硕 士 学 位 论 文

基于社会化标注的查询扩展技术研究

Query Expansion based on Social Annotation

作 者 姓 名：晋 松

学科、 专业：计算机应用技术

学　　号：20809376

指 导 教 师：林鸿飞　教授

完 成 日 期：2010 年 11 月

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

在信息检索任务中，查询扩展技术都表现出具有提高检索效果的能力。大多数基于文档集的查询扩展技术都是基于一个相关性假设，即初次检索结果中排名靠前的一部分文档是与原始查询相关的，并且可以当作是原始查询的上下文信息。因此，这些文档可以用做查询扩展的扩展词来源。但是，当初次检索结果过程中相关性文档不多时，依然利用以上提出的相关性假设，从不相关文档中提取扩展词，这些扩展词就可能与原始查询不相关，从而影响查询扩展技术的检索性能。

许多研究表明利用外部资源作为扩展词的来源，能够有效避免由于初次检索的文档不相关对查询扩展技术性能的影响。随着 Web 2.0 的发展，大量社会化标注信息出现在互联网上。在社会化标注体系中，用户根据自己的兴趣爱好，利用自由的词汇对网络资源进行标注。研究表明，这种社会化标注资源可以用来帮助提高信息检索的效果，但是，关于利用社会化标注资源作为扩展词资源，用以提高查询扩展性能的研究仍比较少见。

本文主要研究利用社会化标注信息作为扩展词资源对传统查询扩展技术的改进。首先，本文探讨并挖掘出社会化标注信息作为扩展词资源的可能性，通过对从社会化标注中挖掘出来的扩展词进行分析，发现社会化标注信息可以为原始查询提供语义相关的扩展词。在此发现基础上，本文提出了三种基于社会化标注资源的扩展词挖掘方法：（1）基于词共现统计的扩展词挖掘方法；（2）基于词依赖的扩展词挖掘方法；（3）基于排序学习的扩展词挖掘方法。在基于词共现统计的扩展词挖掘方法中，充分分析了社会化标注的产生机制，利用标签之间的语义关联性，为原始查询挖掘出语义关联的扩展词。在该方法基础上，我们着重考虑了原始查询中词项之间的依赖关系，并提出了基于词依赖的扩展词挖掘方法。对于挖掘出来的扩展词，我们利用基于排序学习的方法，根据扩展词对检索效果的潜在影响程度，对其进行二次排序，从而挖掘出能够有效提高检索性能的扩展词。

在标准 TREC 数据集的实验表明，本文提出的三种基于社会化标注的查询扩展方法能够有效的提高检索性能，尤其在利用基于排序学习的方法对扩展词进行二次排序之后，相对于原始查询和相关性模型的检索效果，基于排序学习的方法检索性能评价提高了 34.3%和 14.35%。这表明排序学习方法能够为传统的查询扩展技术提高较大帮助。最后，本文的实验表明，社会化标注资源可以作为查询扩展技术中扩展词的来源，并且能够为原始查询提供足够相关的扩展词。

关键词：信息检索；查询扩展；社会化标注；排序学习

Query Expansion based on Social Annotation

Abstract

Automatic query expansion technologies have been proven to be effective in many information retrieval tasks. Most existing approaches are based on the assumption that the most informative terms in top-retrieved documents can be viewed as context of the query and thus can be used for query expansion. One problem with these approaches is that some of the expansion terms extracted from feedback documents are irrelevant to the query, and thus may hurt the retrieval performance.

Using a large external collection as the resource of expansion terms, it is an effective way to avoid the detrimental effect of irrelevant top-retrieved documents. With the rise of Web 2.0 technologies, social annotation has become a popular way to allow users provide different keywords describing the respective Web pages from various aspects. These features may be used to improve IR performance. However, to date, the potential of social annotation for this task has been largely unexplored.

In this paper, we explore the possibility and potential of social annotation as a new resource for extracting useful expansion terms. In particular, we propose three expansion term selection methods based on social annotation resource: (1) the term selection method based on term co-occurrence, (2) the term selection method based on term-dependency, (3) the term selection method based on learning to rank. Under the assumption of different tags describing the same Web resource are semantically related to some extent, the first method selects the relevant expansion terms based on the co-occurrence between the query and expansion terms. The second method selects the relevant expansion terms using the term sequential dependence assumption in original queries. For the third method, we develop a machine learning method for term ranking, which is learnt from the statistics of the candidate expansion terms, using ListNet.

Experimental results on three TREC test collections show that the retrieval performance can be improved when the query expansion methods based on the social annotations are used. Moreover, the learning to rank method has been proven to be effective for query expansion technologies. In addition, we also demonstrate that terms selected by the term-dependency method from social annotation resources are beneficial to improve the retrieval performance.

Key Words：Information Retrieval; Query Expansion; Social Annotation; Learning to Rank

目　　录

1 绪论

1.1 研究背景

伴随着互联网的高速发展，它已成为人们获取信息资源的重要来源。随着互联网信息量爆炸性的增长，想要从海量数据中得到想要的信息变得十分困难。搜索引擎的兴起与发展在一定程度上缓解了人们寻找信息的困难，但是由于无法充分理解用户查询意图，搜索引擎的搜索结果有时也很难令人满意。

目前，大部分搜索引擎只提供基于关键字的搜索。由于教育文化背景的差异，在检索过程中，用户提交的查询也是各不相同的，有些甚至与实际的检索意图存在一定差异。其关键原因在于信息检索领域有两个语言现象一直困扰着传统的词汇不匹配问题，即同义现象和歧义现象。据统计，不同人使用相同关键词描述同一事物的概率小于 20%，这就是当前信息检索领域所谓的“词典问题”。对于现实的搜索引擎来说，如果用户提交的查询所包含的词汇较少时，检索的效果通常会比较差。如果用户使用足够多的词描述所要查询的内容，那么词典问题会在一定程度上得到缓解，检索效果也会得到提升。但是，在实际应用当中，用户很少会向搜索引擎提交包含足够多词的查询，并且，由于语言的歧义性，传统的基于关键词匹配的检索方法，很难能够满足用户的检索需求。同时，在某些情况下，即使用户提交的查询词项在文档中出现，该词项也未必会具有足够大的权重。因此，仅依靠用户提交的短查询很难获得足够的信息检索到用户满意的文档。

为了向用户提供高质量的检索服务，信息检索系统应该能够通过用户提交的查询挖掘出用户的检索意图，根据用户的不同查询意图进行有针对性的信息检索，提高信息检索的准确率和召回率。

1.2 查询扩展技术的作用与意义

在实际的信息检索应用中，用户提交的查询请求往往不能准确全面的反应出用户查询意图，这就会引起信息迷向、信息过载和词不匹配等问题，对检索性能有比较严重的负面影响。如何使用户提交的查询能够准确的反应用户需求已成为信息检索领域中一个重要的研究课题。

Van Rijsbergen 在 1986 年指出：“仅限于原查询词来提高系统的检索性能是有限的，必须对原查询进行修改以提高检索性能”[1]。后来，普遍认为 Van Rijsbergen 提出的对原查询的修改就称为查询扩展。其主要涉及对原始查询关键词的权重分配以及向原始查询中增加相关的词，这种加入到原始查询的词称之为扩展词。近些年来，查询扩展技术

越来越受到商业应用的重视，并获得了巨大的成功。目前，查询扩展技术已成为改善信息检索中准确率和召回率的关键技术之一，倍受学术界的重视和关注。

查询扩展技术指的是利用自然语言处理、计算机信息学等多种技术，将与原始查询相关的词或者语义相关联的概念以逻辑或方式添加到原始查询中，从而得到比原始查询更长的新查询，然后检索文档，以改善信息检索的准确率和召回率，解决信息检索领域长期困扰的词不匹配问题，弥补用户查询信息不足的缺陷。广义上讲，查询扩展技术就是指实现查询扩展的方法和手段，其核心问题是如何设计和挖掘与原始查询相关的扩展词。目前，扩展词的来源有三种：一是来自初次检索中被认为相关的文档中；二是利用某种文本发掘技术（如聚类技术等）从已有文献集或是查询日志中找出与原始查询相关的词作为扩展词；三是来自某种包含词与词之间相关信息的资源，这种资源可以是人工生成的，也可以是利用大规模语料通过统计的方法自动生成。其中，两个人工生成资源的例子为：WordNet 和 HowNet。

1.3 查询扩展技术的研究现状

随着查询扩展技术越来越受到重视，很多研究从不同的领域出发提出了各种查询扩展模型，目前关键词查询扩展技术按照其扩展词的来源不同主要分为基于查询文档集的查询扩展技术和基于外部扩展资源的查询扩展技术。

基于查询文档集的查询扩展技术可以分为两大类：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）。

全局分析方法是最早被提出来的查询扩展优化方法。其基本思想是对整个查询文档集的词进行相关性分析，得到每对词之间的关联程度（如共现率），构造产生一种词表。查询扩展过程中，从词表中选取与原始查询关联程度较高的词作为扩展词进行查询扩展。这里的词表是指一种数据结构，类似于同义词词典，用来表示词与词之间的关系。

常见的全局分析方法包括 LSI（Latent semantic indexing）[2]、基于词之间相似性词典的方法[3]和 Phrasefinder 方法[4]等。全局分析的优势是可以最大限度地探求词间关系，并在词典建立之后以较高的效率进行查询扩展。但是，当文档集合非常大时，建立全局的词关系词典在时间和空间上往往是不可行的，并且在文档集合改变后的更新代价巨大。因此，近期的查询扩展研究主要集中在基于局部文档集的分析上。

局部分析方法是利用初次检索得到的与原查询最相关的 N 篇文章作为扩展用词的来源。目前，流行的局部分析方法主要是局部反馈（Local feedback），也称为伪相关反馈（Pseudo-relevance feedback）。伪相关反馈技术可以应用于多种检索模型中：向量空间模型[5]，概率模型[6]，相关性语言模型[7]，混合模型[8]等等。与此同时，很多的研究工作集中在对

传统伪相关反馈技术的改进上，例如，使用篇章信息代替文档信息[9]，基于局部上下文的分析方法[10] ，基于查询正则化的分析方法[11]，基于潜在概念的分析方法[12]，基于伪相关文档聚类分析方法[13]。局部上下文分析方法的检索效果明显优于传统的全局分析和局部分析方法。但是，当初次查询后排在前面的文档与原查询相关度不大时，局部分析会把大量无关的词加入查询，从而严重降低查询精度，甚至低于不做扩展优化的情形。

局部文档集分析的方法的前提假设是从初次检索到的前 N 篇文档中提取的扩展词是与原始查询相关的。但是，当初次检索的前 N 篇文档相关性不高时，从中提取的扩展词并不是全都对查询扩展有帮助[14]。因此，一些研究开始考虑利用外部资源对查询扩展进行改进，从而降低局部文档集分析的方法对初次检索结果的依赖性。

目前，很多研究将焦点集中在利用外部资源对查询扩展技术进行改进。所谓的外部资源包括：一些词关系词典（例如，HowNet，WordNet）[15]，搜索引擎的用户日志[16]，锚文本信息[17]，维基百科[18]等等。

本文所提到的查询扩展技术就属于基于外部扩展资源的查询扩展技术的一种，我们提出了一种基于社会化标注的查询扩展方法，该方法利用社会化标注资源作为扩展词的来源，为原始查询提供相关性高的扩展词，从而提高检索性能。

1.4 论文的组织结构

本文的主要内容包括：

第一章：绪论部分，主要介绍论文相关研究的背景及意义，以及国内外相关技术的研究现状。

第二章：主要介绍了查询扩展的相关技术及实现方法。首先，介绍信息检索系统中的几种常用模型即：向量模型、概率模型、统计语言模型。其次，介绍查询扩展技术中常用的相关技术。

第三章：介绍了社会化标注的产生机制，并且详细分析标签之间的语义相关性。最后，简单介绍一下社会化标注相关的研究现状。

第四章：系统阐述了基于社会化标注的查询扩展方法。重点介绍了三种挖掘扩展词的方法：基于词共现的挖掘方法，基于词依赖的挖掘方法和基于排序学习的标签挖掘方法。作为一种新的扩展词资源，本章还用实验证明了社会化标注集的可用性。

结论部分：总结全文并对以后的研究方向进行展望。

2 查询扩展的相关技术及实现方法

2.1 信息检索模型

在过去 40 余年的信息检索研究中，检索模型也随着研究的深入不断的进行创新。从最早的布尔模型，向量空间模型，到近年来研究的热点概率模型，统计语言模型，检索模型为信息检索领域的研究提供了坚实的应用基础。接下来，我们简单介绍一下常用的几种检索模型。

2.1.1 向量空间模型

向量空间模型[19]是 20 世纪 70~80 年代绝大多数信息检索研究的基础，由于简单、直观而很引人注目，实现的框架便于进行词项加权，排序和相关反馈等工作。但是，作为一个检索模型，它的缺点在于对加权和排序算法如何影响相关性的说明不够详细。

向量空间模型假设文档和查询都是一个 t 维向量，其中 t 代表索引词项的总数。一篇文档 D_i 表示为索引词项的一个向量：

$$D_i = (d_{i1}, d_{i2}, ..., d_{it}) \qquad (2.1)$$

其中，d_{ij} 表示第 j 个词项的权值。

向量空间模型的一个吸引人的方面是，可以采用简单的图形来对文档和查询进行可视化表示，如图 2.1 所示。虽然这种可视化的表示方式有利于理解，但是会使人产生误解，认为三维空间就能够应用到真实的高维空间中。通常情况下，文档空间的维数 t 都是百万两级的。

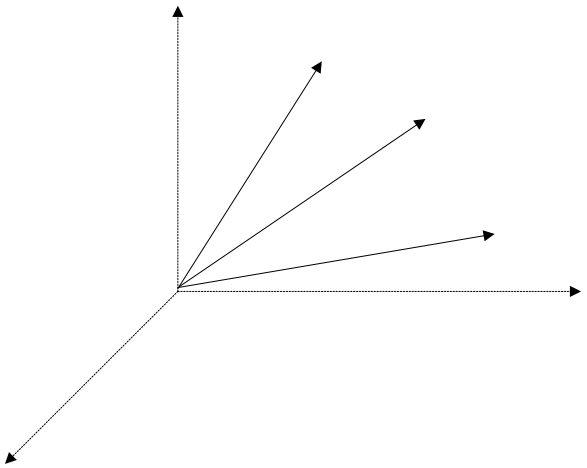

图 2.1 文档和查询的向量表示

Fig. 2.1 Vector representation of documents and queries

基于这种表示，文档可以通过计算文档和查询的相似度进行排序。在众多计算相似度函数之中，最成功的就是余弦相似度计算方法。余弦相似度是指查询向量与文档向量之间的夹角余弦值。当对向量进行归一化之后，所有文档和查询都表示成长度相同的向量，那么两个完全相同的向量夹角的余弦值为 1（夹角为 0 度），两个完全没有公共词项的向量的夹角的余弦值为 0。余弦相似度的计算公式如下：

$$\cos(D_i,Q)=\frac{\sum_{j=1}^{t} d_{ij}\cdot q_j}{\sqrt{\sum_{j=1}^{t} d_{ij}^2\cdot\sum_{j=1}^{t} q_j^2}} \qquad (2.2)$$

其中，上式的分子表示查询和文档所有匹配词项对应权重的乘积之和；分母通过除以两个向量长度的乘积来归一化分子。

现在还没有研究从理论上解释为什么余弦相似度比其他相似度度量方法要好，但是在检索性能的评价上，这种方法要表现得好一些。

2.1.2 概率模型

无论是布尔模型或是向量空间模型，对于模型的验证一般都是经验性的。而概率检索检索模型的产生，则是从理论性的角度推动了信息检索模型的发展。概率检索模型是以概率论为基础，它为表示和操纵信息检索过程中固有部分的不确定性提供了坚实的基础。目前，有多种概率检索模型，每个模型都提出了不同的方法以估计文档和查询的相关性的概率值。本节我们主要介绍一种简单的概率模型——BM25 概率模型。

在介绍 BM25 概率模型之前，首先介绍一下二元独立模型。在二元独立模型中，对每个查询都有两组文档：相关文档集合和不相关文档集合。对于一篇新的文档来说，算法的任务就是判断这个文档是否属于相关文档集。从概率的角度来看，就是看新的文档属于相关文档集的概率是否够高。如图 2.2 所示，当 $P(R|D)>P(NR|D)$ 时，判定文档 D 是相关的，其中 $P(R|D)$ 是相关性的条件概率，$P(NR|D)$ 是不相关的条件概率，这也是著名的贝叶斯决策法则。

下面将要面临的问题是如何计算这些概率。要计算 $P(R|D)$，可以先从 $P(D|R)$ 进行入手，利用贝叶斯法则：

$$P(R|D)=\frac{P(D|R)P(R)}{P(D)} \qquad (2.3)$$

其中，$P(R)$ 是相关性的先验概率，$P(D)$ 是个常数，起到归一化的作用。

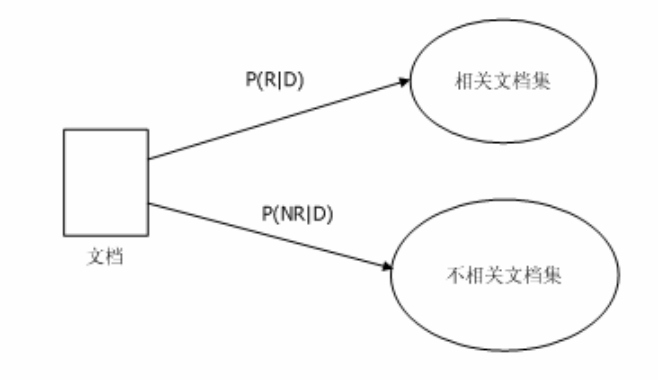

图 2.2　判断一个文档是相关的还是不相关的

Fig. 2.2　Classifying a document as relevant or non-relevant

在模型中，文档表示为词项的组合，相关集合和不相关集合表示为词项概率。利用词项独立性假设，可以通过独立词项的概率乘积 $\prod_{i=1}^{t} P(d_i \mid D)$ 来估计 $P(D \mid R)$。基于此，可以将贝叶斯决策法则改成：

$$\frac{P(D \mid R)}{P(D \mid NR)} = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i} \tag{2.4}$$

其中，$\prod_{i:d_i=1}$ 表示文档中值是 1 的词项概率的连乘，p_i 是词项 i 在相关文档集中出现的概率，s_i 是词项 i 在不相关文档集中出现的概率。接下来做一些数学推导：

$$\prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \left(\prod_{i:d_i=1} \frac{1-s_i}{1-p_i} \cdot \prod_{i:d_i=1} \frac{1-p_i}{1-s_i} \right) \cdot \prod_{i:d_i=0} \frac{1-p_i}{1-s_i} = \prod_{i:d_i=1} \frac{p_i(1-s_i)}{s_i(1-p_i)} \cdot \prod_{i} \frac{1-p_i}{1-s_i} \tag{2.5}$$

第二个连乘对于排序能够忽略掉。由于连乘了很多较小的数值，会导致结果精度问题，对乘积是使用等价的取对数操作，即：

$$\sum_{i:d_i=1} \log \frac{p_i(1-s_i)}{s_i(1-p_i)} \tag{2.6}$$

如果没有相关集合的其他信息，可以额外假设 p_i 是一个常数，s_i 可以被近似估计为整个文档集中的词项出现情况。做出这样的假设，是基于这样一个事实，即相关文档的数量远小于整体文档集合的大小。设定 p_i 的值为 0.5，则得分函数变为：

$$\sum_{i:d_i=1} \log \frac{N-n_i}{n_i} \tag{2.7}$$

其中，n_i 是保护词项 i 的文档数目，N 是整个数据集中文档的数目。

如果知道相关集合的词项出现信息，整体的得分函数为：

$$\sum_{i:d_i=q_i=1} \log \frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)} \tag{2.8}$$

其中，r_i 是包含词项 i 的相关文档数目，n_i 是包括词项 i 的文档数目，N 是整个文档集中的所有文档数目，R 是相关文档集的所有文档数目。

BM25 模型通过加入文档和查询的权重，扩展了二元独立模型的得分函数，即公式（2.8）。这种扩展是基于概率论和实验验证的，并不是一个正式的模型。现在 BM25 的得分函数有很多变形，这里只介绍一种最普遍的形式：

$$\sum_{i\in Q} \log \frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)} \cdot \frac{(k_1+1)f_i}{K+f_i} \cdot \frac{(k_2+1)qf_i}{k_2+qf_i} \tag{2.9}$$

其中，f_i 是词项 i 在文档中的频率，qf_i 是词项 i 在查询项中的频率，k_1、k_2、K 都是经验设定的参数。K 是一个较为复杂的参数，用来利用文档长度归一化 tf 因子。具体的形式如下，

$$K = k_1\left((1-b)+b\cdot\frac{dl}{avdl}\right) \tag{2.10}$$

其中，b 是一个参数，dl 是文档长度，$avdl$ 是数据集中文档的平均长度。常量 b 控制长度归一化的一个小，其中 $b=0$ 对应于没有长度归一化，$b=1$ 表示完全的归一化。在 TREC 数据集的实验中，$b=0.75$ 被证明是有效的。

2.1.3 统计语言模型

在统计语言模型检索框架下，可分为以下两种检索模型：查询似然模型（Query-likelihood model）[20]和风险最小化框架（Risk minimization framework）[21]。本文主要将查询似然模型作为对照组实验进行对比，下面主要介绍一下查询似然模型。

在查询似然模型下，将文档产生查询的概率看作是文档与查询之间的一种相似性，这种假设直观上很好理解：如果一个文档产生一个查询的概率越大，则表明文档与查询越相关。假设文档服从多项式分布且词汇间相互独立，则有：

$$sim(q,d) \propto p(q|d) = \prod_{i=1}^{n} p(q_i|d) \tag{2.11}$$

其中，q_i 表示第 i 个查询词，n 是查询串的长度，$p(\cdot|d)$ 是文档语言模型。由公式（2.11）可以看出，在查询似然模型下，核心问题为如何估计精确的文档语言模型。在估计文档的语言模型过程中，一般需要通过引入集合语言模型 $p(\cdot|C)$ 来对极大似然估计进行平滑，以避免零概率问题。实际应用中，采用检索效果较好且应用广泛的 Dirichlet 先验平滑方法如下[22]：

$$p_\mu(w|d)=\frac{c(w,d)+\mu p(w|C)}{\sum_{w'\in V}c(w',d)+\mu} \quad (2.12)$$

其中，$\sum_{w'\in V}c(w',d)$表示文档包含词项的总数，$c(w,d)$表示词项w在文档d中出现次数。$p(w|C)$表示词项w在文档集C中出现概率，这个值一般是利用极大似然估计算得，μ表示 Dirichlet 先验参数（本文中设定$\mu=1500$）。

目前，对传统的查询似然模型的改进主要集中在两个方面：（1）改进文档语言模型的估计方法；（2）削弱原始模型中词的独立性假设。语言模型框架的简洁性，结合对多种检索应用的描述能力以及相关排序算法的有效性，使得这种方法对于基于主题相关性的检索模型是一个好的选择。

2.2 查询扩展方法的相关技术

查询扩展作为解决用户查询表意不全的一种解决方案，引起了国内外很多学者的关注，提出了许多可行性的研究方法。查询扩展技术按照其扩展词的来源不同主要分为基于查询文档集的查询扩展技术和基于外部扩展资源的查询扩展技术。

2.2.1 基于查询文档集的查询扩展技术

基于查询文档集的查询扩展研究一直都是该领域的研究热点，它可以分为两大类：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）。

全局分析方法就是以全部数据集文档为研究对象，对文档中的词项进行相关分析，计算词项之间的关联程度。当用户提交一个查询时，利用预先计算好的词项的相关性关系，将与原始查询关联程度最高的词项加入到其中，生成新的查询。主要的技术有全局聚类技术[4]、相似性词典[3]、潜在语义索引[2,23]等。全局分析的优势是可以最大限度地探求词间关系，并在词典建立之后以较高的效率进行查询扩展。但是，当文档集合非常大时，建立全局的词关系词典在时间和空间上往往是不可行的，并且在文档集合改变后的更新代价巨大。因此，近期的查询扩展研究主要集中在基于局部文档集的分析上。

局部分析方法是利用两次查询的方法解决扩展问题。它利用初次检索得到的与原查询最相关的N篇文章作为扩展用词的来源，而并非利用先前计算得到的全局词关系词典。局部分析主要技术有局部聚类、相关反馈和局部反馈等，相对于全局分析，局部分析的计算量比较小。目前，在 TREC 数据集上的实验效果显著的局部分析方法是伪相关反馈技术（Pseudo-relevance feedback）。

伪相关反馈技术（Pseudo-relevance feedback）被广泛的应用在信息检索领域，并在许多不同的检索模型下有不同的实现方法。

最早应用在向量空间模型中的是著名的 Rocchio 算法，它是基于最佳查询的算法。该算法的思想是将相关文档向量的平均向量和不相关文档向量的平均向量之间的差异最大化。在获得伪相关文档信息的基础上，Rocchio 算法会对查询向量进行权重重新修改，生成一个新的查询：

$$q_j' = \alpha \cdot q_j + \beta \cdot \frac{1}{|Rel|} \sum_{D_i \in Rel} d_{ij} - \gamma \cdot \frac{1}{|Nonrel|} \sum_{D_i \in Nonrel} d_{ij} \qquad (2.13)$$

其中，q_j 是查询词项 j 的初始权值，Rel 是相关文档集合，$Nonrel$ 是不相关文档集合，$|\cdot|$ 表示一个集合的大小，d_{ij} 是文档 i 中第 j 个词项的权值，α、β 和 γ 是控制每个部分响的参数。

在统计语言模型框架下，伪相关反馈技术主要是用于改进对查询模型 $p(w|d)$ 的估计。在这方面，应用较为广泛的是，相关性模型（Relevance model）[7]和混合模型（Mixture model）[8]，本文的实验采用相关性模型作为伪相关反馈的代表，进行对比实验。下面主要介绍一下相关性模型。

相关性模型（Relevance model）假设每个查询词项由一个词的相关性模型 $p(w|\theta_R)$ 产生。这就需要相关性文档信息来计算词的相关性模型，模型利用初次检索得到的前 N 篇文档作为相关性文档，对相关性模型进行估计：

$$p(w|\theta_R) \approx \sum_{D \in F} p(w|D)p(D|\theta_R) \qquad (2.14)$$

其中，F 表示初次检索返回的文档集合。θ_R 是对原始查询的极大近似，利用贝叶斯公式，对公式（2.14）进行贝叶斯展开：

$$p(w|\theta_R) = \sum_{D \in F} \frac{p(w|D)p(Q|D)p(D)}{p(Q)} = \sum_{D \in F} p(w|D)p(Q|D) \qquad (2.15)$$

由公式（2.15）可知，在相关性模型中词项 w 的概率由该词项在伪相关文档集中的概率 $p(w|D)$ 和查询的后验概率 $p(Q|D)$ 而决定。利用上式可以对原始的查询模型进行线性插值：

$$p(w|\theta_q) = (1-\lambda)p(w|\theta_o) + \lambda p(w|\theta_R) \qquad (2.16)$$

其中，插值权重 λ 表示扩展词在整个查询模型的重要程度。参数的选定根据不同的数据集有所不同，本文的实验中，有详细的实验介绍（见 4.4.7 节）。

2.2.2 基于外部扩展资源的查询扩展技术

目前，很多研究将焦点集中在利用外部资源对查询扩展技术进行改进。所谓的外部资源包括：一些词关系词典（例如，HowNet，WordNet）[15]，搜索引擎的用户日志[16]，锚文本信息[17]，维基百科[18]等等。

从一些词关系词典（例如，HowNet，WordNet）产生之后，它就成为查询扩展的一个研究工具。查询扩展的核心是利用和设计扩展词的来源为用户的查询丰富和扩展。基于词关系词典的方法是利用现有的词关系词典，为原始查询词项选取一定数量的关系紧密的词项加入到原始查询中，形成新的查询。该方法的缺点是过于依赖词关系词典，当一个查询中包含过多“稀疏”词项时，基于词典的查询扩展方法的效果就不够理想了。

搜索引擎的用户日志是反应用户查询与结果之间关系的一种重要资源。近年来，越来越多的研究集中在对搜索引擎的用户日志发掘上，包含用户的搜索行为分析，用户意图分析，结果排序算法研究等等。基于用户日志的查询扩展方法也是研究的一个热点。相关研究主要以用户的搜索查询日志为研究对象，搜索查询日志是用户使用搜索引擎时多次“反馈”结果的积累，对它的分析相当于使用大量用户的相关反馈。基于搜索日志的查询扩展技术的基本思想是，对用户的查询记录建立用户查询空间，同时对文档集合建立文档空间，根据用户真实的点击情况，将两个空间中的词按照用户的点击行为以某种方式连接起来。当新查询到来时，系统将新查询映射到查询空间中，并为其以某种算法找出可能相关的文档集合，进行相关的查询扩展。基于用户搜索日志的查询扩展方法的优点是从查询日志中得到的大量“先验知识”要比利用个别用户的临时判断或在毫无人为参与的情况下得到的相关性结果更为准确，并且将关于用户反馈的学习放在检索之前，省去了初始检索和用户参与的代价。但是，针对一般研究工作者来说，得到充足真实的用户搜索日志是很困难的。

由于获取充足真实的用户搜索日志比较困难，Dang 等[17]提出了一种利用锚文本信息替代用户搜索日志的想法，并且将其运用到查询扩展领域，取得了较好的实验效果。Dang 的研究将一条锚文本$<text_i, url_j>$看作是一条用户点击记录，其中将$text_i$看作是用户提交的查询，url_j看作是用户提交查询后点击的结果，即与$text_i$相关的结果。经过转换之后，就可以对大量网页的锚文本信息进行转换，构造出一种“特殊的”用户搜索日志。实验结果表明，基于“特殊的”搜索日志查询扩展技术可以获得与基于真实搜索日志的查询扩展技术相当的检索性能。这为今后的查询扩展相关的研究工作开辟了一条新的思路。