NATURAL LANGUAGE PROCESSING

AULA 3

Prof. Luciano Frontino de Medeiros

CONVERSA INICIAL

Esta aula tem o objetivo de apresentar o tema da recuperação de informação, que pressupõe uma coleção de documentos com base na qual se quer efetuar uma consulta, considerando-se as palavras ou termos dentro dos documentos para identificar a relevância de um documento sobre o outro. Em seguida, são apresentados os modelos de recuperação de informação. A atribuição de pesos mediante a frequência que um termo aparece também é tratada em um tema específico, bem como o esquema de cálculo de frequência de termos mais utilizados, o TF-IDF. Por fim, uma aplicação do esquema TF-IDF é realizada visando maior compreensão dos conceitos e fórmulas apresentados nesse material.

TEMA 1 – RECUPERAÇÃO DA INFORMAÇÃO (RI)

Em um contexto massivo de dados não estruturados, como o que se apresenta com a web hoje em dia, a busca eficiente de informações torna-se um desafio cada vez mais pertinente. No caso de pesquisas envolvendo documentos de texto, a análise do conteúdo relativo a palavras ou termos de forma quantitativa torna-se um procedimento altamente relevante para a organização e a descoberta de informações.

A recuperação de informação (RI) é uma tarefa da área do processamento de linguagem natural que tem o objetivo de encontrar documentos relevantes para as necessidades de informação de um usuário (Russell; Norvig, 2004, p. 813). Um exemplo típico é o processo de busca na www: com base em um conjunto de termos expressos em uma consulta, um engenho de busca (*search engine*) retorna um conjunto de documentos ordenados por relevância.

Conforme Baeza-Yates e Ribeiro-Neto (2011, p. 1),

> Recuperação de Informação lida com a representação, armazenamento, organização de e acesso à itens de informação tais como documentos, páginas web, catálogos *online*, registros estruturados ou semiestruturados, objetos multimídia. A representação e organização dos itens de informação devem ser tais que possam proporcionar aos usuários o fácil acesso à informação de seu.

RI se refere à localização de materiais (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfazem uma necessidade de informações em grandes coleções geralmente armazenadas em computadores. Definido dessa maneira, RI costumava ser uma atividade na qual

apenas certas pessoas se engajavam: bibliotecários de referência, paralegais e pesquisadores profissionais semelhantes. Agora, centenas de milhões de pessoas se envolvem na recuperação de informações todos os dias, quando usam um mecanismo de pesquisa na *web* ou pesquisam em sua caixa de *e-mails* (Manning; Raghavan; Schütze, 2009, p. 38).

Um sistema de RI pode ser composto por:

1. **Uma coleção de documentos:** o que pode ser tratado como documento – parágrafo, um texto completou um conjunto de textos;
2. **Linguagem de consulta**: conforme um conjunto de regras para a definição da consulta que especifica aquilo que o usuário quer pesquisar;
3. **Um conjunto de resultados**: subconjunto coletado com base na busca efetuada relevante aos termos colocados na consulta;
4. **Apresentação dos resultados**: que pode ser na forma de uma simples lista ordenada por relevância, ou um mapa tridimensional em cores com os documentos relacionados em rede (Russell; Norvig, 2004, p. 813).

Em RI, o processo de *modelagem* é uma atividade complexa que tem o objetivo de produzir uma *função de classificação* (*ranking*), ou seja, uma função que atribua escores (*rank*) para documentos com referência a uma consulta fornecida. Esse processo consiste de duas tarefas principais:

a. A concepção de um *framework* lógico para representar documentos e consultas; e
b. A definição de uma concepção de classificação que calcule o escore de cada documento com relação a uma consulta especificada.

O *framework* lógico é geralmente baseado em *conjuntos, vetores* ou *distribuições de probabilidade*. Isso afeta diretamente o cálculo do escore dos documentos, os quais são então utilizados para ordenar os documentos retornados em resposta a uma consulta fornecida (Baeza-Yates; Ribeiro-Neto, 2011, p. 57).

Um modelo de RI é uma quádrupla $[\boldsymbol{D}, \boldsymbol{Q}, \mathcal{F}, R(q_i, d_j)]$ em que (Baeza-Yates; Ribeiro-Neto, 2011, p. 58):

1. $\boldsymbol{D}$ é um conjunto composto de visões lógicas ou representações dos documentos em uma coleção;
2. $\boldsymbol{Q}$ é um conjunto composto de visões lógicas ou representações das necessidades de informações do usuário;

3. $\mathcal{F}$ é um *framework* para modelar representações de documentos, consultas e seus relacionamentos, bem como relações booleanas, vetores e operações de álgebra linear, e ainda espaços amostrais e distribuições de probabilidade;
4. $\boldsymbol{R}(q_i, d_j)$ é uma função de classificação que associa um número real com uma representação de consulta $q_i \in \boldsymbol{Q}$ e um documento $d_j \in \boldsymbol{D}$. Tal classificação ou escore define um ordenamento entre os documentos com referência à consulta q_i.

A RI clássica engloba três modelos: *booleano*, *vetorial* e *probabilístico*. Enquanto no modelo booleano, os *termos de indexação (index terms)* não possuem qualquer peso, nos modelos vetoriais e probabilísticos eles se constituem simplesmente elementos de um conjunto. Nos dois últimos modelos, os *index terms* têm pesos associados que permitem melhorar o processo de classificação (Baeza-Yates; Ribeiro-Neto, 2011, p. 61).

1.1 Conceitos básicos

1.1.1 Termo de indexação ou palavras-chave

É uma palavra ou um grupo consecutivo de palavras em um documento. De forma geral, um termo de indexação é qualquer palavra em uma coleção. Mas em uma forma mais restrita, os termos de indexação ou palavras-chaves constituem um grupo pré-selecionado de palavras que representam um conceito chave ou tópico em um documento. Enquanto a abordagem geral é própria dos desenvolvedores de engenhos de busca, a abordagem restrita se adequa aos bibliotecários e cientistas da informação (Baeza-Yates; Ribeiro-Neto, 2011, p. 62).

1.1.2 Vocabulário

É o conjunto de todos os termos de indexação em uma coleção. Definindo-se $V = \{k_1, \ldots, k_t\}$ como um vocabulário, os k_i são os termos de indexação que estão contido em V, sendo que o tamanho de V é t. Um vocabulário é um elemento importante de uma coleção, pois contém todos os termos de indexação. À medida que a coleção cresce, o tamanho do vocabulário também tende a crescer.

1.1.3 Documentos e representações de consulta

Com base na definição de vocabulário vista anteriormente, sejam k_l, k_m e k_n os termos de indexação que ocorrem em um documento d_j. Então, diz-se que o padrão $[k_l, k_m, k_n]$ de coocorrência dos termos podem ser observados. Para um vocabulário $V = \{k_1, \ldots, k_t\}$, o número total de padrões de coocorrência em documentos da coleção é igual a 2^t. Por exemplo, o padrão (1, 0, ..., 0) indica apenas a presença do termo k_1, não havendo nenhum outro termo de indexação em um dado documento. Já o padrão (1, 1, ..., 1) indica a presença de todos os termos de indexação. Cada um desses padrões de co-ocorrência de termos de indexação é denominado de *componente conjuntivo de termo*. Com o documento d_j, está associado um componente conjuntivo de termo único $c(d_j)$, o qual descreve quais termos estão presentes no documento e quais não estão.

De forma análoga, a uma consulta q é associado um componente conjuntivo de termo $c(q)$, que descreve por sua vez os termos de indexação que ocorrem na consulta e quais não ocorrem. As consultas e documentos são representados de forma simples pelo componente conjuntivo de termos que refletem os termos que ele contém. É comum encontrar também a denominação *bag of words* ou *saco de palavras* (Baeza-Yates; Ribeiro-Neto, 2011, p. 62).

1.1.4 Matriz termo-documento

É a forma matricial de representação da ocorrência dos termos de indexação em um documento. Isto permite a quantificação da frequência do termo dentro do documento. Por exemplo, uma matriz termo-documento contendo três termos de indexação e dois documentos pode ser representada da forma abaixo, em que cada $f_{i,j}$ significa a frequência do termo k_i no documento d_j:

$$\begin{array}{c} . \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{array}$$

Dessa forma, levando-se em conta o uso da frequência de ocorrência para quantificar a relação termo-documento, fornece-se mais informação do que simplesmente registrar a ocorrência ou não em um documento.

1.1.5 *Stopwords*

Em se tratando de coleções muito grandes, o que é comum de ocorrer na *web*, pode ser recomendado reduzir o conjunto de palavras-chave representativas. Isso pode ser conseguido a partir da eliminação de *stopwords* (como, por exemplo, artigos, advérbios e preposições).

1.1.6 *Stemming*

Outra forma de se reduzir o conjunto de termos de indexação, essa tarefa busca identificar o radical comum de palavras – por exemplo, diferentes conjugações gramaticais de verbos e também a identificação de grupos de substantivos comuns.

TEMA 2 – MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO

Os modelos de recuperação de informação mais utilizados são três: *modelo booleano*, *modelo vetorial* e *modelo probabilístico*. A seguir, cada um desses modelos é apresentado em detalhes.

2.1 Modelo booleano

O *modelo booleano* por palavras-chave é característico dos sistemas mais antigos de RI. Nesse modelo de busca, se a palavra consultada ocorre no documento, seu valor de retorno é verdadeiro, ou falso se a palavra não ocorre. Para o caso de mais palavras-chave a serem consultadas, a linguagem de consulta utilizada é a linguagem de expressões booleanas (Russell; Norvig, 2004, p. 814). Para expressões complexas, as expressões booleanas preveem o uso de operadores lógicos AND, OR e NOT. Assim, um documento tem valor de relevância quando a expressão de busca é avaliada como verdadeira. Por exemplo, a seguinte expressão de busca pode ser utilizada no âmbito do processamento de linguagem natural:

processamento AND *linguagem* AND *natural*

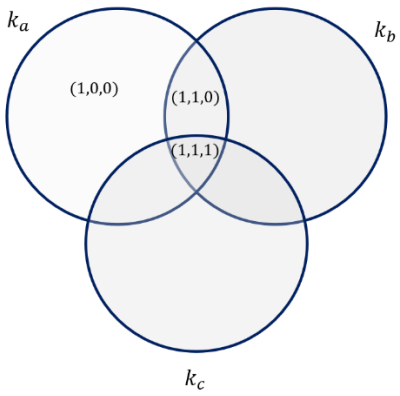
Caso seja necessário fazer a busca de documentos que contenham além das palavras mencionadas, o termo *inteligência artificial*, uma possível expressão seria:

(*processamento* AND *linguagem* AND *natural*) OR (*inteligência* AND *artificial*)

Apesar de ser um modelo simples de implementar, com apenas um valor booleano, fica difícil de se fazer um ordenamento pela relevância de documentos. O modelo booleano recebeu considerável atenção no passado e foi adotado por diversos sistemas comerciais de referência bibliográfica (Baeza-Yates; Ribeiro-Neto, 2011, p. 64).

Para uma definição mais formal, em um modelo booleano todos os elementos da matriz termo-documento são 1 (um), para indicar a presença do termo no documento e 0 (zero) para indicar a ausência. Uma consulta q é uma expressão booleana sobre os termos de indexação. Por exemplo, $[q = k_a \wedge (k_b \vee \neg k_c)]$. Sendo dada tal consulta, um componente conjuntivo de termo que satisfaz tais condições é denominado um componente conjuntivo de consulta $c(q)$. Representando o exemplo anterior na forma normal disjuntiva (contendo agora o operador disjunção $\vee$), a consulta pode ser representada por $[q = (1,1,1) \vee (1,1,0) \vee (1,0,0)]$, conforme a Figura 1. Note que em todas as triplas aparece sempre o elemento k_a (o índice 1 em primeiro lugar nas triplas). Como aparece a disjunção na consulta, k_b aparece na primeira e na segunda tripla, ou não aparece na terceira tripla, quando k_c.

Figura 1 – As três formas conjuntivas para a consulta $[q = k_a \wedge (k_b \vee \neg k_c)]$.



Fonte: Baeza-Yates; Ribeiro-Neto, 2011, p. 65.

Considere a seguir um documento d_j que contém os termos k_a e k_c, mas não contém k_b. A consulta anterior não conseguirá recuperar esse documento, visto que a tripla $(1, 0, 1)$, ou seja, o componente conjuntivo, não está presente na instanciação da consulta.

2.2 Modelos vetoriais

Os modelos baseados em estatística de palavras compõem a maior parte dos sistemas de RI e utilizam estatísticas de contagem de palavras, reconhecendo que o modelo booleano é bastante limitado para uma representação significativa. No modelo vetorial, os pesos utilizados para os termos de indexação são não binários: é possível utilizar um *grau de similaridade* entre cada documento armazenado no sistema e na consulta do usuário. Os documentos podem ser ordenados em ordem decrescente de similaridade no modelo vetorial, considerando que os termos podem casar (*match*) parcialmente. O efeito resultante no ranqueamento é a maior precisão das respostas (Baeza-Yates; Ribeiro-Neto, 2011, p. 77).

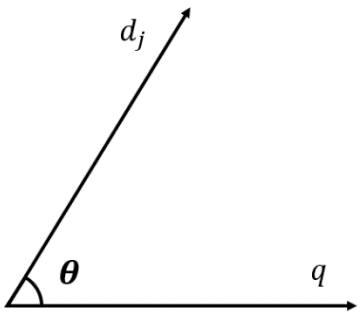
O modelo vetorial utiliza pesos w_ij associados com o par termo-documento (k_i, d_j). Os valores são não negativos e não binários. Os termos de indexação são assumidos como sendo mutuamente independentes e representados como vetores unitários de dimensão t, em que t é o número total de termos de indexação. Portanto, a representação vetorial de um documento d_j e uma consulta q é dada por:

$$\overrightarrow{d_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$$

$$\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{tq})$$

Os pesos $w_{tq} \geq 0$ são os pesos associados com o par termo-consulta (k_i, q).

Figura 2 – O ângulo θ adotado como a medida de similaridade entre o documento d_j e a consulta q



Fonte: Baeza-Yates; Ribeiro-Neto, 2011, p. 78.

O modelo vetorial avalia o grau de similaridade de um documento d_j com uma consulta q de acordo com a correlação entre os vetores d_j e q. Essa correlação pode ser quantificada, por exemplo, utilizando-se do cosseno do ângulo entre os vetores. Denominando a medida de similaridade entre o documento d_j e a consulta q como $sim(d_j, q)$, tem-se a expressão da álgebra linear para o cálculo do cosseno do ângulo (Baeza-Yates; Ribeiro-Neto, 2011, p. 78):

$$sim(d_j, q) = \frac{\overrightarrow{d_j}.\vec{q}}{|\overrightarrow{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (1)$$

Nessa expressão, $|\overrightarrow{d_j}|$ e $|\vec{q}|$ são as normas dos documentos e vetores de consulta respectivamente; e $\overrightarrow{d_j}.\vec{q}$ é o produto interno ou escalar entre os dois vetores. Dessa forma, o valor de $sim(d_j, q)$ está entre 0 e 1. O valor 0 (zero) indicar que não há correlação entre os vetores, enquanto o valor 1 (um) indica a correlação máxima entre eles.

2.3 Modelos probabilísticos

O modelo probabilístico é baseado em um *framework* de probabilidades de termos de indexação nos documentos. A base para esse modelo é descrita a seguir. Dada uma consulta de um usuário, existe um conjunto de documentos que contém exatamente os documentos relevantes e nenhum outro. Esse conjunto de documentos é assumido como sendo um conjunto de resposta ideal. Com base em um conjunto de resposta ideal, é possível recuperar então os documentos relevantes.

Portanto, o processo de consulta torna-se um processo de especificação das propriedades de um conjunto de resposta ideal. O problema consiste em não se saber exatamente quais são essas propriedades. O que se tem em mãos é que existem termos de indexação cuja semântica pode ser utilizada para caracterizar tais propriedades. Como tais propriedades são desconhecidas no momento da consulta, é necessário antecipar quais seriam. Um palpite inicial permitiria gerar uma descrição preliminar das probabilidades de um conjunto ideal de resposta, o qual seria então utilizado para recuperar um primeiro conjunto de documentos (Baeza-Yates; Ribeiro-Neto, 2011, p. 80).

Para aprimorar a descrição das probabilidades, um sistema pode calcular a partir da experiência do usuário na execução de várias consultas, refinando o processo ao longo do tempo. Dessa forma, a descrição tenderá a se aproximar daquela do conjunto ideal de respostas.

De maneira formal, dada uma consulta de um usuário q e um documento d_j em uma coleção, o modelo probabilístico tenta estimar a probabilidade de que o usuário irá encontrar o documento d_j de interesse. O modelo irá assumir que essa probabilidade depende da consulta e da representação do documento apenas (ou seja, a informação que está disponível no sistema). O modelo também assume que existe um conjunto de todos os documentos preferidos pelo usuário como resposta à consulta q executada. O conjunto ideal de respostas é denominado R e deve maximizar a probabilidade total de relevância para o usuário. Em contrapartida, aqueles documentos que não estão no conjunto de predição são denominados de *não relevantes* (Baeza-Yates; Ribeiro-Neto, 2011, p. 80).

Em um modelo probabilístico, uma consulta q é um subconjunto de termos de indexação. Um documento d_j é representado por um vetor de pesos binários indicando a presença ou ausência do termo de indexação, assim expresso:

$$\overrightarrow{d_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$$

em que $w_{ij} = 1$ no caso do termo k_i ocorrer no documento d_j, e $w_{ij} = 0$ em caso contrário. Se R é o conjunto de documentos conhecidos (ou inicialmente pressupostos) como sendo relevantes para o usuário, para a consulta q, e $\bar{R}$ é o complemento de R (ou seja, o conjunto de documentos não relevantes), então $P(R|\overrightarrow{d_j}, q)$ é a probabilidade de que o documento d_j com representação $\overrightarrow{d_j}$ é relevante para a consulta q. Adicionalmente, $P(\bar{R}|\overrightarrow{d_j}, q)$ é a probabilidade de que o documento d_j não seja relevante para a consulta q. A partir daqui, a definição de similaridade $sim(d_j, q)$ do documento d_j para a consulta q é definida como a razão:

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

Pode-se derivar, a partir da regra de Bayes e utilizando logaritmos, a expressão a seguir que permite calcular o grau de similaridade:

$$sim(d_j, q) \cong \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{p_{iR}}{1 - p_{iR}}\right) + \log(\frac{1 - q_{iR}}{q_{iR}}) \quad (2)$$

em que

$$p_{iR} = P(k_i|R, q)$$

$$q_{iR} = P(k_i|\bar{R}, q)$$

Assim, p_{iR} representa a probabilidade de que o termo de indexação k_i esteja presente num documento aleatoriamente selecionado do conjunto de documentos relevantes R (em relação à consulta q); e q_{iR} é a probabilidade de que o termo de indexação não esteja no conjunto R.

O uso de uma *tabela de contingência* pode auxiliar na obtenção de uma expressão que facilita o cálculo das probabilidades. Seja N o número de documentos em uma coleção e n_i o número de documentos que contenham o termo de indexação k_i. Se R é o número total de documentos relevantes para a consulta q, como apontado anteriormente, então pode-se definir r_i como o número de documentos relevantes que contenham o termo k_i. Pode-se gerar uma tabela de contingência que orientará as próximas definições (Tabela 1).

Tabela 1 – Tabela de contingência

Caso	Relevante	Não relevante	Total
Documentos que contenham k_i	r_i	$n_i - r_i$	n_i
Documentos que não contenham k_i	$R - r_i$	$N - n_i - (R - r_i)$	$N - n_i$
Todos os documentos	R	$N - R$	N

A partir daqui, dois casos podem ocorrer. O primeiro se dá *quando a informação estiver disponível* para qualquer consulta, as probabilidades podem ser escritas como:

$$p_{iR} = \frac{r_i}{R}$$

$$q_{iR} = \frac{n_i - r_i}{N - R}$$

A equação (2) pode ser reescrita, portanto, como sendo

$$sim(d_j, q) \cong \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{r_i(N - n_i - R + r_i)}{(R - r_i)(n_i - r_i)}\right) \quad (3)$$

No caso de envolver o cálculo de valores muito pequenos, é conveniente adicionar a constante 0.5 para que a fórmula produza valores bem-comportados:

$$sim(d_j, q) \cong \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)}\right) \quad (4)$$

Essa equação não pode ser utilizada sem a estimativa de r_i e R. Isso leva ao segundo caso, que envolve o cálculo da similaridade *quando a informação de relevância não estiver disponível.* Dessa forma, assume-se $R = r_i = 0$, e a equação (3) pode ser então simplificada:

$$sim(d_j, q) \cong \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \quad (5)$$

Essa fórmula pode levar a comportamentos anômalos no cálculo das probabilidades (obtenção de valores negativos, o que torna difícil a interpretação dos resultados). Uma fórmula alternativa retira o termo n_i do numerador:

$$sim(d_j, q) \cong \sum_{k_i \in q \wedge k_i \in d_j} \log\left(\frac{N + 0.5}{n_i + 0.5}\right) \quad (5.1)$$

TEMA 3 – ATRIBUIÇÃO DE PESOS A TERMOS DE INDEXAÇÃO

A frequência de termos em um documento ou em um conjunto de documentos constitui-se num indicador preponderante para o processo de classificação de documentos e não se constitui numa tarefa trivial. Um termo que apareça em todos os documentos de uma coleção acaba não sendo tão significativo quanto um termo que apareça em poucos documentos. Para obter esse efeito, é necessário fazer a atribuição de *pesos* aos termos de indexação.

De maneira formal, um peso w_{ij} (maior que zero) é associado a um termo de indexação k_i de um documento que é parte de uma coleção d_j. Dessa forma, para um termo k_i que não apareça em um documento, o peso será $w_{ij} = 0$. Portanto, o peso w_{ij} quantifica a importância do termo de indexação para descrever o conteúdo de um documento. A atribuição de pesos para os termos de indexação permite o *ranqueamento* de cada documento parte de uma coleção em uma consulta de usuário (Baeza-Yates; Ribeiro-Neto, 2011, p. 68).

A geração de pesos para termos de indexação está em proporção direta da importância de um termo descrevendo um documento ou os documentos de uma coleção. Por exemplo, considere uma coleção de 1000 documentos que versem sobre Immanuel Kant. Se um usuário especifica a busca apenas como "Immanuel Kant", como deve ser feito o ordenamento dos documentos para que sejam visualizados pelo usuário, de maneira a mostrar aqueles de maior interesse em primeiro lugar? Essa consulta não agrega qualquer valor informacional no contexto dessa coleção em particular, sendo que a quantidade de informação associada a essa consulta é, portanto, *zero*.

Assim, podem-se concluir duas coisas:

a. Diferentes termos não são igualmente importantes e devem ser ponderados de forma diferenciada;
b. Os pesos são influenciados pelos documentos em uma coleção. Para contabilizar a importância dos termos, é necessário calcular os pesos que reflitam a importância dos termos em uma coleção e em cada documento individualmente. Portanto, os pesos dependem diretamente da *frequência de ocorrência* de termos dentro dos documentos.

Formalmente, a frequência de ocorrência f_{ij} está associada ao termo de indexação k_i dentro do documento d_j, ou seja, refere-se ao número de vezes que o termo k_i aparece dentro do documento d_j. Assim, a frequência total F_i do termo k_i em uma coleção é a soma das frequências de ocorrências do termo em todos os N documentos desta coleção, ou seja,

$$F_i = \sum_{j=1}^{N} f_{ij} \tag{6}$$

3.1 Frequência de termos (TF)

A frequência de termos de indexação (*term frequency* – TF) é a primeira forma de atribuição de pesos. O valor ou peso de um termo k_i que ocorre em um documento d_j é proporcional à frequência do termo f_{ij}. Isso pode ser entendido da seguinte forma: quanto mais frequente for um termo k_i que ocorre no texto de um documento d_j, mais alta será o peso da frequência tf_{ij}. Assim, baseando-se na observação de que uma alta frequência de um termo de indexação é importante para descrever os tópicos chaves de um documento, chega-se à formulação da frequência de termos:

$$tf_{ij} = f_{ij} \tag{7}$$

Uma variação utiliza uma forma logarítmica (base 2). Isso é feito para que os pesos da frequência de termos possam ser comparados mais adiante com os pesos da frequência inversa.

$$tf_{ij} = \begin{cases} 1 + \log_2 f_{ij} & se\ f_{ij} > 0 \\ 0 & se\ f_{ij} = 0 \end{cases} \tag{8}$$

3.2 Frequência inversa de documentos (IDF)

A noção por trás da frequência inversa de documentos (*inverse document frequency* – IDF) relaciona-se ao fato de que a indexação dos documentos em uma coleção é obtida de uma forma mais efetiva considerando-se a especificidade dos termos de indexação presentes nos documentos. Se um termo é encontrado em apenas um documento em uma coleção, o potencial de recuperação desse documento para tal termo de indexação é aumentado em função desta especificidade.

A *especificidade* é uma propriedade da semântica dos termos, ou seja, um termo é mais ou menos específico dependendo do seu significado. Por exemplo, o termo *refeição* é menos específico do que os termos *almoço* ou *jantar*. No caso de a indexação ser feita manualmente, espera-se que o termo *refeição* seja utilizado para indexar mais documentos que os termos *almoço* ou *jantar*. Dessa forma, a interpretação da especificidade do termo é baseada na acurácia dos termos para ser um descritor de um documento. Uma alternativa é considerar a especificidade como uma função da utilização do termo,

interpretando-se a especificidade de forma estatística (Baeza-Yates; Ribeiro-Neto, 2011, p. 70).

Por outro lado, o conceito de exaustividade é importante para caracterizar a frequência inversa de termos. A *exaustividade* é uma propriedade relacionada à descrição de documentos. Ela pode ser entendida como a cobertura fornecida para os tópicos principais de um documento. À medida que são adicionados mais termos em um documento, a exaustividade do documento também aumenta. Ou seja, quanto mais vocabulário é adicionado ao documento, mais exaustiva a descrição do documento se torna (Baeza-Yates; Ribeiro-Neto, 2011, p. 70).

Portanto, a premissa por trás da noção de frequência inversa considera a especificidade e a exaustividade. Se o documento se torna muito extenso, a especificidade de termos tende a se tornar menor. De forma particular, se um termo ocorre em todos os documentos de uma coleção, a sua especificidade é mínima, e o termo não será útil para executar a recuperação. Dessa forma, é necessário adotar uma atribuição de pesos considerando-se a especificidade.

No caso de textos em linguagem natural, as frequências relativas das palavras nos documentos podem ser aproximadas por uma função de distribuição matemática denominada de *Lei de Zipf*. Se n_i é a frequência do termo k_i, ordenam-se todas as frequências dos termos nos documentos, de forma que $n(r)$ refere-se à r-ésima maior frequência de termos nos documentos. Se r é a posição (*rank*) com frequência $n(r)$, a lei de Zipf afirma que

$$n(r) \sim r^{-\alpha}$$

em que α é uma constante a ser determinada empiricamente. Dessa forma, a frequência de termos nos documentos pode ser modelada como uma função exponencial do ranqueamento.

Pode-se derivar com base nessa expressão o cálculo da frequência inversa de termos. Se k_i é a r-ésima frequência de termos, a frequência inversa é calculada como

$$idf_i = \log \frac{N}{n_i} \tag{9}$$

em que N é o número de documentos na coleção considerada.

Em grandes coleções de documentos, é comum selecionar os termos relacionados a substantivos, sendo retirados da lista artigos, preposições, conjunções, que são frequentemente denominados de *stop words*.

TEMA 4 – TF-IDF

O esquema mais popular de atribuição de pesos que combina a frequência de termos de indexação com a frequência inversa nos documentos é denominado de TF-IDF. Seja w_{ij} o peso associado ao par (k_i, d_j), definido por

$$w_{ij} = \begin{cases} (1 + \log f_{ij}).\log\frac{N}{n_i} & se\ f_{ij} > 0 \\ 0 & senão \end{cases} \quad (10)$$

Os pesos no esquema TF-IDF são bem efetivos para coleções de documentos de forma geral (Baeza-Yates; Ribeiro-Neto, 2011, p. 73). Existem variações para a expressão dada em (10), indo desde o esquema binário até expressões com normalização, que são resumidos na Tabela 2. Para a frequência de termos (TF), o esquema binário segue o padrão do modelo booleano, em que conta apenas a ocorrência (1) ou não (0) do termo no documento. A frequência direta é a alternativa básica que utiliza contagem de frequência, geralmente utilizada. A variante logarítmica busca atribuir pesos decrescentes à medida que a frequência de termos aumenta. A dupla normalização 0.5 normaliza os pesos pela frequência máxima encontrada no documento ou consulta, e os valores ficam na faixa $\{0.5, 1\}$. A dupla normalização K generaliza para outros valores.

Tabela 2 – Expressões alternativas para tf_{ij}

Esquema de pesos	Peso TF
Binário	$\{0, 1\}$
Frequência direta	f_{ij}
Normalização logarítmica	$1 + \log f_{ij}$
Dupla normalização 0.5	$0.5 + 0.5\frac{f_{ij}}{max_i f_{ij}}$
Dupla normalização K	$K + (1 - K)\frac{f_{ij}}{max_i f_{ij}}$

Fonte: Baeza-Yates; Ribeiro-Neto, 2011, p. 73.

No caso da frequência inversa de termos nos documentos (IDF), cinco alternativas também podem ser utilizadas para o cálculo na expressão (10).

Tabela 3 – Expressões alternativas para idf_i

Esquema de pesos	Peso IDF
Unário	1
Frequência inversa	$\log\frac{N}{n_i}$
Frequência inversa suavizada	$\log\left(1+\frac{N}{n_i}\right)$
Frequência inversa com máximo	$\log\left(1+\frac{max_i n_i}{n_i}\right)$
Frequência inversa probabilística	$\log\left(\frac{N-n_i}{n_i}\right)$

Fonte: Baeza-Yates; Ribeiro-Neto, 2011, p. 74.

Conforme a Tabela 3, a alternativa unária atribui o valor 1 para todos os termos (ou seja, IDF é excluída da atribuição de pesos). A frequência inversa é a expressão geralmente utilizada, podendo ser substituída pela frequência inversa suavizada a qual soma 1 para evitar comportamentos irregulares para valores extremos. A frequência inversa máxima calcula o peso relativamente ao valor do termo com maior frequência nos documentos. A frequência inversa probabilística aparece quando é considerado o modelo probabilístico de recuperação de documentos (Baeza-Yates; Ribeiro-Neto, 2011, p. 74).

O uso de combinações distintas de cálculo de TF e IDF produzirão diferentes conjuntos de pesos TF-IDF. Na Tabela 4, estão descritas três formas recomendadas de cálculo de pesos TF-IDF (Salton, citado por Baeza-Yates; Ribeiro-Neto, 2011, p. 74). O esquema a ser utilizado neste material didático consta na terceira linha.

Tabela 4 – Esquemas TF-IDF recomendados

Esquema de pesos	Peso do termo no documento	Peso do termo na consulta
1	$f_{ij}\times\log\frac{N}{n_i}$	$\left(0.5+0.5\frac{f_{iq}}{max_i\, f_{iq}}\right)\times\log\frac{N}{n_i}$

2	$1 + \log f_{ij}$	$\log\left(1 + \frac{N}{n_i}\right)$
3	$(1 + \log f_{ij}) \times \log\frac{N}{n_i}$	$(1 + \log f_{iq}) \times \log\frac{N}{n_i}$

Fonte: Baeza-Yates; Ribeiro-Neto, 2011, p. 74.

4.1 Correlação de termos de indexação

De maneira geral, os pesos dos termos de indexação são assumidos como mutuamente independentes. Dessa forma, um peso w_{ij} associado a um determinado par termo-documento (k_i, d_j) não tem relação com um peso $w_{(i+1)j}$ associado a outro par (k_{i+1}, d_j). Essa simplificação é útil para a formulação das premissas por trás do cálculo da frequência de termos nos documentos. Entretanto, em um documento podem aparecer termos relacionados a um conjunto semântico específico. Por exemplo, *time* e *jogador* tendem a aparecer em conjunto em documentos sobre futebol. Assim, o aparecimento de um termo atrai o aparecimento de outro. Dessa forma, existe uma correlação, e seus pesos devem refleti-la (Baeza-Yates; Ribeiro-Neto, 2011, p. 67).

Esta correlação pode ser obtida com base no cálculo de uma *matriz de correlação*. Seja $\boldsymbol{M} = [m_{ij}]$ a matriz termo-documento contendo t linhas e N colunas (conforme o tópico 1.1), em que cada $m_{ij} = w_{ij}$, ou seja, cada elemento indexado pelo par (i, j) é dado pelo peso associado com o par termo-documento (k_i, d_j).

Considerando $\boldsymbol{M}^T$ como a transposta da matriz $\boldsymbol{M}$, a matriz resultante da multiplicação é a matriz de correlação termo a termo $\boldsymbol{C} = \boldsymbol{M}^T\boldsymbol{M}$. Cada elemento c_{uv} expressa uma correlação entre termos k_u e k_v, dado por

$$c_{uv} = \sum_{d_j} w_{uj} \times w_{vj} \tag{11}$$

A matriz de correlação termo-a-termo $\boldsymbol{C}$ estabelece um relacionamento entre quaisquer dois termos k_u e k_v baseados nas ocorrências em conjunto, dentro dos documentos da coleção. A Figura 3 demonstra um exemplo de cálculo de uma matriz de correlação utilizando três termos e dois documentos.

Figura 3 – Exemplo de matriz termo-documento (três termos e dois documentos) com sua transposta e o cálculo produzindo a matriz de correlação

$$\underbrace{\begin{matrix} & d_1 & d_2 \\ k_1 & w_{11} & w_{12} \\ k_2 & w_{21} & w_{22} \\ k_3 & w_{31} & w_{32} \end{matrix}}_{\boldsymbol{M}} \times \underbrace{\begin{matrix} & k_1 & k_2 & k_3 \\ d_1 & w_{11} & w_{21} & w_{31} \\ d_2 & w_{12} & w_{22} & w_{32} \end{matrix}}_{\boldsymbol{M^T}}$$

$$\begin{matrix} & k_1 & k_2 & k_3 \\ k_1 & w_{11}w_{11} + w_{12}w_{12} & w_{11}w_{21} + w_{12}w_{22} & w_{11}w_{31} + w_{12}w_{32} \\ k_2 & w_{21}w_{11} + w_{22}w_{12} & w_{21}w_{21} + w_{22}w_{22} & w_{21}w_{31} + w_{22}w_{32} \\ k_3 & w_{31}w_{11} + w_{32}w_{12} & w_{31}w_{21} + w_{32}w_{22} & w_{31}w_{31} + w_{32}w_{32} \end{matrix}$$

$$\boldsymbol{C = M^T M}$$

Fonte: Baeza-Yates; Ribeiro-Neto, 2011, p. 68.

TEMA 5 – APLICAÇÃO DE TF-IDF

Levando-se em conta o formalismo de atribuição de pesos para esquemas TF-IDF vistos anteriormente, neste tema será elaborado um exemplo mostrando as fórmulas estudadas. Supondo uma coleção de documentos bastante simples e por sua vez contendo expressões simples conforme a Figura 4, essa coleção precisa ser analisada para verificar como pode ser atribuída a relevância aos documentos, em função dos termos presentes e da frequência correspondente a cada um.

Figura 4 – Uma coleção contendo quatro documentos, cada um com um conjunto de termos



Em primeiro lugar, cria-se a matriz com as frequências de ocorrência de cada termo dentro de cada documento. A Tabela 5 contém as frequências. Note que algumas aparecem apenas uma vez em um documento, ao passo que outros termos se repetem em todos os documentos.

Tabela 5 – Cálculo das frequências de cada termo dentro dos documentos. Note os índices de cada documento, colocado em cada coluna

No.	Termo	f_{i1}	f_{i2}	f_{i3}	f_{i4}
1	atmosfera	-	-	-	1
2	contém	-	1	-	-
3	enormes	1	-	-	-
4	e	-	1	-	1
5	errantes	-	-	1	-
6	estrelas	2	1	1	1
7	fixas	-	-	1	-
8	galáxias	-	1	-	-
9	planetas	-	1	1	3
10	são	2	-	2	-
11	sóis	1	1	-	-
12	tem	1	-	-	2
13	órbitas	-	-	-	1

Com base nessa tabela, cria-se outra correspondente com o cálculo da frequência de termos TF, que pode ser visto na Tabela 6. Aqui, foi aplicada a expressão (8). Veja que na linha 9, em que consta *planetas*, a frequência de termos é igual a $2{,}585$, que é o resultado de $1 + \log_2 3$.

Tabela 6 – Cálculo de TF da tabela de frequências da figura anterior

No.	Termo	tf_{i1}	tf_{i2}	tf_{i3}	tf_{i4}
1	atmosfera	-	-	-	1
2	contém	-	1	-	-
3	enormes	1	-	-	-
4	e	-	1	-	1
5	errantes	-	-	1	-
6	estrelas	2	1	1	1
7	fixas	-	-	1	-
8	galáxias	-	1	-	-
9	planetas	-	1	1	2,585
10	são	2	-	2	-
11	sóis	1	1	-	-
12	tem	1	-	-	2
13	órbitas	-	-	-	1

O passo seguinte consiste em calcular a frequência inversa de termos nos documentos. Como são quatro documentos, $N = 4$. A Tabela 7 mostra a elaboração da tabela. Note, por exemplo, que o termo *atmosfera* aparece em

apenas um documento, ou seja, $n_1 = 1$; o cálculo considera então $\log_2 N/n_1 = \log_2 4/1 = 2$.

O termo *planetas*, por sua vez, aparece em três documentos, ou seja, $n_9 = 3$; desta forma, $\log_2 N/n_9 = \log_2 4/3 = 0{,}415$. Deve-se notar que a IDF aumenta a relevância do termo que aparece em poucos ou apenas em um documento, enquanto o termo que aparece em vários documentos tende a ter uma valoração menor.

Tabela 7 – Cálculo da IDF

No.	Termo	n_i	idf_i
1	atmosfera	1	2
2	contém	1	2
3	enormes	1	2
4	e	1	2
5	errantes	1	2
6	estrelas	4	0
7	fixas	1	2
8	galáxias	1	2
9	planetas	3	0,415
10	são	2	1
11	sóis	2	1
12	tem	2	1
13	órbitas	1	2

Por fim, tem-se todos os elementos para calcular os pesos TF-IDF. A Tabela 8 mostra os pesos, de acordo com o esquema 3 da Tabela 4. Pode-se constatar agora cada termo com o seu peso TF-IDF, por exemplo, o termo *planetas*, que possui o valor $1{,}073$.

Tabela 8 – Cálculo dos pesos TF-IDF

No.	Termo	d_1	d_2	d_3	d_4
1	atmosfera	-	-	-	2
2	contém	-	2	-	-
3	enormes	2	-	-	-
4	e	-	2	-	2
5	errantes	-	-	2	-
6	estrelas	-	-	-	-
7	fixas	-	-	2	-
8	galáxias	-	2	-	-
9	planetas	-	0,415	0,415	1,073
10	são	2	-	2	-
11	sóis	1	1	-	-
12	tem	1	-	-	2
13	órbitas	-	-	-	2
	Norma	3,162	3,629	3,489	4,141

É interessante notar também que o termo *estrelas* aparece nos quatro documentos. O valor IDF para ele foi, portanto, calculado como zero e, dessa forma, o valor TF-IDF também é zero. Como o termo aparece em todos os documentos, ele não agrega informação para relevância de um documento sobre o outro.

O cálculo dos pesos TF-IDF permite também estabelecer uma relevância entre os documentos. Fazendo-se o cálculo da norma de cada documento, obtém-se um valor que representa a relevância do documento dentro da coleção. A norma $|d_j|$ pode ser calculada com base na raiz quadrada da soma dos quadrados de cada peso:

$$|d_j| = \sqrt{\sum_{d_j} w_{ij}^2}$$

Por exemplo, a norma para o documento d_1 é $|d_1| = \sqrt{2^2 + 2^2 + 1^2 + 1^2} = \sqrt{10} \cong 3{,}162$.

Constata-se, portanto, que d_4 é o documento mais relevante $(4{,}141)$, o que pode ser facilmente identificado devido ao fato de conter mais termos valorados do que os outros documentos.

A matriz de correlação pode ser calculada com base na matriz gerada pela Tabela 8. Na Tabela 9 está colocada a matriz transposta.

Tabela 9 – Matriz transposta da matriz de pesos TF-IDF

1	2	3	4	5	6	7	8	9	10	11	12	13
atmosfera	contém	enormes	e	errantes	estrelas	fixas	galáxias	planetas	são	sóis	tem	órbitas
-	-	2	-	-	-	-	-	-	2	1	1	-
-	2	-	2	-	-	-	2	0,415	-	1	-	-
-	-	-	-	2	-	2	-	0,415	2	-	-	-
2	-	-	2	-	-	-	-	1,073	-	-	2	2

Pela aplicação da fórmula (11), obtém-se a matriz de correlação mostrada na Tabela 10. Como a matriz de correlação é simétrica, as matrizes triangulares originadas acima e abaixo da diagonal principal contêm os mesmos valores. Dessa forma, pode-se analisar apenas uma delas (assinalado em amarelo). A matriz de correlação traz informações relacionadas ao aparecimento conjunto de termos dentro da coleção. Por exemplo, o termo *planetas* aparece correlacionado com a palavra *tem* com o valor de $2{,}146$. Isso pode ser

interpretado de modo que, quando a palavra *planetas* aparece no documento, a palavra *tem* tende a aparecer também. O par *planetas* e *sóis* também apresenta uma correlação de $0{,}415$ (os dois termos aparecem juntos em d_2). Quando um termo não aparece junto com outro, a correlação neste caso é zero.

Tabela 10 – Matriz de correlação do exemplo

		1	2	3	4	5	6	7	8	9	10	11	12	13
		atmosfera	contém	enormes	e	errantes	estrelas	fixas	galáxias	planetas	são	sóis	tem	órbitas
1	atmosfera	4	0	0	4	0	0	0	0	2,146	0	0	4	4
2	contém	0	4	0	4	0	0	0	4	0,830	0	2	0	0
3	enormes	0	0	4	0	0	0	0	0	0	4	2	2	0
4	e	4	4	0	8	0	0	0	4	2,976	0	2	4	4
5	errantes	0	0	0	0	4	0	4	0	0,830	4	0	0	0
6	estrelas	0	0	0	0	0	0	0	0	0	0	0	0	0
7	fixas	0	0	0	0	4	0	4	0	0,830	4	0	0	0
8	galáxias	0	4	0	4	0	0	0	4	0,830	0	2	0	0
9	planetas	2,146	0,830	0	2,976	0,830	0	0,830	0,830	1,496	0,830	0,415	2,146	2,146
10	são	0	0	4	0	4	0	4	0	0,830	8	2	2	0
11	sóis	0	2	2	2	0	0	0	2	0,415	2	2	1	0
12	tem	4	0	2	4	0	0	0	0	2,146	2	1	5	4
13	órbitas	4	0	0	4	0	0	0	0	2,146	0	0	4	4

FINALIZANDO

Neste material foram apresentados os conceitos relativos à recuperação de informação, os modelos de recuperação de informação (booleano, vetorial e probabilístico), os esquemas de atribuição de pesos considerando-se a frequência de termos (TF) e a frequência inversa nos documentos (IDF), bem como o esquema conjunto TF-IDF para atribuição de pesos. Por fim, o cálculo TF-IDF foi demonstrado por meio de um exemplo de aplicação. A recuperação da informação é um dos tópicos importantes de processamento de linguagem natural, pois visa identificar a relevância de termos e documentos em uma coleção.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. 2. ed. Harlow, England: Addison Wesley, 2011.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to modern information retrieval.** 2. ed. Cambridge-UK: Cambridge University Press, 2009.

RUSSELL, S.; NORVIG, P. **Inteligência artificial**. Tradução da 2. ed. Rio de Janeiro: Campus, 2004.