Modelagem Vetorial Estendida por Regras de Associação

B. Pôssas N. Ziviani W. Meira Jr. B. Ribeiro-Neto

Departamento de Ciência da Computação

Universidade Federal de Minas Gerais

Belo Horizonte – MG – Brasil

{bavep,nivio,meira,berthier}@dcc.ufmg.br

Resumo

O objetivo deste trabalho é apresentar uma extensão para o modelo vetorial que leva em consideração a correlação entre os termos de uma consulta, através da utilização de regras de associação, uma conhecida técnica de mineração de dados. Na área de recuperação de informação, o modelo vetorial permite recuperar um conjunto de documentos a partir de uma consulta constituída por um conjunto de termos, onde os termos da consulta e dos documentos são considerados vetores em um espaço vetorial. Apesar de décadas de sucesso da modelagem vetorial, até hoje não é conhecido nenhum mecanismo prático e eficiente que seja capaz de levar em consideração as correlações entre os termos da consulta em cada documento da coleção. A novidade deste trabalho está na proposta de um novo método para a computação da correlação entre os termos da consulta. As modificações necessárias são mínimas, e resultados experimentais mostram que o modelo vetorial estendido leva a uma melhoria consistente na qualidade das respostas para todas as coleções avaliadas.

Abstract

The goal of this work is to present an extension to the vector model that accounts for the correlation among query terms, by using association rules, a popular data mining technique. In Information Retrieval, the vector model allows retrieving a set of documents from a term-based query, where both query terms and documents are vectors in a vector space. Although the vector model has been used succesfully for decades, there are no practical and efficient mechanisms that account for correlations among query terms in each document from the collection until now. The novelty of this work is the proposal of a novel method for computing the correlations among query terms. The changes to the original vector model are minimal, and experimental results show that our extended vector model enhances the precision of the results for all the collections evaluated.

1 Introdução

Na área de recuperação de informação um dos modelos de maior sucesso em termos da habilidade para encontrar respostas relevantes julgadas por pessoas é o modelo vetorial. As principais vantagens do modelo vetorial estão relacionadas com o esquema de pesagem de termos, que melhora a qualidade das respostas e a possibilidade de uma uma recuperação parcial, ou seja, documentos que aproximam as condições da consulta podem ser retornados.

No modelo vetorial, um documento é representado por um conjunto de termos extraídos do próprio documento, com um peso associado representando a importância de cada termo. De forma semelhante, uma consulta é modelada como uma lista de termos, com um peso associado à importância de cada termo na consulta. Tanto para a consulta quanto para os documentos, a representação algébrica corresponde a um vetor em um espaço euclidiano. As respostas para consultas submetidas à sistemas que implementam o modelo vetorial correspondem aos documentos com maior valor de similaridade, computado para cada documento e consulta.

Apesar de sua popularidade e sucesso, o modelo vetorial tem a desvantagem de assumir que os termos são independentes, ou seja, ele assume que os termos aparecem independentemente nos documentos da coleção. Essa suposição frequentemente é adotada para simplificação da implementação e conveniência matemática. Claramente, essa suposição é uma simplificação, dado que o padrão de ocorrência de termos em documentos não é independente. Considere, por exemplo, que os termos "redes" e "computadores" são usados em um documento que trata sobre a área de redes de computadores. Frequentemente, neste documento, o aparecimento de uma dessas palavras implica no aparecimento da outra, significando que elas são correlacionadas.

Neste cenário, pode-se presumir que modelos que incluam informação sobre correlações de termos levarão a resultados de qualidade superior. Entretanto, anos e anos de investigação científica têm mostrado que não é tarefa simples utilizar informação de correlações entre termos para melhorar a qualidade dos resultados de forma consistente. De fato, até hoje não se conhece nenhum mecanismo prático e eficiente que seja capaz de levar em consideração as correlações entre os termos da consulta em cada documento da coleção.

Neste trabalho apresentamos uma extensão para o modelo vetorial, que leva em consideração a correlação entre os termos de uma consulta, através da utilização de uma técnica de mineração de dados conhecida como regras de associação [1]. Em contraste com muitos trabalhos relacionados, a incorporação da correlação entre os termos não modifica a representação algébrica do modelo vetorial. Esse fato possibilita a inclusão dessas correlações nos sistemas de recuperação de informação baseados no modelo vetorial, requerendo apenas uma pequena modificação.

Quando uma extensão a um modelo clássico é proposta, tem-se um novo modelo de recuperação de informação que deve ser validado através de experimentação. Para isso, apresentamos uma comparação entre o modelo proposto e o modelo vetorial, utilizando três coleções de referência bem conhecidas da comunidade científica. Os resultados experimentais mostram que esse novo modelo sempre apresenta resultados significativamente melhores para todas as coleções consideradas com um custo computacional adicional aceitável.

O artigo está organizado da seguinte forma. Na próxima seção apresentamos alguns trabalhos correlatos. Na Seção 3 descrevemos o modelo vetorial.Na Seção 4 apresentamos como os termos podem ser correlacionados através das regras de associação. A Seção 5 apresenta uma descrição formal do modelo vetorial estendido que faz uso das correlações entre os termos. Os resultados experimentais e as coleções de referência são descritos na Seção 6. Finalmente, as conclusões e os trabalhos futuros são apresentados na Seção 7.

2 Trabalhos Relacionados

Salton apresentou uma proposição inicial para o modelo em [17, 18], além de diversas avaliações de pesagem de termos através do sistema de recuperação *SMART* desenvolvido na Universidade de Cornell [14, 23, 15]. A principal limitação do modelo vetorial refere-se à não utilização da correlação dos termos durante o cálculo da similaridade entre os documentos e as consultas.

Diversos autores propuseram diferentes métodos para a consideração da correlação durante o processo de recuperação de informação. Raghavan e Yu [12] usam uma análise estatística das diferentes consultas em relação ao conjunto de documentos relevantes e não relevantes levando a uma função de pesagem não linear.

Uma abordagem probabilística do problema da dependência entre termos foi apresentada por Van Rijsberger e Harper [13, 8]. A suposição básica utilizada basea-se no fato que os termos são distribuídos de maneira dependente dentro do espaço de documentos de uma coleção. Entretanto, a determinação das medidas de correlação não é computacionalmente praticável, mesmo para um pequeno número de termos [16].

O trabalho em [22] apresenta uma tentativa bem sucedida de computar as correlações entre os termos a partir dos sistemas automáticos de indexação. O trabalho mostra que os índices de termos podem ser explicitamente representados em um espaço vetorial 2^n-dimencional, onde n corresponde ao número de termos distintos na coleção, de tal forma que as correlações entre os termos fossem incorporadas no modelo vetorial de uma maneira direta. Nesse trabalho, os vetores de termos são representados a partir de um conjunto de vetores ortogonais básicos chamados min-termos, onde cada min-termo representa um padrão de co-ocorrência dos termos dentro dos documentos. O vetor que representa um termo i é encontrado através da soma de todos os min-termos relacionados a i. Apesar desse modelo apresentar uma efetividade maior na recuperação de informação, ele é bastante complexo e caro computacionalmente, pois existem 2^n min-termos, onde n é igual ao número de termos distintos na coleção.

A maioria dos esforços na área de Mineração de Dados tem como foco o descobrimento de conhecimento em banco de dados estruturados, apesar da enorme quantidade de informação *online* que aparece somente em coleções de textos, como as coleções formadas por documentos coletados na *web*. Várias abordagens em mineração de texto tem utilizado marcações associadas aos documentos [5, 6, 4] ou os termos contidos nesses documentos [10]. Entretanto, a frequência de co-ocorrência e/ou a correlação entre os termos, normalmente exploradas pelas técnicas de mineração de dados, não foram utilizadas nos ambientes de recuperação de informação. A novidade do nosso trabalho está relacionada com a proposição de um novo método para a computação das correlações entre os termos das consultas no modelo vetorial através da utilização de uma técnica de mineração de dados conhecida, as regras de associação.

3 Modelagem Vetorial

Na modelagem vetorial, um documento é conceitualmente representado por um conjunto de termos extraídos do próprio documento, com um peso associado representando a importância de cada termo no documento e na coleção como um todo. De forma semelhante, uma consulta é modelada como uma lista de termos, com um peso associado à importância de cada termo na consulta. Tanto para a consulta quanto para os documentos, a representação algébrica dos conjuntos de termos corresponde a um vetor em um espaço euclidiano t-dimensional, onde t é igual ao número de termos distintos na coleção. Como a co-ocorrência entre os termos não é conhecida *a priori*, o modelo vetorial impoê uma simplificação do problema assumindo que os vetores de termos são ortogonais.

O peso de um termo no vetor que representa um documento pode ser determinado de várias formas diferentes. Uma abordagem comum determina o peso de um termo a partir de dois fatores: (i) o número de vezes que um termo i ocorre em um documento d_j e (ii) o número de vezes que o termo ocorre em toda a coleção. Essa abordagem é caracterizada como $tf \times idf$ [3, 21]. Formalmente, o peso de um termo i em um documento d_j pode ser definido por:

$$w_{i,j} = tf_{i,j} \times idf_i = tf_{i,j} \times \log \frac{N}{df_i}\ , \qquad (1)$$

onde N corresponde ao número de documentos na coleção e idf_j refere-se ao inverso da frequência de ocorrência de um termo na coleção e df_i é igual ao número de documentos contendo o termo i. Esse método assinala pesos altos para termos que são encontrados em um pequeno número de documentos , ou seja, ele privilegia os termos ditos raros.

Uma vez que os pesos dos termos foram determinados, é necessário alguma forma de ordenação (*ranking*) para medir a similaridade entre os documentos e a consulta. Uma medida de similaridade comum, conhecida como medida do cosseno, é definida como produto escalar entre os vetores de documentos $\vec{d_j}$ e o vetor $\vec{q}$ que representa a consulta. Essa medida de similaridade é equivalente ao ângulo formado pelo vetor que representa a consulta e pelo vetor que representa um determinado documento. Precisamente, a similaridade entre um documento d_j e uma consulta q é definida como:

$$sim(q, d_j) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}},$$

onde $w_{i,q}$ é o peso do termo i na consulta q, cuja definição é equivalente ao peso de um termo em um documento, ou seja, $w_{i,q} = tf_{i,q} \times idf_i$. Os fatores $|\vec{d_j}|$ e $|\vec{q}|$ são, respectivamente, a norma dos vetores do documento e da consulta. A ordenação dos documentos não é afetada por $|\vec{q}|$, pois seu valor é o mesmo para todos os documentos. O fator $|\vec{d_j}|$ normaliza a medida de similaridade, evitando que documentos maiores, ou seja, que possuam um número maior de termos, sejam privilegiados.

As principais vantagens do modelo vetorial são: (i) o esquema de pesagem de termos melhora a qualidade das respostas; (ii) a modelagem permite uma recuperação parcial, ou seja, documentos que aproximam as condições da consulta podem ser retornados; e (iii) a utilização da medida do cosseno possibilita uma ordenação dos documentos de acordo com o seu grau de similaridade com a consulta. A principal desvantagem do modelo refere-se a simplificação imposta pela consideração da ortogonalidade entre os termos. As correlações dos termos nos documentos não são consideradas a partir dessa simplificação. Na próxima seção apresentamos como as regras de associação podem ser utilizadas para a determinação da correlação entre os termos.

A implementação da medida do cosseno para uma coleção de documentos utiliza uma estrutura de dados conhecida como arquivos invertidos [3, 21]. Um arquivo invertido é tipicamente composto por um vocabulário e um conjunto de listas invertidas. O vocabulário contém todos os termos distintos que aparecem na coleção de documentos e o número df_i de documentos contendo o termo i. Para cada termo do vocabulário existe uma lista invertida que consciste dos identificadores dos documentos d_j contendo o termo e a frequência $tf_{i,j}$ do termo i no documento d_j. Desta forma, as listas invertidas são compostas por pares de valores $< d_j, tf_{i,j} >$, ordenados de acordo com a frequência do termo. Isso significa que documentos com valores altos para $tf_{i,j}$ podem ser encontrados rapidamente.

Para grandes coleções de documentos o custo de avaliar a medida do cosseno pode ser potencialmente alto, porque o algoritmo atribui uma medida de similaridade para cada documento contendo qualquer dos termos da consulta, requerendo a leitura e o processamento de toda a lista invertida para cada termo da consulta. Esta tarefa pode ter custo alto uma vez que alguns dos termos da consulta podem ocorrer em um número grande de documentos presentes na coleção. Uma das técnicas mais efetivas para a computação de uma aproximação da medida do cosseno sem provocar grandes alterações na ordenação final para cada consulta é proposta em [11]. Ela utiliza limites (*thresholds*) para um rápido reconhecimento de quais documentos influem fortemente na ordenação final para reduzir custos de processamento da consulta. A partir desses limites, se existe um conjunto suficientemente grande de documentos candidatos com altos valores de similaridades parciais com relação a consulta, não é necessário considerar as pequenas similaridades parciais, uma vez que elas não promoveriam uma alteração significativa na ordenação final. Consultas são avaliadas usando apenas 2% da memória usada pela implementação padrão sem degradação na efetividade da resposta obtida. Ambos, acessos em disco e tempo de CPU podem também ser drasticamente reduzidos através por um projeto das listas invertidas que suportem essa técnica.

4 Associação entre Termos

O objetivo desta seção é demonstrar como as correlações entre os termos podem ser computadas de forma automática a partir dos índices dos termos. A forma de correlação utilizada baseia-se na ocorrência simultânea dos termos em um conjunto de documentos, obtida através das associações entre os termos.

A determinação das correlações, ou associações, de termos é um problema semelhante à mineração de regras de associação e os algoritmos utilizados são nosso ponto de partida [9]. Antes de apresentar as idéias principais do algoritmo de determinação de associação de termos, precisamos apresentar uma breve notação.

Seja $T = \{\mathrm{k}_1, \mathrm{k}_2, ..., \mathrm{k}_t\}$ o vocabulário de uma coleção de documentos D, ou seja, o conjunto de todos os termos distintos t que aparecem nos documentos em D. Existe uma ordenação ente os termos do vocabulário que é baseada na ordem lexicográfica dos termos, tal que $\mathrm{k}_i < \mathrm{k}_{i+1}$, para $1 \leq i \leq t-1$. Definimos um $n-term-set$ s como um conjunto ordenado de n termos distintos, tal que $s \subseteq T$ e a ordem dos termos segue a ordenação mencionada. Seja $S = \{\mathrm{s}_1, \mathrm{s}_2, ..., \mathrm{s}_{2^t}\}$ o conjunto de todos os 2^t $term-sets$ que podem aparecer em todos os documentos em D. Cada documento d_j de D é caracterizado por um vetor nesse espaço de $term-sets$.

As associações entre termos são utilizadas para determinar quais termos frequentemente co-ocorrem em documentos. Por exemplo, considere um coleção de documentos D composta pelos documentos d_j, tal que $1 \leq j \leq N$, como na Figura 1. Existem cinco termos distintos nesse conjunto de documentos. Desta forma, temos $T = \{A, C, D, T, W\}$, resultando em no máximo 31 $term-sets$. A coleção D contém 6 documentos, ou seja, $D = \{(A,C,T,W), (C,D,W), (A,C,T,W), (A,C,D,W), (A,C,D,T,W), (C,D,T)\}$. Utilizaremos essa coleção para apresentar algumas definições que são a base para a associação de termos.

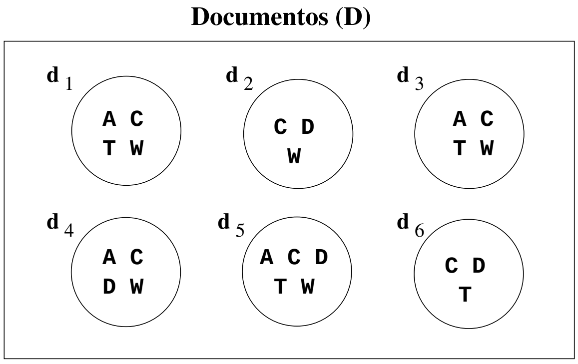

Figura 1: Uma coleção de documentos

Cada $term - set$ s_i, $1 \leq i \leq 2^t$, possui uma lista invertida com os identificadores dos documentos nos quais ele aparece, de acordo com a estrutura descita na Seção 3. Definimos também a frequência ds_i de um $term - set$ s_i como o número de ocorrências de s_i em D, ou seja, o número de documentos d_j, tal que $d_j \in D$, $1 \leq j \leq N$ e $s_i \subseteq T$. A frequência ds_i de um $term - set$ s_i é equivalente ao tamanho da sua lista invertida associada ($| \ l_i \ |$). Finalmente, definimos s_i como um $term - set$ frequente se sua frequênia ds_i for maior ou igual a um limite dado.

A Figura 2 mostra todos os $term - sets$ frequentes para a coleção de exemplo e todas as suas frequências. A frequência mínima utilizada para a seleção dos $term - sets$ frequentes foi igual a 50%, ou seja, para um $term - set$ ser considerado frequente ele deve aparecer em pelo menos 3 dos 6 documentos da coleção. Percebemos que o número de $term - sets$ frequentes, apesar de potencialmente muito grande, é geralmente pequeno e é uma função da frequênia mínima utilizada e das características da coleção de documentos. Desta forma, é possível variar o número de $term - sets$ frequentes através de uma mudança da frequência mínima empregada.

Term−sets Frequentes

Frequência (ds)	Term−sets
100% (6)	C
83% (5)	W, CW
67% (4)	A, D, T, AC, AW, CD, CT, ACW
50% (3)	AT, DW, TW, ACT, ATW, CDW, CTW, ACTW

Figura 2: $Term - sets$ frequentes

Definimos também o fator de coesão cs_i, $1 \leq i \leq 2^t$, como uma medida de correlação entre um $term - set$ e seus subconjuntos. Quanto maior é o fator de coesão, mais representativo é o $term - set$ na coleção. Calculamos o fator de coesão cs_i de um $n - term - set$ s_i $(n > 1)$ como:

$$cs_i = \frac{\sum_{j \in \delta(i)} \frac{ds_i}{ds_j}}{n} \ ,$$

onde $\delta(i)$ é o conjunto dos n $(n-1)-term-sets$ que são subconjuntos de s_i. Formalmente, $\delta(i) = \{s_j | s_j \subset s_i, | s_j |= n-1\}$. Por exemplo, vamos considerar os fatores de coesão para os $term-sets$ ATW e CDW de nossa coleção de exemplo. Os fatores são calculados como se segue:

$$cs_{CDW} = \frac{\frac{ds_{CDW}}{ds_{CD}} + \frac{ds_{CDW}}{ds_{CW}} + \frac{ds_{CDW}}{ds_{DW}}}{3} = 0.78$$

$$cs_{ATW} = \frac{\frac{ds_{ATW}}{ds_{AT}} + \frac{ds_{ATW}}{ds_{AW}} + \frac{ds_{ATW}}{ds_{TW}}}{3} = 0.92$$

Podemos percever que, apesar dos $term-sets$ possuirem a mesma frequência (0.5), seus fatores de coesão são um tanto diferentes. O fator de coesão para ATW é maior porque ele ocorre na maioria dos documentos em que seus subconjuntos aparecem. Como veremos na próxima subseção, a junção dessa medida com a frequência ds_i nos possibilita criar uma medida de relevância de um $term-set$ em toda coleção.

Em resumo, uma associação de termos é composta por um $term-set$, sua frequência e seu fator de coesão. Podemos dividir a geração das associações de termos em duas fases:

1. Enumerar todos os $term-sets$ frequentes. O custo computacional dessa fase de enumeração cresce exponencialmente com o número de termos distintos, pois existem, potencialmente, 2^n $term-sets$ frequentes para n termos.

2. Determinar as associações de termos a partir dos $term-sets$ frequentes.

Na primeira fase, determinamos os $(n+1)-term-sets$ a partir dos $n-term-sets$ da seguinte forma: O processo começão através da determinação de todos os pares de $term-sets$ $(s_i, s_j), 1 \leq i, j \leq 2^t$, que possuam os mesmos $n-1$ termos. Um novo $term-set$ é determinado pela união desses conjuntos $(s_i \cup s_j)$, criando um $(n+1)-term-set$ s_{new}. A lista de documentos onde s_{new} aparece pode ser facilmente determinada através de uma interseção das listas dos dois conjuntos concatenados ($l_{new} = l_i \cap l_j$). Finalmente, se $| l_{new} |$ é maior que a frequência mínima, s_{new} é definido como frequente.

Por exemplo, seja F_3 o conjuntos de todos os $3-term-sets$ para nossa coleção de exemplo, onde F_3={{A C T},{A C W}, {A T W},{C D W},{C T W}}. A fase de enumeração geraria um $4-term-set$, {A C T W}, baseado nos $term-sets$ {A C T} e {A C W}. Primeiro verificamos se todos os outros subconjuntos do novo $term-set$ gerado são também frequentes (neste caso, {A T W} and {C T W}) também são frequentes. Essa condição para um novo $term-set$ ser frequente foi originalmente proposta no algoritmo *Apriori*, apresentado em [2], como uma estratégia de poda.

A determinação das associações de termos pode ser considerada simples, uma vez que ela consciste do cálculo dos subconjuntos de cada $term-set$ frequente e do seu fator de coesão. O custo para a determinação dos $term-sets$ frequentes é exponencial. Para uma consulta contendo n termos, o número potencial de $term-sets$ frequentes equivale a 2^n. Entretanto, sabemos que o número de correlações entre termos de consulta é pequeno para uma grande coleção de documentos, assim como para as coleções de referêcias a serem usadas para a validação do modelo a ser proposto. Por exemplo, para termos 2^n $term-sets$ frequentes são necessárias todas as possíveis combinações de termos em uma grande quantidade de documentos.

Neste trabalho propomos a utilização das associações de termos para estabelecer as correlações entre termos de uma coleção de documentos, e a utilização dessas correlações para melhorar a efetividade de recuperação através da proposição de um novo modelo de recuperação de informação, como será discutido a seguir. Essa abordagem é funcionalmente equivalente a apresentada no trabalho [22](modelo vetorial generalizado), que também determina correlações baseada na ocorrência simultânea dos termos na coleção.

5 Modelagem Vetorial Estendida

Nesta seção apresentamos as características fundamentais do modelo proposto, chamado modelo vetorial estendido, que leva em consideração as correlações entre os termos. A seguir introduzimos os componentes básicos do modelo e apresentamos um algoritmo para a implementação de um mecanismo de busca baseado no modelo vetorial estendido e nas informações de indexação.

As consultas e os documentos ainda são representados por vetores, assim como no modelo vetorial original. Entretanto, os componentes desses vetores não são mais termos, mas sim conjuntos de termos, os $term-sets$. O ponto de partida do modelo vetorial estendido são as associações de termos determinadas para a coleção de documentos, os termos das consulas e a estrutura de indexação utilizada no modelo vetorial padrão.

5.1 Peso dos $Term-sets$

A determinação dos pesos associados aos $term-sets$ de um documento, ou de uma consulta, também pode ser caracterizada de acordo com a abordagem $tf \times idf$ baseada no número de ocorrências de um $term-set$ em um documento, na consulta e na coleção com um todo. Formalmente, o peso de um $term-set$ s_i em um documento d_j é definido como:

$$w_{i,j} = sf_{i,j} \times cs_i \times ids_i = sf_{i,j} \times cs_i \times \log \frac{N}{ds_i}\ , \qquad (2)$$

onde N corresponde ao número total de documento da coleção, $sf_{i,j}$ ao número total de ocorrências do $term-set$ s_i no documento d_j, e ids_i a frequência de ocorrência inversa do $term-set$ s_i na coleção. $sf_{i,j}$ generaliza $tf_{i,j}$ no sentido que conta o número de ocorrências de um $term-set$ s_i dentro de um documento d_j (onde s_i pode conter somente um termo). ids_i também carrega a mesma semântica de idf_i, mas leva em consideração a cardinalidade dos $term-sets$ da seguinte forma: $term-sets$ compostos por muitos termos geralmente possuem uma baixa frequência, resultando em altas frequências inversas. Multiplicando esse valor pelo fator de coesão podemos distinguir entre $term-sets$ cuja frequência seja igual a partir da sua significância. Como no modelo vetorial, essa estratégia associa altos pesos para $term-sets$ que aparecem em um conjunto pequeno de documentos, ou seja, $term-sets$ raros resultam em altos pesos.

Quando utilizamos somente os $term-sets$ de tamanho 1, ou seja, somente os termos, as Equações 1 e 2 tornam-se equivalentes, uma vez que o fator de coesão para esses $term-sets$ é igual a 1. Por isso dizemos que o modelo proposto estende o modelo vetorial. Veremos que a mesma observação se aplica para a função do cálculo de similaridade.

5.2 Cálculo da Similaridade

Como documentos e consultas são representados por vetores, aplicamos um estratégia para o cálculo da similaridade entre a consulta e documentos a partir de uma pequena variação da medida do cosseno. A similaridade entre um documento d_j e uma consulta q é definida como:

$$sim(q, d_j) = \frac{\sum_{s \in F_q} w_{s,j} \times w_{s,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}},$$

onde $w_{s,j}$ corresponde ao peso associado ao $term - set$ s_i no documentos d_j, $w_{s,q}$ ao peso associado ao $term - set$ s_i na consulta q, $w_{i,q}$ ao peso do termo i in na consulta q, $w_{i,j}$ ao peso do termo i no documento d_j, e F_q ao conjunto de todos os $term - sets$ frequentes tal que todos sejam um subconjunto da consulta ($s \subseteq q$). Se forem considerados somentes os $term - sets$ de tamanho 1, o cálculo da similaridade torna-se equivalente ao apresentado para o modelo vetorial. Observe que a normalização (ou seja, os fatores no denominador) é feita usando-se somente os termos que compõem a consulta e os documentos, e não os $term - sets$. Essa simplificação reduz consideravelmente o custo computacional para o cálculo da medida de similaridade, pois o cálculo da norma dos documentos baseado nos $term - sets$ implicaria na geração de todos os $term - sets$ para a coleção. Apesar dessa adaptação a medida do cosseno, a normalização continua válida, uma vez que o objetivo de penalizar documentos grandes ainda persiste. Nossos resultados experimentais confirmam a validade dessa simplificação no cálculo da similaridade entre documentos e consultas.

5.3 Mecanismo de Busca

A estratégia utilizada para o cálculo da medida de similaridade também é uma extensão da implementação usual para o modelo vetorial, conforme descrito na Figura 3.

1. Crie e inicialize uma estrutura de acumuladores (A).
2. Para cada termo t_i na consulta.
   (a) Recupere o índice do termo t_i do disco.
   (b) **Para cada** par $[d_j, f_{i,j}]$ na lista invertida do termo t_i **faça**:
      i. **Se** $A_j \notin A$ **então**
         $A_j = 0$,
         $A = A + \{A_j\}$
      ii. $A_j = A_j + w_{i,j} * w_{i,q}$ de Eq. (1)
3. Enumere os $term - sets$.
   (a) Ao final da geração de um $term - set$ s_i.
      i. **Para cada** par $[d, f_{i,j}]$ na lista invertida do conjunto s_i **faça**:
         $A_j = A_j + w_{i,j} * w_{i,q}$ de Eq. (2)
4. Divida cada acumulador A_j pela norma do documento d_j.
5. Recupere os k maiores valores para os acumuladores
   e retorne os documentos correspondestes.

Figura 3: Algoritmo de busca do modelo vetorial estendido

Estamos considerando que a estrutura de indexação descrita na Seção 3. Primeiro cria-se uma estrutura para o armazenamento dos valores (acumuladores) para as similaridades parciais dos documentos, calculadas para cada conjunto de termos em um documento d_j. Depois, para cada termo na consulta, recupera-se toda a sua lista invertida e para cada entrada nesta lista calcula-se a sua contribuição para a similaridade dos documentos (medida do cosseno). Esses passos correspondem a forma original de cálculo da medida do cosseno para o modelo vetorial. Nos passos 2.(b)ii. e 3.(a)i. da Figura 3 aplicamos a estratégia de poda apresentada em [11] (veja Seção 3), que não é mostrada nesta descrição do algoritmo. A estratégia de poda utiliza uma técnica de avaliação que promove um rápido reconhecimento de quais documentos influem fortemente na ordenação final para reduzir custos, nos permitindo descartar a avaliação completa das listas dos $term - sets$ que não irão contribuir para uma modificação da ordenação final dos documentos. Depois de terminada a computação de cada uma das listas dos termos, enumeramos todos os $term - sets$ e determinamos a sua contribuição para um dos documentos presentes na sua lista (passo 3.(a)i.). Depois de avaliados todos os $term - sets$, normalizamos o valor das similaridades dos documentos, através da sua divisão pela norma. O passo final consciste em selecionar os k maiores valores para os acumuladores e retornar os documentos correspondentes.

A implementação do modelo proposto, assim como a implementação do modelo vetorial, basea-se em uma técnica para a computação aproximada da medida do cosseno, como descrito na Seção 3.

6 Resultados Experimentais

O ambiente de execução dos experimentos é composto por uma máquina AMD k6-II 500 MHz com 256 MBytes de memória principal. Nesta seção apresentamos as coleções de referência utilizadas e os resultados obtidos através da comparação entre o modeo vetorial e o modelo proposto.

6.1 Coleções de Referência

Para a avaliação do modelo proposto, utilizamos três coleções de referência: CACM, CISI e MEDL. A Tabela 1 apresenta as principais características das coleções de documentos avaliadas, cujas descrições são apresentadas a seguir.

	Coleção		
	CACM	CISI	MEDL
# Documentos	3.204	1.460	1.033
# Termos Distintos	14.581	11.164	13.352
# Consultas	52	76	30
# Médio Termos por Consulta	10,13	27,70	11,00
# Médio de Documentos Relevantes por Consulta	15,31	40,97	23,20
Tamanho (MBytes)	2,08	2,13	1,07

Tabela 1: Propriedades das coleções de referência

CACM

Os documentos contidos na coleção *CACM* [7] são compostos por todos os 3.204 artigos publicados no periódico *Communications of the CACM* no período de 1958 a 1979. Além do texto dos documentos, a coleção também provê informação estruturada em sub-campos, a saber: (1) nome dos autores; (2) data de publicação; (3) radicais para palavras do título e do resumo; (4) categorias derivadas de uma classificação hierarquica; (5) referências diretas entre os artigos; e (6) número de co-citações para cada par de artigos. A coleção *CACM* inclui também um conjunto com 52 consultas. Associada a cada consulta existe a sua formulação através de lógica *booleana*, bem como a relação de documentos relevantes associados a cada consulta. Como as consultas são muito específicas, o número médio de documentos relevantes para cada consulta é aproximadamente 15.

CISI

Os documentos contidos na coleção *CISI* foram selecionados de uma coleção anterior criada por Small [19] no Instituto de Ciência da Informação. Os documentos selecionados referem-se a ciência da informação. Os documentos da coleção *CISI* incluem três tipos de sub-campos: (1) nome dos autores; (2) radicais para palavras do título e do resumo; e (3) número de co-citações para cada par de artigos. Essa coleção inclui também um conjunto com 76 consultas, sendo 35 expressas através de lógica *booleana*, e 41 expressas em linguagem natural, bem como a relação dos documentos relevantes associados a cada consulta. Como as consultas são muito gerais, o número médio de documentos relevantes para cada consulta é aproximadamente 41.

MEDL

Os documentos contidos na coleção *MEDL* são compostos por 1033 resumos de artigos médicos encontrados na Biblioteca Nacional de Medicina (*Medline*). Além do resumo dos artigos, cada documento contém os seguintes sub-campos: (1) número de acesso na *Medline*; (2) autor; (3) título; (4) fonte de informação; (5) assuntos de maior e menor importância; (6) referências; e (7) citações. Essa coleção inclui também um conjunto com 30 consultas expressas através de lógica *booleana*, bem como a relação dos documentos relevantes associados a cada consulta. O número médio de documentos relevantes para cada consulta é aproximadamente igual a 23.

6.2 Comparação entre o Modelo padrão e o Modelo Estendido

Como mencionado, cada coleção citada acima tem um conjunto de consultas e, para cada consulta, os documentos relevantes (selecionados por especialistas) são indicados. As medidas padrão de revocação (*recall*) e precisão (*precision*) foram utilizadas para a comparação do desempenho da efetividade de recuperação para o modelo vetorial padrão e para o modelo vetorial estendido. Revocação é definida como uma fração do conjunto de todos documentos relevantes, composta por documentos relevantes que aparecem na resposta gerada pelo algoritmo sendo avaliado. Precisão é definida como uma fração dos documentos recuperados pelo algoritmo sendo avaliado, composta por aqueles documentos que são relevantes, ou seja, que também pertencem ao conjunto de documentos relevantes indicado pelos especialistas. Essas medidas podem ser computadas para uma consulta em especial, mas como medida de comparação utilizamos a média entre todas as consultas para cada coleção.

6.2.1 Medida de Efetividade

A Figura 4 apresenta uma comparação de desempenho da efetividade de recuperação de informação entre os dois modelos para as coleções *CACM*, *CISI* e *MEDL*, respectivamente. O modelo proposto retorna resultados sempre melhores que o modelo vetorial, de forma consistente para todas as coleções avaliadas.

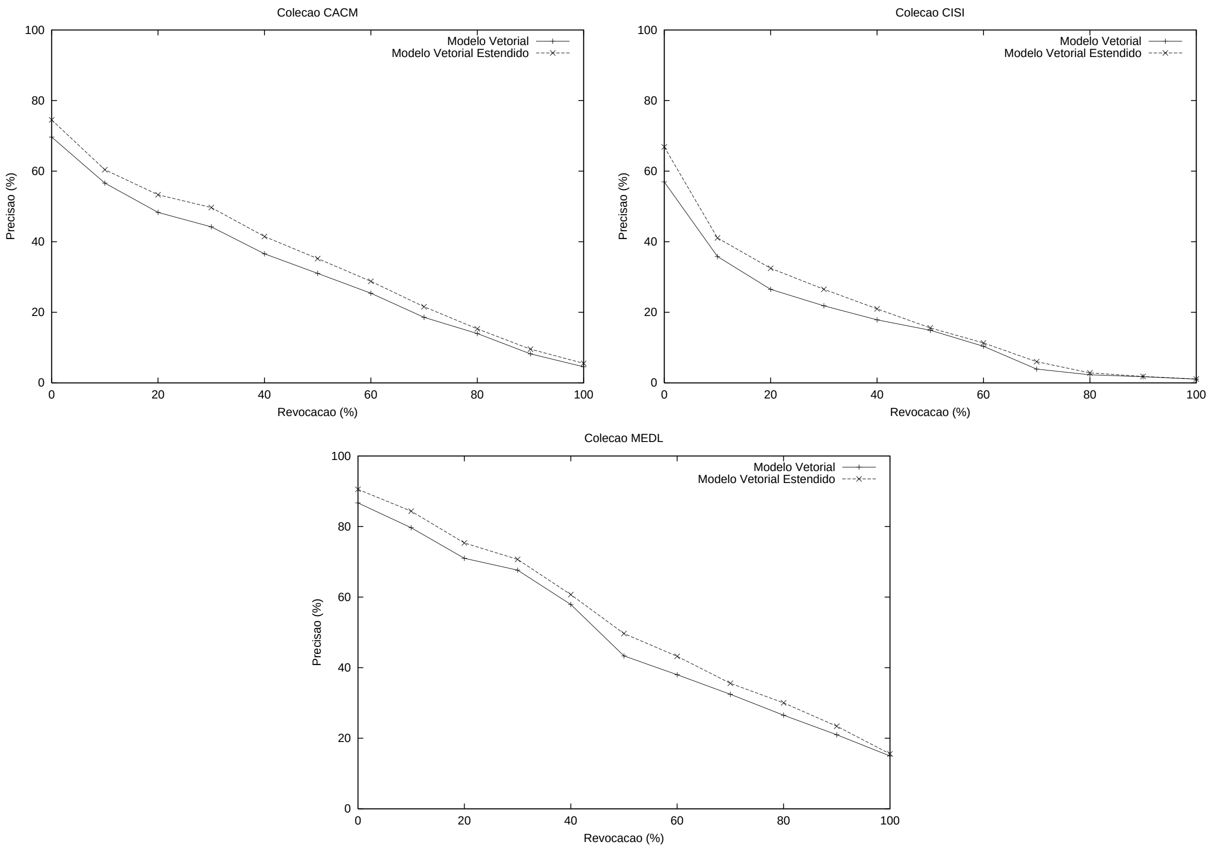

Figura 4: Curva de revocação x precisão

Observamos que, embora os ganhos de precisão não sejam expressivos, eles estão sempre presentes. Ademais, o ponto mais importante a notar é que a utilização de correlações em nosso modelo vetorial estendido jamais leva a degradação da qualidade das respostas. Essa diferença pode ser explicada pela relação entre o tamanho do vocabulário e da coleção avaliada, além do número de termos nas consultas. Ainda, é de se esperar que ganhos mais significativos na precisão média serão observados se considerarmos somente consultas que efetivamente contém termos correlatos.

Nosso modelo vetorial estendido é o primeiro modelo de recuperação de informação que utiliza correlações entre termos de forma eficiente e que produz melhoria consistente na qualidade das respostas, quando avaliado com diferentes coleções de referência.

6.2.2 Medida de Desempenho

Nesta seção comparamos o modelo vetorial estendido com o modelo vetorial a partir dos tempos de resposta para cada uma das consultas, com o objetivo de avaliar sua em termos de recursos computacionais. Uma das principais limitações dos modelos existentes que levam em consideração a correlação entre os termos está relacionada com a grande demanda de recursos computacionais. Muitos desses modelos não podem ser utilizados para coleções de documentos de tamanho médio a grande. A adição da determinação dos conjuntos frequentes de termos e o respectivo cálculo de similaridade para esses conjuntos de termos, não afetam significativamente o tempo de execução das consultas. O acréssimo médio no tempo total de execução das consultas no modelo vetorial estendido ficou entre 26% e 29% em comparação com o modelo vetorial padrão para as três coleções de referência consideradas.

Apesar do tempo de execução médio do modelo proposto se aproximar bastante do tempo médio do modelo vetorial, observamos que consultas com muitos termos não apresentam bons resultados. Isso se deve a etapa de enumeração dos $term-sets$ frequentes que potencialmente pode gerar 2^n $term-sets$, para uma consulta com n termos. Entretanto, o número de correlações entre termos é muito pouco frequente para as coleções avaliadas. Por exemplo, aproximadamente 75% do total de termos distintos aparecem em menos do que 5 documentos.

Figura 5: Tempo de resposta x Tamanho das Consultas

A Figura 5 apresenta a curva do tempo de resposta médio do modelo vetorial estendido a medida que o número de termos na consulta aumenta. A utilização de mecanismos de poda, ou seja, mecanismos que eliminam a necessidade de avaliação de todos os $term-sets$ frequentes, tornam a utilização do modelo vetorial estendido uma realidade prática. Entrentanto, para sistemas de recuperação de informação na *Web*, onde o tamanho médio das consultas é pequeno, cerca de dois termos na média [20], a utilização do modelo vetorial estendido pode apresentar um desempenho muito bom.

7 Conclusões e Trabalhos Futuros

Apresentamos um novo modelo para a recuperação de informação em coleções de textos que leva em consideração as correlações entre os termos através de sua ocorrência simultânea nos documentos. Mostramos que o novo modelo sempre apresenta resultados melhores mantendo um custo computacional adicional aceitável. A determinação das correlações entre os termos pela utilização dos conjuntos frequentes enumerados pelos algoritmos para a geração de regras de associação permite uma extensão direta dos sistemas de recuperação de informação baseados no modelo vetorial. Em contraste com muitos trabalhos relacionados, a incorporação da correlação entre os termos não modifica a representação algébrica do modelo vetorial.

Demonstramos, através das curvas de revocação e precisão, o aumento na efetividade de recuperação do modelo proposto em comparação com o modelo vetorial original para todas as coleções de referência avaliadas. A viabilidade em termos computacionais, medidos através do tempo de resposta para as consultas, permite a utilização do modelo proposto em ambientes cujo tamanho das coleções de textos sejam maiores.

Pretendemos continuar esse trabalho através da avaliação do modelo proposto para coleções maiores e para coleções formados por documentos *Web*. A definição de algoritmos de reconhecem rapidamente quais documentos podem influenciar de uma maneira mais significativa na ordenação final para o modelo vetorial estendido pode reduzir os custos de avaliação das consultas, através do processamento das partes das listas invertidas dos $term - sets$, que possuam altos valores de similaridade parcial.

Referências

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference Management of Data*, pages 207-216, Washington, D.C., May 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *The 20th International Conference on Very Large Data Bases*, pages 487-499, Santiago, Chile, September 1994.

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

[4] R. Feldman, Y. Aumann, A. Amir, W. Klösgen, and A. Zilberstien. Maximal association rules: A new tool for mining for keyword co-occurrences in document collections. In *Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, August 1997.

[5] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In *1st International Conference on Knowledge Discovery and Data Mining*, Montreal, Canada, August 1995.

[6] R. Feldman and H. Hirsh. Exploiting background information in knowledge discovery from text. In *Journal of Intelligent Information Systems*, 1996.

[7] E. Fox. Characterization of two new experimental collections in computer and information science containing textual and bibliographical concepts. Techical report, 1983. `http://www.ncstrl.org`.

[8] D. J. Harper and C. J. Van Rijsbergen. An evaluation of feedback in document retrieval using co-occurrence data. In *Journal of Documentation*, volume 34, pages 189–216, 1978.

[9] R. Agrawaland T. Imielinski and A. Swami. Mining associations between sets of items in massive databases. In *In Proceedings of the 1993 ACM-SIGMOD Int l Conf on Management of Data*, pages 207-216, 1993.

[10] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In *Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, August 1997.

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

[12] V. V. Raghavan and C. T. Yu. Experiments on the determination of the relationships between terms. In *ACM Transactions on Databases Systems*, volume 4, pages 240–260, 1979.

[13] C. J. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. In *Journal of Documentation*, volume 33, pages 106–119, 1977.

[14] G. Salton. *The SMART retrieval system – Experiments in automatic document processing*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. In *Information Processing and Management*, volume 24(5), pages 513–523, 1988.

[16] G. Salton, C. Buckley, and C. T. Yu. An evaluation of term dependencies models in information retrieval. In *The 5th ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1982.

[17] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. In *Journal of the ACM*, volume 15(1), pages 8–36, January 1968.

[18] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. In *Journal of Documentation*, volume 29, pages 8–36, 1973.

[19] H. Small. The relationship of information science to the social sciences: A co-citation analysis. In *Information Processing and Management*, volume 17(1), pages 39–50, 1981.

[20] E. Veloso, E. Moura, P. Golgher, A. Silva, R. Almeida, A. Laender, B. Ribeiro-Neto, and N. Ziviani. Um retrato da web brasileira. In *XXVI Seminário Integrado de Software e Hardware*, Curitiba, PR, Brazil, July 2000.

[21] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes. Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, 1999.

[22] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector space model in information retrieval. In *The 8th ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 18–25, New York, USA, 1985.

[23] C. T. Yu and G. Salton. Precision weighting – an effective automatic indexing method. In *Journal of the ACM*, volume 23(1), pages 76–88, January 1976.