 146–173, June 2000.

[15] Z. Li, Y. Wang, and V. Oria. A new architecture for web meta-search engines. In *Proceedings of the 7th American Conference on Information Systems*, pages 415–421, 2001.

[16] A. A. Macedo, M. G. C. Pimentel, and J. A. Camacho-Guerrero. Latent semantic linking over homogeneous repositories. In *Proceedings of the ACM Symposium on Document Engineering*, pages 144–151, Atlanta, Georgia, USA, November 2001. ACM Press.

[17] A. A. Macedo, M. G. C. Pimentel, and J. A. Camacho-Guerrero. An infrastructure for open latent semantic linking. In *Proceedings of the ACM Hypertext*, pages 107–116, College Park, Maryland, USA, June 2002. ACM Press.

[18] N. Meyrowitz. Intermedia: the architecture and constructionof an object-oriented hypermedia system and applications framework. In *Proceedings of the OOPSLA 1986*, pages 186–201, September 1986.

[19] M. N. Price, G. Golovchinsky, and B. N. Schilit. Linking by inking: trailblazing in a paper-like hypertext. In *Proceedings of ACM Hypertext'98*, pages 30–39, 1998.

[20] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[21] E. Seraphim and R.P.M. Fortes. Ferramenta DB-LiOS para avaliação de reuso de links em WWW. In *Anais do XX Congresso Nacional da Sociedade Brasileira de Computação – XXVII SEMISH*, pages 15–21, 2000.

[22] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, and N. Ziviani. Link-based and content-based evidential information in a belief network model. In *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 96–103, 2000.