4
Recuperação de Informação

No presente capítulo são apresentados os fundamentos da área de Recuperação de Informação utilizados em Mineração de Textos, como por exemplo, os modelos de representação de documentos e as principais operações envolvidas nestes processos.

4.1.
Introdução

Recuperação de Informação lida com a representação, armazenamento, organização e acesso a itens de informação (BAEZA-YATES & BERTIER, 1999) (MANNING, RAGHAVAN, & SCHÜTZE, 2007). O conceito de itens de informação, neste contexto, refere-se ao tratamento diferenciado que estes objetos, geralmente documentos textuais, recebem: todos possuem muita ou pouca relevância. Julga-se um documento relevante quando este supre a necessidade de informação do usuário. Relevância, a característica central de Sistemas de Recuperação de Informação, é o que distingue Sistemas de Recuperação de Informação de Sistemas de Recuperação de Dados.

Recuperação de Dados busca meios eficientes de recuperar objetos baseado em um critério simples: dado o conjunto de termos desejado, encontrar todos os documentos que atendam ao critério booleano determinado. E isto é suficiente para muitas aplicações, como por exemplo, Sistemas Gerenciadores de Bancos de Dados. Mas, para um usuário que deseja informações sobre um determinado tópico, a consulta baseada em termos nem sempre trará somente bons resultados, ou seja, nem sempre será relevante. A Tabela 8 apresenta algumas das diferenças entre Recuperação de Dados e Recuperação de Informação (RIJSBERGEN, 1979).

Tabela 8 - Comparação entre Recuperação de Dados x Recuperação de Informação

Características	**Recuperação de Dados**	**Recuperação de Informação**
Comparação	Exata	Aproximada
Dados	Fortemente estruturados	Fracamente estruturados
Inferência	Dedução	Indução
Modelo	Determinístico	Probabilístico
Ling. Consulta	Artificial	Natural
Esp. da Consulta	Completa	Incompleta

Usuários de Sistemas de RI estão mais interessados na recuperação de informação associada a documentos do que na recuperação dos termos presentes nestes. Com o crescimento do volume de publicações, ao longo dos anos, foram desenvolvidas técnicas específicas para a área de Recuperação de Informação com o intuito de atender às necessidades dos usuários.

A ferramenta mais importante para auxiliar o processo de recuperação de informação é denominada índice. Índices são estruturas de dados associadas à parte textual dos documentos, e, portanto, indicam o local onde a informação desejada pode ser localizada. Segundo (BAEZA-YATES & BERTIER, 1999), há aproximadamente quatro mil anos já são praticadas técnicas de catalogação manual por índices.

Recuperação de Informação, antes, interesse de poucos, agora, é uma das áreas que mais tem recebido atenção de cientistas e pesquisadores. Contribuiu principalmente para isto a explosão demográfica da *Web* que é de longe o maior acervo de dados do mundo (CHAKRABARTI, 2003). E na *Web*, prevalecem os documentos hipertextos que, em sua essência, constituem o objeto de estudo de RI: documentos textuais.

A crescente complexidade dos objetos armazenados e o grande volume de dados exigem processos de recuperação cada vez mais sofisticados. Diante deste quadro, recuperação de informação apresenta a cada dia, novos desafios e se configura como uma área de significância maior (CARDOSO, 2000).

Porém, além de muito sucesso, a *Web* também trouxe novos desafios para a área de RI. Por ser um ambiente onde impera a cultura liberal e informal de propagação de conteúdo, encontrar informação relevante na *Web* tem sido cada

vez mais difícil, motivando também uma grande pesquisa em torno da Recuperação de Informação na Internet, em especial, na *Web*.

4.2. Histórico da área de Recuperação de Informação

A área de Recuperação de Informação pode ser cronologicamente dividida em três fases. A primeira fase compreende as décadas de 50 e 60. A fase seguinte situou-se entre as décadas de 70 e 80. Por último, a terceira fase que compreende a década de 90 aos dias atuais.

4.2.1. 1ª Fase – Décadas de 50 e 60

Sistemas de Recuperação de Informação foram originalmente utilizados para gerenciar a explosão de conteúdo da literatura científica na segunda metade do século XX (RIJSBERGEN, 1979). Bibliotecas estão entre as primeiras instituições a adotarem Sistemas de RI.

Em suas primeiras versões, Sistemas de RI funcionavam como um simples catálogo eletrônico. O processo de indexação era basicamente manual e os documentos eram indexados somente pelos termos principais de um dicionário de sinônimos criado para este propósito: um dicionário *thesaurus* (ver item "3.3.2"). A ideia deste conceito é simples: permitir a indexação somente do termo principal sempre que o próprio termo ou termos sinônimos estiverem presentes em um texto, evitando assim, que a escolha de um sinônimo ou outro possa impedir a localização do documento. Já na década de 60, Sistemas de RI deram início ao processo de indexação automática, porém, somente título e abstract eram processados. Surgiram também os primeiros algoritmos de busca textual.

4.2.2. 2ª Fase – Décadas de 70 e 80

Neste período, houve grandes avanços na área tecnológica, o que resultou em aumento significativo do poder computacional da época, permitindo, também,

a evolução de diversos sistemas, inclusive dos Sistemas de RI. Avanços como a indexação automática de todo o conteúdo e o desenvolvimento de funcionalidades adicionais de pesquisas foram possíveis.

RI – unida a área de Linguística – iniciou os primeiros estudos de Processamento de Linguagem Natural possibilitando a criação de um sistema simples de perguntas-respostas (BAEZA-YATES & BERTIER, 1999). Foi também nesta fase que o modelo de representação de documentos mais utilizado foi criado: o Modelo de Espaço Vetorial.

4.2.3.
3ª Fase – Década de 90 em diante

Nesta fase, o grande crescimento da *Web* e a necessidade de informação relevante neste ambiente colocaram em foco novamente a área de RI. Inicialmente, técnicas tradicionais de Sistemas de RI foram utilizadas, porém, grandes foram os problemas encontrados na adaptação destas técnicas:

- Escalabilidade das soluções: escalabilidade, neste contexto, indica a capacidade de preparo para a manipulação de grandes quantidades dados, seja esta relacionada ao poder de processamento ou armazenamento.
- Velocidade de atualização das páginas-*web*: a incrível velocidade de modificação do conteúdo dos *web sites* torna difícil manter um índice operacional e coerente sem saber a frequência de atualização dos documentos indexados.
- Velocidade de acesso aos documentos: em razão da sua distribuição geográfica mundial, a Web contém documentos nas mais diversas localidades. O acesso e indexação destes documentos exigem a disponibilidade dos mesmos, além do tempo necessário para que toda a informação neles seja transferida de um local para outro.

Atualmente, novas tecnologias estão sendo desenvolvidas para explorar as peculiaridades de um documento hipertexto e toda a sua relação n*a Web.*

4.3.
Recuperação de Informação Clássica

Recuperar informação é o propósito básico de qualquer sistema Recuperação de Informação. Baseada em índices, a recuperação de informação nestes sistemas obedece à arquitetura ilustrada na Figura 21.

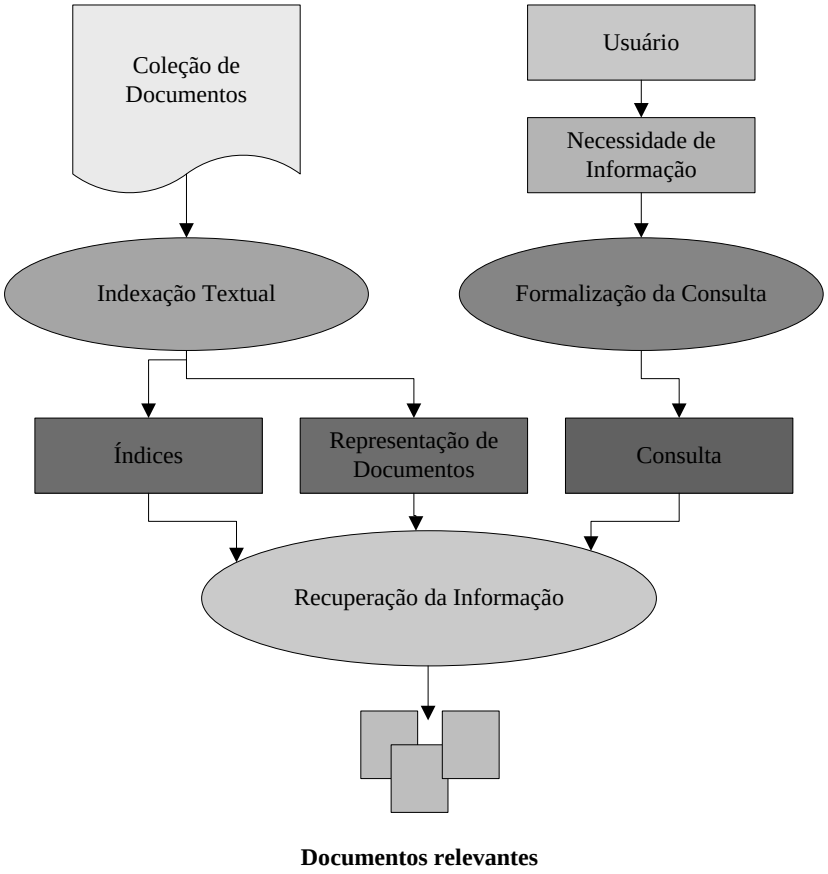

Figura 21 - Sistema Clássico de Recuperação de Informação

Neste modelo, duas entidades justificam a existência de um sistema de RI: a coleção de documentos, estes, geralmente textos, e o usuário com necessidade de informação. Os outros componentes decorrem destes.

A consulta é a representação formalizada da necessidade de informação do usuário em uma linguagem entendida pelo sistema. O processo de especificação da consulta geralmente é uma tarefa difícil. Há frequentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada (CARDOSO, 2000). Essa distância é gerada pelo limitado

conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta.

Assim que formalizada, a consulta é processada junto aos documentos, que estão representados pelos seus respectivos modelos de representação textuais, e, em seguida, a resposta à necessidade de informação na coleção de documentos é exibida ao usuário. O processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos são usados para acelerar esta tarefa. Além disso, a lista de documentos recuperados é classificada em ordem decrescente de um grau de similaridade entre o documento e a consulta (CARDOSO, 2000).

Os modelos de representação textuais utilizados em Sistemas de RI podem ser vistos como uma representação fortemente estruturada dos textos. E, como todo documento é considerado um conjunto de termos ou *tokens*, esta nova representação estruturada é baseada na presença ou ausência destes termos ou *tokens*.

Quando todo o conjunto de *tokens* de um documento é utilizado para representá-lo tem-se uma indexação textual completa ou *full text indexing*. Porém, embora a indexação textual completa seja aquela que forneça a visão lógica mais completa de um documento, nem sempre é possível utilizá-la, em razão do elevado custo computacional para o manuseio desta enorme quantia de dados, tornando necessário que um documento seja representado por um conjunto menor de *tokens*.

Como nem todas as palavras num texto não igualmente importantes para representá-lo semanticamente, para que seja bem representado por um conjunto menor de *tokens*, um documento pode ser submetido a sucessivos métodos de processamento textual, tais como remoção de *stopwords* e *stemming*, que visam eliminar conteúdo irrelevante do texto, permitindo que seja possível a representação lógica do mesmo. A Figura 22 ilustra algumas das possibilidades existentes em um processo de indexação (BAEZA-YATES & BERTIER, 1999).

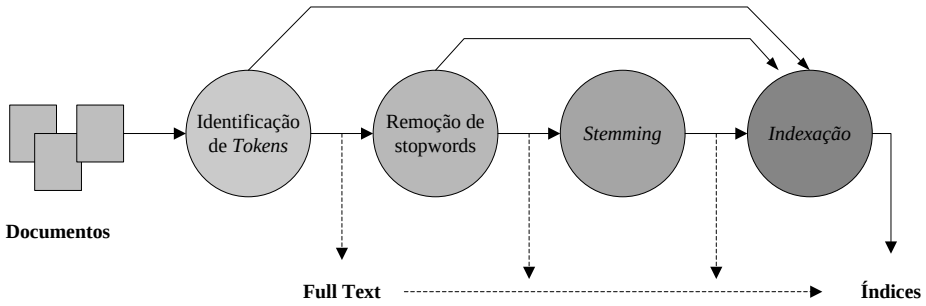

Figura 22 – Etapas possíveis no processo de Indexação de documentos textuais

Por utilizar seus próprios métodos de processamento textual, o interesse de Mineração de Textos na área de RI restringe-se às técnicas de representação e identificação de documentos. Muitos dos métodos de processamento textual utilizados em Mineração de Textos foram baseados naqueles utilizados em RI, e, portanto, foram abordados, sob o enfoque de Mineração de Textos, no item "3.2".

A seguir, serão apresentados dois modelos de representação de documentos utilizados, tanto em RI, como em Mineração de Textos: **Modelo de Recuperação Booleano**[18] e Modelo de Espaço Vetorial.

4.3.1.
Modelos de Representação de Documentos

4.3.1.1.
Modelo de Recuperação Booleano

Um dos primeiros modelos de pesquisa a ser adotado foi o Modelo de Recuperação Booleano ou, simplesmente, Modelo Booleano. Fundamentado na Álgebra Booleana e na Teoria dos Conjuntos, interpreta toda consulta como uma expressão lógica, permitindo até mesmo a utilização dos conectivos lógicos "e", "ou" e "não", e, portanto, possui critério de decisão simples para julgar a

[18] Do termo inglês, *Boolean retrieval model.*

relevância de um documento: documentos relevantes são aqueles que contêm, ou não, os termos que satisfazem a expressão lógica da consulta.

Em virtude do critério de decisão binário deste modelo não existem meios para a realização de igualdade parcial da consulta com os documentos. Portanto, também não existem critérios de graduação de relevância dos documentos encontrados, ou seja, não é possível ordenar documentos de acordo com a relevância individual de cada um.

Este modelo de representação é muito mais utilizado em Sistemas de Recuperação de Dados do que em Sistemas de Recuperação de Informação. É de fácil utilização para usuários que dominam lógica booleana, o que não ocorre na maioria dos casos.

Algumas das vantagens do modelo booleano são a excelente *performance* e a fácil implementação. Possui como principal desvantagem a dificuldade de se expressar a necessidade de informação por meio de uma expressão booleana. Outra característica ruim deste modelo é desconsiderar a frequência de ocorrência dos termos em um texto.

4.3.1.2. Modelo de Espaço Vetorial

O Modelo de Espaço Vetorial busca abordagem geométrica para resolver problemas de representação de documentos. Documentos são representados como vetores em um espaço Euclidiano *t-dimensional* em que cada dimensão corresponde a um *token* da coleção de documentos (REZENDE, 2005), ou seja, cada *token* é um eixo deste espaço Euclidiano.

Neste modelo, vetores são representados pela forma $D_i = (t_1; t_2; t_3; ...; t_n)$, em que D_i é o *i-ésimo* documento de uma coleção, e t_n o *n-ésimo token* da coleção de documentos, ou seja, para cada documento da coleção existem n *tokens*-índices que os representa (SILVA A. A., 2007), conforme ilustrado na Figura 23 . Cada *token* desta coleção de documentos está associado a sua frequência de ocorrência em cada documento, desta forma, para o documento D_i e para o token t_j, $w_{i,j} \geq 0$ representa essa associação e o tamanho do $eixo_j$ no vetor D_i. Quando o *token* j não ocorre no documento D_i, tem-se $w_{i,j} = 0$.

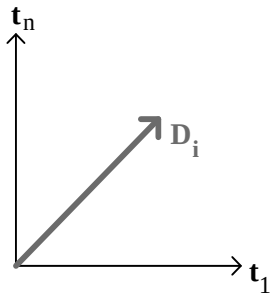

Figura 23 - Representação vetorial do documento Di no espaço n-dimensional (*n =2*)

No Modelo de Representação Vetorial, o processamento de uma consulta é realizado através de um cálculo de similaridade entre cada documento da coleção e a própria consulta, ou seja, toda consulta é também representada de forma vetorial, e através de um cálculo de similaridade entre cada documento da coleção e a consulta, obtém-se uma lista dos documentos relevantes para aquela necessidade de informação.

O Modelo do Espaço Vetorial é o modelo de representação mais utilizado em Mineração de Textos (REZENDE, 2005). Contribuem para isto a sua forma de representação, intuitiva e prática, que torna possível:

- A ponderação de termos na representação dos documentos e processamento das consultas;
- A recuperação de documentos que não possuem todos os termos definidos na consulta;
- Ordenação do resultado baseada na relevância dos documentos.

Desvantagens deste modelo de representação são a necessidade de novo processamento da coleção de documentos quando esta é alterada e a ausência de relação semântica entre os *tokens* de uma coleção.

4.3.1.3. Frequência dos termos

No Modelo de Espaço Vetorial, cada documento é representado por um vetor cujas dimensões são os termos presentes na coleção de documentos. Cada

coordenada do vetor é um termo da coleção de documentos e possui valor numérico que representa a frequência de ocorrência deste termo no documento (LOPES, 2004).

A associação de valores numéricos as coordenadas dos vetores é conhecida como **atribuição de pesos**[19] e visa atribuir maior importância aos termos que são mais relevantes. A seguir, são citadas e explicadas as medidas de atribuição de pesos mais comuns:

- Binária: Quando um termo está presente em determinado documento, é atribuído o valor *true* ou um para indicar esta ocorrência. Quando um termo está não presente em determinado documento, é atribuído o valor *false* ou zero para indicar esta ausência. Por ser muito simples, esta medida de atribuição de pesos é raramente utilizada.
- Frequência do Termo: **Frequência do Termo**[20] ou **TF**[21] é definida como o número de ocorrências de um determinado termo em um documento (SALTON & BUCKLEY, 1988). Em geral, termos presentes em muitos documentos com alta frequência não possuem caráter discriminatório para a diferenciação dos documentos de uma coleção e são considerados como uma *stopword*. É comum normalizar em um documento a frequência de seus termos, pois, sem este artifício, os documentos mais extensos de uma coleção seriam privilegiados no processo de recuperação de informação. Na Equação 5 é ilustrado o cálculo normalizado da frequência do Termo_i no Documento_j que possui *k* termos.

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Equação 5 - Cálculo da medida TF em um documento

[19] Do termo inglês, *weighting*.

[20] Do termo inglês, *Term Frequency*.

[21] Acrônimo de *Term Frequency*.

- TF-IDF: TF-IDF ou *Term Frequency – Inverse Document Frequency* é uma medida de atribuição de pesos que favorece termos que ocorrem em poucos documentos de uma coleção (SALTON & BUCKLEY, 1988). É utilizada para avaliar o quão importante é um termo para o documento em que ele ocorre, em relação a todos os documentos da coleção. A medida TF-IFD de um termo, ilustrada na Equação 6, é a combinação de sua medida local (TF) e global (IDF).

$$\boldsymbol{tf - idf_{i,j} = tf_{i,j} \times idf_i}$$

$$idf_i = \log \frac{|D|}{\left|\{d_j : t_i \in d_j\}\right|}$$

$$|D|, número\ total\ de\ documentos\ da\ coleção;$$
$$\left|\{d_j : t_i \in d_j\}\right|, número\ de\ documentos\ em\ que\ o\ termo\ t_i\ ocorre;$$

Equação 6 - Cálculo da medida TF-IDF em um documento

- Escore de relevância: Proposto por (WIENER, PEDERSEN, & WEIGEND, 1995) é baseado na importância que um termo possui em representar uma determinada categoria da coleção de documentos. Termos que aparecem em muitas categorias obtém valores baixos, por serem pouco discriminantes; termos que aparecem em poucas categorias recebem valores altos, podendo representar a categoria em que possui maior frequência. O escore de relevância de um termo, ilustrado na Equação 7, é dado por:

$$\boldsymbol{r_t} = \log \frac{\frac{w_{ct}}{d_c} + \frac{1}{6}}{\frac{w_{\bar{c}t}}{d_{\bar{c}}} + \frac{1}{6}}$$

Equação 7 - Cálculo do escore de relevância de um termo

Em que:

- w_{ct} é o número de documentos pertencentes a uma categoria *(c)* que contém o termo *t*;
- d_c é o número total de documentos da categoria considerada *(c)*;
- $w_{\overline{ct}}$ é o número de documentos de outras categorias que contém o termo *t*;
- $d_{\bar{c}}$ é o número total de categorias de outros documentos.
- A constante $\frac{1}{6}$ é utilizada para eliminar o problema de divisão por zero, possível de ocorrer quando um termo só está presente na categoria considerada *(c)*.

- Coeficiente de correlação: Desenvolvido por (NG, GOH, & LOW, 1997) para indicar o grau de correlação de uma palavra e um documento. Leva em conta a quantidade total de documentos de uma coleção, a quantidade de documentos em que o termo aparece e a quantidade de documentos em que o termo não aparece. A Equação 8 ilustra a definição do coeficiente de correlação entre o termo *t* e a classe *c*:

$$\boldsymbol{C_{(t,c)}} = \frac{(\boldsymbol{N_{r+}} \times \boldsymbol{N_{n-}} - \boldsymbol{N_{r-}} \times \boldsymbol{N_{n+}}) \times \sqrt{\boldsymbol{N}}}{\sqrt{(\boldsymbol{N_{r+}} + \boldsymbol{N_{r-}}) \times (\boldsymbol{N_{n+}} + \boldsymbol{N_{n-}}) \times (\boldsymbol{N_{r+}} + \boldsymbol{N_{n+}}) \times (\boldsymbol{N_{r-}} + \boldsymbol{N_{n-}})}}$$

Equação 8 - Cálculo do Coeficiente de Correlação

Em que:

- N_{r+} é o número de documentos relevantes para $C_j$que contém o termo *t*;

- N_{r-} é o número de documentos relevantes para C_j que não contém o termo *t*;
- N_{n+} é o número de documentos não relevantes para $C_j$que contêm o termo *t;*
- N_{n-} é o número de documentos não relevantes para C_j que não contêm o termo *t.*

- Ganho de informação: Métrica de atribuição de pesos proposta por (YANG & PEDERSEN, 1997), é um critério que define a qualidade de cada termo. Ele mede a quantidade de pequenos pedaços ou partições de informação obtidos para a predição da categoria através da presença ou ausência de um termo no documento. Este método é comumente utilizado no campo de aprendizagem de máquina e na construção de árvores e regras de decisão. Os autores afirmam que a categorização de textos, normalmente, possui um espaço dimensional muito grande, alcançando até dezenas de milhares de características, e é preciso calcular a qualidade do termo de maneira global. A partir de um conjunto de textos de treinamento, para cada termo único é calculado o ganho de informação. Os termos que não alcançarem um limiar predefinido serão excluídos. A ideia principal deste método é dividir o conjunto de exemplos em partições ou subconjuntos de exemplos, sendo estes subconjuntos compostos de exemplos de uma mesma classe ou similares. Ao grupo aplica-se o cálculo do ganho. O conjunto vai sendo subdividido repetidamente até que um subconjunto contenha apenas exemplos de uma única classe ou o número de exemplos seja inferior a um limite estabelecido. A conclusão é que o ganho de informação reduz os ruídos conforme o conjunto vai sendo subdividido, de forma que, no final, o último subconjunto será composto apenas por exemplos similares. A Equação 9 exibe a fórmula proposta por (MLAENIC & GROBELNIK, 1998) para o cálculo de ganho de informação:

$$G_f = P_w \times \sum_i P_{C_i|w} \times \log \frac{P_{C_i|w}}{P_{C_i}} + P_{\bar{w}} \times \sum_i P_{C_i|\bar{w}} \times \log \frac{P_{C_i|\bar{w}}}{P_{C_i}}$$

Equação 9 - Cálculo do Ganho de Informação

Cálculo do Ganho de Informação

Em que:

- G_f é o ganho de informação de característica *f*. O termo *w* é representado pela característica *f*;
- P_w é a probabilidade de ocorrer o termo *w*;
- $P_{C_i|w}$ é a probabilidade condicional de ocorrer o termo *w* na *i-ésima* classe;
- P_{C_i} é a probabilidade da *i-ésima* classe;
- $P_{\bar{w}}$ é a probabilidade de não ocorrer o termo *w*;
- $P_{C_i|\bar{w}}$ é a probabilidade condicional de não ocorrer o termo *w* na *i-ésima* classe.

4.3.1.4. Cálculo de Similaridade

No Modelo do Espaço Vetorial cada documento é representado por um vetor de *n* dimensões, em que cada dimensão é um termo distinto e presente em algum documento da coleção. A cada termo é atribuído um peso como forma de identificar a importância deste no documento e para isto são utilizadas as medidas de atribuição de pesos mencionadas acima.

Uma das técnicas mais utilizadas para obter o grau de similaridade entre documentos ou entre documentos e consultas decorre naturalmente deste modelo de representação: é através do cosseno do ângulo formado pelos vetores de representação destes objetos (BAEZA-YATES & BERTIER, 1999).

O cálculo do cosseno do ângulo entre dois vetores é ilustrado na Equação 10. Quanto mais perto de um o valor do cosseno, mais ortogonais são os vetores comparados, o que significa que existem poucos termos comuns entre os documentos. Quanto mais perto de zero o valor do cosseno, mais paralelos são os vetores comparados, o que significa que existem muitos termos comuns entre os documentos.

$$\mathbf{cs}\,\boldsymbol{\theta} = \frac{\overrightarrow{\boldsymbol{v_1}} \cdot \overrightarrow{\boldsymbol{v_2}}}{\|\overrightarrow{\boldsymbol{v_1}}\| \times \|\overrightarrow{\boldsymbol{v_2}}\|}$$

Equação 10 - Cálculo de similaridade entre documentos por meio do cosseno