博 士 学 位 论 文

基于网络挖掘与机器学习技术的相关反馈研究

The Research of machine learning techniques and external Web resources for relevance feedback

作 者 姓 名：叶正

学科、 专业：计算机软件与理论

学　　号：10709017

指 导 教 师：林鸿飞

完 成 日 期：2011 年 4 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：基于网络挖掘与机器学习技术的相关反馈研究

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

随着互联网上各种信息的爆炸式增长，人们可获取和利用的信息越来越多的同时，也往往使得人们淹没在信息的海洋中，时常很难找到所需要的信息，这就是人们常说的信息过载（Information Overload）现象。在此背景下，信息检索技术得到迅速的发展，其中互联网搜索引擎是信息检索技术最重要和常见的应用之一。大量的实验表明相关反馈技术是提高信息检索系统性能的有效手段。本文在前人的研究基础上，主要研究了如何挖掘网络资源和使用机器学习技术进一步提高基于查询扩展的相关反馈技术的性能。基于这两方面，本文所做的主要研究工作包括：

（1）针对当前大多数相关反馈算法，候选扩展词权重的计算都是使用扩展词在文档级别的统计信息。然而，一篇反馈文档，即使是人工判定为相关的，都可能包含多个主题（topic），显然并不是每个主题都有益于相关反馈算法。本文认为在较小的粒度上使用相关反馈算法更为合理，研究了如何应用主题模型（topic model）从反馈文档中推导出查询相关主题，并应用于相关反馈算法中。

（2）传统相关反馈模型中，对不同的反馈文档都是同等的对待，而实际上，不同的反馈文档的质量各不相同，对相关反馈算法的作用也不一样。针对以上问题，本文重新讨论和修改了 Rocchio 相关反馈模型，并将其应用于概率检索模型中，提出了一个新的相关反馈机制，即质量偏重反馈模型。

（3）研究了通过对高质量网络资源的挖掘来加强相关反馈算法的性能。针对伪相关反馈文档集质量难以得到保证这一问题，本文尝试了使用外部资源（相对于检索文档集）来解决该问题，并提出不同算法利用外部资源。具体，本文提出了一种生成式模型，从社会化标注标签（social annotation tags）中选取高质量的扩展词进行查询扩展，以弥补首次检索中获取的反馈文档质量较低的问题。

（4）研究了在相关反馈扩展词选择的过程中，如何考虑不同上下文信息对候选扩展词权重的影响。传统相关反馈模型中，候选扩展词的选择通常是基于其在反馈文档集中的统计信息得到，查询的上下文信息在传统相关反馈模型中通常被忽略。因此，相关反馈过程中可能选用偏离查询主题的扩展词，这就导致检索性能下降。本文中，提出了基于贝叶斯网络的相关反馈方法，该模型可以考虑多种不同的上下文信息。

关键词：文本信息检索；检索模型；相关反馈；文本挖掘；机器学习

The Research of machine learning techniques and external Web resources for relevance feedback

Abstract

With the explosive growth of information on the Internet, there is an increasing need for information systems to help users find the resource they need. Information retrieval system is to response this challenge of information overload in general. Its main application, search engine, has achieved great success during the past decade. Extensive experiments have proven that relevance feedback technique is one of the most effective techniques for ad hoc information retrieval. In this dissertation, we mainly explored utilizing machine learning and Web mining techniques to further enhance relevance feedback methods. In particular, the main work of this dissertation can be summarized as follows:

(1) For most of the current relevance feedback models, the expansion terms are selected based the document level statistics. However, for a given feedback document, even it is humanly judged to be relevant, may consist of different topics. Obviously, not all these topics are useful for relevance feedback models. We argued that it is more reasonable to conduct relevance feedback on a fine-grained level. Following this argument, a novel topic-based relevance feedback model is proposed in this dissertation, in which three different methods for approaching the query-related topic are explored.

(2) In traditional relevance feedback models, each feedback document is treated equally. In fact, the feedback documents are different in quality, therefore will influent the relevance feedback process differently. In order to address this problem, we revisit Rocchio's algorithm by proposing to integrate this classical feedback method into the divergence from randomness(DFR) probabilistic framework for pseudo relevance feedback(PRF). Such an integration is denoted by RocDFR in this paper. In addition, we further improve RocDFR's robustness by proposing two quality-biased feedback methods, called QRocDFR and ReRocDFR.

(3) Most existing relevance feedback approaches are based on the assumption that the most informative terms in top-ranked documents from the first-pass retrieval can be viewed as the context of the query, and thus can be used to specify the information need. However, there may be irrelevant documents used in PRF (especially for hard topics), which can bring noise into the feedback process. The recent development of Web 2.0 technologies on Internet has provided an opportunity to enhance PRF as more and more high-quality resources can be freely obtained.

(4) Most current PRF approaches estimate the importance of the candidate expansion terms based on their statistics on document level. However, in traditional PRF approaches, the context information is always ignored in traditional query expansion models. Therefore, off-topic terms can also be selected, which may result in a decrease of retrieval performance. In this paper, we propose a context-based feedback framework based on Bayesian network, in which multiple context information can be taken into account.

Key Words：Text Information Retrieval; Retrieval Model; Relevance Feedback; Machine Learning

目　　录

1 绪论

1.1 课题背景

随着互联网技术的迅速发展，网络已经融入越来越多人的日常生活的方方面面，而且人们对网络的依赖，特别是对网络信息的获取日益倚重。其中，随着互联网上各种信息爆炸式地增长，人们可获取和利用的信息越来越多的同时，也往往使得人们淹没在信息的海洋中，时常很难找到所需要的信息，这就是人们常说的信息过载（Information Overload）现象。在此背景下，信息检索技术得到迅速的发展，其中互联网搜索引擎是信息检索技术最重要和常见的应用之一，并在最近十几年取得了迅猛的发展，其中最具代表性的有 Google、百度等。

文本信息检索（Information Retrieval, IR）技术主要研究海量文本信息的表示、存储和检索。信息检索是一门交叉学科，研究领域涉及到计算机科学、信息科学、数学、图书馆学、认知心理学、语言学、统计学等一系列学科。现代信息检索的研究最早可以追溯到 20 世纪 50 年代，一些学者提出如何在计算机里存储和查询文本信息的基本想法。其中最著名的方法之一是由 H.P. Luhn [1]在 1957 年提出的，使用词作为索引单元来表示文档，并用查询和文档中索引单元的重叠度来衡量查询和文档的相关度。之后，随着计算机硬件性能的提高，学者们为了进一步提高检索系统性能，提出了一系列经典的检索模型，其中最具代表性的检索模型有布尔模型、向量空间模型、概率模型以及语言模型等。

然而，实际应用中，用户提交的查询请求通常只包含很少几个关键词，而且有时查询本身也会有歧义，往往不能很好地描述用户的查询需要（Information Need）。此外，由于不同用户教育和文化背景等的差异，也会导致查询与相关文档的词不匹配问题，从而导致用户不能获取想要的信息，而传统基于词匹配的检索系统对此通常无能为力。为了解决上述问题，信息检索领域的研究者们在上述基本检索模型的基础上，提出了基于查询扩展（Query Expansion，QE）的相关反馈（Relevance Feedback，RF）技术来解决此问题，在信息检索领域通常简称为相关反馈、查询扩展或者查询重构（Query Reformulation），这也是本文的主要研究课题。相关反馈技术根据用户提供的反馈信息，重构用户给出的原始查询（original query），通常包括加入新的查询词和修改查询词权重这两个紧密联系的过程。

相关反馈是解决词不匹配问题的有效技术手段，它以用户的初始查询为基础，根据返回的相关或不相关文档，通过一定的策略在查询中加入一些相关词并更改查询词的权

重，提供更多有利于判断文档相关性的信息，从而提高检索性能。近些年来，查询扩展技术越来越受到商业应用的重视，而且相关反馈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，备受学术界的重视和关注。

1.2 相关反馈研究现状

大体上，根据反馈文档不同的获取方式，基于查询扩展的相关反馈技术可以分为三大类：显式相关反馈（Explicit Relevance Feedback）、隐式相关反馈（Implicit Relevance Feedback）和伪相关反馈（Blind or Pseudo Relevance Feedback）。

显式反馈方法需要评判人员或者用户显式地参与，对给定查询返回的文档集进行相关性判断，然后反馈给检索系统，这样系统就可以根据这些反馈信息优化查询结果。对于给定查询返回文档的相关性判断，可以是二元相关性或者多等级相关性。二元相关性仅指出文档对于给定查询是相关（Relevant）还是不相关（Irrelevant），而多等级相关性则需要对相关程度进行判定，譬如不相关（not relevant），有些相关（somewhat relevant），相关（relevant）和非常相关(very relevant)。显式相关反馈算法的主要优点是[2]：1）用户只需要判断文档的相关性，而不要理解查询扩展技术的细节；2）把检索任务分成一系列小的步骤，更易于理解；3）提供了一个可控的过程，即强调相关的词项（term），降低不相关词项的权重。

相对显式相关反馈，隐式相关反馈不需要用户直接参与。隐式相关反馈的主要目标是通过对用户行为的研究，譬如用户的点击数据（click-through data）[3]，浏览了哪些文档，不同文档使用的时间长度和鼠标滚动行为等[4]，根据用户的行为判断可能相关文档，从而优化最终的查询结果。但由于用户行为的监控通常涉及到隐私等敏感问题，往往很难得到用户的支持。

伪相关反馈技术，则是在完全不需要用户参与的情况下，提供了一种在假设相关的反馈信息基础上自动反馈的方法。通常，该类方法假设给定查询首次检索返回排名靠前的文档是相关的，然后在一组假设相关的文档集的基础上进行查询扩展，从而提高检索系统的性能。伪相关反馈技术由于不需要用户参与，也不涉及到用户的隐私信息，在学术界吸引众多研究者的关注和研究。显而易见，当反馈文档确定后，伪相关反馈算法可以很容易应用到相关反馈中，因此本文主要在伪相关反馈数据集上进行实验。

早期使用 Smart 系统[5]和概率检索模型[6]的实验表明，查询扩展技术在一些小的数据集上可以取得非常好的总体性能。随后，一系列研究也以进一步验证查询扩展技术的有效性。

根据扩展词的来源不同，基于查询扩展的伪相关反馈技术又可以细分为两类：基于

全局数据集的分析方法（简称全局分析方法）和基于局部文档集分析方法（简称局部分析方法）。

其中，全局分析方法是最早提出来的伪相关反馈方法。其基本思想是对整个待查询的文档集中词进行相关性分析，得到词与词之间的关联程度（如共现率），并构造产生一种词表。查询扩展过程中，从词表中选取与原始查询关联程度最高的一些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基于词之间相似性词典的方法[8]、LSI（Latent Semantic Indexing）[7]和 Phrasefinder 等方法[9]。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，其主要不足是当文档集合过大时，建立全局的词关系词典在时间和空间上往往消耗很大，并且在加入新的文档后，往往更新代价巨大。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信息检索系统首次检索得到的与原查询最相关的 N 篇文章作为扩展词的来源。目前，流行的局部分析方法主要是局部反馈（Local Feedback），其主要思想是分析首次检索中返回文档集中词的重要性，然后把最有区分力的词加入到原始查询中。其主要优点是该过程简单高效且不需要用户参与，但反馈效果受反馈文档质量影响很大。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原始查询，从而严重降低查询精度，甚至低于不做扩展优化的情形。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、概率模型、统计语言模型等等（见本文第 2.5 节）。虽然不同的检索模型理论上各不一样，但本质都是如何修改原始查询表示以及不同查询词的权重分配等问题。

目前，查询扩展相关研究工作主要集中在对传统局部反馈技术的改进和不同检索理论中查询扩展技术的应用，例如，基于模型的反馈方法[10]、相关性模型[11]、基于马尔科夫随机域的潜在概念扩展[12]、基于局部上下文的分析方法[13]、基于查询正则化的分析方法[14]、基于伪相关文档聚类分析方法[15]。实验结果表明，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。

局部文档集分析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 篇文档相关性或质量不高时，从中提取的扩展词并不是全都对伪相关反馈算法有帮助[14]。因此，一些研究人员研究了如何从反馈文档集中寻找一个质量较高的查询子集。

此外，一些研究开始考虑利用外部资源进行伪相关反馈算法，从而降低局部文档集分析的方法对初次检索结果质量的依赖。目前，一些研究将焦点集中在利用外部资源作

为新的扩展源进行伪相关反馈。这里所谓的外部资源包括：词关系词典（例如，HowNet，WordNet）[16]、搜索引擎的用户日志[17]、锚文本信息[18]、维基百科[19-20]等等。

1.3 课题动机及方法

在传统基于查询扩展的伪相关反馈技术中，主要存在以下问题：1）由于伪相关反馈文档集中的文档只是假设其与查询相关，没有经过用户的人工判断，因此可能引入噪音；2）传统反馈技术中，往往假设整个文档都有益于反馈算法，而实际上即使一篇文档被人工判断为与查询相关，可能文档中只有部分段落或主题有益于反馈算法。

本文在前人的研究基础上，针对以上问题主要研究了如何挖掘网络资源和使用机器学习技术进一步提高基于查询扩展的伪相关反馈技术的性能。具体，随着互联网的发展，特别近些年 Web2.0 迅猛发展，互联网上可以免费获得越来越多的高质量资源，譬如维基百科（Wikipedia）[1]、百度百科[2]等网络百科全书、delicious[3]等类似社会化标注系统，如何处理和利用这些网络资源，从中挖掘出一些对信息检索有价值的知识模式，是一个非常有意义且重要的研究课题。另一方面，传统信息检索模型一般只能对少数几种特征进行建模，而且这些特征大多数都是同质的，但实际上信息检索是一个比较复杂的过程，可能有很多不同质的特征都会影响信息检索的性能，这时传统信息检索模型往往无能为力。而近些年，机器学习技术的发展，为我们进一步提高信息检索的性能提供了契机。

本文在此背景下，研究了如何利用网络挖掘与机器学习技术来进一步提高文本信息检索的性能，特别是如何提高查询扩展技术在信息检索中的应用。

1.4 主要工作及组织结构

本文研究工作正是在上述背景下展开，针对传统（伪）相关技术的不足，提出一系列算法进一步提高检索系统的性能。本文主要工作及其组织如下：

1. 第二章：信息检索概述。首先介绍信息检索中主要问题以及评测方法，然后回顾信息检索中主要经典检索模型，包括向量空间模型、概率检索模型和语言检索模型，及在这三种不同理论下导出的查询扩展方法。

2. 第三章：基于主题的相关反馈模型。大多数传统相关反馈算法对候选扩展词权重的计算，都是使用扩展词文档级别上的统计信息。但是，一篇文档，即使是人工判定为相关的，都可能包含多个主题（topic），显然并不是每个主题都有益于相关反馈算法。

[1] http://www.wikipedia.org

[2] http://baike.baidu.com/

[3] http://www.delicious.com/

本文认为在更小的粒度上使用相关反馈算法更为合理，并提出了一种基于主题的相关反馈算法。具体来说，本文使用 Latent Dirichlet Allocation (LDA)模型把反馈文档集表示成不同的主题，在此基础上提出三种不同策略近似得到与查询相关的主题，并应用于相关反馈算法中。实验表明本文所提出基于主题的相关反馈算法能取得较好的检索效果，相对语言模型中的相关性模型（RM3）能在统计意义上大幅提高检索性能。

3. 第四章：基于质量偏重的相关反馈模型。传统类似 Rocchio 相关反馈算法中，对不同的反馈文档都是同等的对待。而实际上，不同的反馈文档的质量各不相同，对相关反馈算法的用处也不一样。针对以上问题，本文重新讨论和修改了 Rocchio 相关反馈模型，并将其应用于概率检索模型中，提出了一种新的相关反馈机制，称作 RocDFR。在 RocDFR 基础上，引入了文本质量评价因子，在扩展词的选取和评分过程中区别对待不同质量的查询词，提出基本文档质量偏重反馈模型（QRocDFR）和基于回归的文档质量偏重反馈模型（ReRocDFR）。实验结果表明，将 Rocchio 反馈方法应用于概率模型中，即 RocDFR，能取得良好的检索性能，且较语言模型中最有代表性的反馈方法相关模型（RM3）性能有大幅提升。而且两种质量偏重反馈模型在 RocDFR 模型的基础上，能使得检索性能进一步提高。

4. 第五章：社会化标注在相关反馈中的应用。针对伪相关反馈文档集，质量难以得到保证这一问题，我们尝试了使用外部资源（相对于检索文档集）来解决，并提出不同算法利用外部资源。具体，本文提出了一种生成式反馈模型，从社会化标注标签（social annotation tags）中选取高质量的扩展词进行相关反馈，以弥补首次检索中获取的反馈文档质量较低的问题。该模型主要优点：a) 本文所提出模型显式地解释了每个候选扩展词的生成过程；b) 本文所提出模型可以利用社会标注中人工标注的标签与标签间的语义关系，且大量实验结果表明社会化标注有着较好的标注质量，可以作为新的扩展源应用于相关反馈中。

5. 第六章：基于贝叶斯网络的相关反馈模型。通过假设基于查询扩展的伪相关反馈（PRF）在首次检索中的排名靠前的文档是相关的，于是伪相关文档中提供信息最多的关键词可用于修改原始查询以提高检索性能。目前常用估算待扩展词重要性的 PRF 方法是基于该词文档级的统计信息。然而，传统 PRF 方法中，查询的上下文信息在传统相关反馈模型中通常被忽略，因此，相关反馈过程中可能选用偏离查询主题的扩展词，这就导致检索性能的下降。本文中，我们提出了一种基于贝叶斯网络的相关反馈方法，该模型可以考虑多种不同类型的上下文信息。为了验证本文所提出相关反馈方法的效果，本文实验中考察了 2 种不同类型的上下文信息。实验结果表明本文所提算法能明显提高检索系统的性能。

6. 第七章：总结与展望。总结全文的主要研究工作，同时对后续的研究作了进一步展望。