第 3 部分

数据库与存储技术

基于广义词汇共现模型的信息检索

乔亚男　齐 勇　侯 迪

（西安交通大学电信学院计算机系，西安，710049）

摘　要：本文一是提出了广义词汇共现模型（General Term Co-occurrence Model，GTM），该模型统一了传统的词汇共现模型密切关注的两个方面，可以将分别注重于两个方面的诸多传统词汇共现模型叠加成各种复合模型，在两类模型之间建立一个平衡，最终在综合性能指标上达到最佳；二是提出了以 GTM 为基础的查询词临近性（Query Term Proximity，QTP）辅助信息检索模型，对于有多个查询词的查询，和未经 QTP 辅助的原始信息检索模型相比，可以有效地提升查准率。

关键词：信息检索；词汇共现；查询词临近性

1．引言

如何有效地定义文档和查询之间的相关性，以便计算出相应的相关度，是信息检索研究的核心问题。针对这个问题，每一种解决方案就对应一种信息检索模型。自 20 世纪 70 年代以来，研究者们提出了很多不同侧重点的信息检索模型，包括经典的布尔模型、向量空间模型[1, 2]、概率模型[3-5]、统计语言模型[6, 7]和它们的一些衍生模型。对于这些应用已十分广泛的信息检索模型，文档和查询之间相关度的计算通常都是基于各种类型的词汇统计量，如文档频率（Term Frequency，TF）、反文档频率（Inverse Document frequency，IDF）和文档长度等[8]。但是对于查询词临近性（Query Term Proximity，QTP）[9]这个统计量，这些传统的信息检索模型并没有广泛使用。如果一个查询包含了两个查询词，检索系统针对该查询返回了两篇文档 A 和 B，A 文档中这两个查询词距离很远而 B 文档非常近，从常理来讲 B 文档满足该查询的可能性要远大于 A 文档。例如，考虑查询“information retrieval”，对于下面这两个文档：

Document A: “ … information … retrieval … ”

Document B: “ … information retrieval … ”

很明显，文档 B 在检索结果中的排名理应大于文档 A。因为对于文档 A 来说，“information”和“retrieval”这两个查询词被其他的词所隔开，则在文档 A 中这两个词的本意就未必是“information retrieval”了。

可以看出，QTP 作为一种应用比较少的词汇统计量，在传统的信息检索模型中如果能加以考虑，应当可以显著地提高查询结果的精确度。

需要指出的是，QTP 研究和词汇共现（Term Co-occurrence）研究既有联系也有区别。词汇共现研究是基于统计的计算语言学研究领域的重要课题之一。以一个词为中心，可以

基金资助：国家自然科学基金（90612014），教育部博士点基金（20060698018）

联系作者：乔亚男，E-mail: new_siberia@163.com

找到一组经常与之搭配的词，称之为共现词汇集，它描述了这个词的语义上下文和语境。针对共现词汇集的生成和应用的研究称为词汇共现研究，而相应的词汇共现模型（Co-occurrence Model）则反映了共现词汇研究的基本框架和采用的相关技术。词汇共现模型是建立在这样一个基本假设的基础之上的：如果在大规模语料（训练语料）中，两个词经常共同出现（共现）在同一窗口单元中，则认为这两个词在意义上是相互关联的，而且，共现的频率越高，其相互间的关联越紧密[10]。词汇共现研究试图通过对大量文档集的挖掘和分析，得到词汇之间潜在的语义关系（如同义词、近义词和相关词）并量化为相关度，分析结果可以用于查询扩展（Query Expansion）和查询相关词推荐等研究领域，最终的目的是获得词汇间的关系；而 QTP 研究的重点在于通过分析查询词在被检索文档中的位置和它们之间的距离来估计文档和该查询的相关性，最终的目的是得到更精确的文档检索结果。尽管目的不同，但两者用于分析词汇间关系的模型和算法都是相通的，词汇共现的一些研究结果（尤其是词汇共现模型）可以直接在 QTP 研究中得到应用。

本文对 QTP 和词汇共现两个研究领域中现有的词汇共现模型进行了分析，根据这些模型不同的关注点将它们划分为了两个类型，依此为基础提出了广义词汇共现模型（GTM），该模型统一了传统的词汇共现模型密切关注的两个方面，可以将分别注重于两个方面的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。同时，以 GTM 为基础的 QTP 辅助信息检索模型也有效地改善了传统的信息检索模型在多查询词查询环境下的查询效果。

本文其他部分的内容的组织方式如下：第 2 节简要叙述相关研究者的主要工作；第 3 节给出广义词汇共现模型的形式化定义，并对广义词汇共现模型的多种表现形式进行讨论；第 4 节定义了 QTP 辅助的信息检索模型；第 5 节针对第 3 节和第 4 节中的内容进行了具体的实验，首先对传统的词汇共现模型和广义词汇共现模型进行性能比较，接着比较了单纯的向量空间模型和经 QTP 辅助的向量空间模型在多查询词查询环境下的查准率；第 6 节为结论和进一步的工作。

2．相关研究现状

在信息检索领域研究的早期，信息检索系统一般都采用基本的布尔模型，而用户在进行查询的时候使用的查询词通常很少，但自然语言中同义词繁多，再加之英语中即使同一个词在不同地区也有不同的拼法（如 refrigerator 和 fridge 等），直接导致当时的信息检索系统的查全率非常低。为了解决这个问题，部分研究者尝试在用户查询中添加查询词的同义词或近义词来进行查询扩展从而提高查全率，取得了非常好的效果。随后又有研究者试图在用户查询中进一步添加和查询词有语义关系的非同义近义词，尽管研究证明这种方法对于信息检索系统性能改善并不大[11]，但建立在查询扩展研究基础上的词汇共现研究却应运而生并在众多领域发挥了重要作用。

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，如果在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？针对这两个问题，研究者们提出了多种不同的评价模型，但基本都是

针对这两个问题中的某一个问题的处理进行孤立地改进，而简单地忽略或者简化另一个问题的处理，没有进行全面综合的考虑，势必影响了评价模型的性能和适用范围。

项（Term）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被统称为文本的项。由此可见，项和一般意义上的词（Word）是既有联系又有区别，项偏重于描述抽象的概念（concept），是跨语言的，在一种语言中一个项是一个词，同一个项在另外一种语言中可能就是一个词组，甚至是一个短语。其实，“词汇共现”、“词汇共现模型”这个说法改成“项共现”、“项共现模型”似乎更准确一些。

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两个项共现。对大量长短不一，并且每篇文档都有明确专业背景的文档集进行分析时，有时也可以将整个文档视为一个窗口。同一个窗口单元中若干个项的有序排列称为项组（Term Array），如果是两个项的有序排列也称为项对（Term Pair）。词汇共现模型的研究通常都被归结为两个方面：一是针对一个窗口单元，如何计算特定项对的相关度，二是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。

对于第一类问题，也就是同属于一个窗口单元的两个项相关度的计算方法，最常用的是“常数模型”：在某些对精确度要求不高的场合，可以简单地认为同一个窗口内的两个词无论位置如何，它们基于这个窗口的相关度都是恒定的，为一个固定的常数，也就是只考虑项对是否同时出现在同一个窗口单元中，而忽略了窗口内部的细节。使用“常数模型”来计算项对的相关度有其自身的优势，如算法直观，计算量不大，但无疑是不够精确的。比如并没有考虑到所考察的项之间的距离，无论项对是紧邻的还是被很多词隔开，只要是在同一个窗口单元中，都一致被认为是意义上相关的。然而，事实上，即使在同一个窗口单元中，词之间的相关性也不会完全一致。为了使词汇共现模型能够反映出词之间的距离信息，在“常数模型”的基础上需要进行修正。很多研究者注意到了这个问题，提出了很多改进模型，绝大部分都认为项对基于窗口的相关度随着这两个词之间的距离的增大而减小（即“递减模型”），区别只在于距离和相关度之间具体的函数关系上，如多项式递减模型[12]、指数递减模型[13]与吸引和排斥模型（指数模型的变形）[14]以及项场模型[15]等。除了项场模型之外，这些改进模型都属于“递减模型”的范畴，认为项对的相关度随着两个项的距离的增加而减小，直到距离超出窗口的大小，相关度减至 0，区别只在于距离和相关度具体的函数关系。而在项场模型中，两个项距离比较小的时候相关度基本不随着距离的改变而改变，只有当距离大于一个临界值的时候，相关度才随着距离的增加而减小。相对于“递减模型”，在小距离的情况下项场模型相对更符合语言学的常理。

对于第二类问题，也就是对于整个文档中的多个窗口单元特定项对相关度的计算方法，最常用的就是“频次模型”，认为对整个文档来说两个项的相关度只和共现窗口的数量有关[16]。以频次模型为基础，研究者们又提出了余弦（COSINE）、掷色（DICE）、TANIMOTO、Z-Score、T-Score[17]和互信息（Mutual Information）等改进模型，这些模型多数基于信息论中衡量两个事件集合相关性的一些常用统计量。要注意的是，并不是每一个单独对第二个问题进行分析的研究成果都显式地引用了共现窗口的概念，如“频次模型”的常见形式其实是隐式地将整个文档当作一个窗口。

3．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 $r([a,b]_i)$ 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。

例如，COSINE 共现模型的表示式为：

$$COSINE(X,Y)=\frac{F(X,Y)}{\sqrt{F(X)\times F(Y)}} \tag{4}$$

其中 $F(X,Y)$ 为 X 和 Y 的共现窗口的出现次数，$F(X)$ 和 $F(Y)$ 分别为 X 和 Y 出现的次数。

与之搭配的第一类词汇共现模型选择线性递减模型，同时各窗口同权，即

$$SIM_\theta(X,Y)=1-\frac{D_\theta(X,Y)}{W} \tag{5}$$

$SIM_\theta(X,Y)$ 为共现窗口 θ 内 X 和 Y 的相关度，$D_\theta(X,Y)$ 为共现窗口 θ 内 X 和 Y 的距离，W 为约定的共现窗口大小。

很明显，线性递减模型中 $SIM_\theta(X,Y)$ 的值域为 $(0,1)$，数学期望为 0.5。

综上，根据式（3）将式（4）和式（5）合并，就有

$$SIM(X,Y)=\frac{\sum_{\theta=1}^{n}\frac{SIM_\theta(X,Y)}{0.5}}{\sqrt{F(X)\times F(Y)}}=\frac{2}{\sqrt{F(X)\times F(Y)}}\sum_{\theta=1}^{n}SIM_\theta(X,Y) \tag{6}$$

式（6）即为 COSINE 模型和线性递减模型复合而成的广义词汇共现模型。

4．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。

5．实验和结果分析

在本节中，我们进行了两组实验，分别对广义词汇共现模型和 QTP 辅助的信息检索模型的性能和特性进行了分析。

5.1 广义词汇共现模型

实验使用的数据集为 20 Newsgroups。它是一个信息检索领域非常流行的实验数据集，包含从 20 个新闻组收集来的大约 20 000 篇文档，分为 20 个类别，包括软件、硬件、摩托

车、运动、医药和政治等。和其他常用的数据集相比，20 Newsgroups 比较接近 Internet 中实际的语言文字环境，包含着一定数量的垃圾信息，因此可以对广义词汇共现模型在实际应用中的行为进行比较准确的模拟。

实验的工作平台为 IBM Eclipse 3.2.0 和 Apache Lucene 2.2。Eclipse 是 IBM 公司推出的集成开发环境，通常用于 Java 开发，对 C 和 C++的开发也有很好的兼容性。Apache Lucene 是一个纯 Java 的高性能全文搜索引擎开发库，非常适合于跨平台全文搜索引擎的开发[18]。

我们挑选了 6 对关键词，分别使用频次模型、线性递减模型、COSINE 模型以及广义词汇共现模型（线性递减模型和 COSINE 模型复合）来计算它们的相关度，得到的结果如表 1 和表 2 所示。

通过对 4 种模型对 6 对关键词相关度排序的比较，我们可以看出 4 种模型各自的侧重点。频次模型是最基本的词汇共现模型，算法简便效率很高，但相关度计算结果很容易受到文档测试集内容的干扰。例如，20 Newsgroups 文档集中有关计算机软硬件的文章较多，关键词对 1 出现的频率就明显多一些，这对于其他领域的关键词对无疑是不公平的，线性递减模型也有类似的问题；而 COSINE 共现模型不但考虑了共现的频率，还将了两个关键词各自单独出现的频率一并考虑，就大大减少了文档测试集中某一领域文档偏多时对该领域关键词对相关性的高估。

此外，通过表 1 和表 2 的比较我们还发现，窗口大小 30 和 40 两种情况下其余三个模型的相关度排序都没有变化，而 COSINE 模型却有很明显的变化。容易受窗口大小的影响就是 COSINE 共现模型的一个不足之处，所以在单纯使用 COSINE 共现模型时，窗口的大小选择要十分慎重，而且要根据具体问题和文档测试集的不同仔细调整。

表 1 6 对关键词在 4 种模型下的相关度（窗口大小：40）

序号	关键词 1	关键词 2	频次模型	线性递减模型	COSINE	复合模型
1	keyboard	computer	22	12.50	0.1499	0.1704
2	ill	sick	7	2.95	0.1525	0.1286
3	Clinton	president	1	0.98	0.0692	0.1349
4	head	eye	2	0.975	0.0285	0.0278
5	problem	solution	7	4.975	0.1401	0.1991
6	long	short	14	8.575	0.1205	0.1477
相关度排序			1,6,5,2,4,3	1,6,5,2,3,4	2,1,5,6,3,4	5,1,6,3,2,4

表 2 6 对关键词在 4 种模型下的相关度（窗口大小：30）

序号	关键词 1	关键词 2	频次模型	线性递减模型	COSINE	复合模型
1	keyboard	computer	18	12.20	0.1227	0.1663
2	ill	sick	6	2.775	0.1307	0.1209
3	Clinton	president	1	0.975	0.0692	0.1349
4	head	eye	1	0.825	0.0142	0.0235
5	problem	solution	7	4.975	0.1401	0.1991
6	long	short	12	8.40	0.1034	0.1447
相关度排序			1,6,5,2,4,3	1,6,5,2,3,4	5,2,1,6,3,4	5,1,6,3,2,4

将线性递减模型和 COSINE 模型复合而成的广义词汇共现模型就规避了上面两个缺点，没有领域偏向性，相关度排序也不易受到窗口大小的影响，与其他三种模型相比，给出的相关度排序更加准确和稳定。究其原因，广义词汇共现模型的优越之处在于它可以将分别着眼于两类词汇共现模型问题的传统模型综合起来，强化两者的优点的同时弱化其缺点，在两类模型之间建立一个平衡，最终在综合性能指标上达到最佳。

5.2 QTP 辅助的信息检索模型

在本实验中，我们使用的数据集是我们实验室编纂的《人民日报英文版》数据集，共 71 158 篇，纯文本大小为 328MB，包括人民日报英文版 1999 年—2007 年共 8 年间刊登的大部分有关时事、经济和科技方面的文章。我们使用的查询都是双查询词查询，两个关键词之间都有一定的语义联系。首先使用单纯的向量空间模型方法进行检索作为 Baseline，然后再使用 QTP 辅助的向量空间模型（词汇共现模型采用广义词汇共现模型）进行检索，两者的检索结果都保留评分靠前的前 20 篇文档人工判断是否和查询相关(人工判断时使用 2 元相关度：相关为 1 不相关为 0)，分别计算出两个模型的 Precision@5、Precision@10 和 Precision@20，实验结果如表 3 所示。

由表 3 可以看出，针对有语义联系的双查询词查询，经过 QTP 辅助的向量空间模型与未经 QTP 辅助的单纯的向量空间模型相比，在查准率上有着明显的性能提升，符合第 1 节中给出的推测。

表 3 未经辅助和经过 QTP 辅助的信息检索模型间的性能比较

	向量空间模型	QTP 辅助的向量空间模型	性能提升
Precision@5	0.933	0.967	3.57%
Precision@10	0.800	0.950	18.75%
Precision@20	0.717	0.842	17.44%

6. 结论和进一步的工作

本文一是对传统词汇共现模型分析词汇共现问题的两个不同的角度进行了研究，提出了广义词汇共现模型，该模型将这两个分析角度从形式上加以统一。实验证明两类传统词汇共现模型复合而成的广义词汇共现模型可以吸收两类模型的优点，弱化它们的缺点，最终在综合性能指标上优于两个源模型；二是以 GTM 为基础给出了一个 QTP 辅助的信息检索模型，使用该模型在多查询词查询环境下可以有效地提升查准率。

对于第一个问题，我们下一步计划对多种传统模型进行适当的组合，叠加成不同种类的复合模型，对它们进行具体的性能测试和比较；对于第二个问题，对 QTP 辅助的概率模型和统计语言模型的性能分析则是我们未来工作的重点，为我们对该问题的进一步研究打下充分的理论和实验基础。

参考文献

[1] G.Salton, A.Wong, and C.S.Yang. A vector space model for information retrieval. Communications of the ACM, 1975, 18(11): 613-620.

[2] Gerard Salton and Christopher Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management. 1988, 24(5):513-523.

[3] N.Fuhr. Probabilistic models in information retrieval. The computer Journal. 1992, 35(3):243-255.

[4] S.E.Robertson, C.J.V.Rijsbergen and M.F.Porter. Probabilistic models of indexing and searching. In SIGIR '80: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval; Kent, UK: Butterworth & Co. 1980.35-56.

[5] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. ACM Trans. Inf. Syst. 1991, 9(3):187-222.

[6] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval; New York, NY, USA: ACM. 2001.111-119.

[7] Victor Lavrenko andW. Bruce Croft. Relevance based language models. In SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval; New York, NY, USA: ACM. 2001.120-127.

[8] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval; New York, NY, USA: ACM. 2004.49-56.

[9] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval; New York, NY, USA: ACM. 2007.295-302.

[10] 王中等. 基于并行计算的词共现衰减因子模型实现及评价. 河北工业大学学报，2004，33(1)：22-26.

[11] HJ Peat, P Willett. The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems. Journal of the American Society for Information Science, 1991,42(5):378-383.

[12] 鲁松，白硕. 自然语言处理中词语上下文有效范围的定量描述. 计算机学报，2001，24(7)：742-747.

[13] J Gao, M Zhou, JY Nie, H He, W Chen. Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations. In SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval; New York, NY, USA: ACM. 2002.183-190.

[14] 郭锋等. 基于词汇吸引与排斥模型的共现词提取. 中文信息学报，2004，18(6)：16-22.

[15] QIAO Ya-nan, QI Yong and HE Hui.The Research on Term Field Based Term Co-occurrence Model. In SKG '07: Proceedings of the Third International Conference on Semantics, Knowledge and Grid; Washington, DC, USA: IEEE Computer Society. 2001.471-474.

[16] David Yarowsky. ONE SENSE PER COLLOCATION. In HLT '93: Proceedings of the workshop on Human Language Technology; Morristown, NJ, USA: Association for Computational

Linguistics.1993.266-271.

[17] 罗盛芬等. 基于字串内部结合紧密度的汉语自动抽词实验研究. 中文信息学报，2003，17(3)：9-14.

[18] Apache Lucene Homepage: http://lucene.apache.org/.

General Term Co-occurrence Model Based Information Retrieval

QIAO Yanan, QI Yong, HOU Di

(Dept. of Computer Science & Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Key words: Information Retrieval; Term Co-occurrence; Query Term Proximity

Abstract: This paper first proposes General Term Co-occurrence Model，which unites two types of traditional term co-occurrence models, and could derive a series of compound models of them for various conditions. It balances the advantages and shortcomings of original models then achieves best comprehensive performance. Secondly this paper proposes the Query Term Proximity Allied Information Retrieval Model, which can improve the precision ratio effectively for multi-keywords query compared with original information retrieval model.