一种基于 HTML 位置信息的查询扩展技术

陈志玮[1]，肖诗斌[1]，施水才[1]，王昕[2]

（1. 北京信息科技大学中文信息研究中心，北京 100101；2. 中船建筑工程设计研究院，北京 100101）

摘 要： 查询扩展是指对用户提供的有关实体属性查询的描述进行语义上同义或近义方面的扩展。针对信息检索中文档与查询之间的词不匹配问题，本文提出一种基于 HTML 位置信息的查询扩展方法。由于 HTML 文件中存在位置信息（即 Tag 标签信息），所以，选择 HTML 文件进行查询扩展，相对于选择纯文本文件来说效果更好。本文中利用现有的各大搜索引擎的搜索结果组成训练语料，且利用词项与所有查询词在局部文档集合中的共现程度来评估扩展词的质量。最后，使用标准的向量空间模型(VSM)作为检索算法，将使用位置信息进行查询扩展与不加查询扩展及使用查询扩展的效果进行比对。该查询扩展技术对于查询短小、文档集内容比较分散的情况应尤为适用，可以极大地提高查询效果。同时，利用 HTML 中的位置信息，能够更好得对查询进行扩展。

关键词： 信息检索；查询扩展；共现

Query Expansion Using Tags in the HTML File

Chenzhiwei[1], Xiaoshibin[1], Shishuicai[1], Wangxin[2]

(1. Chinese Information Processing Research Center, Beijing Information Science & Technology University, Beijing 100101; 2..Zhong Chuan Architecture Design& Research Institute, Beijing 100101)

Abstract: Query Expansion is adding related words and phrases to the original query which was supplied by the user. Techniques for automatic query expansion have been extensively studied in information retrieval research as a solution to the word mismatch problem between queries and documents. Using tags which were in the HTML file, this paper proposed a expansion method. Because of tags in the HTML file, using HTML files was better than using plain text to query expansion. In this paper, the training collection was made up of the search result from several search engines. And it utilized the local co-occurrence information in top-ranked documents and global information in training collection to select most appropriate expansion terms. Then it used the vector space model(VSM) to index the documents of the testing collection, and compared the result of query expansion model using location information with the model which was not using query expansion and the query expansion model. Query expansion is applicable when the query terms are short and the content of the documents are dispersed, and it can effectively improve the query result. After that, the query expansion which uses the location information in the HTML file can more effective.

Keywords: Information Retrieval; Query Expansion; Co-Occurrence

---

基金资助：国家自然科学基金项目(60272084)；北京市教育委员会科技发展计划重点项目(KZ200310772013)；北京市教委项目(KM200510772008, KM200610772008)

作者简介：陈志玮（1982-），女，江西永丰人，硕士生，chen.zhiwei@trs.com.cn

1 引言

在信息检索中，由于用户的信息需求不可能充分表达，因此，目前基于关键字的搜索引擎都存在一些问题：一方面，对于某个查询可能返回成千上万的网页列表，其中很多网页与查询的相关性并不高；另一方面，又存在有很多与查询相关的网页中可能并不包含相应的关键字。而查询扩展是解决词不匹配问题的有效技术手段。查询扩展对用户提交的查询信息进行同义或近义的扩展，提高信息检索的智能性。因此，查询扩展能在一定程度上弥补用户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。

2 查询扩展概述

查询扩展早在20世纪70年代就被提出来了，它大致可以分成三类：基于语义知识辞典的方法、基于全局语料集分析的方法（简称全局分析方法）和基于局部文档集分析的方法（简称局部分析方法）。基于语义知识辞典的方法在进行查询扩展时，通过现有人工构筑的语义知识辞典来进行扩展词的选取；全局分析方法通过分析词与词之间在待检索的整个语料集中的关联关系来选择扩展词；局部分析方法则根据初始检索结果的前面若干篇文档中（而不是整个语料库）每个词的重要性来选取扩展词。从总体的扩展效果上来看，全局分析方法一般要好于基于语义知识辞典的方法，但这两类方法选取的扩展词语语义过于宽泛，模糊性比较大，因而难以取得较好的扩展效果。相对而言，局部分析的方法能有效地将那些与查询表示的主题相关的词加入初始查询中，因而其扩展的效果要好于前两类查询扩展方法。

3 一个查询扩展系统的设计

3.1 训练语料

自1998年到现在，出现了一个搜索引擎空前繁荣的时期。虽然现今的搜索引擎还面临着诸多的难题和问题，但不可否认，搜索引擎确实能够帮助用户准确的找到所要的相关的文档。如在搜索引擎google、baidu、yahoo中，虽不能保证搜索出来的所有网页都是相关的，但排在前列文档的相关程度是不容置疑的。而目前大多数的Web网页都是HTML（HyperText Markup Language）格式的。HTML是一种标识语言，它常常将重要的内容赋以标题、黑体、大字等形式，用以吸引网页浏览者的注意力。因此，与纯文本相比，选择HTML网页文件作为训练语料集，能够利用其中的位置信息，更准确得扩展查询词。因此，本文中使用多个搜索引擎，将从每个搜索引擎中搜出与原查询最相关的n篇Web网页，即排在最前面的n篇Web网页作为扩展的训练语料集。

3.2 扩展词选择

3.2.5 3.2.1 共现评估函数

在局部分析的查询扩展方法中，一个普通采用的策略是利用词项在局部文档集中的词频信息来度量该词项的重要性，将词频最高的若干个词（停用词除外）作为扩展词。然而，这种策略存在一个严重的问题：当S中含有很少的相关文档时，选出的大多数扩展词很有可能是那些不相关文档中的词，导致最终的检索性能反而不如未进行扩展时的检索性能。研究表明，利用词项之间的共现信息来选取扩展词，能够取得更好的扩展效果。共现即指两个词项在一定的文本窗口中同时出现。本文中文本窗口指一篇文档，而词项w和整个查询Q在整个局部文档集S中的关联程度则通过由[4]得的评估函数f(w,Q|C,S)来比较。

首先初始查询Q在待检索语料集C中执行初始检索，然后选出检索结果的前n篇文档组成局部文档集S。初始查询Q分成若干个查询词q，查询词q与词项w在文档D中的共现频度为：

$$coof(w,q \mid D) = \log(tf(w \mid D)+1.0) * \log(tf(q \mid D)+1.0) \quad (1)$$

其中tf(*|D)表示一个词项w在D中的出现次数。

定义词项w和查询词q在局部文档集S中的共现度cood(w,q|S)为w和q在S中的所有文档内的平均共现频

度。

$$cood\ (w,q\mid S)=\frac{\sum_{D\in S} coof\ (w,q\mid D)}{n} \tag{2}$$

针对每个查询词 q，我们可以简单地根据 cood(w,q|S)从 S 中选出与查询词 q 共现度最大的若干个词作为扩展词，但这种方法存在两个问题：一是单个的查询词并不能很好地表征查询所蕴涵的主题；二是每个查询词到底应分别扩展多少个词，没有一个很好的判断依据。一种更好的策略是评估词项 w 和整个查询 Q 的关联程度。令 cohd(w,Q|S)表示词项 w 与查询 Q 在局部文档集 S 中的关联度，定义为：

$$\text{cohd}\ (w,Q\mid S)=\prod_{q\in Q}(cood\ (w,q\mid S)+1.0)^{idf\ (w|C)idf\ (w|C)} \tag{3}$$

其中 idf(*|C)为倒转文档频率，表示语料集 C 中出现某个词项的文档数目。定义为：

$$idf\ (\cdot\mid C)=\frac{\log(\ N)}{\log(\ df\ (\cdot\mid C)+1.0)} \tag{4}$$

对式(3)的等号两边取对数，不影响扩展词的选取。因此，我们最终得到如下的评估函数 f(w,Q|C,S)：

$$f(w,Q\mid C,S)=\sum_{q\in Q} idf\,(q\mid C)idf\,(w\mid C)\log(cood\,(w,q\mid S)+1.0) \tag{5}$$

3.2.6　3.2.2 位置信息

本文中，将 HTML 网页文件按照重要程度分为标题类（TITLE）、粗体类（STRONG、EM、B、I、H1、H2、H3 等）以及普通文本类（BODY）三个级别，根据检索词的标记类别累计不同的权值。标题一般是网页制作者对网页内容的高度概括，应该赋以最高的权值；黑体、大字等粗体类的文字一般是网页制作者希望强调的内容，应该赋以较高的权值；而网页中普通的文字，应该赋以较低的权值。因此，可将上面的共现频度 coof(w,q|D)改为：

$$\begin{aligned}\text{coof}\ (\text{w},\text{q}\mid \text{D})=&\log(\text{tf}_1(\text{w}\mid \text{D})+\alpha\text{tf}_2(\text{w}\mid \text{D})+\beta\text{tf}_3(\text{w}\mid \text{D})+1.0)\\&*\log(\text{tf}_1(\text{q}\mid \text{D})+\alpha\text{tf}_2(\text{q}\mid \text{D})+\beta\text{tf}_3(\text{q}\mid \text{D})+1.0)\end{aligned} \tag{6}$$

其中，$tf_1(w|D)$为词项 w 在 D 中的普通文本类中出现的次数。$tf_2(w|D)$为词项 w 在 D 中的粗体类中出现的次数。$tf_3(w|D)$为词项 w 在 D 中的标题类中出现的次数。在本文中，α 的值为 0.1，β 的值为 0.2。评估函数 f(w,Q|C,S)的计算如上 3.2.1 所示。

3.2.7　3.2.3 命名实体

命名实体（Named Entity）是文本中重要的信息元素，是正确理解文本的基础。狭义地讲，命名实体是指现实世界中的具体的或抽象的实体，如人、组织、公司、地点等。广义地讲，命名实体还可以包含时间、数字表达式等。

在查询扩展时，首先需将文档进行分词处理并除去禁用词，从而得到词项 w；最后，利用公式计算词项 w 与查询 Q 的关联程度，得到扩展词。利用命名实体的识别，可以使词项 w 不再是单纯的词组，还可以是人名、组织名、地名、会议名称、药名等等。命名实体相对于一般词组而言歧义比较少，在表达用户的信息需求时，更加准确。因此，扩展命名实体，应该会有更好的查询扩展效果。

3.3　对比算法

本文中，我们使用标准的向量空间模型(VSM)作为检索算法。利用该算法，对测试语料进行排序。将使用位置信息进行查询扩展与不加查询扩展及使用查询扩展的效果进行比对。查询 Q 与文档 D 间的相似度定义为向量夹角的余弦函数值，公式如下：

$$\text{Sim}(\ \text{Q}\ ,D\ )=\frac{\sum_{k=1}^{M} W\ (t_k\mid Q\ )*W\ (t_k\mid D\ )}{\sqrt{(\sum_{k=1}^{M} W\ (t_k\mid Q\ )^2)*(\sum_{k=1}^{M} W\ (t_k\mid D\ )^2)}} \tag{7}$$

其中，$W(t_k|Q)$为词项 t_k 在查询 Q 中的权重，定义如下：

$$W(t_k \mid Q) = \log idf(t_k \mid C) \tag{8}$$

$W(t_k|D)$为词项 t_k 在文档 D 中的权重，定义如下：

$$W(t_k \mid D) = \log(1 + tf(t_k \mid D)) * \log idf(t_k \mid C) \tag{9}$$

M 为特征向量的维数，即文档中包含词项 t 的个数。

t_k 为向量的第 K 维，即文档中包含的第 k 个词项 t。

该设计首先从各个搜索引擎中采集前几页的网页，作为查询扩展的基础。然后经过分词处理，除去禁用词，识别实体后，使用上面 3.2.1 中的评估函数（即公式(1)(2)(3)(4)(5)），选择前 30 个与原查询关联度最高的词成为扩展词。对于基于 HTML 位置的扩展模型，利用上面 3.2.2 中的评估函数（即公式(6)(2)(3)(4)(5)），同样选择前 30 个与原查询关联度最高的词成为扩展词。最后，利用上面 3.3 中的对比算法（即公式(7)(8)(9)），给测试语料进行排序，将使用位置信息进行查询扩展与不加查询扩展及使用查询扩展的效果进行比对评价。

4 结论

本文提出得查询扩展方法，利用词项在局部文档集中与初始查询的共现程度，来评估扩展词的质量，并综合考虑了词项在语料集中的全局信息，并利用位置信息和命名实体，使得选取的扩展词与初始查询所表征的主题或概率具有更好的相关性。

参考文献：

[1] Van Rijsbergen. Improving the Effectiveness of Information Retrieval with Local Context Analysis[J]..ACM Transactions on Information Systems，1979：79-112

[2] Sparck Jones. Automatic Keyword Classification for Information Retrieval[J]. Butterworths London，1971

[3] 丁国栋，白硕，王斌. 一种基于局部共现的查询扩展方法[J]. 第二届全国信息检索与内容安全学术会议，2005

[4] 崔航，文继荣，李敏强.基于用户日志的查询扩展统计模型[J].软件学报，2003

[5] 孙建军，成颖等.信息检索技术[M].第一版，北京：科学出版社，2004. 286-340