3
Recuperação de Informações Textuais

"Tudo deveria se tornar o mais simples possível, mas não simplificado."

Albert Einstein

Sistemas tradicionais de indexação costumam utilizar-se de termos-índice, que podem ser palavras ou grupos de palavras chave para realizar a recuperação da informação. A maneira mais usual de se recuperar uma informação indexada é obter os documentos no qual o termo chave buscado se encontra. O conceito básico envolvido nessa tarefa é o de que a semântica do documento e da informação precisa ser naturalmente expressa através do conjunto de termos-índice (BAEZA-YATES, 1999). Essa tarefa se torna problemática na medida em que o casamento entre a requisição do usuário e cada documento é calcado em um espaço impreciso de termos, considerando que grande parte da semântica da requisição do usuário ou do documento é perdida quando expressa na forma de conjunto de palavras. Nesse contexto o problema central se torna dizer que documentos são ou não relevantes para uma dada consulta e a melhor forma de fazer isso é apresentando o resultado de forma ordenada, evitando que o usuário necessite se estender aos demais documentos caso encontre o que queira ou perceba que os documentos retornados estão se afastando do que ele realmente procura. Pode-se dizer que o algoritmo de ordenação dos resultados é parte crucial dos sistemas de recuperação de informação no que tange a sua qualidade.

São três os modelos clássicos de recuperação de informação: Booleano, vetorial e probabilístico. Ao modelo booleano se aplica a teoria dos conjuntos podendo-se utilizar inclusive lógica nebulosa (BOY, 1986) e (CROSS, 1994) como forma de extensão ao modelo clássico; ao modelo probabilístico se aplica a teoria das probabilidades, como teorema de Bayes e as redes de inferência (WILSON, 2000); e ao modelo vetorial se aplica a teoria da álgebra linear ou ainda modelos mais complexos como índice de semântica latente (LETSCH, 1997) e (LITTMAN, 1997) e redes neurais (WANG, 2002). Segundo (BAEZA,-YATES, 1999) um modelo de recuperação de informação pode ser formalmente caracterizado da forma a seguir:

Modelo geral de recuperação de informações

Um modelo de recuparação de informações é representado pela quádrupla $[D, Q, F, R(q_i, d_j)]$ onde:

(1) D é o conjunto de representações para os documentos eu um coleção;

(2) Q é o conjunto de representações para as necessidades de informação do usuário, conhecido na literatura como *query* e aqui chamado de consulta;

(3) F é *framework* para o modelo de representação dos documentos, consultas e seus relacionamentos;

(4) R é uma função de ordenação (*ranking*), que associa um número real com uma consulta $q_i \in Q$ e uma representação de documento $d_j \in D$. Dessa forma R define a ordem em que os documentos aparecem como resultado de uma consulta submetida pelo usuário.

Os modelos clássicos consideram cada documento como sendo um conjunto de palavras que o representam, que no presente trabalho serão denominadas termos-índice. Considerando que a semântica de um documento esteja alicerçada nos termos que o compõe pode-se dizer que os termos-índice resumem o conteúdo de um documento. Porém muito da semântica é dada pela ordem em que esses termos se sucedem, além da organização estrutural do documento como um todo, contudo, é possível se dizer ao menos que os termos-índice são capazes de fornecer indícios sobre o tema (ou temas) sobre o qual os documentos versam. Serão apresentadas algumas formas de se capturar esses indícios com menor ou maior grau de potencial semântico.

Todos os modelos são baseados na seguinte definição (BAEZA-YATES, 1999):

Representação de um documento por um vetor de termos-índice

Seja t um número de termos-índice em um sistema e k_i um termo-índice genérico. Considera-se $K = (k_1, \ldots, k_t)$ o conjunto de todos os termos-índice, que para efeito de terminologia será chamado de léxico. Um peso $w_{i,j} \geq 0$ está associado a cada um dos termos-índice k_i do documento d_j. Dessa forma, um termo-índice pertencente ao léxico e não presente a um documento d_j tem seu peso associado $w_{i,j} = 0$. Com isso é possível representar um documento d_j através de um vetor $\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$. Além disso, é possível caracterizar uma função g_i como sendo aquela que retorna o peso associado a determinado termo-índice k_i em um vetor t-dimensional, ou seja, $g_i(d_j) = w_{i,j}$.

Cabe ressaltar que a forma mais simples de se representar um documento usando um vetor de termos-índice é através da introdução de todos os termos do documento no vetor que o representa. Porém, é intuitivo que alguns termos caracterizem melhor do que outros, determinado documento. Por exemplo, termos que aparecem em todos os documentos, ou até mesmo em uma grande quantidade deles, acabam por se tornar incapazes de serem utilizados com o objetivo de diferenciação entre documentos, portanto podem ser desconsiderados no vetor representação. Em termos sintáticos, os mais comuns a serem incluídos na categoria dos termos menos úteis à caracterização de um documento são os artigos, advérbios, adjetivos e conjunções. A decisão de que termos devem ser úteis para tal caracterização é uma tarefa não muito trivial e se torna ainda mais complexa se considerarmos a existência de importância relativa entre as palavras, isto é, quais palavras, presentes no léxico, são mais importantes para cada documento. Modelos que tratam a recuperação da informação de forma não-binária (algumas dessas formas serão abordadas mais a frente) são capazes de pesar a importância das palavras e a técnica mais conhecida para isso é a que utiliza as métricas de freqüência do termo[19] (*tf*) e freqüência documental inversa[20] (*idf*), que será abordada mais adiante.

[19] Do inglês, *term frequence*
[20] Do inglês, *inverse document frequence*

Para se recuperar a informação desejada é necessário buscar a associação existente entre a consulta proposta e os documentos armazenados. Com o objetivo de se quantificar o grau de associação é comum a utilização de métricas de distância ou de similaridade (DUDA, 2000). Algumas dessas métricas serão apresentadas no capítulo 4. Em cada um dos modelos vistos a seguir será abordada a medida de similaridade mais utilizada e adequada ao respectivo modelo.

Os modelos discutidos aqui são base para o emprego de técnicas de inteligência computacional e, em particular o modelo vetorial será utilizado como subsídio dentro do presente trabalho.

3.1
Modelo Booleano

O modelo booleano é um simples modelo de recuperação baseado na teoria da álgebra booleana. Dessa forma esse modelo se torna bastante intuitivo e suas consultas podem ser especificadas por expressões booleanas que possuem semântica extremamente precisa baseada em relações de pertinência e operações lógicas para composições das expressões supramencionadas. Por outro lado, por ter sua estratégia de recuperação baseada em um critério binário, ou seja, o documento é ou não relevante à consulta apresentada sem graduação de importância, esse modelo se assemelha mais a um modelo de recuperação de dados do que propriamente recuperação de informação. Formalmente, o modelo booleano tem sua função similaridade caracterizada pela seguinte definição (BAEZA-YATES, 1999):

Os pesos dos termos-índice no modelo booleano assumem valores binários, ou seja, $w_{t,j} \in \{0,1\}$ e uma q nesse modelo é simplesmente uma expressão convencional booleana. Seja $\vec{q}_{dnf}$ a forma normal disjuntiva da consulta q, isto é, representada com disjunções de vetores conjuntivos. Além disso, seja $\vec{q}_{cc}$ um dos componentes conjuntivos da disjunção $\vec{q}_{dnf}$. A similaridade de um documento d_j a uma consulta q é dada por:

$$sim(d_j, q) = \begin{cases} 1, se\ \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall k_t, g_t(\vec{d}_t) = g_t(\vec{q}_{cc})) \\ 0, caso\ contrário \end{cases}$$

Se $sim(d_j, q) = 1$ então o modelo booleano prediz que o documento d_j é relevante à consulta q e caso contrário prediz que não é relevante.

Para aprofundamento em operações booleanas, forma normal disjuntiva e outras questões de lógica booleana e teoria de conjuntos consulte (MENDELSON, 1997).

3.2
Modelo Probabilístico

O modelo probabilístico (ROBERTSON, 1976) tenta abordar o problema de recuperação da informação sob a ótica da teoria da probabilidade e para isso usa a idéia básica de que, a partir de uma consulta proposta exista um sub-conjunto de documentos que contenham todos os documentos relevantes e apenas estes, conhecido como resposta ideal à consulta. A resposta ideal não é previamente conhecida e o processo utiliza uma etapa inicial que permite gerar uma descrição probabilística preliminar do conjunto resposta ideal, que é usado para recuperar o primeiro conjunto de documentos. A partir desse ponto interações com o usuário são iniciadas com o propósito de aumentar a descrição probabilística da resposta ideal. Isso é feito com o usuário indicando quais documentos são relevantes à sua consulta. O sistema usa essas informações para melhorar a descrição da resposta ideal. A repetição do processo refina o sistema de forma a aumentar a proximidade das respostas com a resposta ideal.

O modelo se baseia na premissa de que a probabilidade de relevância dependa apenas das representações dos documentos e da consulta e também na premissa de que existe um subconjunto dos documentos armazenados que o usuário prefira como resposta a determinada consulta.

De maneira mais formal pode-se definir a similaridade no modelo probabilístico da forma a seguir (BAEZA-YATES, 1999):

Similaridade no modelo probabilístico de recuperação de informações

Os pesos dos termos-índice no modelo probabilístico podem assumir valores binários, ou seja, $w_{i,j} \in \{0,1\}$ e $w_{i,q} \in \{0,1\}$. Uma consulta q é um subconjunto de termos-índice. Seja R o conjunto de documentos conhecidos (ou designados na etapa inicial do processo) como sendo relevantes e $\bar{R}$ o complemento de R, isto é, o conjunto de documentos não-relevantes. Seja $P(R|\vec{d}_j)$ a probabilidade de que o documento d_j seja relevante à consulta q e $P(\bar{R}|\vec{d}_j)$ a probabilidade de que o documento d_j seja não-relevante à consulta q. A similaridade $sim(d_j, q)$ do documento d_j à consulta q é dada por:

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\bar{R}|\vec{d}_j)}$$

ou usando a regra de Bayes:

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) \times P(R)}{P(\vec{d}_j|\bar{R}) \times P(\bar{R})}$$

Para aprofundamento em teoria da probabilidade e teorema de Bayes consulte (MEYER, 1976) ou (MONTGOMERY, 2006).

3.3
Modelo vetorial

Reconhecendo as limitações do uso de pesagem binária o modelo vetorial (SALTON, 1968) propõe um casamento parcial entre documentos e consultas, ou seja, estabelece um grau de similaridade entre cada documento armazenado e uma consulta proposta, através do uso de pesos não-binários aos termos-índice. Por

esse motivo a resposta a uma determinada consulta é a ordenação dos documentos armazenados em ordem decrescente de similaridade à respectiva consulta.

Assim como um documento, uma consulta, nesse modelo, também é representada como um vetor t-dimensional. Dessa maneira, o grau de similaridade entre os documentos e a consulta proposta pode ser obtido através de uma relação entre os vetores $\vec{d_j}$ e $\vec{q}$. Esta correlação é normalmente calculada pelo cosseno do ângulo formado entre os vetores no espaço t-dimensional (LAY, 2005), conforme a figura 15.

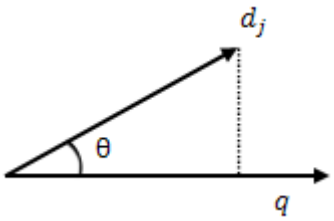

Figura 15 – Cosseno de θ, adotado como medida de similaridade.

A similaridade, dada pelo cosseno de θ, pode ser formalmente definida por (BAEZA-YATES, 1999):

Similaridade no modelo vetorial de recuperação de informações

Os pesos dos termos-índice no modelo vetorial, $w_{i,j}$, assumem valores positivos e não-binários associados com um par (k_i, d_j) e uma consulta q nesse modelo também é associada a um peso. Seja $w_{i,q}$ o peso associado com o par $[k_i, q]$, onde $w_{i,q} \geq 0$. Então o vetor consulta $\vec{q}$ é definido como $q = (w_{1,q}, w_{2,q}, \cdots, w_{t,q})$, onde t é o número total de termos-índice presentes no léxico. Sendo ainda o vetor documento d_j representado por $\vec{d_j} = (w_{1,j}, w_{2,j}, \cdots, w_{t,j})$, a similaridade de um documento d_j a uma consulta q é dada por:

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

onde $|\vec{d_j}|$ e $|\vec{q}|$ são respectivamente as normas do vetor documento e vetor consulta.

Obviamente os valores de similaridade variam entre zero e um e, dessa forma, representam o grau de associação entre os documentos armazenados e a consulta proposta possibilitando a recuperação da informação através de casamentos parciais entre documento e consulta, diferente do modelo booleano, onde o casamento era exato. Devido ao fato de todo documento armazenado possuir certo grau de similaridade com qualquer consulta proposta, mesmo que esse grau seja igual a zero, pode ser interessante se estabelecer um limiar para apresentação dos documentos resultantes da consulta.

Com base no conceito de similaridade é possível modelar o problema de recuperação de informação como um problema de agrupamento (*clustering*). Para isso, caracteriza-se a coleção de documentos como sendo um conjunto C, e consulta proposta como sendo a especificação de um conjunto A, tendo-se em mente que os documentos que pertencerem a A sejam aqueles recuperados pela consulta submetida. Então o problema passa a ser determinar que documentos pertencem ao conjunto A e quais não pertencem e, do ponto de vista do agrupamento, isso é realizado a partir da definição de quais características melhor descrevem os elementos do conjunto A. Para responder essa questão as técnicas de agrupamento se baseiam em dois aspectos complementares. O primeiro é a relação de similaridade entre os elementos de um mesmo grupo, no caso, evidenciado pelas características que melhor descrevem os documentos pertencentes ao conjunto A, ou similaridade intra-grupo. O segundo é a diferenciação entre os elementos de grupos distintos, no caso, evidenciado pelas características que melhor distinguem os documentos pertencentes ao grupo A do restante dos documentos pertencentes à coleção C, dissimilaridade inter-grupo. O modelo vetorial estabelece valores para tratar os dois aspectos. O primeiro é a **freqüência do termo** (tf), ou seja, quantas vezes o termo k_i aprece no documento d_j. Esse parâmetro é usado para quantificar a importância de um termo no conteúdo do documento. A segunda é uma medida de freqüência inversa do termo k_i perante todos os documentos, **freqüência documental inversa** ou simplesmente fator idf. Essa medida se baseia na idéia de que um termo que aparece em muitos documentos não é útil para diferenciá-los. Para obtenção de bons resultados de agrupamento é necessário que seja feito um balanceamento das

duas medidas. A fórmula de balanceamento mais utiliza é apresentada na definição a seguir (BAEZA-YATES, 1999), porém algumas variações desta já foram sugeridas (SALTON, 1988).

Balanceamento entre freqüência do termo e freqüência documental inversa

Seja N o número total de documentos de um *corpus* e n_i o número de documentos em que o termo k_i aparece. Seja a $freq_{i,j}$ o número de vezes que o termo k_i aparece (é mencionado) no documento d_j. Então a freqüência normalizada do termo k_i no documento d_j é dada por:

$$f_{i,j} = \frac{freq_{i,j}}{freq_{l,j}},$$

onde l é o termo que mais aparece no documento j. Se um termo k_i não aparece no documento d_j então $f_{i,j} = 0$. Além disso, a freqüência documental inversa para o termo k_i é dada por:

$$idf_i = \frac{\log N}{n_i}.$$

O balanceamento é dado por:

$$w_{i,j} = f_{i,j} \times \frac{\log N}{n_i}.$$

Pode-se utilizar, de forma mais simples, o mesmo princípio do balanceamento da freqüência do termo pela freqüência documental inversa do termo sem o uso da normalização e do logaritmo. Dessa forma a freqüência do termo seria dada pelo número de vezes que o termo aparece no documento sobre o número de termos do documento e a freqüência documental pelo número de documentos que contém o termo sobre o número total de documentos do *corpus*. Porém essa abordagem tende a trazer resultados menos precisos.

3.4
Extensões aos modelos

O modelo booleano pode ser modificado empregando-se duas outras abordagens, o modelo booleano estendido e o modelo de conjuntos *fuzzy*.

O modelo booleano estendido (SALTON, 1983) funciona como uma combinação dos modelos booleano e vetorial, pois agrega ao modelo booleano a funcionalidade de casamento parcial entre consulta proposta e documentos armazenados. Pelo modelo clássico um documento que contenha apenas k_x ou que contenha apenas k_y é tão irrelevante quanto outro que não contenha nenhum dos dois, quando submetida a consulta $k_x \wedge k_y$. O modelo estendido trabalha com valores normalizados de pesos. A normalização pode ser feita com base nos fatores *tf-idf*, por exemplo, e dessa forma os vetores resultantes não ficam sempre ortogonais, sendo relevantes ou não-relevantes conforme o casamento exato, do modelo booleano clássico. Dessa maneira é possível se estabelecer um casamento parcial entre consulta proposta e documentos armazenados. Para aprofundamento dessa estratégia consulte (SALTON, 1983) ou (BAEZA-YATES, 1999).

A outra abordagem alternativa para o modelo booleano clássico é a utilização de conjuntos *fuzzy*. Mais uma vez a idéia é fugir do aspecto puramente binário existente no modelo clássico. Pela teoria de conjuntos nebulosos um elemento não está associado a um conjunto apenas com grau de pertinência zero ou um e sim varia dentro desse intervalo, $[0,1]$. Como resultado o casamento entre consulta proposta e documentos armazenados passa critérios aproximados ou vagos e o casamento passa a poder ser estabelecido de forma parcial, e não mais simplesmente como relevante ou não-relevante. Para aprofundamento desse estratagema consulte (OGAWA, 1991) ou (BAEZA-YATES, 1999).

Com relação ao modelo probabilístico a extensão sugerida passa pela utilização de Redes Bayesianas. Uma é conhecida como rede de inferência[21] (TURTLE, 1990) e a outra é uma extensão dessa última conhecida como rede de crença[22] (RIBEIRO-NETO, 1996). Para aprofundamento da utilização de ambos os modelos em recuperação de informações consulte (BAEZA-YATES, 1999).

Para o modelo vetorial, genericamente conhecido como modelo algébrico, destacam-se três estratégias principais, modelo generalizado do espaço vetorial (WONG, 1985), modelo de indexação por semântica latente (FURNAS, 1988) e modelo de redes neurais.

[21] Do inglês, *inference network*.
[22] Do inglês, *belief network*.

O modelo generalizado do espaço vetorial (WONG, 1985) parte do princípio de que os vetores de termos-índice sejam linearmente independentes, porém pertencentes a um espaço vetorial não necessariamente ortogonal. Esse princípio adota como fundamento básico a idéia de que a co-ocorrência de termos-índice dentro de um documento em uma coleção induz à existência de dependência entre esses termos, diferentemente do princípio, introduzido por simplificação, de independência entre termos-índice nos modelos clássicos. A visão clássica era de que os termos-índice seriam mutuamente independentes, ou seja, conhecendo-se o peso $w_{i,j}$ associado ao par (k_i, d_j) nada é possível afirmar sobre o peso $w_{i+1,j}$ associado ao par (k_{i+1}, d_j). Porém, intuitivamente, termos como *data* e *mining*, por exemplo, utilizados em documentos que cubram a área de mineração de dados, claramente evidenciam um caráter de correlação e conseqüentemente de dependência. Fatos como esses abrem um campo também para a utilização de processamento de linguagem natural (KAO, 2004), mais especificamente nesse caso para o estudo de termos compostos, ou, como conhecido amplamente na literatura, colocações (veja seção 2.2.2.1 – Identificação de colocações). A correlação existente entre termos é utilizada, em combinação com os pesos, no cálculo da ordenação dos documentos resultantes de uma consulta proposta (BAEZA-YATES, 1999).

A premissa principal do modelo de semântica latente (FURNAS, 1988) é a de que a idéia passada por um texto está mais relacionada a conceitos descritos nele do que aos termos-índice usados na descrição, e dessa forma, uma consulta deve ser mapeada em um espaço dimensional inferior, que está associado com os conceitos, e não com os termos-índice. A argumentação utilizada para defender esse ponto de vista é a de que a recuperação de informações em um espaço de dimensões reduzido deve apresentar melhores resultados. Isso faz sentido se considerarmos que existem muitas palavras distintas que traduzem um mesmo conceito. Porém, podem existir casos em que uma mesma palavra possa traduzir mais de um conceito, dependendo do contexto em que se encontre e do emprego utilizado. O principal exemplo desse último caso é a ambigüidade do sentido das palavras, abordado no presente trabalho. Existe uma grande dificuldade em se mapear no mundo concreto algo que está implícito em um universo extremamente abstrato como é o caso dos conceitos que são representados pelas palavras. Na

tentativa de modelar essa transição do universo abstrato para um universo mais concreto o modelo de semântica latente introduz uma formulação interessante para o problema de recuperação de informações através da utilização de decomposição de valores singulares, visando a eliminar os ruídos pela remoção das redundâncias (BAEZA-YATES, 1999).