Métodos Estatísticos para Recuperação de Informação

João Ferreira
ISEL
jferreira@deetc.isel.ipl.pt

Alberto Rodrigues da Silva
INESC-ID, IST
alberto.silva@acm.org

José Delgado
Instituto Superior Técnico
Jose.Delgado@tagus.ist.utl.pt

SUMÁRIO

É abordado o problema dos modelos de recuperação de informação, sob o ponto de vista estatístico, no sentido de estabelecer relações entre os diferentes algoritmos e apresentar uma visão unificada dos diferentes modelos com base em métodos estatísticos. É proposta uma notação comum para os mesmos conceitos apresentados por modelos diferentes evitando-se assim a grande diversidade das notações existentes, identificando-se os requisitos para um índice flexível capaz de fornecer matéria-prima para todos os modelos de pesquisa com base nas propriedades estatísticas dos documentos.

PALAVRAS CHAVE

Recuperação de Informação, Modelos, Vectorial, Probabilístico, Inferência, Modelo Linguístico

1 Modelos de comparação

O objectivo dos modelos de comparação criados é definir um conjunto de regras para comparar os termos representativos dos documentos com os das perguntas e assim encontrar um conjunto de documentos que satisfaçam a necessidade de informação expressas na pergunta. Muitos dos sistemas de pesquisa variam na forma como comparam os representativos e o seu nome encontra-se ligado à designação do modelo empregue como demonstrado na Figura 1.

O presente trabalho pretende explorar uma visão unificada dos diferentes modelos, identificando os dados (matéria-prima) comuns, propondo uma indexação flexível e capaz de fornecer a matéria-prima para todos os modelos.

2 Métodos com base na Semelhança

2.1 Método Booleano

Embora não seja o método que melhores resultados produz, é este o mais usado nos sistemas comerciais existentes. A pergunta é feita com um conjunto de termos ligados através das proposições lógicas $(\wedge, \vee, \sim)$, indo o sistema procurar documentos onde se encontrem estes termos de acordo com as proposições usadas. Um dos principais problemas deste método é a

enorme quantidade de documentos que é devolvida, apresentada sem respeitar qualquer ordem. O método de *Fuzzy* tenta resolver este problema, com a introdução de operadores lógicos para incluir associação parcial dos termos às classes [1].

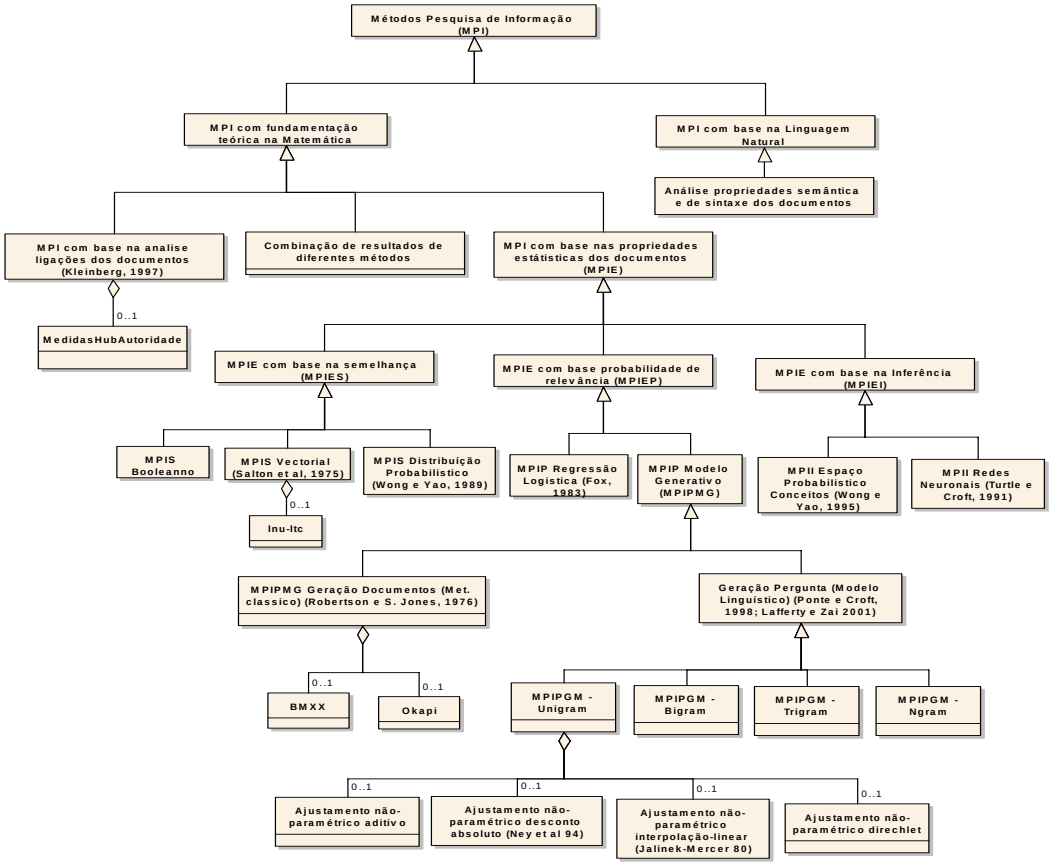

Figura 1: Descrição dos principais modelos de pesquisa de informação.

2.2 Método Vectorial

No método vectorial cada documento é representado por um vector num espaço N-dimensional $D_i = (w_{1i}, ..., w_{ni})$ onde são guardados os pesos de cada termo. Um documento é relevante, para uma determinada pergunta, se o seu peso apresentar um valor superior a um determinado nível previamente definido:

$$sim\left(\vec{D_i}, \vec{Q}\right) = \sum_{t=1}^{N} w_{it} w_{tq} \; ; w_{tq} = \frac{(k_3+1)* f_{tq}}{k_3 + f_{tq}} \log_2\left(\frac{D - d_t + 0.5}{d_t + 0.5}\right); w_{it} = \frac{(1+\log(f_{it}))/(1+\log(\overline{f_t}))}{(1-s)*\overline{n_{tc}} + s * n_{ti}} \qquad (1)$$

2.3 Método de Distribuição Probabilística

Neste método os documentos são representados por uma distribuição multinomial dos termos [2]. Para maior detalhe consultar [3].

3 Métodos Probabilísticos com base na relevância

O objectivo deste tipo de métodos é ordenar documentos com base na probabilidade de relevância em relação a uma necessidade de informação de um utilizador.

Considerem-se 3 variáveis aleatórias: pergunta q, documentos d_i $(1 \le i \le D, i \in \aleph)$ e relevância R $\in \{0,1\}$. Tendo como objectivo ordenar os documentos e considerando a probabilidade de relevância de um documento dada uma pergunta: $P(R \mid q, d_i)$

Nas sub-seccções seguintes apresentam-se os três casos seguintes: (1) Regressão Logística Linear; (2) Método Generativo com base na geração de documentos (teoria clássica); (3) Método Generativo com base na geração de perguntas.

3.1 Regressão Logística Linear

A relevância depende das semelhanças entre a pergunta e os documentos, definindo parâmetros característicos dos documentos e das perguntas (e.g., número de termos semelhantes, comprimento da pergunta e do documento, frequência dos termos, etc…).

Assim o método de regressão permite estimar a probabilidade de relevância de um documento em relação a uma pergunta, baseado num conjunto de parâmetros estimados a partir de um conjunto de treino, da seguinte forma:

$$\log \frac{P(R \mid q, d_i)}{P(\overline{R} \mid q, d_i)} = \alpha + \sum_{i=1}^{6} \beta_i \times x_i \qquad \begin{array}{l} X_1 = \frac{1}{M} \sum_{k=1}^{M} \left(\log f_{qk} \right); X_2 = \sqrt{n}; X_3 = \frac{1}{M} \sum_{k=1}^{M} \log \left(\mathrm{f_{ki}} \right) \\ X_4 = \sqrt{n_i}; X_5 = \frac{1}{M} \sum_{k=1}^{M} \log idf_k; X_6 = \log M \end{array} \qquad (2)$$

Sendo que, M é o número de termos comuns entre a pergunta e o documento; X_1 é a média da frequência absoluta da pergunta; X_2 é o comprimento da pergunta (nº de termos após a remoção das *stop words* e executada a radicalização dos termos); X_3 é a média da frequência absoluta de um documento; X_4 é o comprimento do documento (nº de termos após a remoção das *stop words* e executada a radicalização dos termos); X_5 é a média da frequência inversa de um documento. Parâmetros determinados do conjunto de treino; α é o termo de intercepção da regressão; β_i $\left(1 \le i \le 6, i \in Z\right)$, são calculados a partir dos dados de treino, sub-colecção com julgamentos de relevância previamente conhecidos.

3.2 Métodos Generativos

Apresentam-se dois casos: (1)Geração de documentos: $P(q,d_i|R)=P(d_i|q,R)P(q|R)$, Teoria clássica

[4], donde se destaca a formula Okapi; (2) Geração de perguntas: $P(q,d_i|R)=P(q|d_i,R)P(d_i|R)$ [5,6].

3.2.1 Método Clássico (Geração de documentos)

$$P(R=1|q,d_i)=\frac{P(q,d_i|R=1)P(R=1)}{P(q,d_i)} \underset{ordem}{\propto} \log_2 \frac{P(d_i|q,R=1)P(q,R=1)}{P(d_i|q,R=0)P(q,R=0)} \approx \log_2 \frac{P(d_i|q,R=1)}{P(d_i|q,R=0)}$$

Assumindo que os documentos têm atributos, termos independentes $d_i=(t_{i1},t_{i2},...,t_{in_i})$ cujos valores são $(a_1,a_2,...a_{ni})$ resulta que

$$\log_2 \frac{P(d_i|q,R=1)}{P(d_i|q,R=0)}=\sum_{t=1}^{n_i}\left(\log_2 \frac{P(t_{it}=a_t|q,R=1)}{P(t_{it}=a_t|q,R=0)}-\log_2 \frac{P(t_{it}=0|q,R=1)}{P(t_{it}=0|q,R=0)}\right)=\sum_{t=1}^{n_i}\left(\log_2 \frac{p_t\left(1-\overline{p_t}\right)}{\overline{p_t}\left(1-p_t\right)}\right)$$

$$=\sum_{t=1}^{n_i} w_{it}\text{; com } p_t=P(t_{it}=a_t|q,R=1);\ \overline{p_t}=P(t_{it}=a_t|q,R=0)\text{, definindo } w_{it}=\log_2 \frac{p_t\left(1-\overline{p_t}\right)}{\overline{p_t}\left(1-p_t\right)}$$

Sendo que:(1) p_t é a probabilidade do termo t, ocorrer num documento relevante; (2) $\overline{p_t}$ é a probabilidade do termo t, ocorrer num documento não-relevante.

	Re*levante*	*Não* – Re*levante*	
$t_{it}=a_t$	r_t	d_t-r_t	d_t
$t_{it}=0$	$R-r_t$	$N-d_t-R+r_t$	$N-d_t$
	R	$N-R$	N

Tabela 1:Tabela de contingência para cada termo t_{it}

Esta aproximação só é possível se conhecermos os julgamentos de relevância para todos os documentos na colecção (r e R). r_t é o número de documentos relevantes para o termo t. $p_t=\frac{r_t}{R};\overline{p_t}=\frac{d_t-r_t}{N-R}$ vem que $w_{it}=\log_2\left(\frac{\left(r_t+0.5\right)\left(N-d_t-R+r_t+0.5\right)}{\left(d_t-r_t+0.5\right)-\left(R-r_t+0.5\right)}\right)$.Para evitar singularidades na fórmula de w_{it}, Robertson e Jones [7], introduziram 0.5.

Na ausência de informação relevante, $\overline{p_t}\approx\frac{d_t}{N}$ pode ser estimado pela porção dos documentos que tenham o termo t usando a colecção completa, pois o número de documentos relevantes é pequeno comparado com o número de documentos da colecção. $p_t=\text{constante}$, assume-se constante pois não há forma de estimar o resultando $w_{it}\approx\log_2\frac{N-nt_i}{nt_i}\approx\log_2\left(\frac{N}{nt_i}\right)=idf_i$, para N>>$nt_i$ (nt_i - nº termos do documento i); Esta fórmula foi melhorada por Robertson [7], tendo como base o método de Poisson (Okapi). Robertson assume que a frequência de um termo numa colecção pode seguir duas distribuições de Poisson. Uma

distribuição dos termos dos documentos que representam conceitos ('elite' (E)) e outra distribuição dos restantes:

$$p(f_{it} \mid Q,R) = p(E \mid Q,R)\,p(f_{it} \mid E) + P(\bar{E} \mid Q,R)\,p(f_{it} \mid \bar{E}) = p(E \mid Q,R)\frac{\mu_E{}^{f_{it}}}{(f_{it})!}e^{-\mu_E} + P(\bar{E} \mid Q,R)\frac{\mu_{\bar{E}}{}^{f_{it}}}{(f_{it})!}e^{-\mu_{\bar{E}}}$$

μ-média doc. 'elite'; λ-média doc. não 'elite'; Dada a complexidade da função, Robertson substituiu parâmetros da distribuição por outros baseados na frequência de termos, com comportamentos semelhantes, introduzindo uma constante k_1 (determinada experimentalmente), a qual influência a forma da curva, resultando $w'_{it} = \frac{f_{it}(k_1+1)}{k_1 + f_{it}} w_{it}$ (F2.20). $w'_{it} = \frac{f_{it}(k_1+1)}{k_1 + f_{it}} w_{it}$ (3). k_1 determina como o peso dos termos reagem à variação da frequência dos termos f_{it}. Se k_1 é elevado os pesos são aproximadamente lineares com f_{it}. Na TREC verificou-se que os melhores valores para $k_1 \in [1.2, 2]$, isto mostra que o comportamento dos pesos não é linear com a frequência dos termos f_{it}. Após 3 ou 4 ocorrências de um termo, as ocorrências adicionais têm um impacto reduzido. Falta ainda introduzir as variações de tamanho dos documentos, pois a equação anterior assume que todos os documentos têm o mesmo tamanho. As diferenças entre os comprimentos dos documentos têm duas visões principais: (1) *scope* – documentos longos cobrem mais tópicos que os pequenos; (2) *verbosity* – documentos longos cobrem os mesmos tópicos, usando mais termos. A realidade demonstra ser uma mistura destas duas abordagens [8]. Estas constatações levaram a outro factor de correcção introduzido na fórmula dos pesos. $NF = (1-b) + b\frac{dl_i}{\overline{dl}}$, sendo b uma outra constante determinada experimentalmente. Se b=1, estamos perante uma aproximação pura da *verbosity*. Assim temos a fórmula:

$$w'_{it} = \frac{\frac{f_{it}}{NF}(k_1+1)}{k_1 + \frac{f_{it}}{NF}} w_{it} = \frac{f_{it}(k_1+1)}{k_1\left((1-b)b\frac{dl_i}{\overline{dl}}\right) + f_{it}} w_{it} = \frac{f_{it}(k_1+1)}{K + f_{it}} w_{it} \qquad (4)$$

Existe um segundo factor de correcção (muitas vezes ignorado), dependente do comprimento do documento e do número de termos na pergunta $NF_2 = k_2 \times nq\frac{\overline{dl} - dl_i}{\overline{dl} + dl_i}$ sendo $k_2 \in [0, 0.3]$ para as colecções da TREC. O produto entre os pesos dos termos dos documentos e das perguntas origina as (*Best Match*) BMxx (fórmulas implementadas no sistema *Okapi*).

3.2.2 Método Linguístico

O primeiro método linguístico foi publicado por Ponte e Croft [5], baseado na intuição de que as perguntas não são criadas sem o conhecimento dos documentos e que os

utilizadores têm uma ideia dos termos que ocorrem nos documentos relevantes. A ideia base é estimar a probabilidade de a pergunta ser feita dado um documento, baseado no método linguístico e usar esta probabilidade para ordenar os documentos em vez da probabilidade de relevância. Os métodos linguísticos definem um mecanismo probabilístico para gerar um conjunto de palavras/termos [9,10].

$$O(R=1|q,d_i) \propto \frac{P(q,d_i \mid R=1)}{P(q,d_i \mid R=0)} = \frac{P(q,d_i,R=1)P(d_i \mid R=1)}{P(q,d_i,R=0)P(d_i \mid R=0)} \propto P(q \mid d_i,R=1)\frac{P(d_i \mid R=1)}{P(d_i \mid R=0)}$$

(Assumindo $P(q \mid d_i, R=0) \approx P(q \mid R=0)$)

Assumindo uma distribuição uniforme temos que $O(R=1|q,d_i) \propto P(q \mid d_i, R=1)$, havendo necessidade de calcular $P(q \mid d_i, R=1)$ que é feito em dois passos: (1) estimar o modelo de linguagem baseado no documento d_i; (2) calcular a probabilidade da pergunta de acordo com o método estimado.

$$\log_2 p(q \mid d_i) = \sum_{t=1}^{n} \log_2 p(w_t \mid d_i);\ q = (w_1, w_2, ..., w_n); \{p(w_t \mid d_i) \rightarrow \text{Método-Linguístico do documento}\}$$

Ficando o problema da pesquisa/recuperação de informação (i.e., ordenação dos documentos por grau de relevância para o utilizador) reduzido à estimativa do valor de $p(w_t \mid d_i)$, w_t – peso termo t da pergunta.

A maior parte dos métodos de estimativa com base numa colecção de teste, tenta descontar a probabilidade $p(w_t \mid d_i)$ das palavras vistas nos documentos (colecção de teste) e tenta aumentar a partir de zero a probabilidade $p(w_t \mid d_i)$ de palavras não encontradas nos documentos, usando um método de interpolação.

$$p(w_t \mid d_i) = \begin{cases} p_{ml}(w_t \mid d_i) & \text{se } w_t \in d_i \\ \alpha_{d_i} p_{ml}(w_t \mid C) & \text{caso contrário} \end{cases} \quad \{ml\text{-método linguístico}\}$$

4 Métodos probabilísticos com base na inferência

4.1 Redes Neuronais

As redes neuronais utilizam os métodos de activação expansiva, como forma de expandir o vocabulário de pesquisa de acordo com o contexto e assim complementar o conjunto de documentos seleccionados [11,12].

A técnica usual é construir, manual ou automaticamente, dicionários de termos que especifiquem relações entre os termos, ou dicionários de palavras que contenham definições, e outra informação referente aos termos usados. Nesta expansão são estabelecidas relações entre os documentos. A dificuldade deste método consiste na determinação das relações ou associações que realmente permitem melhorar os resultados da pesquisa. Este método tem sido bem sucedido em domínios especializados.

As técnicas de expansão baseiam-se na existência de funções que especificam as relações particulares entre termos e conceitos.

Os termos são representados por nós numa rede e as relações etiquetadas por arcos entre os nós. Neste método de activação expansiva o processo começa por colocar um peso inicial num nó (determinado empiricamente) e os pesos resultantes são obtidos da aplicação de técnicas probabilísticas. A mesma rede é constituída para as perguntas. A ligação entre estas duas redes é estabelecida ao nível dos conceitos. Este método é bastante exigente a nível computacional e tem-se tornado um método importante à medida que os computadores se vão tornando mais rápidos.

$$P\left(r_j \mid R_i\right) = T \times I;\ \mathrm{T}=\frac{\mathrm{f_{it}}}{\mathrm{f_{it}} + 50 + 150\dfrac{nt_i}{n_{tc}}};\ I = \frac{\log\left(\dfrac{N+0.5}{d_t}\right)}{\log\left(N+1\right)};\ P\left(r_j \mid c_i\right) = b_d + \left(1-b_d\right)T \times I \qquad (5)$$

b_d - valor mínimo da inferência, (sendo 0,4 um valor típico)

5 Referências

[1] Bookstein A.(1985).Probability and fuzzy-set applications to information retrieval. Annual Review of Information Science and Technology 20 117-151.

[2] Wong, S. K. M. and Yao, Y. Y. (1989). A probability distribution model for information retrieval. Information Processing and Management, 25(1):39–53.

[3] Ferreira,J. (2005). Tese de Doutoramento. Metodologia para Concepção e Construção de Sistemas de Recuperação de Informação.

[4] Robertson S. E. e Sparck Jones K. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science 27 129-146.

[5] Ponte, J. and Croft, W. B. (1998). A language modeling approach to information retrieval. In Proceedings of the ACM SIGIR'98, pages 275–281.

[6] Zhai, Lafferty J. (2001). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval, *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* SIGIR 2001.

[7] Robertson S E et al. Okapi at TREC-3 (1995). In: Overview of the Third Text REtrieval Conference (TREC-3). Edited by D K Harman. Gaithersburg, MD: NIST, April 1995

[8] Singhal A. Buckley C. e Mitra M. (1996). Pivoted document length normalization. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval 21-29.

[9] Jelinek, F. (1997). Statistical methods for speech recognition. MIT Press.

[10] Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? In Proceedings of IEEE, volume 88.

[11] Kwok, K. L. (1995). A network approach to probabilistic information retrieval. ACM Transactions on Office Information System, 13:324–353.

[12] Lippmann R. P. (1987). An introduction to computing with neural nets. IEEE ASSP Magazine 4(22).