Capítulo 6

Para ranquear os resultados de uma busca é necessário algum sistema de pontuação. Um documento que menciona um termo da consulta mais vezes é mais interessante, logo, deve receber um pontuação maior. Essa pontuação de um termo t em um documento d será chamado de peso de t em d. Pensando em consultas de texto livre, que é uma forma mais comum na *web*, seria fácil computar a pontuação de um documento, sendo a soma dos pesos dos termos da consultapresentes no documento. A forma mais natural de atribuir este peso é pela frequência do termo, denotada tft,d, mas pode-se usar qualquer função de ponderação. Essa forma é conhecida como modelo saco-de-palavras (do inglês, *bag of words model*) que leva em consideração apenas o número de ocorrências dos termos, mas não a ordem destes. Mesmo que a ordem das palavras divirja entre dois documentos, se as frequências dos termos forem similares, esses devem ser similares.

Alguns termos podem ser muito frequentes na coleção, logo, não terão muita influência no ranqueamento dos documentos, sendo necessária algum tipo de atenuação. Utilizando a frequência do termo na coleção, um termo muito presente em poucos documentos e um termo pouco presente em muitos documentos teriam a mesma frequência, então, a atenuação não surtiria efeito. Portanto, é comum usar a frequência de documentos (dft), ou seja, o número de documentos da coleção em que o termo está presente. Assim, definimos frequência inversa do documento (idf):

idft = log

Um termo que está em todos os documentos terá idf = 0. Quanto mais raro o termo na coleção, maior será o seu idf. Multiplicando tft,d por idft temos uma balanceamento entre a presença de um termo em um documento e a sua raridade na coleção. Um documento que diz muito sobre um assunto que aparece pouco na coleção deve ser mais relevante para este assunto. Definimos então o balanceamento tf-idf:

tf-idft,d = tft,d · idft

Trabalhando com o documento como um vetor de pesos para os termos do dicionário, temos a pontuação de um documento d em relação à uma consulta q:

Score(q,d) = tf-idft,d

Esta representação de documentos como vetores em um espaço vetorial é conhecida como modelo vetorial (do inglês, *vector space model*) e pode ser usado em outras operações de recuperação de informação. Além disso, outras funções de ponderação podem ser usadas no lugar de tf-idf. Para cada documento d temos, então, um vetor com componentes sendo as ponderações. Assim, o espaço vetorial terá um eixo para cada termo.

Dois documentos serão similares se tiverem vetores com componentes proporcionalmente parecidas, mesmo que as componentes de um documento sejam maiores do que a do outro. Para computar esta similaridade, é comum o uso do cosseno entre os dois vetores:

sim(d1,d2) = /

Temos a divisão do produto vetorial entre os vetores sobre o produto de seus comprimentos euclidianos. Esta divisão é necessária para normalizar os resultados. Se os vetores já estiverem normalizados, apenas o produto vetorial é necessário.

A coleção será representada pelo conjunto destes vetores: uma matriz MxN, sendo M o número de termos do dicionário e N o número de documentos. Da mesma forma que um documento, uma consulta pode ser representada por um vetor. Sendo assim, a pontuação de um documento em relação a uma consulta será:

score(q,d) = /