基于半监督学习的查询扩展模型①

苏俊杰，陈 俊

(装备指挥技术学院，北京 101416)

摘 要：查询扩展是针对信息检索中常见的“词不匹配”问题提出的一种优化方法。通过分析现有查询扩展方法的不足，提出一种基于半监督学习的查询扩展模型，该模型将查询扩展看作一个分类问题，并采用直推式支持向量机对样本进行训练。实验结果表明该方法进一步提高了搜索引擎的查全率和查准率。

关键词：信息检索；查询扩展；直推式支持向量机；半监督学习；SVM

Query Expansion Model Based on Semi-Supervised Learning

SU Jun-Jie, CHEN Jun

(Institute of Command and Technology of Equipment, Beijing 101416, China)

Abstract: Query expansion is a optimization method for “word mismatch” issues in information retrieval domain. By analyzing the shortcomings of existing methods, query expansion model based on semi-supervised learning is proposed, the model seems query expansion as a classification problem, and using transductvie support vector machine to train the samples. Experiments show that the recall and precision rates of search engine are further improved by this method.

Key words: information retrieve; query expansion; transductive support vector machines; semi-supervised learning; SVM

1 研究背景

目前的信息检索系统，几乎都是基于关键字或者关键词的机械式符号相匹配的方式进行检索的，因此，查询词只有在被检索的文档中出现，文档才能被检索出来。但在人类语言中常常出现一词多义的现象再加之有时候人们不能清晰地表达自己的查询意图，使得很多与查询相关的重要文档不能被检索出来。查询扩展技术的出现就是为了解决信息检索中“词不匹配”(word mismatch)的问题，在初次检索基础上添加一些与查询项相关的词或词组形成新的查询，然后利用新的查询对文档进行再次检索。大量实验表示，通过查询扩展技术的应用，搜索引擎的性能得到了进一步提高。

2 传统查询扩展方法分析

查询扩展优化作为解决表达差异的一种有效方法，即在原查询词的基础上加入与用户用词相关的词或词组，组成新的、更准确的查询词序列，这在一定程度上能弥补用户的查询表达与可能的候选段落的差别，尽可能以较小的遗漏检索出候选文档[1]。

查询扩展算法的关键是建立扩展词表。目前扩展词表的建立方法通常有 2 种：1) 基于语义的扩展词表构造方法，该方法通过构建一些大规模的手工词典，例如 WordNet、HowNet 等来进行查询扩展；2) 基于大规模语料库统计信息的构造方法[2]。

基于语义的查询扩展方法通过词典指示词项之间的关系，可以方便的对同义词和歧义词等做出选择，但手工词典的建立较为困难，尤其是适用于中文信息检索的词典，很难对词项之间关系做出全面的描述。基于大规模预料库的查询扩展方法避免了手工词典的建立，由首次检索得到的全部或局部文档作为语料来源，也是目前最常用的查询扩展方法。但通过分析发

① 收稿时间:2011-07-01;收到修改稿时间:2011-09-01

现，该方法在进行扩展词项选择时通常仅考虑词项之间的统计信息，例如共现、概率、互信息等，影响因素单一并没有考虑页面的结构信息及页面重要程度等因素。如果查询项 q 和扩展词项 c 同时出现在标题或首段，当词频相近时我们会认为该词项比出现在其他位置(例如文章中部或尾部)的词项具有更高的相关性，而且重要程度较高的页面，其扩展词项应具有更高的权重。基于大规模语料库的查询扩展算法的另外一个缺点是该算法通常引入一些参数进行调节，当算法中参数较多时，根据经验来对参数进行设置和调节显得尤为困难。

3 基于半监督学习的查询扩展模型

查询扩展从语料库中选取与查询项 q 最相关的词项加入到原始查询中，我们将候选词项组成的集合称为 C，则 C 中的词项与 q 之间存在两种关系，即{相关，不相关}。查询扩展可以看成一个分类问题，将 C 中的词项划分为两类，则可采用分类的方法来解决查询扩展问题。近年来，基于半监督学习（Semi-supervised Learning）的分类方法逐渐受到人们的青睐，在图像、文本等许多分类领域得到了应用，并取得了良好的效果，成为当前研究的热点。基于以上分析，本文提出一种基于半监督学习的查询扩展模型。

3.1 直推式支持向量机基本原理

直推式支持向量机（Transductive Support Vector Machines，TSVM ）[3]是一种半监督学习方法，在直推式学习中，训练样本集中只需要少量的有标签样本，而无标签样本是相对较多的。该方法克服了监督学习算法需要构建大量有标签样本的缺点，而且通过引入无标签样本，能够更好地刻画出整个样本空间的特性，从而使经过训练所得到的分类器具有较好的适用性。

TSVM 形式化的定义如下：

给定一组有标签的训练样本集R_1：

$(x_1,y_1),\ (x_2,y_2)\dots(x_l,y_l)\ ,\ x_i\in R^n,\ y_i\in\{-1,+1\}$

和另一组具有相同分布的无标签训练样本集R_2：

$x_1^*, x_2^*\cdots x_m^*$

使得联合样本集$R_1\cup R_2$最大间隔的被分开，则 TSVM 可以被描述为以下优化问题：

Minimize：

$V(y_1^*,\dots,y_m^*,\ w, b, \varepsilon_1,\dots,\varepsilon_m^*)=\frac{1}{2}w\cdot w+C\sum_{i=1}^{l}\varepsilon_i+C^*\sum_{j=1}^{m}\varepsilon_j^*$

Subject to:

$\forall_{i=1}^{l}: y_i[w\cdot x_i+b]\geq 1-\varepsilon_i$

$\forall_{j=1}^{m}: y_j^*[w\cdot x_j^*+b]\geq 1-\varepsilon_j^*$

$\forall_{i=1}^{l}: \varepsilon_i>0$

$\forall_{j=1}^{m}: \varepsilon_j^*>0$

$\forall_{j=1}^{m}: y_j^*\in\{-1,+1\}$

其中C和C^*是用户指定的用于调节的参数，C 代表有标签样本的影响因子，C^*代表无标签样本的影响因子，ε_i和ε_j^*分别表示有标签样本和无标签样本的损失项。

通过在训练样本中引入无标签样本的信息，TSVM 的泛化性得到进一步的加强。但由于在训练过程中需要预先设置正负样本的比例，而且该算法需要多步迭代才能得出最佳的分类模型，使得算法的计算开销非常大，在一定程度上影响了它的应用。针对以上缺点，许多学者对其作出了改进，陈毅松等人提出了一种渐进直推式支持向量机（Progressive Transductive Support Vector Machines，PTSVM）[4]，PTSVM 算法的主要思想是，在 TSVM 的训练过程中不需要事先确定无标签样本中正负样本的比例，而是在训练过程中每次迭代只选取少量“重要”的无标签样本赋予临时标签，然后利用临时标记的样本进行下一次迭代的模型训练，并对先前错误标记的样本采用标签重置法使其具有一定的纠错能力。PTSVM 不仅性能优于 SVM，而且计算开销比 TSVM 要小。本文采用 PTSVM 作为样本分类模型。

3.2 基于 PTSVM 的查询扩展模型

基于 PTSVM 的查询扩展模型将信息检索中的查询扩展看作一个分类问题，采用分类模型将候选词项分为“相关”和“不相关”两类，将标记为“相关”的词项加入到原始查询中，以解决词的不匹配问题。然而查询扩展又有其特殊性，不能够用普通的 SVM 模型来解决。基于 PTSVM 的查询扩展模型应主要解决以下两个问题：1) 样本向量的表示；2) 目标函数的定义。

3.2.1 样本向量表示

传统的查询扩展模型以初次检索返回的前 n 篇文档作为扩展单位，一方面，当前 n 篇文档中包含的不相关文档较多或不相关文本较长时可能会带来主题偏移的问题，影响搜索引擎的查准率；另一方面，该方

法很难体现文档的结构、重要程度等一系列复杂的特征。本文在进行样本向量表示时以单个文档作为扩展单位，可以很好地解决以上两个问题。

PTSVM 将样本表示为向量的形式，样本向量通常使用最能体现样本特征的值来表示，如，n 为样本的特征数，即向量的维数。样本的特征表示很大程度上影响分类模型的性能。本文将样本表示为(查询项，候选词项)的形式，具体特征如下表所示：

表 1 基于 PTSVM 的查询扩展模型所使用的特征

特征值	含义
tf_q	查询项 q 在文档 d 中的词频
idf_q	查询项 q 的逆文档频率
tf_i	词项t_i在文档 d 中的词频
idf_i	词项t_i的逆文档频率
$c(q,t_i)$	查询项 q 与词项t_i在该文档中的互信息值
$log\ (c(q,t_i)+1)$	互信息值的 log 刻度
$c(q,t_i)/\|d\|$	查询项 q 与词项t_i的绝对互信息值，$\|d\|$表示文档的长度
log (BM25 score)	BM25 是常用的网页排序算法，表示该文档在初次检索中的顺序。
$length(q)$	查询项长度
$first\_location(q)$	查询项 q 在文档中首次出现的位置
$first\_location(t_i)$	词项t_i在文档中首次出现的位置
$total\_hits$	该文档点击数，表示文档的重要程度

样本被表示为上表描述的 12 维向量，随着其他一些相关的特征的出现，可以不断增加样本的维数，以更好地刻画样本特征，并用新的特征向量重新训练分类模型。PTSVM 的另一个优点是样本维数增加对算法性能的影响很小。

3.2.2 目标函数定义及算法流程

支持向量机的目标函数是分类模型进行优化的依据。普通支持向量机对不同级别训练样本数量的差异会产生模型偏置问题。例如在查询扩展模型中，只有与原始查询最相关的词项才能标记为（+1 表示相关，-1 表示不相关），在文档中大多数词项与查询项不相关，因此不相关词项构造的样本数会远远大于相关词项构造的样本数，而原始的目标函数并没有对此进行区分，目标函数为了达到最优，使得分类器对训练样本数较少的相关词项训练实例区分不好，甚至视若无睹，从而违背了的问题的初衷。

基于以上考虑，我们定义了一个新的目标函数，如下所示

Minimize：

$V(y_1^*,\dots,y_m^*)$，

$w,b,\varepsilon_1,\dots,\varepsilon_m^*)=\frac{1}{2}w\cdot w+\mu_k(C\sum_{i=1}^{l}\varepsilon_i+C^*\sum_{j=1}^{m}|\theta_j|\varepsilon_j^*)$

其中μ_k为不同等级样本数量之间的调节参数，用来解决模型的偏置问题。μ_k的参数值定义为：

$$\mu_k=\frac{max\ (instances\in lable_i)}{instances\in lable_k}$$

$instances\in lable_i$表示标记值为$lable_i$的样本数。

基于 PTSVM 的查询扩展流程如图 1 所示：

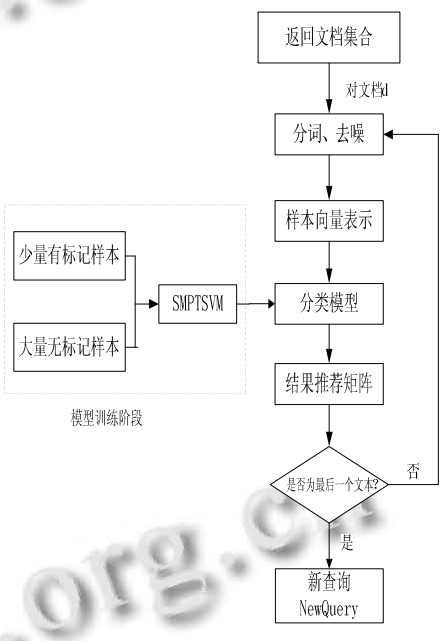

图 1 基于 PTSVM 的查询扩展流程图

为进一步降低算法的复杂度，在训练中可根据样本向量中$c(q,t_i)$特征值对样本进行初次过滤。样本经分类后词项分为两类，将标签值为+1 的词项存储为推荐矩阵的形式，如图 2 所示：

文档 / 词项	d_1	d_2	d_3	…	d_k	Total
C_1	+1		+1			7
C_2		+1				4
⋮						
C_n			+1		+1	2

图 2 候选词项推荐矩阵

在推荐矩阵中，$(C_i, d_j) = +1$表示在文档d_j中C_i被推荐为 q 的查询相关项。$Total(C_i)$表示在文档集中C_i被推荐为查询相关项的次数。设置一个推荐阈值σ，当$Total(C_i) \geq \sigma$时，C_i为最终的查询相关项，σ可用来调节查询相关项的个数。若$Total(C_i) = Total(C_j)$，当 $\sum_{i=1}^{t} Rank(d_i) < \sum_{j=1}^{t} Rank(d_j)$ 时，$weight(C_i) \geq weight(C_j)$，其中$Rank(d_i)$表示$C_i$推荐文档在初次检索后的返回顺序，$weight(C_i)$为词项的权重。

3.2.3 实验设计与结果分析

文中采用 TREC2004 Ad Hoc Genomics 检索任务的语料完成实验，数据大小为 5.16G，其中包括 3481008 个文档。实验中使用编号 1-50 的 Topics 中的 Title 域信息作为用户的查询语句，其平均查询长度为 8 个单词。为了表明该模型的有效性，与传统的查询扩展方法中常采用的 LocalFeedback 算法性能进行对比。

采用 MAP(Mean of Average Precision)作为主要评测指标，并以 Prec@5、Prec@10 作为辅助性的评测指标。Prec@n 表示针对某个查询 q，检索出 n 篇文档时的准确率。MAP 表示查询集中每个查询的平均准确率的算术平均值。

实验以初次检索中返回的前 20 篇文档作为扩展词项来源并采用人工的方法对结果进行评价。实验结果如表 2 所示，其中“No Expansion”表示不进行查询扩展。

表 2　实验结果

评价标准 / 扩展方法	MAP	Prec@5	Prec@10
No Expansion	0.201	0.450	0.438
LocalFeedback	0.257	0.498	0.480
本文方法	0.283	0.514	0.501

从上表可以看出，本文提出的查询扩展模型与传统的查询扩展方法相比搜索引擎的查全率和查准率有了较大幅度的提高，从而证明了该模型的有效性。

4　结语

本文提出了一种基于 PTSVM 的查询扩展模型，该模型利用分类的思想来解决查询扩展中相关词项的选择问题，使用该模型的优点是通过引入一些复杂的特征充分刻画出查询项与候选词项之间的相关性关系，使得相关词项的选择更加准确。通过实验对模型进行了验证，结果表明，该方法进一步提高了搜索引擎的搜索性能。下一步将重点对相关词项的权值进行优化。

参考文献

1 翟国忠.查询扩展技术研究[学位论文].武汉:华中师范大学,2007.

2 李卫疆,赵铁军,王宪刚.基于上下文的查询扩展.计算机研究与发展,2010,47(2):300–304.

3 Joachims T. Transductive inference for text classification using support vector machines. Proc. of the 16th International Conference on Machine Learning, San Francisco, CA, 1999:200–209.

4 陈毅松,汪国平,董士海,基于支持向量机的渐进直推式分类学习算法.软件学报,2003,14(3):451–460.