基于广义词汇共现模型的信息检索

【总结：①这篇文章提出了一种广义词汇共现模型，属于信息检索领域，大体工作就是针对词汇共现研究的两个问题，分析了已有的模型，提出一个模版，模版中有两个槽，这两个槽可以填充已有的相关度计算模型。②查询词临近性 QTP，就是考虑查询词在文档中的距离，并结合广义词汇共现模型计算的相关度，计算出最终的查询词之间的相关度，仅适用于 2~3 个查询词】

作者：乔亚男，齐勇，侯迪

单位：西安交通大学电信学院计算机系

参考：乔亚男,齐勇,侯迪等.基于广义词汇共现模型的信息检索[C].//2008 中国计算机大会论文集.2008:139-147.

领域：信息检索

摘要：1）提出了广义词汇共现模型 GTM，2）提出了以 GTM 为基础的查询词临近性（Query Term Proximity，QTP）辅助信息检索模型。

背景和术语：

1、词汇共现评价的主要问题：在词汇共现研究的过程中，研究者们通常从两个角度进行分析：第一，如果两个词同时出现于一个窗口单元，如何评价这两个词在这个窗口单元中含义的关联程度？第二，如果一个文档中有多个这样的词汇共现窗口单元，如何评价这两个词在这个文档或文档集中含义的关联程度？

2、项（Term）和词（Word）：项是词汇共现模型研究中最基础的概念。文本的内容特征常常用它所含有的基本语言单位（字、词、词组或短语等）来表示，这些基本的语言单位统称为文本的项。

3、窗口单元：词汇共现模型事先约定一个窗口单元的大小，当两个项同属于一个窗口单元的时候认为这个两个项共现。同一个窗口单元中若干个项的有序排列成为项组（Term Array），如果是两个项的有序排列也称为项对（Term Pair）。

4、词汇共现模型的研究主要方面：

1、针对同一个窗口单元，如何计算特定项对的相关度；

常数模型；（相关度为常数，不考虑项之间的距离）

递减模型（多项式递减模型、指数递减模型、吸引与排斥模型、项场模型等）；（相关度随项间距离递减）

2、针对整个文档中的多个窗口单元如何计算一个特定项对的相关度。

频次模型（相关度只和共现窗口的数目有关）

改进频次模型：余弦、掷色、TANIMOTO、Z-Score、T-Score 和互信息模型

广义词汇共现模型：

对于文档 D 中由两个项 a 和 b 组成的项对 $[a,b]$，设 D 中有 m 个该项对的个共现窗口，分别标记为 $[a,b]_1,[a,b]_2,\cdots,[a,b]_{m-1},[a,b]_m$，则各个窗口中项对 $[a,b]$ 的相关度可以分别表示为 $r([a,b]_1),r([a,b]_2),\cdots,r([a,b]_{m-1}),r([a,b]_m)$。如何计算 $r([a,b]_\theta)$ 就是前文提到的第一类词汇共现模型问题。

类似地，对于文档 D 中由两个项 a 和 b 组成的项对 $[a,b]$，该项对对于整个文档来说总体的相关度可以表示为 $R_D([a,b])$。如何计算 $R_D([a,b])$ 就是前文提到的第二类词汇共现模型问题。在单独研究第二类词汇共现模型问题时，一般不显式地用到共现窗口的概念，而使用 a 和 b 共同出现的频率这个类似的概念来代替，即：

$$R_D([a,b])=G(a,b,f(a,b)) \tag{1}$$

$f(a,b)$ 为 a 和 b 共同出现的频率，也就是说，$R_D([a,b])$ 由 a 和 b 相对独立的特性以及 a 和 b 共同出现的频率决定。

如果将共现窗口的概念显式地引入第二类词汇共现模型问题中，并使用前面第一类词汇共现模型问题的表达方式，我们可以得到：

$$R_D([a,b])=G(a,b,\int_D \frac{r([a,b]_\theta)}{E(r([a,b]))}g(\theta)d\theta) \tag{2}$$

这就是广义词汇共现模型中项对相关度的表示式。θ 为 D 中某一出现 $[a,b]$ 的特定窗口；$r([a,b]_\theta)$ 为位于 θ 窗口的项对 $[a,b]$ 的相关度；$E(r([a,b]))$ 为项对 $[a,b]$ 相关度的数学期望，对最终计算出的相关度值进行归一化；$g(\theta)$ 相当于共现窗口的权，$0\leqslant g(\theta)\leqslant 1$。

特别地，对于平均分布（所有窗口同等看待，权值均为 1），就有：

$$R_D([a,b])=G(a,b,\int_D \frac{r([a,b]_\theta)}{E(r([a,b]))}g(\theta)d\theta)=G(a,b,\sum_{\theta=1}^{n}\frac{r([a,b]_\theta)}{E(r([a,b]))}) \tag{3}$$

n 为整个文档中共现窗口的数量。

式（3）中令 $r([a,b]_\theta)=1$，显然有 $E(r([a,b]))=1$，因此

$$R_D([a,b])=G(a,b,\sum_{\theta=1}^{n}\frac{r([a,b]_\theta)}{E(r([a,b]))})=G(a,b,n)=G(a,b,f(a,b))$$

这正是第二类词汇共现模型问题中项对相关度的表达式，也就是说，第二类词汇共现模型问题就是广义词汇共现模型中将 $r([a,b]_\theta)$ 的计算简化为“常数模型”后的特殊情况。

从整个词汇共现模型的架构上来说，第一类词汇共现模型问题是从“共现”的本质出发的，着重研究共现窗口的内部结构，是微观的；第二类词汇共现模型问题则从宏观的角度看问题，着重研究整个文档中各个共现窗口之间的关系。广义词汇共现模型统一了这两类问题，将第一类词汇共现模型问题视为“细胞”，第二类词汇共现模型视为细胞所组成的“组织”，可以充分利用这两类问题已有的研究结果，将诸多传统模型叠加成很多复合模型，以便适应各种不同的应用环境。

QTP 辅助的信息检索模型：

对于一个文档 D 和查询 $Q(q_1,q_2)$，设由传统信息检索模型所求出的相关度为 $RSV(D,Q)$，我们定义经 QTP 辅助的相关度为：

$$RSV_{qtp}=\lambda RSV(D,Q)+(1-\lambda)SIM(q_1,q_2) \tag{7}$$

其中 RSV_{qtp} 为经过 QTP 辅助的相关度，$SIM(q_1,q_2)$ 为使用广义词汇共现模型计算得出的项对 q_1 和 q_2 在文档 D 中的相关度，λ 为权值，可以根据具体的应用环境动态地调整 $RSV(D,Q)$ 和 $SIM(q_1,q_2)$ 的权重，一般 λ 应控制在 0.8 以上。

式（7）是针对双查询词查询的，对于一个有 n 个查询词的查询 $Q(q_1,q_2,\cdots,q_n)$，不考虑顺序，n 个查询词之间两两配对共有 $\frac{n(n-1)}{2}$ 种不同的组合，需要求出这 $\frac{n(n-1)}{2}$ 种组合各自的相关度，然后再归一化：

$$RSV_{qtp}=\lambda RSV(D,Q)+(1-\lambda)\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}SIM(q_i,q_j) \tag{8}$$

可以看出，如果查询词的数量过多，式（8）的计算量会比较大，因此 QTP 辅助的信息检索模型比较适合查询词比较少的情况，一般以两个或三个关键词为宜。