第43卷 第8期 西 安 交 通 大 学 学 报 Vol. 43 №8
2009年8月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Aug. 2009

具有孤立项过滤的信息检索查询词的分析方法

乔亚男，齐勇，侯迪
（西安交通大学计算机科学与技术系，710049，西安）

摘要：针对传统查询词临近性(QTP)分析方法无法有效提高查准率的问题，提出了一种孤立项过滤的信息检索查询词分析方法.该方法根据词汇相似度较高的查询词对之间具有强可替代性这一事实，从查询词及其实例中分解出查询内的孤立项和文档内的孤立项，在分析查询词临近性之前预先进行孤立项过滤，使之不参与QTP统计量的计算，由此减小了过分强调临近性对查准率的影响.实验结果表明，对于词汇相似度差异比较显著的查询，进行孤立项过滤的查询词临近性分析方法的平均检索精确度比传统分析方法提高14%.

关键词：信息检索；查询词临近性；孤立项；词汇相似度
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2009)08-0006-05

An Analysis Method of Query Terms for Information Retrieval Based on Isolated Terms Filtrating

QIAO Yanan，QI Yong，HOU Di
(Department of Computer Science and Technology，Xi′an Jiaotong University，Xi′an 710049，China)

Abstract：To address the issue that most of traditional query term proximity (QTP) statistics are lack of effective precision improvement，an analysis method of query terms based on isolated terms filtrating is proposed. The method is based on the fact that the query terms with higher similarity have higher substitutability. Isolated terms in-queries and isolated terms in-documents are extracted from query terms and their instances. The filtrating of isolated terms is earlier than the analyzing of QTP，so that the isolated terms are ignored in the calculating of QTP statistics and that the impact of overemphasis of QTP on precision is reduced. Experimental results on queries with significantly different proximities show that the performance of the QTP with isolated terms filtrating is higher (about 14%) than that of the QTP without isolated terms filtrating is.

Keywords：information retrieval；query term proximity；isolated term；term similarity

查询词临近性(QTP)[1]是有关文档与查询之间相关度的一个重要统计量，它基于以下假设：如果一个查询包含了2个查询词，检索系统针对该查询返回了2篇文档，文档A的查询词间距较远而文档B的较近，则文档B满足查询的可能性要大于文档A.在信息检索模型中加入QTP，一定程度上可提高查询结果的精度[2-4].

QTP有2种类型4种形式：类型为基于跨度(Span)的和基于距离的(Distance)；形式为跨度ξ、最小覆盖min$c$、最小距离min$d$和平均距离avg$d$[1].这4种形式对信息检索模型查询精度的改善有着明显的差异，文献[1]对其检索性能进行了比较，指出：除了最小距离可以有效提高检索性能以外，其他的对检索性能的改善十分有限.将基于跨度的与基于距离的参数进行比较，后者的参数性能明显优于前者.与未引入QTP参数的信息检索模型相比较，基于跨度的参数在很多情况下反而降低了查准率.基于跨度的参数是对文档查询词集中程度

收稿日期：2008-09-22. 作者简介：乔亚男(1979—)，男，博士生；齐勇(联系人)，男，教授，博士生导师. 基金项目：教育部高等学校博士学科点专项科研基金资助项目(20060698018)；国家高技术研究发展计划资助项目(2006AA01Z101).

的量化，从常理来讲，查询词在文档中越集中，文档与查询相关的可能性也应该越大，但实验结果却与人们的直观感觉相违背.

本文解释了 QTP 大多数形式不能获得良好性能的原因，在此基础上给出了孤立项的概念，提出了具有孤立项过滤的信息检索查询词的分析方法，以此提高 QTP 对信息检索模型性能的改善幅度.

1　基本概念

1.1　查询词及其实例

符合用户需求的文档应当出现用户给定的全部或一部分查询词，且同一个查询词可能出现多次. 在文档中，一个特定查询词的每一次出现，称其为该查询词的一个实例. 在进行 QTP 相关研究时，根据研究对象是查询词还是实例，可以将 QTP 参数分为查询内 QTP 和文档内 QTP.

定义 1　查询内 QTP：研究对象是查询词，认为文档中同一查询词的不同实例是等价的，如 min$d$ 和 avg$d$.

定义 2　文档内 QTP：研究对象是实例，认为文档中同一查询词的不同实例之间彼此独立，如 ξ 与 min$c$.

表 1 给出了 QTP 常用的 4 种形式的相关属性. 有 n 个关键词查询，对于跨度、最小覆盖和平均距离，这 n 个关键词都在考虑的范围内，而最小距离只需要考虑 2 个关键词. 对文档中每个查询词的全部实例，跨度均要考虑，最小覆盖和平均距离只需考虑部分实例，而对每个查询词，最小距离只需考虑一个实例.

表 1　QTP 常用的 4 种形式

形式	查询词涉及的范围	实例涉及的范围	总体上涉及范围的程度
ξ	全部	全部	大
min$c$	全部	部分	中
min$d$	2 个词	1 个词	小
avg$d$	全部	部分	中

1.2　查询词间的词汇相关度

QTP 研究的查询涉及多个查询词. 2 个查询词的情况比较简单，查询词间的临近性完全由这 2 个词决定，如查询 information 和 retrieval，它隐含着要求文档中 information 和 retrieval 的词距接近，所以使用 QTP 和不使用 QTP 对于性能提高是显而易见的. 对于 3 个或 3 个以上的查询词，情况会复杂许多，如查询 cold、symptom、medicine，从查询词的角度，要求这 3 个词之间的词距都接近，或者仅要求其中的某一个词对的词距接近，而从查询词的实例角度，只需要考虑某一个实例的临近性，或者考虑每一个实例的临近性，但截止目前，有关这些问题的 QTP 研究均未给出明确答案. 实际上，在 QTP 的 4 种形式中，每一种都对应有这 2 类问题的 1 种答案，本文正是试图从查询词相关度的角度分析这个问题.

词汇或概念的相关度 R 是进行语言学研究的重要依据之一，常用的词汇相关度算法有 GlossVector[5]、Jcon[6] 和 Lin[7] 等. 为了确立一个统一的衡量标准，本文使用 GlossVector、Jcon 和 Lin 的线性组合作为最终相关度，即

$$R=\frac{\lambda_1}{e_1}+\frac{\lambda_2}{e_2}+\frac{\lambda_3}{e_3}$$

式中：λ_1、λ_2、λ_3 分别表示算法 GlossVector、Jcon、Lin 的相关度；e_1、e_2、e_3 分别表示算法 GlossVector、Jcon、Lin 在特定范围内的相关度数学期望，此处为同一个查询中每一词对由相应算法算得的相关度的算术平均值.

以查询 cold、symptom、medicine 为例，使用 4 种算法，即 GlossVector、Lin、Jcon 和组合法（本文算法），分别计算两两词之间的相关度，结果如表 2 所示.

表 2　词汇相关度

词对	R			
	GlossVector	Lin	JCon	本文算法
symptom，cold	0.384 2	0.298 7	0.076 7	0.901 8
symptom，medicine	0.710 1	0.340 9	0.093 2	1.234 8
medicine，cold	0.393 5	0.261 5	0.075 9	0.863 5

由表 2 可以看出，symptom 和 medicine 的相关度明显高于 cold 与 symptom、medicine 之间的相关度. symptom、medicine 的概念是平行的，相关度比较高，经常应用于类似的场合；cold 与 symptom、medicine 之间是从属关系，在文档中 symptom、medicine 通常依附于 cold. symptom 和 medicine 不仅可以与 cold 搭配，也可以与 cardiopathy 和 insomnia 等搭配，然而一旦在某一文档中与 cold 搭配，该文档无疑就与查询 cold、symptom、medicine 高度相关. 对于这种查询，cold 是这 3 个词的语义中心，因此 cold 与 medicine 和 cold 与 symptom 均要

求其临近性.

对于 symptom 和 medicine 这 2 个平行概念却有所不同. 相关度越高的词,它们之间的可替代性也越强,因为它们都可以与语义中心建立独立的语义联系. 在文档的某一部分可能专门论述 cold 的 symptom,而在另一部分专门论述 cold 的 medicine,在该篇文档中 symptom 和 medicine 并不临近,但无疑与查询 cold、symptom、medicine 高度相关. 此时,如果强求 symptom 和 medicine 的临近性,会低估该文档与查询的相关度,进而降低信息检索模型的查准率.

综上所述,查询 cold、symptom、medicine 最合理的形式应为

(cold PROX symptom) AND
(cold PROX medicine)

其中,PROX 操作符表示临近性. 该查询需要这样一种文档,即在该文档中 cold 和 symptom 临近并且 cold 和 medicine 临近,对于 symptom 和 medicine,并不要求其间的临近性.

从另一个角度看,PROX 操作符所做的是语义弥补的工作. 如果 2 个查询词的语义相关度较小,这 2 个词之间的相互可替代性就比较小,在检索时就需要使用 PROX;相反,如果语义相关度较大,2 个词之间就具有一定的可替代性,即使这 2 个词并不临近,也足以描述查询的语义,此时使用 PROX 就会适得其反.

1.3 QTP 的 4 种形式分析

下面以临近性和查询词语义相关度的观点,来分析 QTP 4 种形式之间辅助检索性能差异的原因.

所有查询词及其实例都会影响跨度计算,也就是说,跨度考虑了所有查询词及其实例的临近性. 类似地可知,最小覆盖和平均距离考虑了所有查询词及其部分实例的临近性. 由 1.2 节的分析可知,如果部分查询词之间的语义相关度较大,在信息检索模型的相关度公式中引入跨度、最小覆盖和平均距离不一定能提高查准率,甚至会降低查准率. 最小距离只考虑了 2 个查询词及其各自的一个实例,查询词间语义相关度较大的可能性很小,因此 QTP 辅助信息检索的效果比较好. 文献[1]的实验结果有力地证明了上述的分析结果.

2 孤立项

2.1 孤立项概念

根据上述分析可以推断,QTP 辅助信息检索性能难以提高的原因在于,QTP 的 4 种基本形式均未考虑查询词之间的语义相关度,因为相关度小的查询词才需要考虑临近性. 便于讨论,本文将查询中一个相关度较小的查询词对彼此互称为孤立项.

2.2 查询内孤立项

如果一个查询包含了 n 个关键词 $t_1,t_2,\cdots,t_n$,$n\geqslant 3$,根据关键词彼此间的语义距离可将这 n 个关键词划分为 2 个组(T_1 和 T_2),使得组间的关键词语义相关度尽可能小,组内关键词的语义相关度尽可能大,则 T_1 和 T_2 内的关键词彼此互称为查询内孤立项.

如查询{cold, symptom, medicine},根据它们之间的语义相关度可分为

$$T_1=\{\text{cold}\};\quad T_2=\{\text{symptom, medicine}\} \quad (1)$$

{cold}和{symptom, medicine}彼此之间就互称为查询内孤立项.

2.3 文档内孤立项

对于一个查询 $t_1,t_2,\cdots,t_n$,其中的 n 个关键词在文档 D 中共有 m 个实例 $I_1,I_2,\cdots,I_m$. 确定一个窗口最大值 w,再根据实例之间的距离将它们划分为 k 个组,划分原则如下.

(1)组内相临实例之间的最大距离$\leqslant w$.

(2)组与组之间的最小距离$>w$.

由此,可将满足下列条件的组内项确定为文档内孤立项. 组内项只属于查询分组中的一个组,即要么全部属于 T_1,要么全部属于 T_2.

例如,查询 t_1,t_2,t_3,t_4,其分组结果为

$$T_1=\{t_1,t_2\};\quad T_2=\{t_3,t_4\}$$

对于文档

$$D=\{t_1,X,X,t_3,X,X,X,t_1,X,t_4,X,X,X,X,t_1,X,t_2\}$$

设 $w=2$,则文档 D 的分组结果为

$$D=\{(t_1,X,X,t_3),X,X,X,(t_1,X,t_4),X,X,X,X,(t_1,X,t_2)\}$$

根据条件判断可知,第 3 组(t_1,X,t_2)中的项$\{t_1,t_2\}\in T_1$,因此该组内项为文档内孤立项,又称之为孤立组.

3 基于孤立项过滤的查询词分析

3.1 孤立项分解算法

由查询内孤立项的定义知,对于用户给出的关键词,需根据词间语义相关度分组(2 组),相应的孤立项分解算法的具体步骤如下.

(1)选择一种语义相关度算法 R.

(2)将 n 个关键词中的每一个视为 1 组,即共分

为 n 个组，其可作为初始分组.

(3)使用算法 R 计算词间的语义相关度，共得到 $n(n-1)/2$ 个相关度.

(4)从 $n(n-1)/2$ 个相关度中选取语义相关度最大的一个结果(如果同时有多个相等的最大相关度，则随机选取其中的一个)，并将涉及这个结果的 2 个关键词 t_i 和 t_j 合并为一组$[t_i,t_j]$，再将所有涉及 t_i 和 t_j 的相关度算法$R[t_i,t_k]$和$R[t_j,t_k]$(共 $2(n-2)$个)置换为 $R[t_{ij},t_k]$，则有

$$R[t_{ij},t_k]=\frac{1}{2}(R[t_i,t_k]+R[t_j,t_k])$$

此时剩余$(n-1)$组关键词，其有

$$n(n-1)/2-(n-2)-1=(n^2-3n+2)/2=(n-1)(n-2)/2$$

个相关度.

(5)重复步骤 4，直至剩余 2 组关键词，这 2 组关键词就是最后的分组结果.

例如，查询 t_1,t_2,t_3,t_4,t_5，彼此间的相关度见表 3. 根据孤立项分解算法，t_1,t_2 首先合并为一组，合并后的分组相关度如表 4 所示. 将 t_3,t_5 合并为一组，合并后的分组相关度如表 5 所示. 最后，t_1,t_2 和 t_3,t_5 应合并为一组，此时已剩下 2 组，分组完成，得到 $T_1=\{t_1,t_2,t_3,t_5\}$，$T_2=\{t_4\}$. 表中的 N/A 表示不适用.

表 3　查询 t_1,t_2,t_3,t_4,t_5 彼此间的相关度

	R				
	t_1	t_2	t_3	t_4	t_5
t_1	N/A	0.700	0.450	0.300	0.370
t_2		N/A	0.490	0.560	0.560
t_3			N/A	0.280	0.690
t_4				N/A	0.430
t_5					N/A

表 4　t_1,t_2 合并后分组的相关度

	R			
	t_1,t_2	t_3	t_4	t_5
t_1,t_2	N/A	0.470	0.430	0.465
t_3		N/A	0.280	0.690
t_4			N/A	0.430
t_5				N/A

表 5　t_3,t_5 合并后分组的相关度

	R		
	t_1,t_2	t_3,t_5	t_4
t_1,t_2	N/A	0.468	0.430
t_3,t_5		N/A	0.355
t_4			N/A

3.2　孤立项过滤

查询内孤立项过滤类似于一种检索表达式重整. 例如查询{cold, symptom, medicine}，在不考虑 QTP 的情况下，它的查询表达式为

cold AND symptom AND medicine

在考虑 QTP 的情况下，查询表达式为

(cold PROX symptom) AND (symptom PROX medicine) AND (medicine PROX cold)

如果在考虑 QTP 的基础上将孤立项的因素考虑进去，根据查询的分组结果(见式(1))，上述查询表达式中的(symptom PROX medicine)属于同组关键词，应该过滤掉，最终的查询表达式应为

(cold PROX symptom) AND
(medicine PROX cold)

文档内孤立项过滤的情况比较简单. 根据前文所述的文档内孤立项的识别方法，在计算 QTP 参数时均将孤立组中的项排除在外. 孤立组中关键词的实例也不再视为关键词，其仅在计算词间距时与其他非关键词一样当作文档中一个普通词汇来考虑.

3.3　算法分析

3.1 节中的孤立项分解算法本质上就是 K 均值算法的变形. 在孤立项分解算法的步骤 2 中，将每一个词视为 1 组，其实就是将每一个词视为 K 均值算法中的聚类中心. 由于孤立项分解算法的最终目标是将关键词分为 2 组，因此在此并不需要用类似 K 均值算法中的测度函数来判定收敛情况[8-10].

对于同一个信息检索，给出的关键词数量越多，对用户意图的描述就越精确，查准率也就越高. 但是，关键词太多，孤立项过滤便会影响查准率，也就是说孤立项过滤算法的敏感性有所降低. 1 个关键词、2 个关键词的查询是不需要孤立项过滤的，孤立项过滤适用于 3 个关键词的查询，但对 4 个或 4 个以上关键词的查询，其敏感性比较低.

此外，为了验证孤立项过滤机制在 QTP 辅助信息检索中的有效性，本文对未使用和使用孤立项过滤机制的 QTP 辅助信息检索性能进行了比较. 在选取查询词时，挑选了相关度差异较大的词汇. 查询 t_1,t_2,t_3，先使用 1.2 节中的算法计算出两两词汇的相关度，即

$$R_1=R_{t_1,t_2};\quad R_2=R_{t_2,t_3};\quad R_3=R_{t_1,t_3}$$

然后使用类似 χ^2 检验的方法度量这 3 个相关度的差异性，即

$$\chi^2=\sum_{i=1}^{3}\frac{(\max(R_1,R_2,R_3)-R_i)^2}{\max(R_1,R_2,R_3)}$$

4 实 验

实验采用《人民日报》英文版数据集和 Reuters 标准数据集[11]，同时选择 χ^2 值较大的词汇的 30% 作为实验用查询词，使用 Okapi BM25 公式对实验数据集进行检索. 对于查询的前 50 个检索结果，分别计算其 4 个原始的 QTP 参数($\min d$, $\mathrm{avg}d$, ξ 和 $\min c$)和经过孤立项过滤的 QTP 参数($\min d'$, $\mathrm{avg}d'$, ξ' 和 $\min c'$). 为了便于比较，这 4 个参数使用下列公式进行叠加，即

$$Q=\frac{1}{\lg(\min d+1)}+\frac{1}{\lg(\mathrm{avg}d+1)}+\frac{1}{\lg(\xi+1)}+\frac{1}{\lg(\min c+1)}$$

$$Q'=\frac{1}{\lg(\min d'+1)}+\frac{1}{\lg(\mathrm{avg}d'+1)}+\frac{1}{\lg(\xi+1)}+\frac{1}{\lg(\min c'+1)}$$

式中：Q 表示原始 QTP 参数的叠加统计量；Q' 表示经孤立项过滤后的 QTP(IQTP)参数叠加统计量.

将 50 个检索结果根据各自 QTP 值和利用 BM25 公式计算得出的加权平均数(BM25 权值为 0.80)重新排序，并对前 20 个结果进行人工相关度判断. 最后，分别计算前 5 个、前 10 个、前 20 个检索结果(P@5、P@10、P@20)的 QTP 和经加权的精度值(OWP@20). 3 个关键词和 4 个或 4 个以上关键词的查询结果如表 6 和表 7 所示.

表 6 3 个关键词的实验结果

QTP 参数	P@5	P@10	P@20	OWP@20
Q	0.440	0.520	0.550	0.525
Q'	0.760	0.620	0.540	0.603
性能提升率/%	+72.73	+19.23	−1.82	+14.86

表 7 4 个或 4 个以上关键词的实验结果

QTP 参数	P@5	P@10	P@20	OWP@20
Q	0.760	0.840	0.767	0.779
Q'	0.880	0.840	0.767	0.806
性能提升率/%	+15.79	+0.00	+0.00	+3.47

从表 6 可以看出，对于 P@5 和 P@10，IQTP 的性能均明显优于 QTP，对于 P@20 却略显不足. P@5、P@10、P@20 均未考虑检索结果的内部排序，如 P@10 只考虑排序后前 10 个检索结果的平均精确度，并认为这 10 个检索结果彼此之间是等价的；OWP@20 考虑了 20 个检索结果的内部顺序，而且排序靠前的赋予了较高的权重. 由于 IQTP 将相关度较高的检索结果排在前，因此 OWP@20 下的性能提高了 14.86%. 相应地在 OWP@5 和 OWP@10 下，IQTP 的优势会更加显著.

从表 7 可以看出，无论哪种算法，检索精确度都明显高于表 6. 对于 4 个和 4 个以上的关键词，IQTP 的性能仍然具有一定的优势，但不如 3 个关键词的情况. 这验证了 3.3 节中孤立项过滤算法的敏感性问题.

5 结 论

本文提出了一种基于孤立项过滤的信息检索查询词的分析方法，其在查询词临近性分析方法的基础上，采用查询词之间的相关度，并认为相关度较高的查询词之间有着较强的可替代性，如果一概考虑词汇间的临近性反而会降低检索精度. 实验结果表明，针对两两关键词的相关度差异比较显著的查询，采用孤立项过滤的查询词临近性分析方法，其性能明显优于未采用孤立项过滤的情况.

参考文献：

[1] TAO Tao, ZHAI Chengxiang. An exploration of proximity measures in information retrieval [C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2007: 295-302.

[2] BUTTCHER S, CLARKE C, LUSHMAN B. Term proximity scoring for ad-hoc retrieval on very large text collections [C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2006: 621-622.

[3] BUTTCHER S, CLARKE C L A. Efficiency vs. effectiveness in terabyte-scale information retrieval [C]// Proceedings of the 14th Text Retrieval Conference (TREC). Gaithersburg, MD, USA: National Institute of Standards and Technology, 2005: 1-6.

[4] RASOLOFO Y, SAVOY J. Term proximity scoring for keyword-based retrieval systems [C]// Proceedings of the 25th European Conference on IR Research. Heidelberg, Germany: Springer, 2003: 207-218.

[5] PATWARDHAN S, PEDERSEN T. Using WordNet-based context vectors to estimate the semantic related-

ness of concepts [C]// Proceedings of EACL 2006 Workshop on Making Sense of Sense — Bringing Computational Linguistics and Psycholinguistics Together. Cambridge, MA, USA: MIT Press, 2006: 1-8.

[6] JIANG J, CONRATH D. Semantic similarity based on corpus statistics and lexical taxonomy [C]// Proceedings of International Conference of Research in Computational Linguistics. Heidelberg, Germany: Springer, 1997: 1-15.

[7] LIN D. An information-theoretic definition of similarity[C]// Proceedings of International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 1998: 296-304.

[8] 张云，冯博琴，麻首强，等. 蚁群-遗传融合的文本聚类算法[J]. 西安交通大学学报，2007，41(10)：1146-1150.
ZHANG Yun, FENG Boqin, MA Shouqiang, et al. Text clustering based on fusion of ant colony and genetic algorithms [J]. Journal of Xi′an Jiaotong University, 2007, 41(10): 1146-1150.

[9] 冯中慧，鲍军鹏，沈钧毅. 一种增量式文本软聚类算法[J]. 西安交通大学学报，2007，41(4)：398-401.
FENG Zhonghui, BAO Junpeng, SHEN Junyi. Incremental algorithm of text soft clustering [J]. Journal of Xi′an Jiaotong University, 2007, 41(4): 398-401.

[10] LEWIS D D, YANG Y, ROSE T, et al. RCV1: a new benchmark collection for text categorization research [J]. Journal of Machine Learning Research, 2004, 5: 361-397.

[11] ROBERTSON S E, WALKER S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval[C]// Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berlin, Germany: Springer-Verlag, 1994: 232-241.

（编辑　苗凌）