第35卷 第18期 计 算 机 工 程 2009年9月
Vol.35 No.18 Computer Engineering September 2009
·软件技术与数据库· 文章编号：1000—3428(2009)18—0054—03 文献标识码：A 中图分类号：TP309

基于文档空间向量距离的查询扩展

王秀娟[1]，郑康锋[2]
(1. 北京工业大学计算机学院，北京 100124；2. 北京邮电大学信息工程学院，北京 100876)

摘 要：查询扩展是信息检索中优化查询的一种有效方法。在分析几种基于互信息的查询扩展方法的基础上，将检索词在文档空间中的距离引入到互信息计算中，提出基于向量距离的改进互信息的查询扩展方法。实验结果表明，该方法能够有效提高信息检索中的查询效果。
关键词：信息检索；查询扩展；互信息；向量距离

Query Expansion Based on Vector Distance in Documents Space

WANG Xiu-juan[1], ZHENG Kang-feng[2]
(1. College of Computer Sciences, Beijing University of Technology, Beijing 100124;
2. School of Information Engineering, Beijing University of Posts and Telecommunications, Beijing 100876)

【Abstract】 Query expansion is an effective method to the queries in information retrieval. After analyzing several methods of query expansion based on mutual information, this paper introduces the vector distance between terms in documents space to the improvement of mutual information and brings forward a query expansion method based on vector distance in documents space. Experimental results show that the method can effectively improve the performance in information retrieval.
【Key words】 information retrieval; query expansion; mutual information; vector distance

信息检索是信息化时代帮助人们快速获得所需信息的有效途径，但是，由于用户不能够准确构造表达信息需求的检索式，导致检索效率低下甚至失败。例如，用户选择使用的词可能与检索库中出现的词不匹配[1]，如何解决词的不匹配现象已经成为信息检索领域的重要研究方向。查询扩展(query expansion)是解决该问题的有效方法之一。它利用各种统计信息，对原始查询进行有利于检索的扩展，从而使得查询可以包含更多的相关信息，有效解决大部分词的不匹配问题，达到提高信息检索性能的目的。

1 查询扩展方法

查询扩展的方法有很多，大致可以分为2类：

(1)人工进行查询扩展，即人工挑选与查询词相关的其他特征词，将其加入到原始查询中构成新的查询；这类查询扩展稳定性和全面性不能够保证，实现局限性比较大。

(2)利用某种资源自动对查询进行扩展[2-5]。这类查询扩展需要利用某种包含有词与词间相关信息的资源来进行，通常是利用大规模的语料，通过统计的方法，自动获得词与词间的相关信息。

2 基于改进互信息的查询扩展

2.1 基本的互信息查询扩展方法

基本的词共现模型直接使用词与词之间的互信息作为词与词相关性的度量。互信息(Mutual Information, MI)度量2个事件 x 和 y 发生的相互依赖程度，是这2个事件的发生概率 $p(\cdot)$ 的函数，其定义公式参见下式：

$$MI(x,y)=\lg\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

在信息检索中，查询 Q 一般都是由多个关键词组成的，由于多词的互信息统计很难计算，因此常用两词的互信息来代替多词之间的互信息。借用式(1)可以计算2个关键词之间的互信息。这时，$p(x)$ 表示关键词 x 单独出现的概率；$p(x,y)$ 表示关键词 x 和 y 共同出现的概率，并且存在：

$$p(x)=C(x)/C(\cdots) \tag{2}$$

$$p(xy)=C(x,y)/C(\cdots) \tag{3}$$

则

$$MI(x,y)=\lg\frac{C(x,y)/C(\cdots)}{(C(x)/C(\cdots))\cdot(C(y)/C(\cdots))} \tag{4}$$

其中，$C(x)$ 表示 x 在检索文档集中的出现次数；$C(y)$ 表示 y 在检索文档集中的出现次数；$C(x,y)$ 表示 x 和 y 在检索文档集中的共同出现次数；$C(\cdots)$ 表示检索文档集中的总词数。

文献[6]提出了用窗口法来统计关键词的共现概率。在这种方法中，连续出现的一些关键词组成了窗口单元，对窗口单元内的关键词进行概率统计。这时，有

$$p(x,y)=CWin(x,y)/CWin(\cdots) \tag{5}$$

其中，$CWin(x,y)$ 表示 x 和 y 在窗口中共同出现的次数；$CWin(\cdots)$ 表示窗口中所有关键词两两共现的次数。窗口单元可以是固定长度的，比如取连续5个关键词组成的窗口，或者窗口也可以是文本中的自然段以及整篇文本等。

查询 Q 可以看作是一些检索词的集合 $T_q=\{t_i\}$，$i=1,2,\cdots,N_q$，N_q 表示查询 Q 中的检索词个数，则查询 Q 与词 t 之间的互信息 $MI(Q,t)$ 可以表示为

基金项目：北京工业大学博士科研启动基金资助项目(52007011200703)
作者简介：王秀娟(1979－)，女，讲师、博士，主研方向：信号与信息处理；郑康锋，讲师、博士
收稿日期：2008-05-25 **E-mail：**xjwang@bjut.edu.cn

$$MI(Q,t)=MI(T_q,t)=\sum_{i=1}^{N_q}MI(t_i,t),\ t_i\in T_q \tag{6}$$

在对查询 Q 进行扩展时，首先计算它与所有可能的扩展词之间的互信息，将得到的互信息按降序排列，取排在最前面的 N 个词来进行查询扩展。

2.2 基于带衰减因子的互信息的查询扩展

为了使词共现模型能反映出词之间的距离信息，文献[6]在基本的词共现模型的基础上进行了改进。文章认为，在同一个窗口单元(句子)中，词与词之间的相关性是随着词之间距离的增加而减少的。并且假定，词与词之间的相关性随着词间距离指数衰减。因此，在原有互信息计算的基础上，加入了一项反映词间距离信息的衰减项，形成了带衰减因子的词共现模型，其计算公式如下式所示：

$$SIM(x,y)=\frac{CWin(x,y)}{CWin()}\cdot\lg\left(\frac{CWin(x,y)/CWin(\cdots)}{(C(x)/C(\cdots))\cdot(C(y)/C(\cdots))}\right)\cdot e^{-\alpha(D(x,y)-1)} \tag{7}$$

其中，$D(x,y)$ 是词 x 和词 y 之间的平均距离。这里的距离是指在一个窗口中，所考察的 2 个词之间出现的其他词的个数。α 是常数，表征词与词之间相关性随距离衰减的剧烈程度。

考察词间距离是否能够代表关键词之间的相关度差别。在很多情况下，相隔较近的关键词不见得相关性肯定高。例如，对于下面一段已分词后的文本内容：

NBA 之所以有今天的辉煌与乔丹密不可分，NBA 成功地把乔丹塑造成了篮球场上的战神，使乔丹名利双收，同时乔丹也让全世界认识了 NBA 。

在这段文本中，按照笔者对文本的理解，显然“乔丹”和“NBA”的相关度应该很高，但是如果以自然句为窗口考察两者的互信息，按照前文所述的方法，NBA 与乔丹之间的距离对于它们的互信息计算反而造成了不利影响。

而对于下面一段文本：

美国国务院发言人鲍彻 17 日宣布，美国、日本和韩国的高级官员将于 18 日在华盛顿举行磋商，为可能于下周在北京举行的有关朝鲜核问题的三方会谈协调立场。

“朝鲜”与“核问题”的距离信息的加入显然可以为它们的互信息计算加分。所以认为词间的物理距离小不能必然说明词间的相关程度高，应该考虑引入其他的距离计算方法。

2.3 基于文档空间的关键词距离计算的查询扩展

在向量空间模型中，文档是由关键词来表征的。在以每个关键词为基的 m 维关键词空间中，每个文档都映射到其中的一个点，则对于整个检索文档库可以形成词-文档矩阵；反之，也可以理解为，以每篇文档作为 n 维文档空间中的一个基，则每个词都可以映射为该向量空间中的一个点。事实上，信息检索系统在对检索文档库建立倒排索引以后，可以得到关键词与文档之间的信息，如下所示：

关键词	数据
检索词 1	<文档 ID，文档中检索词 1 频率>
	<文档 ID，文档中检索词 1 频率>
	…
	<文档 ID，文档中检索词 1 频率>
检索词 2	<文档 ID，文档中检索词 2 频率>
	<文档 ID，文档中检索词 2 频率>
	…
	<文档 ID，文档中检索词 2 频率>
…	…

借鉴在检索词所构成的向量空间中检索词的权重值计算公式 $tf\cdot idf$，通常可以定义第 i 个检索词 t_i 在第 j 篇文档 d_j 上的权重值 w_{ij}，计算公式如下式所示：

$$w_{ij}=(0.5+0.5\cdot\frac{tf_{ij}}{\max tf_i})\cdot\lg\left(\frac{n_i}{Len_j}\right) \tag{8}$$

其中，tf_{ij} 表示在文档 d_j 中检索词 t_i 的出现频率；$\max tf_i$ 为检索词在所有文档中的最高出现频率；n_i 是检索词 t_i 在全部文档集中出现的总频率；Len_j 是文档的长度，等于文档 d_j 中出现的所有检索词数。

这样，每个检索词都转换成 n 维文档空间中的一个 n 维向量，它们之间的距离计算则是笔者已经熟知的。在实验中分别采用欧氏距离和向量的夹角余弦 2 种距离来进行实验，具体计算如式(9)和式(10)所示：

$$D_1(t_i,t_j)=[\sum_{k=1}^{n}|w_{ik}-w_{jk}|^2]^{1/2} \tag{9}$$

$$D_2(t_i,t_j)=\left(\frac{\sum_{k=1}^{n}w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{n}w_{ik}^2)(\sum_{k=1}^{n}w_{jk}^2)}}\right)^{-1} \tag{10}$$

检索词在文档向量空间中的空间距离能够反映它们在文档集中的分布，距离越小的检索词分布越趋于相似，它们之间的相似度越高。反之，距离越大的检索词分布趋于不同，它们之间的相似度也越低。

需要指出的是，可以选择使用检索文档库中的全部文档集来表示检索词，在执行检索之前进行全局分析进而完成查询扩展。但是这种情况有不利的一面，如果检索文档库太大的话，则构建的检索词向量维数会很大，带来巨大的计算负荷；也可以选择在利用原始查询得到初步检索结果后，使用返回的小部分文档集来表示检索词，进行局部分析进而完成查询扩展。

仍然把检索词在文档向量空间中的空间距离 $D(t_i,t_j)$ 引入到检索词 t 和查询 Q 的互信息 $MI(Q,t)$ 计算中，得到

$$MI(Q,t)=MI(T_q,t)=\sum_{i=1}^{N_q}\big(MI(t_i,t)\cdot f(D(t_i,t))\big),\ t_i\in T_q \tag{11}$$

其中，$f(D(t_i,t))$ 是关于查询词 t_i 和检索词 t 之间距离的反比函数，在本文的实验中取以下 2 种函数：

$$f(D)=\frac{1}{D+\delta} \tag{12}$$

$$f(D)=e^{-\alpha D} \tag{13}$$

其中，δ 是为了防止分母出现为 0 而设置的一个常数；α 仍表示衰减因子。

3 实验结果

在实验中取整篇文档作为窗口，用 2003 年国家“863”信息检索测试提供的检索库和查询，构建了一个小规模检索库，共计 2 074 篇文档，以这些文档表示检索词进行基于全局分析的扩展。实验中利用了“863”计划项目提供的其中 5 条查询进行查询扩展，以不加查询扩展的检索为 baseline，实现了本文提出的查询扩展方法。

实验步骤如下所述：

(1)对检索文档库中全部文档进行基于词典的中文分词。

(2)统计词频，建立倒排索引，得到词典中每个检索词在文档向量空间中的表示。

(3)统计词典中所有检索词两两之间的共现信息，计算其向量距离，并得到它们基于向量距离的互信息。

(4)对每条查询语句进行基于词典的中文分词。

(5)针对每条查询，将查询中包含的每个查询词与词典中其他检索词之间基于向量距离的互信息按降序排列，取排在最前面的 L 个检索词加入到原始查询中。

实验包括 4 个方面：

(1)选择距离测度。首先确定使用指数衰减方式，且固定 $\alpha=1, L=1$，通过 2 种不同的距离测度分别进行查询扩展，扩展后的检索结果如图 1 所示。

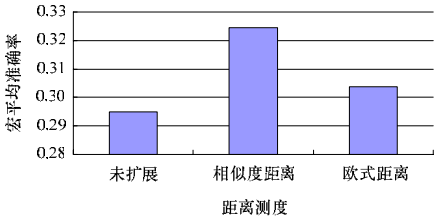

图 1　不同距离测度的扩展结果

从图中可以看到，相似度距离的扩展结果较优。

(2)选择衰减方式。此时以欧式距离测度来确定检索词之间的向量距离，固定 $L=1$，通过 2 种衰减方式进行查询扩展，其中，指数衰减中的 α 取 1，检索结果如图 2 所示。

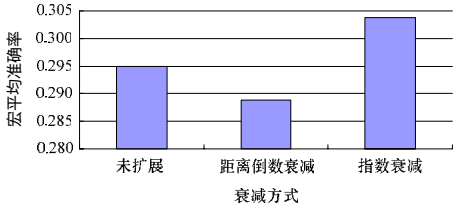

图 2　不同衰减方式的查询扩展结果

从实验结果可以看到，指数衰减方式优于距离倒数衰减。

(3)对衰减因子的影响。采用相似度距离测度，指数衰减方式，固定 $L=2$，对不同的 α 值进行实验，实验结果如图 3 所示。

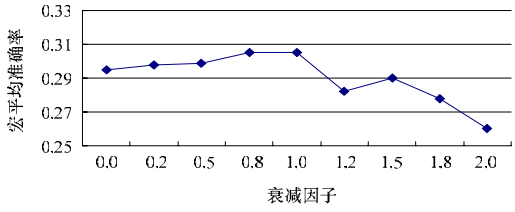

图 3　衰减因子的影响

这里只给 α 赋予了部分值，从这些值的结果可以看出，$\alpha=0.8$ 或 1 时结果近似最优。

(4)扩展词数的影响。采用相似度距离测度，指数衰减方式，固定 $\alpha=1$，对不同的 L 值进行实验，实验结果见图 4。

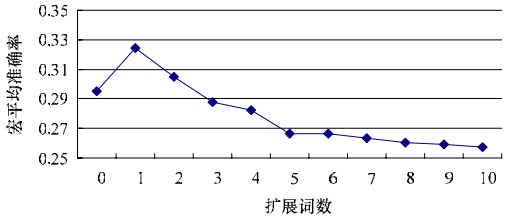

图 4　扩展词数的影响

从实验结果可以看到，扩展词数并非越多越好，过多的扩展词的加入反而会使得系统的检索性能降低。

4　结束语

通过上述实验结果可以看到，在选择了相似度距离、指数衰减方式之后，本文所提出的基于文档空间的向量距离的查询扩展方法能够取得较好的查询扩展效果。本文的不足之处在于没有与文献[6]所提出的查询扩展方法进行对比。

参考文献

[1] Kantor P. Information Retrieval Techniques[J]. Learned Information, 1994, 29(2): 53-90.

[2] Mandala R, Tokunaga T, Tanaka H. Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion[C]//Proc. of the 22nd ACM-SIGIR Conference. [S. l.]: ACM Press, 1999: 191-197.

[3] Buckley C, Mitra M, Walz J, et al. Using Clustering and Super Concepts Within SMART: TREC-6[J]. Information Processing and Management, 2000, 36(1): 109-131.

[4] Hearst M A. Improving Full-text Precision on Short Queries Using Simple Constraints[C]//Proc. of the 5th Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, NV, USA: [s. n.], 1996.

[5] Salton G. Automatic Text Processing — — The Transformation, Analysis and Retrieval of Information by Computer[M]. [S. l.]: Addison-Wesley Publishing Co., 1989.

[6] 贺宏朝. 利用人工和自动生成的资源进行中文信息检索查询扩展[D]. 天津: 天津大学, 2002.

编辑　顾逸斐

(上接第 53 页)

7　结束语

本索引结构充分利用了 DTD 所提供的简化 XML 文档结构，进行查询的预处理，不仅能够将非法表达式查询控制在 DTD 查询阶段，而且通过让 XML 节点也带上 DTD 信息，提高了表达式查询效率。而且在进行结构连接时，使用成熟的包含关系和普通的连接算法，跳过了不必要的节点，减少了扫描代价，实验证明其效率较高。

参考文献

[1] Goldman R, Widom J. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases[C]//Proc. of the 23rd International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997.

[2] Milo T, Suciu D. Index Structures for Path Expressions[C]//Proc. of the 7th International Conference on Database Theroy. Jerusalem, Israel: [s. n.], 1999: 277-295.

[3] 易　平, 胡运安, 陈福生, 等. 基于 PATRICIA-TRIES 的 XML 路径索引设计[J]. 小型微型计算机系统, 2006, 27(3): 474-480.

[4] Poola L K, Haritsa J R. SphinX: Schema-conscious XML Indexing[R]. Pennsylvania State University, Tech. Rep.: TR-2001-04, 2001.

[5] 万常选, 刘云生, 徐升华, 等. 基于区间编码的 XML 索引结构有效实现结构连接[J]. 计算机学报, 2005, 28(1): 113-127.

[6] 万常选. XML 数据库技术[M]. 北京: 清华大学出版社, 2006.

编辑　任吉慧