Expansão de Consultas Utilizando Indexação Semântica Latente

Cristiane Amorim Mendes[2] Edleno Silva de Moura[1] Nivio Ziviani[2]

[1] Departamento de Ciência da Computação
Universidade Federal do Amazonas
Manaus AM - Brasil

[2] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte MG - Brasil

Resumo

Abstract

1 Introdução

Um dos principais problemas encontrados por projetistas de sistemas de recuperação de informação é o fato das consultas submetidas aos sistemas serem normalmente compostas de poucos termos, os quais passam pouca informação a respeito da necessidade de informação dos usuários. Além disso, a falta de conhecimento a respeito do funcionamento dos mecanismos de busca por parte dos usuários dificulta a formulação de consultas que produzam os resultados satisfatórios [9, 14].

Uma maneira eficaz de contornar os problemas apresentados acima é utilizar métodos que expandem automaticamente as consultas dos usuários [11, 8]. O objetivo dos algoritmos de expansão de consultas é formular uma consulta mais elaborada a partir da consulta feita inicialmente pelo usuário, transformando-a em uma nova consulta mais elaborada. A expansão de consultas é útil porque a linguagem natural permite que as pessoas utilizem palavras e expressões diferentes para indicar um mesmo objeto. Por exemplo, se um usuário faz a consulta "Ministério da Educação", um algoritmo de expansão de consultas poderia incluir a palavra "MEC" na consulta expandida, pois muitos documentos fazem referência ao Ministério da Educação através da sigla MEC. Da mesma forma, a consulta por "redes de televisão" poderia ser expandida com termos como TV, Globo, Record, SBT, dentre outros.

O objetivo deste trabalho é propor, implementar e avaliar um novo método de expansão de consultas que melhora a qualidade das respostas dos sistemas de recuperação de informação. O método proposto utiliza um modelo conhecido como Indexação Semântica

Latente (ISL) [6] para identificar automaticamente relações entre palavras existentes na coleção e as palavras encontradas nas consultas dos usuários. Esta informação é utilizada para expandir a consulta, a qual é processada utilizando-se o modelo de espaço vetorial [13]. Nos experimentos realizados com 3 coleções de documentos, o novo método de expansão causou um ganho de até 15 % na qualidade das respostas em relação à recuperação feita apenas com as consultas originais. Além disso, o método proposto não degrada o desempenho dos sistemas em termos do tempo gasto para processar as consultas, o que facilita seu uso em sistemas onde o desempenho é um fator crítico.

Este artigo está organizado da seguinte forma. Na seção 2 apresentamos os trabalhos relacionados. Nas seções 3 e 4 descrevemos o modelo vetorial e o modelo ISL, respectivamente. Na seção 5 apresentamos o método para a expansão de consultas e os resultados obtidos. Na seção **??** apresentamos uma solução para identificar termos relacionados em grandes coleções e os resultados experimentais. Finalmente, na seção 7 apresentamos as conclusões e as direções para trabalhos futuros.

2 Trabalhos Relacionados

As técnicas de expansão de consultas encontradas na literatura podem ser divididas em dois tipos: as que são realizadas com ajuda dos usuários, conhecidas como realimentação de relevantes; e as que são feitas sem qualquer ajuda do usuário, conhecidas como expansão automática. Estudos encontrados na literatura mostram que técnicas de realimentação de relevantes normalmente produzem resultados melhores que técnicas de expansão automática [12]. Contudo, as técnicas de realimentação de relevantes costumam não ser bem sucedidas em sistemas de busca disponíveis na Web, pois seus usuários raramente fornecem informação para realimentar o sistema [9]. Nestes casos, a melhor solução para expandir as consultas é através de técnicas de expansão automática.

A expansão automática de consultas pode ser classificada como local ou global [16]. Uma expansão de consultas é local quando se realiza uma análise das relações entre os termos nos documentos que constituem o conjunto-resposta da consulta original. Na expansão local é necessário que primeiro se processe a consulta original, depois se faça algum tipo de análise sobre o conjunto-resposta para formular uma nova consulta, que deve também ser processada pelo sistema. Estas características fazem com que a expansão local seja extremamente cara em termos computacionais, principalmente se o método de análise das relações entre os termos utilizado tiver custo computacional elevado. Já a expansão global é realizada estudando-se as relações entre os termos em toda a coleção de documentos. Assim, não há a necessidade de se processar a consulta duas vezes, e a análise das relações entre os termos da coleção pode ser pré-computada uma única vez.

O método de expansão proposto aqui utiliza o modelo de indexação semântica latente (ISL) [6] para encontrar relações semânticas entre os termos da coleção. O modelo ISL utiliza uma técnica estatística para identificar conceitos existentes na coleção, conhecida como decomposição do valor singular [7]. O processamento de consultas com o modelo ISL pode ser considerado um método de expansão global de consultas, onde os documentos e as consultas são representados pelos conceitos semânticos encontrados na coleção. Uma consulta formulada pelo usuário primeiro é transformada para o espaço de conceitos e depois processada. Assim, o conjunto-resposta pode ser composto por documentos que contenham palavras que não foram incluídas pelo usuário na consulta.

Apesar da sofisticação matemática do modelo ISL, os resultados práticos obtidos têm sido modestos quando consideramos os custos envolvidos na identificação dos conceitos semânticos. Além disso, existe um custo extra que o uso do modelo adiciona ao processamento da consulta [6, 2, 4, 1]. Um dos principais problemas encontrados no uso do modelo ISL é que a consulta original do usuário pode ser completamente descaracterizada no processo de transformação da mesma para o espaço de conceitos. Este risco ocorre principalmente em consultas curtas onde os termos originais podem estar relacionados a muitos conceitos distintos.

Nossa proposta tira proveito das relações semânticas entre termos obtidas através do modelo ISL e combina esta informação com a consulta original através de um esquema que evita uma completa descaracterização da mesma. Resultados experimentais mostram que nosso método de expansão de consultas produz resultados melhores do que os resultados obtidos com o ISL. Além disso, o tempo para processar consultas é consideravelmente menor que o tempo gasto para se processar consultas com o modelo ISL puro, o que viabiliza a utilização do método em sistemas onde o desempenho é crítico, tal como os sistemas de recuperação de informação disponibilizados na Web.

O método de expansão de consultas proposto neste trabalho é baseado na correlação entre termos da coleção. Em [11] é apresentado um método de expansão global de consultas que também explora a correlação entre termos da coleção, onde os experimentos realizados mostram que a consulta expandida em geral produz resultados melhores que os obtidos pela consulta original. Nosso trabalho difere do trabalho apresentado em [11] nos seguintes pontos principais: Primeiro, nós obtemos a correlação entre termos utilizando as co-ocorrências dos termos no espaço de conceitos gerado pelo modelo ISL, enquanto em [11] as correlações entre termos são obtidas através do estudo da co-ocorrência dos termos nos documentos. Segundo, nós propomos uma nova forma de se estimar a correlação entre uma consulta e os termos da coleção, a qual modela uma informação que não é captada pela fórmula de correlação apresentada em [11]. Por último, enquanto a técnica apresentada em [11] insere os termos da expansão diretamente na consulta, nós propomos e avaliamos uma nova técnica onde os termos da expansão são transformados em um único conceito para então serem adicionados a consulta.

3 Modelo Vetorial

Na nossa proposta, o processamento das consultas após a expansão é feito utilizando-se o modelo de espaço vetorial [13]. O modelo de espaço vetorial é um dos modelos mais conhecidos e empregados em recuperação de informação. Neste modelo, os documentos da coleção e as consultas dos usuários são representados como vetores de termos dentro de um espaço vetorial. Um documento d é representado em um espaço n-dimensional através de um vetor $\vec{d} = \langle w_{d,1}, \ldots, w_{d,n} \rangle$, onde n representa o número de termos distintos da coleção, $w_{d,i}$ representa o peso do termo i no documento d. De forma análoga, uma consulta q é representada neste espaço através de um vetor $\vec{q} = \langle w_{q,1}, \ldots, w_{q,n} \rangle$, onde $w_{q,i}$ representa o peso do termo i na consulta q. Estes pesos são utilizados para quantificar a relevância de cada termo para as consultas e para os documentos no espaço vetorial.

A similaridade entre um documento d e uma consulta q no modelo vetorial é expressa pela correlação entre os vetores $\vec{d}$ e $\vec{q}$, a qual é igual ao cosseno do ângulo formado por estes dois vetores e é expressa pela equação 1.

$$sim(\vec{d}, \vec{q}) = \frac{\vec{d} \bullet \vec{q}}{|\vec{d}| \times |\vec{q}|} = \frac{\sum_{i=1}^{n} w_{d,i} w_{q,i}}{\sqrt{\sum_{i=1}^{n} w_{d,i}^2} \sqrt{\sum_{i=1}^{n} w_{q,i}^2}} \tag{1}$$

onde $|\vec{d}|$ representa a norma do vetor documento e $|\vec{q}|$ a norma do vetor consulta.

Os pesos dos termos nos documentos e na consulta podem ser calculados de várias maneiras [5, 15], dependendo das características da coleção e do tipo de recuperação desejada. Os pesos adotados em nossos experimentos foram adotados como $w_{d,t} = tf_{d,t} \times idf_t$, onde $tf_{d,t}$ é a freqüência do termo t no documento d e idf_t é a a freqüência inversa do documento, a qual é dada pela fórmula $idf_{(t)} = \log\left(\frac{N}{f_t}\right)$, onde N é o número de documentos da coleção e f_t é o número de documentos onde o termo t ocorre.

Para determinar os documentos relevantes a uma consulta q, a similaridade $sim(\vec{d}, \vec{q})$, entre cada documento d da coleção e a consulta q é calculada e comparada a um limiar. Normalmente, uma lista dos z documentos com maior similaridade é retornada ao usuário. Mais detalhes sobre este modelo podem ser encontrados na literatura [12, 15].

O algoritmo de processamento de consultas usado para recuperar e ordenar os documentos de acordo com o modelo de espaço vetorial utiliza um acumulador A_d para cada documento d da coleção, o qual armazena o resultado da similaridade descrita na Equação 1. Para cada termo t, a contribuição $simparc_{q,d,t}$ é adicionada ao acumulador do documento d. Esta contribuição, chamada de similaridade parcial de q e d em relação ao termo t, é definida pela equação 2.

$$simparc_{q,d,t} = w_{d,t} \cdot w_{q,t} \tag{2}$$

O algoritmo utiliza a informação armazenada em uma estrutura de dados conhecida como arquivo invertido [15] para processar as consultas. Para cada termo da coleção o arquivo invertido contém uma lista com todos os documentos onde o termo ocorre com suas respectivas freqüências, a qual é conhecida como lista invertida. A lista invertida é recuperada do disco e manipulada para calcular as similaridades entre os documentos e a consulta. A aplicação do modelo de espaço vetorial resulta em um conjunto de documentos ordenados de forma decrescente de acordo com o grau de similaridade entre a consulta e os documentos recuperados. Uma descrição do algoritmo de processamento de consultas é apresentada na figura 1.

1. Para cada documento d na coleção, criar $A_d \leftarrow 0$.
2. Para cada termo t da consulta,
   (a) Recuperar a lista invertida do termo no disco.
   (b) Para cada entrada $< d, f_{d,t} >$ da lista invertida,
   $A_d \leftarrow A_d + simparc_{q,d,t}$.
3. Dividir cada acumulador $A_d \neq 0$ pela norma do documento $\mid \vec{d} \mid$.
4. Identificar os k valores mais altos de acumuladores, onde k é o número de documentos retornados para o usuário.

Figura 1: Algoritmo para o cálculo da similaridade no modelo vetorial.

4 Indexação Semântica Latente

No modelo ISL a importância dos documentos e dos termos em relação à consulta é estimada considerando a informação extraída da co-ocorrência dos termos na coleção. O modelo assume que existe uma estrutura latente no uso dos termos nos documentos que é parcialmente obscurecida pela variabilidade da escolha de termos [2]. A estrutura latente é obtida através de uma técnica da álgebra linear conhecida como decomposição do valor singular [7].

O primeiro passo para usar o modelo ISL é construir uma matriz de incidência A para representar a base de dados, onde os T termos da coleção representam as linhas e os N documentos as colunas. Cada coluna desta matriz corresponde ao vetor de um documento representado no modelo vetorial [3], onde a posição $A(i,j)$ contém o peso do termo i no documento j, para $1 \leq i \leq T$ e $1 \leq j \leq N$.

A decomposição do valor singular é utilizada para decompor a matriz A em um produto de três outras matrizes, a saber:

$$\underset{T\times N}{A} = \underset{T\times R}{U_o} \cdot \underset{R\times R}{\Sigma_o} \cdot \underset{R\times N}{V_o^T} \tag{3}$$

tal que $U^T U = V^T V = I_N$ e $\Sigma = \text{diagonal}(\sigma_1, \sigma_2, \ldots, \sigma_N), \sigma_i > 0$ para $1 \leq i \leq R$, $\sigma_i = 0$ para $i \geq R+1$. As primeiras R colunas das matrizes ortogonais U e V definem os autovetores ortonormais associados aos R autovalores não zeros de AA^T e $A^T A$, respectivamente. As colunas de U e V são referenciadas como os vetores singulares esquerdo e direito, respectivamente, e os valores singulares de A são definidos pelos elementos da diagonal de Σ, que são as raízes quadradas dos T autovalores não nulos de AA^T.

Se somente os K maiores valores singulares de Σ_o são mantidos com suas correspondentes colunas em U_o e V_o e o restante removido, a matriz resultante A_K é a matriz de posto K que mais se aproxima da matriz original A:

$$\underset{T\times N}{A} \approx \underset{T\times N}{A_K} = \underset{T\times K}{U} \cdot \underset{K\times K}{\Sigma} \cdot \underset{K\times N}{V^T} \tag{4}$$

A quantidade K de valores singulares determina o número de dimensões do espaço conceitual em que os termos e os documentos serão representados. Em aplicações de recuperação de informação a escolha deste valor tem sido feita empiricamente [3]. Neste caso, K é determinado pelo valor que fornece os melhores resultados de precisão na base de dados. A redução de dimensão permite remover muitos dos "ruídos" da coleção, mas um número muito reduzido pode implicar em perda de informação importante.

A Figura 2 mostra uma representação matemática para a decomposição do valor singular. As matrizes U_K e V_K^T resultantes da decomposição do valor singular, fornecem as coordenadas no espaço conceitual dos termos e dos documentos, respectivamente.

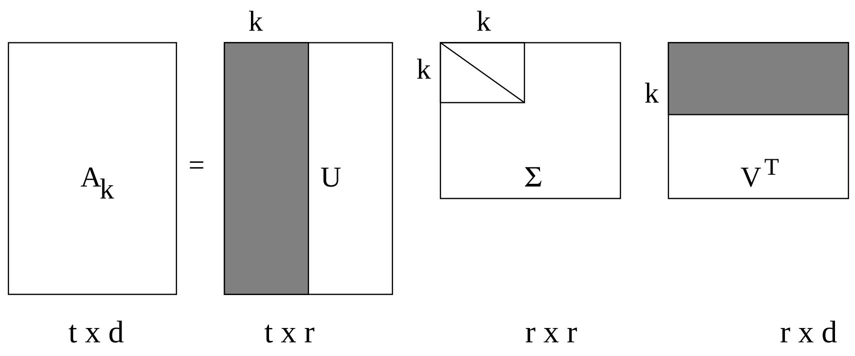

Figura 2: Representação matemática da matriz A_k.

A recuperação de documentos com ISL é feita transformando-se as consultas para este espaço de conceitos através da equação 5, seguida da aplicação do modelo vetorial. Os vetores dos documentos são representados no espaço de conceitos através da matriz V_K^T, onde a j-ésima coluna ($1 \leq j \leq N$) representa um vetor que descreve o j-ésimo documento da coleção no espaço de conceitos.

$$q = q^T U_K \Sigma_K^{-1} \tag{5}$$

Esta transformação da consulta é um dos principais problemas práticos do modelo ISL. Quando o modelo vetorial está sendo utilizado diretamente sobre a matriz A, o principal custo envolvido no processamento da consulta consiste na leitura das listas invertidas relacionadas aos termos fornecidos na consulta. Como as consultas submetidas para a maioria dos sistemas de recuperação de informação modernos são normalmente compostas de poucos termos, apenas algumas listas invertidas são lidas do disco.

Já no processamento do modelo de espaço vetorial aplicado sobre a matriz V_K^T, a consulta é transformada do espaço de termos para o espaço de conceitos, resultando em uma nova consulta, a qual normalmente possui um número de conceitos que é muito maior do que o número de termos da consulta original (tipicamente centenas de vezes maior). As listas invertidas de cada conceito desta nova consulta devem então ser lidas para que se descubra os documentos mais relacionados a mesma. Assim, o custo no processamento das consultas com o modelo ISL torna-se extremamente alto.

Outra desvantagem do modelo ISL é que muitas vezes a transformação da consulta para o espaço de conceitos descaracteriza a consulta original do usuário. (Incluir um exemplo).

EDLENO: CRISTIANE DEVE COLOCAR UM EXEMPLO DE RESULTADO ESTRANHO DO LSI AQUI. TEM QUE SER UM RESULTADO ONDE AS RESPOSTAS NAO TEM RELACAO COM A CONSULTA ORIGINAL PORQUE A CONSULTA NO ESPACO DE CONCEITOS FOI DESCARACTERIZADA.

5 Uma Nova Proposta para a Expansão de Consultas

Nesta seção nós apresentamos nosso método de expansão de consultas. A idéia consiste em explorar informação semântica obtida a partir do modelo ISL, a qual é utilizada para expandir a consulta original fornecida pelo usuário.

A seguir apresentamos os principais passos que constituem o método proposto.

Passo 1: Tratamento da Matriz de Incidências.

O primeiro passo do método consiste em fazer um tratamento na matriz de incidências A para eliminar termos muito comuns e termos que sejam raros na coleção. A eliminação de termos muito comuns na coleção é realizada porque a ocorrência de um destes termos carrega uma quantidade pequena de informação, devendo assim ser desprezada [13, 12, 15]. A eliminação de termos muito raros da coleção deve ser realizada porque tais termos muitas vezes ocorrem devido a erros de ortografia ou erros de digitação cometidos durante a elaboração dos documentos [13]. Tais erros são comuns em coleções formadas a partir de *sites* da Web, onde o controle de qualidade do material publicado normalmente é precário. Outra razão importante para a eliminação de termos que ocorrem pouco na coleção é o fato da informação estatística a respeito destes termos não ser suficiente para determinar suas relações com outros termos da coleção. Experimentos realizados por nós com o modelo ISL mostraram que as associações entre termos que ocorrem menos do que 5 vezes na coleção e os demais termos foram em geral absurdas. A eliminação dos termos raros é realizada somente durante o pré-processamento realizado para a obtenção das relações semânticas entre os termos. Durante o processamento das consultas estes termos são processados normalmente pelo sistema para evitar que consultas por estes termos fiquem sem respostas.

Passo 2: Utilizando o modelo ISL para obter informação sobre termos.

A matriz U_K gerada pelo modelo ISL representa as relações entre os termos encontrados na coleção e os conceitos encontrados pelo modelo ISL. A j-ésima coluna de U_K representa um vetor que descreve o j-ésimo termo da coleção no espaço de conceitos. A partir deste fato é possível obter informação sobre como cada termo se relaciona com os demais termos da coleção.

Uma forma de mensurar esta relação é calcular a correlação existente entre todos os vetores de termos representados em U_K. Dados dois termos i e j quaisquer da coleção, a correlação entre estes termos é dada pelo cosseno do ângulo entre os vetores destes termos, ou seja:

$$Correlacao(t_i, t_j) = \frac{\vec{t_i} \cdot \vec{t_j}}{|\vec{t_i}| \times |\vec{t_j}|} \tag{6}$$

onde $\vec{t_i}$ e $\vec{t_j}$ são os vetores dos termos i e j no espaço de conceitos gerado pelo modelo LSI, os quais são representados pela i-ésima e pela j-ésima colunas de U_K, respectivamente. Deste modo, para cada termo da coleção criamos uma lista de termos relacionados com o mesmo, a qual indica qual a correlação entre este termo e cada um dos demais termos da coleção.

Passo 3: Relacionando os termos da coleção às consultas.

Em nosso método, a expansão é feita adicionando-se um termo artificial que representa um conceito relacionado semanticamente à consulta, denominado termo semântico t_s. O termo semântico t_s é obtido a partir das correlações entre os termos conforme mostrado na equação 6, fazendo com que o vetor que representa o termo t_s seja escrito como a combinação linear $c_1\vec{t_1} + c_2\vec{t_2} + ... + c_T\vec{t_T}$, onde para cada i, $1 \leq i \leq T$, c_i representa o peso do termo t_i na composição de $\vec{t_s}$. Quando a consulta original possui apenas um termo, uma escolha natural para o valor de c_i é usar diretamente a correlação entre t_i e o termo da consulta. Quando a consulta tiver mais do que um termo, é necessário que se estabeleça uma métrica para computar c_i em função de todos os termos da consulta. Isso é feito de tal forma que o valor de c_i represente um valor médio da correlação entre t_i e os termos da consulta, mas seja baixo se houver algum termo na consulta cuja correlação com t_i seja baixa. Em um caso extremo, se houver algum termo na consulta que tenha correlação igual a zero com t_i, é desejável que a correlação de t_i com a consulta como um todo seja zero. O instrumento que melhor se aproxima destas características é a média geométrica. Assim, dada uma consulta q, definimos c_i como:

$$c_i = \Big(\prod_{t \in q} Correlacao(t_i, t)\Big)^{\frac{1}{l}} \tag{7}$$

onde l é o número de termos da consulta.

Uma outra maneira de calcular esta correlação é apresentada em [11], onde propõe-se que o valor de c_i seja dado pela média aritmética das correlações entre o termo i e cada termo da consulta, ou seja:

$$c_i = \frac{\sum_{t \in q} Correlacao(t_i, t)}{l} \tag{8}$$

onde l é o número de termos da consulta.

A diferença entre as equações 7 e 8 está na interpretação do significado da correlação entre um termo e a consulta. A equação 7 indica que a correlação entre a consulta e um dado termo t_i só existe se t_i estiver relacionado a todos os termos da consulta. A equação 8 indica que a correlação entre a consulta e um dado termo t_i existe se t_i estiver relacionado a qualquer um dos termos da consulta. A seção 6 serão apresentados experimentos comparando estas duas formas de se interpretar a correlação entre as consultas e os termos.

Passo 4: Expandindo a consulta.

Após o computo de t_s, a expansão é realizada adicionando-se t_s à consulta e à coleção. O termo t_s passa então a representar uma nova dimensão do espaço vetorial cujo significado está associado ao conceito extraído da consulta usando-se a decomposição do valor singular. Esta nova dimensão substitui os termos utilizados para compor o termo t_s, isto é, aqueles cujo valor é diferente de zero na combinação linear que gerou o vetor t_s. O cálculo do peso do termo semântico t_s em cada documento d da coleção é dado por:

$$w_{d,t_s} = \Big(\sum_{i=1}^{T} c_i f_{i,d}\Big) Idf(t_s) \tag{9}$$

onde $1 \leq i \leq T$, $f_{i,d}$ é a freqüência do termo i no documento d e $Idf(t_s)$ é calculado em função do número de documentos onde ocorre pelo menos um dos termos utilizados para compor t_s.

Outra forma de se fazer a expansão é adicionar diretamente os termos que compõem t_s à consulta. Neste caso cada termo que compõe t_s seria tratado individualmente, o que significa na prática substituir o idf utilizado na equação 9 pelo idf individual de cada termo que compõe t_s. Assim a equação para computar o peso de t_s seria:

$$w_{d,t_i} = \Big(\sum_{i=1}^{T} c_i f_{i,d}\Big) Idf(t_i) \tag{10}$$

onde $1 \leq i \leq T$, $f_{i,d}$ é a freqüência do termo i no documento d e $Idf(t_i)$ é o idf no termo i na coleção.

A forma de realizar expansão através da criação de um conceito relacionado à consulta faz com que o impacto final de cada termo na expansão dependa de sua correlação com a consulta e do número de documentos onde ocorre pelo menos um dos termos da expansão, o qual é usado para computar o $idf(t_s)$. Isso significa que se a expansão for muito genérica então o impacto final de todos termos incluídos na expansão será baixo. Além disso, uma correlação alta entre termos que ocorrem poucas vezes na coleção e os termos da consulta muitas vezes não significa que há uma relação semântica entre estes termos. Se um termo raro é incluído erroneamente na expansão de consultas, então o impacto deste termo numa expansão direta pode prejudicar bastante o resultado final da ordenação do conjunto-resposta. Este problema ocorre na expansão direta devido a utilização do *idf* individual de cada termo na expansão, pois termos raros tem *idf* alto. Na expansão através do termo semântico o problema é amenizado porque todos os termos, raros ou não, são reunidos em um único conceito.

Na seção 6 serão apresentados experimentos comparando estas duas formas de se fazer a expansão de consultas.

Na prática, a inclusão direta ou indireta dos termos da expansão significa que para processar a consulta utilizando-se o modelo vetorial será necessário processar as listas invertidas de todos os termos cuja correlação com os termos da consulta é maior que zero. Este aumento no número de listas invertidas processadas pode degradar significativamente o desempenho dos sistemas que utilizam qualquer uma das duas técnicas de expansão mostradas. A seguir apresentamos algumas formas de evitar esta degradação no desempenho sem que haja perdas na qualidade da ordenação das respostas apresentadas pelos sistemas.

Passo 4: Descartando termos do vetor $\vec{t_s}$.

Como o número de termos da expansão pode ser potencialmente grande o custo para processar uma consulta pode tornar-se proibitivo. Uma solução para este problema é utilizar estratégias de poda para descartar automaticamente termos que foram incluídos em t_s, mas que não causam impacto na ordenação final da resposta. Nossa proposta para computar este impacto é inspirada nas idéias apresentadas por Persin [10], cuja proposta é constituída de uma estratégia que computa o impacto que o processamento de cada novo termo da consulta pode causar no conjunto resposta, procurando descartar termos que tenham pouco ou nenhuma influência na ordenação da resposta.

Para a determinação da poda é necessário calcular previamente o valor da a similaridade parcial mínima entre um termo da consulta q e qualquer documento da coleção onde haja ocorrências de algum termo da consulta. í Vale lembrar que a similaridade parcial expressa a medida de contribuição de um termo durante o processamento no modelo vetorial, conforme mostrado na seção 3. A partir da equação 2, o cálculo da similaridade parcial mínima pode ser feito através da equação:

$$SimMinParc(q) = 1,0 \times idf(min_q) \times 1,0 \times idf(min_q) \tag{11}$$

onde assumimos que a freqüência mínima do termo no documento e na consulta é igual a 1 e $idf(min_q)$ é o menor *idf* dentre os $idfs$ dos termos da consulta. O cálculo do valor de $SimMinParc(q)$ é usado para descartar termos que não causariam impacto na ordenação final da resposta se os mesmos fossem incluídos na composição do termo semântico. O valor de $SimMinParc$ representa a menor similaridade parcial que pode ocorrer entre um termo da consulta original e um documento. Para detectar se um termo t_i utilizado para compor o termo semântico t_s não modifica a ordenação das respostas, calcula-se o maior valor de similaridade parcial que este termo pode ter com algum documento da coleção e compara-se este valor a $SimMinParc(q)$.

Para cada termo t_i cujo valor c_i na composição de de t_s seja maior que zero, tem-se que

$$SimParcMax(q, t_i) = fmax(t_i) \cdot idf(t_i) \cdot c_i \cdot idf_t \tag{12}$$

onde $fmax(t_i)$ é a maior freqüência com que o termo t_i ocorre na coleção, $idf(t_i)$ é o *idf* do termo t_i na coleção, e c_i é a correlação entre o termo t_i e a consulta, calculada conforme a equação **??**. Se $SimParcMax(q, t_i) < SimMinParc(q)$, então a inclusão do termo t_i não modificaria a ordenação das respostas e, portanto, o termo é automaticamente excluído da lista de componentes de t_s, fazendo-se $c_i = 0$. Note que os valores de *idf* e freqüência máxima dos termos são normalmente armazenados na memória junto com o vocabulário, o que faz com que o tempo gasto neste pré-processamento seja baixo em relação ao tempo total gasto para processar a consulta.

Conseqüentemente, somente o conjunto de termos que não foram descartados pela regra de poda é utilizado na composição final t_s. Em alguns casos, o número de termos restantes na expansão pode ainda ser alto demais mesmo depois da realização da poda descrita acima. Trabalhos encontrados na literatura mostram que uma saída razoável nestes casos é estabelecer um limite máximo para o número total de termos, mesmo que haja a possibilidade de perda na qualidade das respostas neste caso. Em nossos experimentos nós utilizamos o limite máximo de 100 termos na expansão, um valor próximo ao sugerido em [11].

Passo 5: Calculando um valor aproximado do Idf de t_s

Como dito anteriormente, o valor de $idf(t_s)$ depende do número de documentos onde pelo menos um dos termos utilizados para compor t_s ocorre. O custo para se computar este valor é muito alto para que a operação seja realizada durante o processamento da consulta. Para evitar este custo, nossa implementação prática utiliza um valor aproximado de *idf* de t_s, o qual é dado pelo menor *idf* dentre todos os *idfs* dos termos que foram utilizados para compor o t_s. Esta aproximação é razoável, uma vez que espera-se que haja uma forte correlação entre os termos que compõem t_s, com conseqüente intersecção entre suas listas invertidas. Logo, a união das listas invertidas leva a um número de documentos próximo ao número de documentos do termo de menor *idf*.

6 Resultados Experimentais

Cristiane vai escrever primeira versao desta secao

Algumas sugestoes para ajudar na escrita:

Secao deve incluir experimentos com colecoes Web e com a CACM

A secao comeca descrevendo detalhadamente as colecoes que serao usadas nos experimentos, o ambiente de experimentacao e os objetivos gerais dos experimentos realizados

Serao feitas comparacoes com o modelo ISL, com o vetorial sem expansao e, se for conveniente, com outra tecnica de expansao.

As comparacoes vao levar em conta a qualidade das respostas e o tempo total para processar as consultas

6.1 Impacto no tempo de processamento

Em nossos experimentos, o número médio de termos nas consultas expandidas com a utilização da técnica de poda proposta acima foi de XXX termos. O gráfico abaixo mostra o impacto do método de expansão no tempo de processamento das consultas. Como pode ser observado, o tempo médio para processar as consutas expandidas ficou próximo ao tempo de processamento das consultas no sistema original sem a utilização da expansão.

AQUI VEM UM GRAFICO COM UMA COMPARACAO NO TEMPO DE PROCESSAMENTO DA CONSULTA

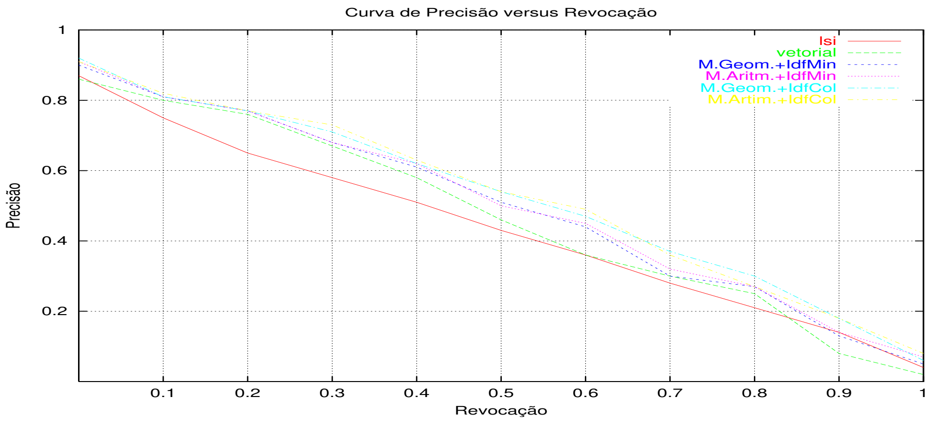

Figura 3: Curva de precisão e revocação para a coleção MED utilizando-se o modelo ISL(lsi), o modelo vetorial(vetorial) e as quatro técnicas de expansão estudadas

7 Conclusões

Referências

[1] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335–362, 1999.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[3] S. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236, 1991.

[4] S. T. Dumais. Latent semantic indexing (lsi) and trec-2. In *TREC -2*, pages 105–116, 1993.

[5] W. B. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Englewoods Cliff, N.J., 1992.

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

[7] G. Golub, V. Klema, and G. Stewart. Rank degeneracy and least squares problems. Technical report, University of Maryland, College Park, MD, 1976.

[8] J. Greenberg. Automatic query expansion via lexical-semantic relationships. *Journal of American Society For information Science and Technology (JASIS)*, 52(5):402–415, 2001.

[9] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214, 1998.

[10] M. Persin, J. Zobel, and R. Sacks-Davis. Filtered document retrieval with frequency-sorted indexes. *Journal of the American Society of Information Science*, 47(10):749–764, 1996.

[11] Y. Qiu and H. Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.

[12] Modern Information Retrieval. *R. Baeza-Yates and B. Ribeiro-Neto*. Addison Wesley, 1999.

[13] G. Salton. *Automatic Information Organization and Retrieval.* McGraw-Hill, New York, NY, 1968.

[14] 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 the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96–103, 2000.

[15] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes*. Morgan Kaufmann Publishers, San Francisco, California, USA, 2nd. edition, 1999.

[16] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (ACM TOIS)*, 18(1):79–112, 2000.