Indexação e recuperação da informação com função de crença

Wagner Teixeira da Silva
Ruy Luiz Milidiú

Resumo

Um modelo usando funções *de* crença para *indexar e recuperar documentos á* proposto. *Tal modelo é* baseado em um vocabulário *controlado, semelhante a um tesauro, e na frequência dos termos em cada documento. Cada descritor nesse vocabulário é um termo escolhido entre seus sinônimos. Um descritor pode ter um subconjunto de descritores mais* gerais, um subconjunto de descritores *mais específicos e um subconjunto de descritores relacionados. Assim, descritores não* são mutuamente exclusivos *e modelos probabilísticos convencionais não são adequados. Contudo, uma função de crença pode* ser definida *sobre um subconjunto dos descritores atômicos.* Taís *descritores são aqueles* sem te*rmos mais específicos (denotados* por Ω*). Subconjuntos de* Ω *podem* ser *vistos corno temos mais gerais, ou como termos relacionados. Desde modo, uma função* de crença sobre Ω pode estimar o *conteúdo semântico de um documento. Uma consulta ponderada (à base de* documentos) *pode* ser vista como outra função *de crença. Desde que ambas as funções são definidas* sobre Ω*, é possível computar o* grau de condordância *ente elas. Equivalentemente, é possível determinar o grau de concordância entre a consulta* e os documentos e ordená-los segundo esse *valor.*

Palavras-chave
Indexação automática; *Ordenação de documentos, Recuperação da informação; Modelo de recuperação;* Teoria de *função de crença; Modelo com função de crença; Modelo baseado* em *frequência;* Relevância de documentos.

Ci. Int, Brasília, *20(2): 155-164,* jul/dez. 1991

INTRODUÇÃO

A Teoria de Funções de Crença[1] vem ganhando espaço no ambiente de inteligência artificial graças a sua maleabilidade e poder para representar e atualizar conhecimento impreciso. Nesse trabalho, o conteúdo semântico de documentos e a consuta do usuário à base de documentos são representados por funções de crença distintas. Taís funções são comparáveis, e o *grau* de concordância entre elas pode ser mensurado. Isto toma possível estimar a relevância d*os* documentos recuperados por uma dada consulta

Indexar *documentos é* uma tarefa difícil. Quando este trabalho é feito por especialistas que mantêm suas próprias listas de terrnos-chave, eles tendem a indexar documentos em relação aos seus interesses atuais, ao invés de fazê-lo em relação aos interesses gerais da *comunidade* usuária, com respeito a todos os assuntas aos quais os documentos se relacionam. Deste modo, uma indexação automática pode *ser* mais vantajosa *do que aquela* realizada por especialistas no assunto.

Muito esforço tem sido envidado na busca de um conveniente sistema cie representação do conteúdo semântico de documentos. Contudo esta tareia não é fácil. Várias propostas vêm sendo feitas, tais como:

1) usar *termos* associados manualmente por um leitor que lê o documento;

2) usar termos que aparecem no título do documento, com o conceito KWIC *(Key* word in *context) ou KWOC (Key word out of context);*

3) usar termos que aparecem nos nomes das seções do documento;

4) usar termos estatisticamente significantes - ou seus sinónimos - que aparecem no documento, ou somente no resumo do documento (Doyle[2], 1975, *Heaps[3],1978,* Salton & McGill[4],1983);

5) usar frases estatisticamente significantes ou suas frases semanticamente *equivalentes* que apareçam no documento (Jones, Gassie & Radhakrishnan[5],1990, Fagan[6],1987).

Aqui é proposto um modelo de função de crença *que* possibilita indexação automática e escalonamento de documentos. Dado um vocabulário controlado com um subconjunto de descritores e um documento, este modelo associa uma massa básica de crença para cada descritor que *apareça no* documento. Apenas os desertores no vocabulário são considerados. A massa básica de crença é associada ao descritor a partir de sua frequência ou das freqüên*cias* de *seus* sinônimos no documento. Os descritores com massa básica de crença maior que zero representam o conteúdo semântico do documento no qual eles aparecem.

Baseados na representação semântica de cada documento, o modelo proposto calcula o grau de concordância de cada documento com uma dada consulta ponderada do usuário. Este grau de concordância é usado para escalonar *os* documentos. Supostamente, quanto maior o grau de concordância de um documento, mais relevante ele será para o usuário consultante. Neste sentido o grau de concordância *é* um estimador para o grau de relevância.

Funções de crença parecem mais convenientes do *que* a usual probabilidade para modelar a indexação *e escalonamento* de documentos com um tesauro. Pois a construção de um espaço cie probabilidade com um vocabulário de tipo tesauro ê uma tarefa não trivial, já *que os* descritores no tesauro não são, semanticamente, mutuamente disjuntos. Contudo, os descritores neste vocabulário sem termos mais específicos constituem um quadro de discemimento, quando cada um deles é tomado como "o melhor representante semântico de documentos".

Um termo mais geral pode ser visto como um rótuto de um subconjunto do quadro de

discernimento. As frequências relativas dos termos do documento são somadas às frequências relativas dos seus descritores representantes. Deste modo, naturalmente, todas as medidas de crença sobre o quadro de descritores atómicos estão definidas.

TEORIA DE FUNÇÕES DE CRENÇA

Nesta seção, são resumidos os principais conceitos da Teoria de Funções de Crença[1], ou teoria de Dempster - Shafer (DS), os quais são usados na descrição do modelo de recuperação de informação proposto. Um enfoque mais recente dessa teoria é dado em Shafer, Shenoy & Mellouli[7], 1987, Silva[8], 1991, e em Silva & Milidiú[9,10,11],1990.

CRENÇA E INCERTEZA

Seja Θ um conjunto finito de proposições mutuamente exclusivas, das quais a única verdadeira é desconhecida. Todo subconjunto **A** de Θ é visto como uma disjunção das proposições $\mathbf{p} \in \mathbf{A}$. Logo cada $\mathbf{A} \subset \Theta$ é uma proposição com as seguintes propriedades:

1) Se existe $\mathbf{p} \in \mathbf{A}$ tal que p é a proposição verdadeira, então A também é uma proposição verdadeira.

2) Se $\mathbf{A} \subset \mathbf{B}$, então $\mathbf{A} \rightarrow \mathbf{B}$

A incerteza sobre qual das proposições elementares de Θ é a verdadeira pode ser expressa pela função de densidade de probabilidade de um subconjunto aleatório X de Θ, não vazio. A probabilidade Pr [X = **B**] mede a parte da massa de crença unitária de alguém na hipótese de a disjunção **B** ser verdadeira, sem levar em conta as massas de crença Pr [X = **A**], alocadas nos A que são subconjuntos próprios de **B**. Assim Pr [X = **B**] é uma massa especifica de **B**, e de nenhum de seus subconjuntos próprios.

MASSA BÁSICA DE CRENÇA

Dentro do enfoque evidenciai da Teoria de Função de Crença ou Teoria de DS, a função densidade de probabilidade do subconjunto aleatório X é renomeada por uma função $m: 2^{\Theta} \rightarrow [0,1]$, chamada de **massa básica de crença*** (mbc), que tem as seguintes propriedades:

$$m(\Phi) = 0; \text{ e}$$
$$\sum_{\mathbf{A} \subset \Theta} m(\mathbf{A}) = 1$$

* A função m é mais conhecida na literatura com o nome de probabilidade básica associada (bpa - *basíc probability assignmenf).* Contudo, "massa básica de crença" parece mais adequado por realçar o caráter epistêmico desta função.

Tal função distribui a massa de crença unitária entre os subconjuntos de Θ. Assim, m(**A**) = Pr(X = **A**), para cada $\mathbf{A} \subset \Theta$.
A quantidade m(A) mede a crença básica em **A**.

Do ponto de vista epistêmico a função m é obtida por um mapeamento entre uma peça de evidência **E** e o espaço de proposições, onde cada elemento $\mathbf{A} \subset \Theta$ tem um grau de compatibilidade m(**A**) com a peça de evidência **E**. Alternativamente a mbc m pode ser obtida pela combinação de uma coleção de outras mbc, veja o item "combinação de crenças".

CRENÇA E PLAUSIBILIDADE

Enquanto m(**A**) = Pr (X = **A**) é uma medida especifica de **A**, e de nenhum subconjunto próprio de **A**, como poderiam ser interpretadas as probabilidades $Pr(X \subset \mathbf{A})$ e $Pr(X \cap \mathbf{A} \neq \Phi)$? Estas medidas são conhecidas no jargão da Teoria de DS, como grau de crença em **A**, denotada por Cr(**A**), e grau de Plausibilidade em **A**, Pl(**A**), respectivamente.

A função de crença Cr é dada pela fórmula (2.1) e atribui graus de crença a todos os subconjuntos não vazios de Θ. Cr(**A**) é interpretado como o total de massa de crença atribuída ao subconjunto A, isto é, Cr(**A**) é a soma de todas as massas de crença atribuídas a A e aos seus subconjuntos próprios.

$$Cr(\mathbf{A}) = Pr[X \subset \mathbf{A}] = \sum_{\mathbf{B} \subset \mathbf{A}} 2\, Pr[X = \mathbf{B}] = \sum_{\mathbf{B} \subset \mathbf{A}} 2\, m(\mathbf{B}), \text{ para cada } \mathbf{A} \subset \Theta \quad (2.1)$$

A função de plausibilidade - Pl, dada por (2.2), atribui graus de plausibilidade a todo subconjunto de 0. Pl(**A**) mede a massa de crença que não contradiz a hipótese de **A** ser verdadeira.

$$Pl(\mathbf{A}) = Pr[\mathbf{B} \cap \mathbf{A} \neq \Phi) = \sum_{\mathbf{B} \cap \mathbf{A} \neq \Phi} Pr[X = \mathbf{B}],$$

$$= \sum_{\mathbf{B} \cap \mathbf{A} \neq \Phi} m(\mathbf{B}), \quad (2.2)$$

$$= 1 - \sum_{\mathbf{B} \subset \mathbf{A}} m(\mathbf{B})$$

$$= 1 - Cr(\bar{\mathbf{A}}), \text{ para cada } \mathbf{A} \subset \Phi. \quad (2.3)$$

De (2.1) e *(2.2)* é fácil ver que $Cr(\mathbf{A}) \leq Pl(\mathbf{A})$ e que $Cr(\mathbf{A}) + Pl(\bar{\mathbf{A}}) = 1$. Já a expressão (2.3) relaciona_Pl(**A**) com a dúvida em **A**, isto é, com Cr(**A**). Nesta interpretação, Pl(**A**) representa o erro que alguém comete por refutar **A**.

Uma interpretação destas funções em termos evidenciais é bastante intuitiva. A mbc m(**A**) é interpretada como o suporte de crença que as evidências dão especificamente ao subconjunto **A**, e a nenhum de seus subconjuntos próprios. A função de crença Cr(**A**) mede o grau de crença com que as evidências suportam o subconjunto **A**, ou equivalentemente é o total de crença atribuída ao subconjunto **A**. Vendo **A** como uma proposição, Cr(**A**) é a soma das crenças básicas em todas as proposições que implicam **A**. A função Pl(**A**) mede o grau de plausibilidade de **A** suportado pelas evidências. Os valores extremos de Cr e Pl para uma dada proposição **A** são interpretados como:

. Cr(**A**) = O - não há evidências que favoreçam **A**.
. Cr(**A**) = 1 - todas as evidências favorecem **A**.
. Pl(**A**) = 1 - não há evidências que neguem **A**.
. Pl(**A**) = O - as evidências negam **A** conclusivamente.

Observa-se que se Pl (**A**) = 0, então Cr(**A**) = 1, todas as evidências negam a proposição **A**. Enquanto Pl(**A**) = 1 implica que $Cr(\bar{\mathbf{A}}) = 0$, não há evidências que neguem a proposição **A**.

Uma função de crença *Cr.* $2^{\Theta} \rightarrow [0,1]$ tem as seguintes propriedades:

.$Cr(\Phi) = 0$,
.$Cr(\Theta) = 1$,
. Se $A_1, A_2, \ldots A_n$, são subconjuntos de Θ, então

$$Cr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{I \subset \{1,\ldots,n\}} (-1)^{|I|+1} Cr\left[\bigcap_{i \in I} A_i\right]$$

Uma função de crença Cr é mais bem representada por

$$M = (\Theta, \mathbf{F}, \mu,)$$

onde

. Θ é um conjunto de proposições mutuamente exclusivas;
. $\mathbf{F} = \{\mathbf{A} \subset \Theta \mid m(\mathbf{A}) > 0\}$ é o conjunto dos elementos focais de Cr, chamado de conjunto focal;
. μ: $F \rightarrow [0,1]$ é uma restrição da mbc m ao conjunto **F**, isto é $\mu(\mathbf{A}) = m(\mathbf{A})$, para cada $\mathbf{A} \in \mathbf{F}$.

A união dos elementos de **F** constituem o núcleo de Cr, isto é,

$$\mathbf{N} = \bigcup_{\mathbf{A} \in \mathbf{F}} \mathbf{A}$$

e para toda função de crença Cr, a massa de crença unitária se concentra em seu núcleo, isto é, Cr(**N**) = 1.

COMBINAÇÃO DE CRENÇAS

Sejam X_1, X_2 subconjuntos aleatórios não vazios do quadro de discernimento Θ. Suponha que X_1 e X_2 sejam probabilisticamente independentes e que $Pr[X_1 \cap X_2 \neq \Phi] \} 0$. Sejam Cr_1 e Cr_2, funções de crença com mbc m_1 e m_2 dadas por

$m_1(\mathbf{A}_1) = Pr[X_1 = \mathbf{A}_1]$, para cada $\mathbf{A}_1 \subset \Theta$; e,
$m_2(\mathbf{A}_2) = Pr[X_2 = \mathbf{A}_2]$, para cada $\mathbf{A}_2 \subset \Theta$.

Então, define-se a função da crença $Cr = Cr_1^+ + Cr_2$, como resultante da combinação de Cr_1 e Cr_2, sendo sua mbc m dada por

$m(\mathbf{A}) = Pr[X_1 \cap X_2, = \mathbf{A} \text{ I } X\mathbf{1} \text{ n } X_2 \neq \Phi]$, para cada $\mathbf{A} \subset \Theta$

A combinação de duas funções de crença pode ser reformulada, em termos evidenciais, do seguinte modo. Sejam Cr_1 *e* Cr_2 funções de crença sobre **Θ**, representadas por **M1** e $\mathbf{M_2}$, respectivamente. O grau de concordância entre elas é dado por

$$\mathbf{K}^{-1} = \Sigma \mu_1(\mathbf{A_1}).\mu_2(\mathbf{A_2})$$
$$\mathbf{A_1} \in \mathbf{F_1},$$
$$\mathbf{A_2} \in \mathbf{F_2}, \text{ e}$$
$$\mathbf{A_1} \cap \mathbf{A_2} \neq \Phi$$

Se $\mathbf{K}^{-1} \rangle 0$, e as funções de crença Cr_1 e Cr_2 são suportadas por evidências independentes, então elas são combináveis; e a mbc m da função resultante $Cr = Cr1 + Cr_2$ é dada por

$$m(\mathbf{A}) = \mathbf{K}.\Sigma \mu_1(\mathbf{A_1}).\mu_2\{\mathbf{A_2})$$
$$\mathbf{A}, \in \mathbf{F},.$$
$$\mathbf{Aj} \in \mathbf{F_2}, \text{ e}$$
$$\mathbf{A}, \cap \mathbf{A_2} = \mathbf{A}$$

Onde **K** mede a extensão do conflito, e o seu inverso mede o grau de independência entre as funções de crença Cr_1 e Cr_2. Se $\mathbf{K}^{-1}$ =1, então dado qualquer elemento focal de Cr_1 como verdadeiro, isto não informa qual elemento focal de Cr_2 é verdadeiro. Enquanto se $\mathbf{K}^{-1}$ = O, e se um elemento focal $\mathbf{A_1}$ de Cr_1 é verdadeiro, então nenhum elemento focal de Cr_2 pode ser verdadeiro.

O termo $\mathbf{K}^{-1}$ também pode ser interpretado como o grau de concordância entre *as* funções de crença Cr_1 e Cr_2. À medida em que $\mathbf{K}^{-1}$ varia de 0 (zero) a 1 (um), a quantidade de elementos focais de Cr_1, com interseção diferente de vazia com os elementos focais de Cr_2, varia também de zero ao total de elementos focais em Cr_1. Isto é, quanto mais $\mathbf{K}^{-1}$ se aproxima de 1 (um), mais as funções de crença Cr_1 *e* Cr_2 são concordantes.

A função de crença, Cr, resultante da combinação de Cr_1 e Cr_2 é chamada de soma ortogonal, e é denotada por $Cr_1 \copyright Cr_2$. O nome soma ortogonal lembra *a* necessidade de independência entre as evidências que suportam as funções de crença combinadas. A regra para formação da soma ortogonal é chamada de Regra de Dempster.

MODELO COM FUNÇÃO DE CRENÇA

ESTRUTURA DE UM TESAURO

Tesauro é uma lista de termos grupados segundo seus significados. Em geral, um tesauro é construído para fins de indexação. Assim, pode haver coincidência entre o termo escolhido pelo indexador e o termo semanticamente equivalente ou relacionado, procurado pelo pesquisador.

Um tesauro pode ser visto como uma rede, onde os nós são representados por termos e os arcos definem as relações semânticas existentes entre os termos. Por exemplo, a figura 2 mostra uma rede parcial do tesauro da figura 1. As diferentes relações entre os termos são estabelecidas por diferentes (graficamente) tipos de arcos.

Um arco contínuo vai de um "termo mais geral" (TG) para um "termo mais específico" (TE): **E** (Aves de granja) = {Galinha, Pato, Ganso} é o conjunto de termos mais específicos do que Aves de granja, em primeiro nível. Já **G** (Pato) = {Aves de granja, Aves domésticas} é o conjunto de termos mais gerais do Pato até o segundo nível.

Um arco tracejado liga "termos relacionados" (TR): **r** (Galinha) = {Crista, Ovos, Penas} é o conjunto dos termos hierarquicamente relacionados à Galinha. O termo Galinha herda os termos relacionados Ovos e Penas do seu termo mais geral Aves.

Um arco com traços e pontos liga termos sinónimos: **S** (Animais domésticos) = {Animais domésticos, Animais de fazenda} é o conjunto de termos sinónimos de Animais domésticos. Um termo **t** é sempre sinônimo de si mesmo.

Um arco pontilhado liga um termo atómico (um termo que não possui termos mais específicos) a conceitos mais específicos que não figuram no tesauro, mas cuja união forma o termo atômico. O termo Pato pode ser expresso pelo conjunto { Pato ∩ Ovo, Pato ∩ Penas, Pato - ‹Pato ∩ Ovo, Pato ∩ Penas›}. Esta relação não é usada em termos operacionais, mas em termos conceituais fica mais claro o papel do termo relacionado em notação de conjunto.

Animais
Use para Animal.
TE animais domésticos.
Animais de fazenda
Use Animais domésticos.
Animais domésticos
Use para Animais de fazenda
TG Animais.
TE Aves de granja, Gado.
Aves
TG Animais.
TE Aves domésticas.
TR Ovos, Penas.
Aves de granja.
Use para Aves de fazenda, Aves de ter-
reiro
TG Aves, Animais domésticos
• Aves domésticas.
TE Galinhas, Patos, Gansos.
Aves de fazenda
Use Aves de granja.
Aves de terreiro
Use Aves de granja
Aves domésticas
TG Aves, Animais domésticos.
TE Aves de granja.
Cabras
Use para cabra.
TG Gado.
Gado
T6 Mamíferos, Animais domésticos.
TE Vacas, Cabras, Ovelhas.
Galinhas
Use para Galinha, Frango, Frangos.
TG Aves de granja.
Gansos
Use para Ganso.
TG Aves de granja
Leite
TR mamíferos.
Mamíferos
TG Animais.
TE Gado.
TR Leite.
Ovelhas -
TG Animais domésticos, Gado.
Ovos
TR Aves, Aves de granja.
Patos
Use para Patos.
TG Aves de granja.
Penas
TR Aves, Aves de granja.
Vacas
Use para Vaca.
TG Gado.

Figura 1 - Pequeno Tesauro

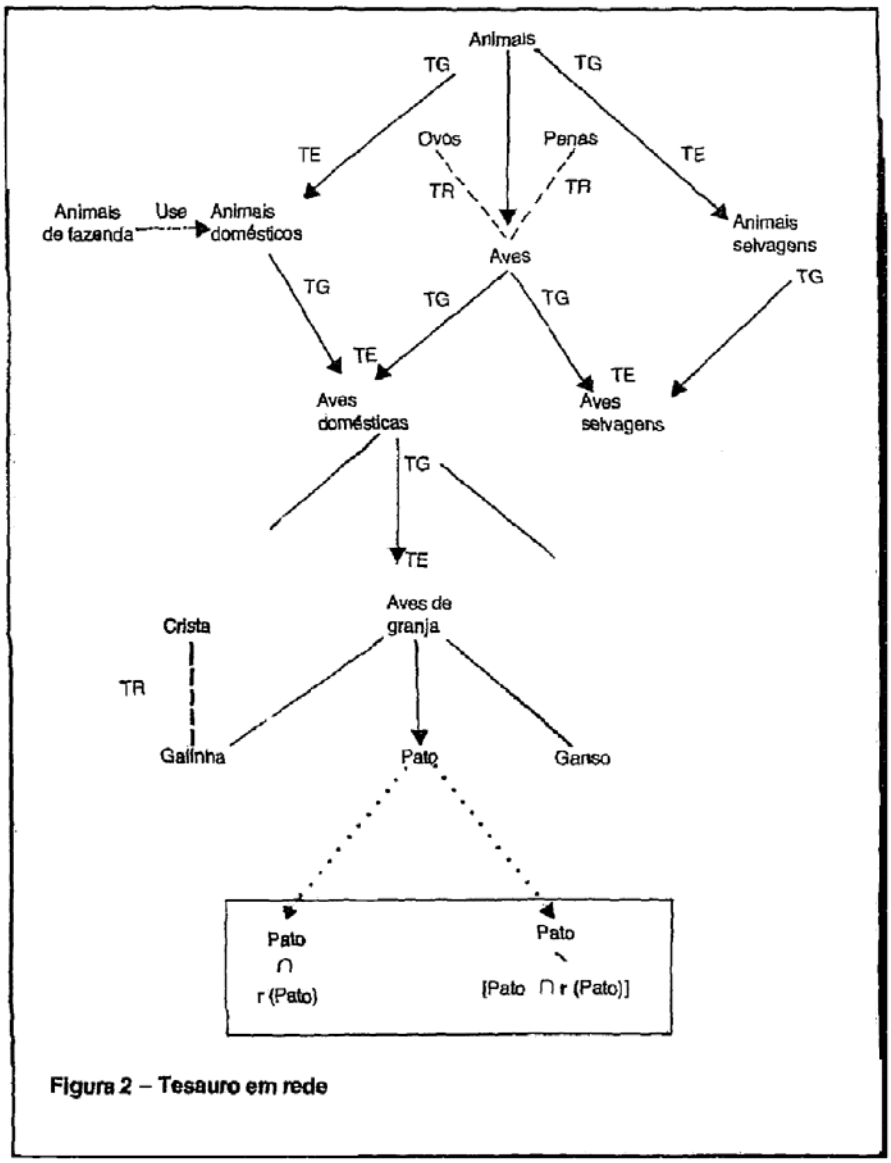

Figura 2 – Tesauro em rede

Se **S** (Pato) representa o conjunto de sinônimos do termo Pato, e **r**(Pato) = {Ovos, Penas} representa o conjunto de termos relacionados a Pato, então **S** (Pato) ∩ **r** (Pato) contém ao menos o conjunto {Pato n Ovo,Pato n Penas}.

Alguns termos no tesauro têm uma conotação especial. Tais termos são chamados de descritores. Um descritor é um termo índice que representa os seus termos sinónimos. No tesauro da figura 1, o termo Animais domésticos é um descritor. A figura 2 liga OS termos Animais domésticos e Animais de fazenda por um arco com traços e pontos, isto significa que eles são sinónimos. Em notação de conjunto **S** (Animais domésticos) = { Animais de fazenda, Animais domésticos}.

Para efeitos operacionais, pode-se minimizar o esforço de indexação usando apenas um dos termos do conjunto de sinônimos como termo índice. Tal termo índice é escolhido entre os termos sinônimos, e apenas ele figura como descritor em um documento indexado. Denotando **S(x)** como o conjunto de termos sinônimos do termo **x**, e considerando que **x** é sinônimo de si próprio, tem-se que **x∈S(x)**. Se **(α)** é um descritor, então, a representa todos os seus sinónimos em S (a), e a é o único elemento em **S (α)** que é descritor.

A estrutura principal da rede é dada por arcos contínuos. Eles ligam descritores, indo de descritores mais gerais para descritores mais específicos. Assim, a estrutura hierárquica de um tesauro pode ser feita toda em função dos seus descritores.

Como os arcos contínuos na rede são orientados, pode-se falar de descendentes e ancestrais de um dado descritor **x**. Em notação de conjunto **E (x)** representa o conjunto de descendentes de **x**, isto é, o conjunto de descritores mais específicos do que **x. G (x)** representa o conjunto de ancestrais de **x**, isto é, o conjunto de descritores mais gerais do que **x**.

Se um termo **r** está relacionado a um termo **g**, em geral, ele está também relacionado a todos os termos mais específicos do que **g**. Assim, o termo Ovos e o termo Penas estão relacionados ao termo Aves. Seria redundante e oneroso relacionar Ovos e Penas explicitamente para cada termo mais específico do que Aves. Assim, qualquer termo r relacionado a um termo mais geral **g** está automaticamente relacionado aos termos mais específicos do que **g**. O inverso também é verdadeiro, mas a relação é mais fraca. Por exemplo, o termo Crista é relacionado ao termo Galinha, sabe-se que essas aves são as únicas dotadas de crista, logo o termo Crista não pode estar relacionado diretamente ao termo Aves. Contudo Aves está relacionado à Crista através de Galinha.

Em notação de conjunto, **r**(**t**) é o conjunto de descritores relacionados aos ancestrais de **t**, ao próprio **t**, ou a descendentes de t. Assim as relações hierárquicas de **t** são dadas por r(**t**), o qual pode ser definido por **r**(**t**) ={x|**z TRx,z∈**(**G** (**t**) **U** {**t**} ∪ **E** (**t**))}.

Por exemplo, **r**(Galinha) = {Ovos, Penas, Crista}, e **r**(Aves) contém {Ovos, Penas, Crista}.

Contudo, um termo pode estar relacionado a termos fora da estrutura hierárquica. Assim, denotando por **R**(**t**) o conjunto de todos os termos relacionados a **t**, direta ou indiretamente, considerando todos os relacionamentos, pode-se definir **R**(**t**) = **r(t)U{y|y∈(G(w)UE(w)),w∈r(t)}.**

Por exemplo, **R**(Leite) = { Mamíferos}, mas R(Leite) = {Animais, Mamíferos, Gado, Vaca, Ovelha}.

A relação semântica **TR** não é uma relação de equivalência, como se poderia supor. Por exemplo: Ovos **TR** Aves, Aves**TR** Penas, mas não é sempre verdade que Penas **TR** Ovos, porque nem todo animal que bota ovos tem penas. Como ê o caso de jacarés, tartarugas etc.

Os termos Galinha, Pato.e Ganso, nas figuras 1 e 2, são considerados termos atómicos, isto é, eles não têm termos mais específicos: **E**(Galinha) = Φ, **E** (Pato) = Φ e **E**(Ganso) = Φ. Cada termo atômico **x** pode ser representado por um conjunto, cujos elementos são a interseção de **x** com seus termos relacionados, **x** n **r(x)** unidos com o conceito específico de **x**, **x** -(**x** ∩ **r**(**x**)). Assim, **x** = (**x** ∩ **r**(**x**)) **U** [**x** - (**x** ∩ **r**(**x**))]. Os arcos pontilhados na figura 2 ligam termos atómicos aos seus conceitos específicos ou a sua interseção com os seus conceitos relacionados. Tal detalhamento de um termo atômico não é explicitamente feito no tesauro, ele apenas se presta para uma formulação do tesauro em termos da teoria ingénua dos conjuntos. Deste modo, um conceito atômico po-

de ainda ser particionado pela sua interseção com conceitos relacionados unidos ao seu conceito específico.

Um termo qualquer do tesauro pode ser visto como um rótulo de um subconjunto de termos atómicos. Por exemplo: Aves de granja = { Galinha, Pato, Ganso }, e Animais domésticos ={ Galinha, Pato, Ganso, Vacas, Cabras, Ovelhas}, segundo o tesauro da **figura 1**.

Naturalmente que Aves de granja está contido em Animais domésticos. Deste modo, o tesauro pode ser visto como um conjunto de termos atômicos, e cada termo **x** não atômico é um subconjunto dos seus termos atómicos descendentes.

Um tesauro **T** pode ser representado por um conjunto minimal de descritores atómicos. Seja D o conjunto de descritores de **T**, e $\Omega=\{\alpha|\alpha\in D, E(\alpha)=\Phi\}$ um subconjunto de **D** constituído apenas de descritores atómicos. Seja $\Theta = U \{ \alpha \cap r(\alpha)) U [\alpha - (\alpha \cap r (\alpha))] | \alpha \in \Omega\}$, um refinamento de Ω, isto é, Ω é uma partição de Θ. Os elementos de Θ são disjuntos, dado que eles refinam elementos de Ω que são também disjuntos. Como qualquer termo **t** de **T** tem um descritor representante em ***D***, e cada descritor a em ***D*** pode ser representado por subconjunto de descritores atómicos de Ω, cada um dos quais é representado por um subconjunto de Θ, então cada termo **t** de **T** pode ser representado por um subconjunto de Θ. Em particular qualquer descritor a de ***D*** é representado por um subconjunto de 0. Assim, tanto Θ quanto a sua partição Ω podem representar o tesauro **T**. Mas, dado que os elementos de Θ não são necessariamente descritores, o conjunto minimal de descritores que representa **T** é Ω.

DEFINIÇÕES BÁSICAS

Nesta subseção é feita uma definição resumida dos principais elementos usados em conjunção com tesauro e enfatizada a notação adotada nesta seção.

- t_1 Use para $t_2, t_3,..., t_k$ denota que t_1 e $t_{2,...,}t_k$ são termos sinônimos e t_1 é escolhido como um descritor para representá-los.
  Exemplo 1: Aves de granja, Aves de terreiro, e Aves de fazenda, na figura 1, são termos sinônimos.

t_2 Use t_1 denota que t_1 é o descritor de t_2. Obviamente, t_1 e t_2 são também sinônimos.

- **TE** é uma sigla para Termo mais Específico.
  Exemplo 2: Galinhas, Gansos e Patos são termos mais específicos que Aves de granja, de acordo com o tesauro na figura 1.

- **TG** é uma sigla para Termo mais Geral. Exemplo 3: conforme a figura 1, Aves, Aves domésticas e Animais domésticos são termos mais gerais que Aves de granja.

- **TR** é uma sigla para Termo Relacionado.
  Exemplo *4:* conforme a figura 1, Ovos e Penas são termos relacionados a Aves. Desde que Aves é um termo mais geral que Aves de granja, então Ovos e Penas são termos relacionados a Aves de granja também. Isto é, Aves de granja herda os termos relacionados de Aves. Em geral um termo **e**, mais específico do que um termo mais geral **g**, herda os termos relacionados a **g**. O inverso também é verdadeiro. Mas no último caso **g** é mais fracamente relacionado aos termos relacionados a **e**.

- **T** denota conjunto de termos do tesauro.
  Exemplo 5: todos os termos do tesauro da figura 1 formam **T**.

- **D** denota o conjunto de documentos.

- **T (d)** denota o subconjunto de termos de **T** que aparecem no documento **d** ∈ **D**.

- **S(t)** denota o conjunto de termos em **T** que são sinônimos do termo **t**. Assume-se que **t** e **S (t)** e que **S (t) = S(x)**, para cada **x e S(t).**
  Exemplo 6: **S** (Aves de granja) = { Aves de fazenda. Aves de terreiro. Aves de granja }.

- *Descritor* **(S(t))** denota a função que retoma o descritor de **S(t)**, isto é, o termo índice que representa o sinónimo de **t.**
  Exemplo 7: Aves de granja = *Descritor (**S*** (Aves de fazenda)).

- **D =**$\{ \alpha | \alpha \in T, \alpha =$ *Descritor* **(S (α))**$\}$ denota o subconjunto de termos em **T** que são descritores. Supõe-se que cada conjunto de sinônimos **S(t),** para qualquer **t** em **T**, tenha somente um descritor, o qual pertence a **S(t)**. Exemplo 8: o conjunto **T**, representado pelo tesauro na figura 1, tem um subconjunto de descritores dado por ***D*** = *{* Animais, Animais domésticos, Aves, Aves de granja, Cabras, Gado, Galinhas, Gansos, Leite, Mamíferos, Ovos, Patos, Penas, Vacas}.

- **G(t)** denota osubconjunto de descritores em *D* que são mais gerais do que**t**, em qualquer nível de altura. Exemplo 9:**G**(Aves de granja) = {Animais, Animais domésticos, Aves, Aves domésticas }.

- **E (t)** denota o subconjunto de descritores em *D* que são mais específicos do que **t**, em qualquer nível de profundidade.
  Exemplo 10: **E** (Aves de granja) = { Galinhas, Gansos, Patos }.

- **E (t, p)** denota o subconjunto de descritores mais específicos do **t** até o nível de profundidade **p**.
  Exemplo 11: pela figura 2, **E** (Animais, 1) ={ Animais domésticos, Animais selvagens, Aves}

- **r(t) = { x**$|$ **z TR x, z ∈ (G (t) U{ 1} U E(t))},** conjunto de descritores hierarquicamente relacionados a **t**.

- **R(t)= r(t) U (y** $|$ **y ∈G(w) UE(w)), w∈r (t) }**denota o subconjunto de descritores em **T** que são relacionados a **t**, exceto **t**.
  Exemplo 12: **R** (Aves de granja) = {Ovos, Penas, Crista}, **R** (Crista) = {Aves, Aves domésticas, Aves de granja, Galinha}.

- **Ω = {** α $|$ **α∈** ***D*****, E (α) = Φ**}, subconjunto de descritores atômicos de D. Ω é uma partição de Θ.

- Ω* é uma σ- álgebra com base Ω.

- **Θ= U { α ∩** $_r$ **(α)) U [α - (α ∩r (α))]** $|$ α **∈Ω}**,Refinamento do Conjunto Ω. Cada descritor de Ωrepresenta um subconjunto de Θ

- **f(d, t)** denota a frequência do termo **t** no documento **d**.

- $F_d = \{\alpha | \exists t \in S(\alpha), f(d, t) > 0$ e $\alpha \in D\}$ denota o conjunto de descritores de *D* que estão presentes no documento **d**.
  Exemplo 13: na tabela 1 , tem-se a frequência dos descritores Animais domésticos, Aves de granja, Gado, Leite, Ovos. Quando a frequência de um descritor em **d** é zero, então este descritor não está presente em **d**. Deste modo, o conjunto F_d, para cada documento d é dado por
  - F_{d1} = {Animais domésticos, Aves de granja, Gado, Leite, Ovos};
  - F_{d2} = {Aves de granja, Gado, Leite, Ovos };
  - F_{d3} = {Animais domésticos, Aves de granja, Ovos };
  - F_{d4} = {Animais domésticos, Gado, Leite}.

Os conjuntos **S(t), E(t), G(t) e R (t)** são estreitamente relacionados. Seja **g ∈ G (t)** um termo mais geral que **t**. Então **t∈ E(g) e S (t) ⊂ E (g)**. Semanticamente se pode dizer que **E(t) ⊂ S (t) ⊂ G (t)** e que **S(t) ∩ R (t) ≠Φ**. A seguir, isto será visto com mais detalhes.

Os termos em **E (t)** são mais específicos do que os termos em **S(t)**, os quais por sua vez são mais específicos do que os termos em **G (t)**, sendo todos semanticamente relacionados. Por esta razão, **E(t)** é semanticamente um subconjunto de **S(t)** e **S(t)** é semanticamente um subconjunto de **G(t).**

A ocorrência dos termos de **R(t)** no documento d sugerem a presença do conceito semântico representado por **t**. Isto é devido a **t** e os termos em **R(t)** serem semanticamente relacionados. Este é o caso para todos os termos **t** em **S(t)**. Neste sentido, a interseção semântica de **S(t)** e **R(t)**, quando **R(t) ≠Φ,** não pode ser vazia.

FREQUÊNCIA DOS DESCRITORES

Seja α∈*D* um descritor. Desde que α∈ **S** *(α), e* apenas um termo em **S**(α) é descritor, então ele tem de ser a. Sabe-se que qualquer termo em **S**(α) pode aparecer no documento **d**, mas somente o descritor a representa os elementos **S**(α) e o conceito semântico que eles suportam. Deste modo, a frequência de a como representante dos elementos de **S**(α) em cada documento **d ∈ D** é dada por

$$\mathbf{m_d}\ (\alpha) = \frac{\Sigma\{f\ (d,\ t)\ |t \in S\ (\alpha)\}}{\Sigma\{f(d,t)|t \in T(d)\}} \qquad (3.1)$$

A quantidade m_d (α) é zero somente quando f(d, t) = 0 para todos os termos t e S *(α).*

Exemplo 14: na figura 1, tem-se que **S** (Aves de granja) = { Aves de fazenda. Aves de terreiro. Aves de granja}. Suponha que as frequências dos termos sinônimos de Aves de granja no documento **d** sejam **f(d),** Aves de granja) = *25,* **f(d**, Aves de fazenda) = 10, e **f(d**, Aves de terreiro) = 5. Suponha agora que a soma de **f(d, t)** para todos os **t** em **T(d)** sejam iguais a 200. Desde que Aves de granja é O descritor para **S** (Aves de granja), então a frequência relativa do descritor Aves de granja no documento **d** é dada por m_d (Aves de granja) = 40/200 = 0,2.

REPRESENTAÇÃO DE DOCUMENTOS NOS MODELOS EXISTENTES

De acordo com Wong & Yao[12], os enfoques feitos até o momento, para representar o conteúdo semântico de um documento, podem ser sumarizados em poucos modelos. Eles identificaram três modelos principais: booleano, vetorial e probabilístico.

Para α ∈*D*, seja m_d (α) uma representação semântica do conteúdo do documento **d ∈ D**. No modelo \booleano um descritor a está presente ou ausente no documento **d**, se m_d (α) |0 ou m_d (a) = 0, respectivamente. No modelo vetorial m_d (α) é considerado como a frequência absoluta de α, isto é, m_d não é normalizado como é feito na equação (3.1). No modelo probabilístico[12], m_d é considerado como uma distribuição de probabilidade em *D, o* que implica a suposição de que os elementos de *D* sejam mutuamente exclusivos. O que, frequentemente, não é o caso, porque pode haver um descritor **g∈D** e um conjunto **E(g)**, de termos mais específicos de **g**, com **E(g)** ∩ *D* ≠ Φ. Isto significa que **E(g)** ∩ D é um subconjunto semântico de **g**, e, contudo, m_d **(g)** pode ser diferente de **Σ {m_d (x) I x ∈ E (g) ∩ D}**. Conseqüntemente, m_d pode não ser uma distribuição de probabilidade em D. O modelo probabilístico de Wong and Yao[12] faz uma suposição muito forte que raramente pode ser mantida: definir uma σ- álgebra e um espaço de probabilidade sobre *D* não é urna tarefa trivial.

REPRESENTAÇÃO DE DOCUMENTOS

SejaΘ um quadro de discernimento,eΩ, o subconjunto de descritores atómicos do tesauro **T**, uma partição de Θ, conforme definido nas subseções "Estrutura de um tesauro" e "Definições básicas". Uma função de crença Cr definida sobre *Θ,* com conjunto focal **F⊂ Ω*** é carregada por Ω. Se todas as funções sob consideração são carregadas por Ω, então Ω pode ser considerado operacionalmente como um quadro de discernimento.

Dado que a σ-álgebra Ω* dos subconjuntos de Θ é formada por todas as uniões possíveis de elementos de Ω e que qualquer termo **t** em **T** rotula um subconjunto de Ω, então o subconjunto de descritores de **T** é também um subconjunto de Ω*. Exemplo 15: considere o pequeno tesauro na figura 1. É fácil ver que os elementos atómicos são dados por Ω = {Cabras, Galinhas, Gansos, Leite, Ovelhas, Ovos, Patos, Penas Vacas}. O descritor Aves de granja é um rótulo para O subconjunto { Galinhas, Gansos, Patos } e Animais domésticos é um rótulo para { Cabras, Galinhas, Gansos, Ovelhas, Patos, Vacas }, e assim por diante.

Como qualquer elemento de *D* é representável por algum elemento em Ω*, pode-se tomar m_d: Ω* →[0,1], como definido na equação (3.1), para ser a mbc da função de crença Cr_d. Doravante, pode se definir as funções de crença e plausibilidade em termos de sua mbc m_d, para cada **S⊂** Ω, como:

$$Cr_d(s) = \Sigma\{m_d(t)|t \in F_d, t \in (E(s) \cup \{s\})\} \qquad (3.2)$$

$$Pl_d(s) = \Sigma\{m_d(t)|t \in F_d, te\ \mathbf{(G\ (s)\ U\ \{s\}\ U\ R\ (s))}$$

Na definição de Cr_d **(s)** são somadas as crenças básicas dos descritores **t** presentes no documento **d**, tal que **t** seja o próprio **s**, ou um termo mais especifico do que **s**.

Na definição de Pi_d (**s**), são somadas as crenças básicas dos descritores **t** presentes em **d**, tal que **t** seja um termo mais geral do que **s**, o próprio **s**, um termo mais específico do **s**, ou um termo relacionado a **s**.

A maneira como Cr_d e Pl_d foram definidas ê operacionalmente equivalente a trabalhar diretamente como os subconjuntos deΘ.

Exemplo 16: considere a tabela 1, cada linha representa a mbc m_d de cada documento **d ∈** { **d_1, d_2, d_3, d_4** }. Os descritores não presentes na tabela têm crença básica igual a zero. A tabela 2, derivada da tabela 1, ilustra os valores da função Cr nos descritores lá presentes. A tabela 3, derivada da tabela 1 ou alternativamente da tabela 2, ilustra os valores da função Pl nos descritores presentes nela.

Quando m_d é uma distribuição de probabilidade em ***D*, Cr_d (α) = Pl_d (α)**. Neste caso, os elementos de *D* são semântica e mutuamente disjuntos. Isto implica **R (α) = Φ e E (α) = G (α) = G (α) = Φ**, para cada ***α ∈ D***. De outro modo, **Cr_d(α) ≤ Pl_d(α)**, para **α ∈ *D*** e **d ∈ D**.

A função de crença Cr_d pode representar o conteúdo semântico do documento **d**. Cr_d(α) é o grau de crença no descritor a, dado que ele é o melhor representante semântico de **d**. No mesmo espírito, **Pl_dα** é o grau de plausibilidade do descritor *α,* dado que ele é o melhor representante semântico de **d**. Neste sentido, Cr_d pode representar o conteúdo semântico do documento **d**. Esta representação pode ser feita, economizando espaço e tempo, armazenando o par (**F_d, m_d**) em cada documento **d**, onde **F_d** é o conjunto focal e m_d é a mbc de Cr_d. O conjunto focal **F_d** representa os descritores presentes no documento **d**, e a mbc **m_d** pode ser restringida a **F_d**, a fim de economizar espaço. Esta representação não implica perda de generalidade, porque, se a mbc **m_d** é conhecida, então sua correspondente função de crença **Cr_d** pode ser obtida da equação (3.2).

Tabela 1 - Frequência relativa dos descritores nos documentos d_1, d_2, d_3, d_4.

	Animais domésticos	Aves de granja	Gado	Leite	Ovos
d_1	0,125	0,335	0,24	0,075	0,225
d_2	0,0	0,378	0,3	0,255	0,067
d_3	0,133	0,654	0,0	0,0	0,213
d_4	0,3	0,0	0,4	0,3	0,0

Tabela 2 - Medidas para Cr_d, derivadas de m_d na tabela 1.

	Animais domésticos	Aves de granja	Gado	Leite	Ovos
d_1	0,7	0,335	0,24	0,075	0,225
d_2	0,678	0,378	0,3	0,255	0,067
d_3	0,787	0,654	0,0	0,0	0,213
d_4	0,7	0,0	0,4	0,3	0,0

Tabela 3 - Medidas para Pl_d, derivadas de m_d na tabela 1.

	Animais domésticos	Aves de granja	Gado	Leite	Ovos
d_1	1.0	0,685	0,44	0,44	0,685
d_2	1,0	0,445	0,555	0,555	0,445
d_3	1,0	1,0	0,133	0,133	1,0
d_4	1,0	0,3	1,0	1,0	0,3

Exemplo 17: considere a tabela 1, lá cada linha, sem os termos com zero frequência, representa a função de crença Cr_d, para $\mathbf{d \in \{d_1, d_2, d_3, d_4\}}$,O documento $\mathbf{d_3}$, por exemplo, é representado por ({ Animais domésticos, Aves de granja, Ovos} { O,133; 0,654; 0,213}).

Não há conflito entre a representação proposta e a representação do modelo probabilístico, exceto que aqui, a interpretação é feita à luz da Teoria de Funções de Crença Shaferianas. A mbc m_d é vista como uma densidade de probabilidade de um subconjunto aleatório de Ω (Shafer & Logan[13]). Nesse sentido, as medidas de utilidade e similaridade, propostas por Wong & Yao[12], entre uma dada consulta e a representação de cada documento continuam a ser válidas.

REPRESENTAÇÃO DA CONSULTA DO USUÁRIO

A fim de escalonar os documentos, deve-se estimar o grau de relevância de cada documento $\mathbf{d \in D}$, com respeito a uma dada consulta. Isto é feito considerando-se a consulta ponderada do usuário e a representação semântica de cada documento.

Suponha que cada documento **d** pertencente a **D** seja semanticamente representado por uma função de crença Cr_d e que cada consulta do usuário seja dada pelo para $\mathbf{(F_q, w, p, r)}$. Onde $\mathbf{F_q}$ é um subconjunto de descritores do conjunto *D;* $\mathbf{w:F_q \rightarrow [0, \infty)}$ é a uma função peso que expressa a crença do usuário em cada descritor $\mathbf{t \in F_q}$ como representante do conteúdo semântico dos documentos a serem recuperados; **p** é o nível de profundidade a ser considerado para termos mais específicos do que **t** que aparecem nos documentos; **r** é um indicador booleano. Tal indicador diz se termos relacionados aos descritores da consulta devem ser levados em conta ou não. Suponha também que cada consulta do usuário seja transformada peio sistema de recuperação **(SR)** em uma função de crença $\mathbf{Cr_q}$ com conjunto focal $\mathbf{F_q}$ e mbc m_q. Para cada $\alpha \in \mathbf{F_q}$, a mbc mq é dada por

$$m_q(\alpha) = \frac{\mathbf{w(\alpha)}}{\mathbf{\Sigma}\ \{\ \mathbf{w(t)}\ |\ t \in \mathbf{F_q}\ \}}$$

O valor Cr_q (α) é a crença do usuário em que *a* seja o melhor representante semântico do conteúdo dos documentos a recuperar.

ESCALONAMENTO POR GRAU DE CONCORDÂNCIA

Ambas as funções Cr_d e Cr_q são definidas sobre Ω. De um modo geral, sem considerar **p e r**, elas são combináveis se há ao
menos um $\mathbf{s \in F_q}$ e um $\mathbf{t \in F_d}$, tal que

1) **t** seja um descritor mais geral do que **s**, $\mathbf{t \in G(s)}$;

2) **t** seja igual a **s**, $\mathbf{t = s}$;

3) **t** seja um descritor mais especifico do que **s,** $\mathbf{t \in E(s)}$;

4) **t** seja um descritor relacionado a **s**, $\mathbf{t \in R(s)}$ ou ainda,

Nesta situação a combinação de Cr_d e Cr_q tem a constante de normalização dada por

$$K^{-1}_{dq} = \Sigma\{m_d(t).m_q(s) \mid S \in F_q . t \in F_d,\ t \in (G(s) \cup \{s\} \cup E(s) \cup R(s))\}$$

Consequêntemente, estas funções são combináveis se $K^{-1}_{dq} > 0$.

Uma intuitiva interpretação para esta definição de $\mathbf{K_{dq}}$ é a seguinte: o produto $\mathbf{m_d(t).\ m_q(s)}$ e considerado no somatório somente quando **s** é um termo que aparece na consulta do usuário $\mathbf{(s \in F_q)}$, **t** aparece no documento d, e uma das quatro condições anteriores ocorre.

Na Teoria de Funções de Crença Shaferiana[1], a constante de normalização $\mathbf{K_{dq}}$ serve como uma medida da extensão do conflito entre $\mathbf{Cr_d}$ e $\mathbf{Cr_q}$, e a medida $\log^{k}_{dq}$ é chamada de peso do conflito

entre Cr_q e Cr_d. Quando a constante $K_{dq}^{-1} = 1$, não há nenhurn conflito entre Cr_q e Cr_q. Quando **O** ‹ K_{dq}^{-1} ‹1, então Cr_d e Cr_q estão parcialmente em conflito. Mas quando $K_{dq}^{-1} = 0$, então Cr_d e Cr_q estão em completo conflito. Neste caso, Cr_d e Cr_q não são combináveis. Pelo exposto K_{dq}^{-1} pode ser tomado como uma medida de concordância Cr_d e Cr_q.

Na definição de K_{dq}^{-1} está sendo considerado que o usuário tenha interesse em recuperar documentos com descritores mais específicos do que os descritores **s** dados na consulta, bem como documentos com descritores relacionados aos descritores **s**, direta ou indiretamente através dos descritores relacionados aos descritores mais específicos do que os descritores **s** da consulta. Mas isto não é sempre o caso. O usuário pode não querer recuperar documentos com termos relacionados ou mesmo com termos mais específicos do que os termos constantes da consulta.

Uma situação desejável seria permitir que o usuário especificasse o seu desejo de incluir ou não termos relacionados e também se ele deseja ou não termos mais específicos do que os termos que figuram na consulta. Na hipótese de o usuário desejar termos mais específicos, então que ele possa dizer até qual nível de especificidade deva ser incluído na consulta. A tudo isto, a representação semântica de documentos por função de crença atende.

Uma definição do grau de concordância entre cada documento **d** e a consulta **q**, representadas respectivamente por Cr_d e Cr_q, de modo a atender às necessidades precedentes é dada por ***A*(d, q, p, r),** conforme figura 3. Onde **E(s, p)** é o conjunto de descritores mais específicos do que **s** até o nível **p**. Quando **p** = 0, nenhum descritor mais especifico é considerado, isto é, **E(s, 0)= Φ**, para todo **s** em ***D***. Quando **p = 1**, então são considerados os descendentes de **s** até o primeiro nível, por exemplo, **E** (Aves, 1) = { Aves domésticas, Aves selvagens}. **E**, quando **p** = ∞, então todos os descendentes de **s** são considerados. Por exemplo, {Aves de granja, Galinha, Pato, Ganso} ⊂ **E** (Aves domésticas).

$$A(d,q,r)=\begin{cases}\text{Se r = 1. então (considera termos relacionados)}\\ \Sigma\{m_d(t).\ m_q(s) \mid s\in f_q.\ t\in f_d.\\ \quad t\in(G(s)\cup\{s\}\cup E(s,p)\cup R(s)\}\\ \text{Se r = 0, então (não considera termos relacionados)}\\ \Sigma\{m_d(t).\ M_q(s) \mid s\in f_q.\ t\in f_d.\\ \quad t\in(G(s)\cup E(s,p)\cup R(s))\}\end{cases}$$

Rgura 3 — Grau de concordância entre a consulta q e o documento d.

O grau de concordância *A*(d, q, r), quando **p = ∞ e r = 1**, tem as seguintes propriedades:

1) 0 ≤ A (d, q, ∞, 1) ≤ 1

2) *A*(d, *q*, °°, *1)* = Σ{m_q(ms). Pl_d(s) s ∈ F_q }, isto é, ***A*(d, q)** *é* proporciona! ao conteúdo semântico de cada descritor presente na consulta **q** e no documento **d**. Este conteúdo semântico *é* estimado pelo grau de plausibilidade.

3) ***A* (d, q, ∞, 1) = 1 - Σ{m_q(s). Cr_d(s) *A*(d, q, ∞, 1)** é igual à massa de crença unitária, menos a média ponderada do grau de crença dos termos que neguem os descritores presentes na consulta **q**. Fica, portanto, em ***A*(d, q, ∞, 1)** apenas a massa de crença não discordante entre os descritores da consulta **q** e do documento **d**. Daí ã denominação da massa em ***A* (d, q, ∞, 1)** de grau de concordância (um nome mais apropriado seria grau médio de plausibilidade).

Naturalmente, ***A* (d, q, p, r)** mede o grau de concordância entre as funções Cr_d e Cr_q, qualquer que sejam os valores arbitrados para **p** e **r**, restrito ao nível de especificidade **p**, e a consideração ou não dos termos relacionados, conforme **r** seja um ou zero.

Exemplo 18: considere a tabela 1, e a consulta **q** do usuário, dada pela função de crença Cr_q com mbc m_q, dada por m_q (Aves de granja) = **1.** Pode-se ver da figura 1 que **G** (Aves de granja) contém o termo Animais domésticos; **E** (Animais domésticos, *&i* contém **U** { Aves de granja, Gado }; **R** (Ovos) contém Aves de granja; e **R** (Leite) contém Gado. Assim, a concordância ***A* (d, q, ∞, 1)** entre a consulta **q** e cada documento na tabela 1 é dado por

***A* (d_1, q,∞,1) = 1X0,125+1X0,335+1 X 0,225 =0,685**
***A* (d_2, q, ∞, 1) = 1 X 0,378 + 1 X 0,067 = 0,445**
***A* (d_3, q, ∞, 1) = 1 X 0,133 + 1 X 0,654 + 1 X 0,213 =1**
***A* (d_4, q, ∞, 1) = 1 x0,3 =0,3**

Usando a propriedade (2), *A* (d, q, ∞, 1) pode ser computado usando a tabela 3 no lugar da tabela 1:

***A* (d_1, q,∞,1) = m_q** (Aves de grania) – **Pl_{d1}**
(Aves de gran]a) = 1 X 0,685 = 0,658
***A* (*d_2*, q, ∞, 1) = m_q** (Aves de granja) • **Pl_{d2}**
(Aves de granja) = 1 X 0,445 = 0,445
***A* (d_3, q, ∞, 1) = m_q** (Aves de granja) - **Pl_{d3}**
(Aves de gran)a) = 1 X 1,0 = 1,0
***A* (d_4, q, ∞, 1) = m_q** (Aves de granja)- **Pl_{d1}**
(Aves de granja) = 1 X 0,3 = 0,3

Assim, ***A* (d_4, q, ∞, 1) ‹*A* (d_2, q, ∞, 1) ‹*A* (d_1 q, ∞, 1)‹*A* (d_3, q,∞, 1).** O grau de concordância ***A* (d_4, q, ∞, 1) = 0,3**, porque os descritores do documento **d_4** são Animais domésticos. Gado e Leite. Desses o único que tem relação semântica com Aves de granja é Animais domésticos. O grau de concordância ***A* (d_3, q, ∞, 1) = 1**, porque os descritores de **d_3** são Animais domésticos. Aves de granja e Ovos. Todos eles são relacionados ao descritor Aves de granja, único descritor na consulta **q**.

COMPARAÇÃO COM ESCALONAMENTO BASEADO EM SIMILARIDADE

Sejam **q** uma dada consulta do usuário e d um documento. Sejam m_q e m_d as correspondentes densidades de probabilidades dos subconjuntos aleatórios **Q** e **D** de **Ω**. Então, de acordo com Wong & Yao[11], uma medida de similaridade (SIM) entre m_q e m_d pode ser definida como

$$SIM(m_q, m_d) = 1 - \beta(m_q, m_d, 1/2, 1/2)$$

onde

$$\beta(m_q.\ m_d, \alpha, \boldsymbol{\lambda}.) = H(\alpha.\ m_q + \boldsymbol{\lambda}.\ m_d) - [\alpha H(m_q) + \lambda.H(m_d)]$$

e

$$H(m_x) = -\ \boldsymbol{\Sigma}\ m_x(\mathbf{t}).\ \log m_x(\mathbf{t}) \qquad t\in F_x$$

Exemplo 19: considere a densidade de probabilidade m_d, para cada **d ∈ {d_1, d_2, d_3 d_4 }**, como mostrado na tabela 4, e a consulta **q** do usuário representada pela linha **q** na tabela 4. Então a medida de similaridade de cada documento é mostrada na tabela 4 sob a coluna SIM. Esta tabela também mostra o grau de concordância de cada documento abaixo da coluna ***A***.

Na tabela 4 pode ser visto que o escalonamento dos documentos baseados na medida de similaridade (SIM) é diferente do escalonamento baseado em graus de concordância (***A***), onde a consulta do usuário tem **p** = ∞, e **r** =1

Tabela 4 - Frequência relativa dos descritores, graus de concordância e similaridade dos documentos, para a consulta na linha q.

	Animais doméstico	Aves de granja	Gado	Leite	Ovos	*A*	SIM
$\mathbf{d_1}$	0,125	0,335	0,24	0,075	0,225	0,685	0,543
$\mathbf{d_2}$	0,0	0,378	0,3	0,255	0,067	0,445	0,548
$\mathbf{d_3}$	0,133	0,654	0,0	0,0	0,213	1,0	0,801
$\mathbf{d_4}$	0,3	0,0	0,4	0,3	0.0	0,3	0,0
q	0,0	1,0	0,0	0,0	0,0		

A ordem dada pelo escalonamento baseado em *A* é $\mathbf{d_3}$, $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_4}$; e a ordem baseada em SIM é dada por $\mathbf{d_3}$, $\mathbf{d_2}$, $\mathbf{d_1}$, $\mathbf{d_4}$. Esta diferença de escalonamento *é* devida à fraca sensibilidade de SIM à carga semântica dos descritores. Obviamente o documento $\mathbf{d_1}$ é preferido ao $\mathbf{d_2}$, desde que $\mathbf{Pl_{d1}}$ (Aves de granja) = 0,685, $\mathbf{Pl_{d2}}$ (Aves de granja) = 0,445, conforme a tabela 3, e a consulta **q** tenciona recuperar somente documentos com o descritor Aves de granja. O descritor Aves de granja representa melhor o conteúdo de $\mathbf{d_1}$ do que o conteúdo de $\mathbf{d_2}$, porque no primeiro sua plausibilidade é maior.

Tomando o grau de plausibilidade como um indicador da carga semântica dos descritores, pode ser visto, na tabela 4, que a medida de similaridade ignora essa massa de plausibilidade. Considerando que o usuário está interessado em documentos cujo conteúdo semântico seja todo explicado pelo descritor Aves de granja e que os documentos $\mathbf{d_1}$ e $\mathbf{d_2}$, na tabela 3, apresentam plausibilidades dadas por Pl_{d1} (Aves de granja) = 0,685, e Pl_{d2} (Aves de granja) = 0,445, é óbvio que o documento d_1 é mais relevante para o usuário do que o documento $\mathbf{d_2}$. Contudo, as medidas de similaridades são respectivamente 0,543 e 0,548, para estes documentos, ignorando completamente a carga semântica do descritor Aves de granja Este exemplo ilustra o quanto a medida de similaridade é insensível à carga semântica dos descritores.

Todavia, o mesmo não ocorre com os graus de concordância para os mesmos documentos, em relação a mesma consulta **q**, eles são respectivamente 0,681 e 0,445, mostrando uma perfeita sensibilidade è carga semântica dos descritores.

Exemplo 20: considere a mbc m_d, para $\mathbf{d \in \{d_1, d_2, d_3, d_4\}}$ e a consulta do usuário representada pela linha q na tabela 5. O grau de concordância e a medida de similaridade para cada documento estão, respectivamente, abaixo das colunas ***A*** e SIM.

Da estrutura semântica dos termos no tesauro, a presença dos descritores Animais domésticos e Ovos em um documento reforça a carga semântica do descritor Aves de granja. Desde que Ovos é relacionado a Aves de granja, e Aves de granja é um termo especifico de Animais domésticos e Ovos. Desta maneira, aumentar a massa de crença em um descritor **x** implica reforçar a carga semântica dos descritores semanticamente relacionados a **x**.

A consulta do usuário na tabela 5 tenciona recuperar documentos cujo conteúdo é mais representável pelo descritor Aves de granja. Espera-se que qualquer medida estimada de relevância (graus de concordância, medida de utilidade, medida de simlaridade etc.) dê maior escalão para documentos com maior massa de crença nos descritores semanticamente relacionados a Aves de granja Na tabela 5, a ordem dos documentos baseada no escalonamento com *A é* $\mathbf{d_3}$, $\mathbf{d_1}$, $\mathbf{d_4}$, $\mathbf{d_2}$; e a baseada no escalonamento com SIM $\mathbf{d_1}$, $\mathbf{d_3}$, $\mathbf{d_2}$, $\mathbf{d_4}$. Este fato mostra, novamente, que SIM ignora *a* carga semântica dos desertores. SIM meramente considera m_d e m_q com distribuições de probabilidade e computa a similaridade entre elas, sem considerar a carga semântica dos descritores.

A análise da tabela 5 reforça a ausência de sensibilidade da medida de similaridade em relação à carga semântica dos descritores semanticamente relacionados aos descritores que aparecem na consulta do usuário. O grau de concordância, contudo, mostra ser muito sensível à carga semântica desses descritores.

Tabela 5 - Frequência relativa dos descritores, graus de concordância e similaridade dos documentos, para a consulta na linha q.

	Animais domésticos	Aves de granja	Gado	Leite	Ovos	*A*	SIM
$\mathbf{d_1}$	0,125	0,335	0,24	0,075	0,225	0,699	0,815
$\mathbf{d_2}$	0,0	0,378	0,3	0,255	0,067	0,578	0,713
$\mathbf{d_3}$	0,133	0,654	0,0	0.0	0.213	0,827	0,788
$\mathbf{d_4}$	0,3	0,0	0,4	0,3	0,0	0,580	0,518
q	0,2	0.6	0.2	0,0	0,0		

CONCLUSÃO

Aqui, é descrito um Modelo de Função de Crença para indexação automática e recuperação de informações. Tal modelo parece mais adequado que o modelo probabilístico para indexação automática *e* escalonamento de documentos. O modelo proposto avalia quanto um descritor, em um tesauro, representa o conteúdo semântico de um dado documento. O principal esforço no modelo é a construção do tesauro, o qual pode também ser construído automaticamente[14]. Dado um tal tesauro, a indexação de documentos pode ser feita automaticamente usando o Modelo de Função de Crença, e muito esforço pode ser economizado, desta maneira.

No Modelo de Função de Crença, cada documento é caracterizado por uma função de crença, e uma consulta do usuário é também representada por uma outra função de crença, ambas sobre o mesmo quadro de discernimento. Deste modo, é possível computar o grau de concordância entre elas. Dada uma consulta do usuário, o grau de concordância é sensível à carga semântica dos descritores presentes no documento e na consulta do usuário. Esta sensibilidade é responsável pela melhor relevância dos documentos recuperados.

Mostrando a flexibilidade do modelo de função de crença, o grau de concordância entre uma consulta do usuário e o conteúdo de um documento sofre a influência do nível de especificidade dos termos considerados e da inclusão ou não de termos relacionados. Esta flexibilidade é um outro trunfo deste modelo.

O presente trabalho pode ser estendido a várias outras aplicações, incluindo indexação com misturas de palavras-chave e frases-chave sem a presença de um tesauro. Neste caso, uma frase é considerada mais específica do que as palavras que a formam. O conjunto de frases mais

específicas constitui o quadro de discernimento. Todas as frases com as mesmas subfrases constituem um subconjunto do quadro e assim por diante. Assim, o Modelo de Função de Crença pode ser generalizado.

REFERÊNCIAS BIBLIOGRÁFICAS

1. SHAFER, *G.,A mathematical theory of evidence.* Princeton University Press, 1976.

2. *DOYLE,L.B. Information retrieval and processing.* New York, John Wiley & Sons, 1975. p. 301-303.

3. HEAPS, H.S. *Information retrieval, computational and theoretical aspects.* New York, Academic Press, 197a p.264-292.

4. SALTON, G., McGILL, M. *Introduction to modern Information retrieval.* New York, McGraw Hill, 1983.

5. JONES, L.P., GASSIE JR. E. W., RADHAKRISHNAN, S. INDEX: The Statistical Basis for an Automatic Conceptual Phrase-Indexing System. *Journal of the American Society for Information Science,* v. 41, n. 2, p. 87-97, 1990.

6. FAGAN, J. Automatic phrase indexing for document retrieval: an examination of syntactic and non-syntactic methods. In: Proceedings of the Tenth Annual International ACMSIGIR - Conference on Research and Development in Information Retrieval. Proceedings..., p. 91-101, New Orleans, 1987.

7. SHAFER, G., SHENOY, P.P., MELLOULI, K. Propagating Belief Functions in Qualitative Markov Trees. *International Journal of Approximate Reasoning,* n. 1, p. 349-400, 1987.

8. SILVA, Wagner T. da. *Algoritmos para Raciocínio Evidenciai usando Funções de Crença.* Departamento de Informática, PUC/Rio, Tese de Doutorado, Abril, 1991.

9. SILVA, Wagner T, da, MILIDIÚ, Ruy L. Combinando funções de crença com uma função Bayesiana. In: *Anais do XXII Simpósio Brasileiro de Pesquisa Operacional.* Fortaleza, out, 1989.

10. SILVA, Wagner T. da, MILIDIÚ, Ruy L. Combinando Crenças Politômicas em um Espaço de Proposições Comum. In: *Anais do VI Simpósio Brasileiro em Inteligência Artificial* -VISBIA, Rio de Janeiro, Nov. 1989.

11. SILVA, Wagner T. da, MILIDIÚ, R.L. Algoritmos para Combinação de Crenças Politômicas. *Revista Brasileira de Pesquisa Operacional,* v. 10, n. 2, dezembro 1990.

12. WONG, S.K.M., YAO, Y.Y. A probability distribution model for Information retrieval. *Information Processing & Management,* v. 25, n. 1,p. 39-53, 1989.

13. SHAFER, G., LOGAN, R., implementing Dempster's Rule for hierarchical evidence, *Artificial Inteligence,* n. 33, p. 271-298, 1987.

Artigo aceito para publicação em 2 de setembro de 1991.

Wagner Teixeira da Silva

Doutor em Informática: Ciência da Computação pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), 1991. Professor do Departamento de Ciência da Computação da Universidade de Brasília

Ruy Luiz Milidiú

Doutor em Pesquisa Operacional pela the University of Califórnia, Berkeley, 1985. Professor do Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

Information indexing and retrieval with a belief function model

Abstract

A belief function model for automatic indexing and ranking of documents with respect to a given user query is proposed here. The model is based on a controlled vocabulary, like a thesaurus, and on term frequency in each document. Each descriptor in this volcabulary is a term among its synonyms chosen to be the index term. A descriptor can have a subset of broader descriptors, a subset of narrower descritors, and a subset of related descritors. Thus descriptors are not mutually exclusive and naive probabilistic models are not adequate. However, a belief function can still be definied over a subset of atomic descriptors. These atomic descriptors are those without narrower terms (denoted Ω*). Subsets of* Ω *can be viewed as broader terms, or as related terms. Hence, the belief function over* Ω *can estimate the semantic content of a document A weighted user query can bem seen as another belief function too. Since both functions are definied over* Ω*, we can compute the conflict between them. The inverse of this computed conflict is a measure of agreement between the document and the user query. Here we propose that the set of documents be ranked by their agreement with the given user query.*

Key words
Automatic indexing; Ranking of documents; Information retrieval; Retrieval model; Belief function theory; Belief function model; Frequency based model; Relevance of documents.