TREC2002 介绍及清华大学实验研究[1]

张敏，马亮，马少平，陈群秀
北京清华大学计算机系 智能技术与系统国家重点实验室 100084
Email: zhangmin@s1000e.cs.tsinghua.edu.cn

摘要：文本信息检索会议(TREC)是由美国国家标准技术局和美国国防高级研究计划局组织召开的一年一度的国际标准评测会议，在文本信息检索领域规模最大最具权威性并深有影响。本文主要介绍 TREC2002 中的三个主要项目：Web 检索、新信息抽取和自适应文本过滤。介绍从项目背景、主要任务、测试数据、评价方法和评测结果几个方面展开。同时对在该评测会议中取得好成绩的清华大学在三个项目中使用的主要研究思想和方法进行介绍。

关键词：文本信息检索，Web 信息检索，新信息抽取，自适应文本过滤

TREC2002 and THU Experiments

Min Zhang, Linag Ma, Shaoping Ma, Qunxiu Chen
State Key Lab of Intelligent Technology and System,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084
Email: zhangmin@s1000e.cs.tsinghua.edu.cn

Abstract: Text Retrieval Conference (TREC) is an annual international standard evaluation conference, sponsored by the National Institute of Standards and Technology and the Defense Advanced Research Projects Agency. It is the most authoritative and influential conference with the largest scale in text retrieval field. This paper introduces the three main tracks in TREC2002, namely web, novelty and adaptive filtering. The track background, main tasks, test set, evaluation and results are described, respectively. Furthermore, the paper gives the main idea and approaches in experiments of Tsinghua University which got good results in the three tracks.

Keywords: Text Information Retrieval, Web IR, Novelty Extraction, Adaptive Text Filtering

一. TREC 简介

信息时代，网络得到广泛应用，人们可获得的信息总量以几何级数增长。如此海量的信息使得可用资源大量丰富，但与此同时，获取有价值的信息就像大海捞针一样，信息的利用难度也大大增加。在这种情况下，对高速度、高质量的信息检索的需求变得空前迫切。而文

[1] 本项目受到国家重点基础研究（973）（G1998030509），自然科学基金项目（No.60223004）以及国家 863 高科技项目（No. 2001AA114082 和 No. 2001AA114040）资助.

本检索则仍然是其中最基础和最常用的部分。

在信息检索中，人们通常用检索精度和召回率两个标准来衡量系统的性能。但是对于召回率的评价有一个前提，即对于用户查询，在整个文档集合中相对应的相关文档都已确定。而这一点经常无法满足，尤其是在 Web 环境下，相关文档的完整集合不可能找到。因此如何对结果进行合理评价是一个重要问题。另外，人们通常通过一个标准测试集，来对不同系统不同检索方法进行公平的评价，而如何构造测试集合，也是一个重要的问题。文本信息检索国际评测会议 TREC（Text Retrieval Conference）（http://trec.nist.org）则为解决这两个问题提供了有效的途径，

TREC 是由 NIST（National Institute of Standards and Technology，美国国家标准技术局）和 DARPA（The Defense Advanced Research Projects Agency，美国国防高级研究计划局）赞助并组织的文本信息检索领域一个国际性标准评测会议。TREC 的评测本着公平和公开的原则，数据集规模很大，结果评价方法可行而有效，因此在学术界有着相当的权威性，并成为目前信息检索领域规模最大的深有影响的标准评测会议。从 1992 年至今每年一次，已经举办了 11 届。评测分为一些不同的项目（"tracks"），包括跨语言检索(Cross language Track)、信息过滤(Filtering)、问答(Question and Answering)等，从 2000 年开始，增加了 Web 信息检索(Web Track)，2002 年又增加了查找新信息的 Novelty Track。

TREC 举办以来，很多著名大学和研究机构都曾参加了项目评测，大大推动了文本信息检索技术的发展。TREC2002 是最近举办的一届，分为 7 个项目，包括跨语言检索，过滤，交互检索，新信息抽取，问答，视频检索和 Web 检索。全世界共有来自 21 个不同国家的 93 个学术、商业和政府的研究机构参加了 TREC2002 的评测。例如 BBN Technologies，CMU（包括三个组），Columbia University（包括两个组），City University of London，IBM-Haifa，Johns Hopkins University，MIT，Microsoft Research Ltd.，NTT，Tokyo University of Science，UC Berkeley，UMass 等，国内有清华大学、复旦大学、中国科学院计算所，微软亚洲研究院四个研究团体，其中清华大学是第一次参加该评测会议。

TREC 提供统一的用 SGML 标注的文档集合，向参加评测者发布同样的用户查询，并在规定时间内收集参赛者的结果，将可能相关的文档（通常是提交的每组结果中的前 *n* 篇文档）组织成评判池（judging pool）进行评价[1]。这种方法假设出现在 Pool 里面的文档才有可能是相关文档，然后通过人工判断的方法，对 Pool 中的每篇文档的相关性进行评价，形成最终的相关文档集合。虽然这一假设不完全正确，但是对评测不同方法的有效性是相当公平的，而这种 Pooling 技术也使得在大规模文档集合中寻找每个查询的相关文档集成为可能。

在下面的章节中，我们将对 TREC2002 的 Web、Novelty 和 Adaptive Filtering 三个测试项目分别进行介绍，包括项目背景、任务要求、测试数据、评价方法，以及清华大学在其中的基本研究方法，最后报告 TREC2002 中上述三个项目的评测结果。

二．Web 文本信息检索任务 Web Track

2.1 项目相关背景

WWW 开始于二十世纪八十年代末，当时没有人能够想象得到它在今天所具有的影响。如今它的广泛使用以及呈指数的增长速度有目共睹。仅仅是能够访问的文本的数量就需用 T（Terabyte）为数量级计算，其中大部分是超文本。Web 上丰富的资源对信息检索提出了急切的需求，商用的搜索引擎（Web 上的信息检索系统）已经得到了广泛的使用。同时，超文本信息检索在学术界也成为研究的热点。

对用户行为的研究对当前检索系统返回结果中仅仅提供相关文档提出了置疑。在 Web 搜索环境下，用户总是键入很短的查询语句，一般为 1 到 3 个单词[2][3]， 并很少考虑如何精确地表示查询。在很多情况下，用户最初输入查询时并不清楚他们到底需要什么样的信息[4]。在这些现实的环境中，很难准确判断返回文档的相关性。

于是 Bharat 和 Henzinger 提出主题提炼(Topic Distillation)的概念，即返回与查询主题相关的高质量结果项[5]。一个查询可能是模糊的或者包含多个主题，这时主题提炼的目标是返回与主要查询主题相关的文档，不试图去解释该查询，而把选择权留给用户，让用户在浏览中找到真正需要的主题。这也是 Web 检索项目任务设计的一个主要背景。

2.2 任务描述

TREC2002 中的 Web 检索项目分为两个任务：查找网页(Named page finding)和主题提炼(topic distillation)。

主题提炼的任务就是根据用户的简短的查询要求，为相应的主题寻找一个能够提供最多相关信息的关键资源列表。这里的“关键资源”可以是：

- 关于该主题的一个站点的首页；
- 关于该主题的一个子站点的入口页；
- 在内容上与用户查询非常相关的一个 html, doc, pdf, ps 页面；
- 一个提供了对该主题非常有用的链接列表的页面；
- 一个相关的服务页面，例如，http://www.nasa.gov/search/ 就是关于主题“NASA”的一个关键资源。

例如，查询“TREC-9”，检索结果可能是如下的一个列表：

1．hq.nasa.gov/catering/houston/menu.html
2．trec.nist.gov/contact.html
3．trec.nist.gov/pubs/trec-9/
4．trec.nist.gov/pubs/trec-9/csiro.pdf
5．www.trec.nist.gov/pubs/trec-9/
6．www.whitehouse.gov/awards/best/conference/2000/trec-9/

则对上述结果是否是关键资源的评价过程：

(1) 从 url 中即可知道第一个结果不可能是和 trec-9 相关的关键资源；

(2) trec.nist.gov 上的第二个结果和 trec-9 相关，但不是关键资源；

(3) 第三个结果 trec.nist.gov/pubs/trec-9/是一个关键资源；

(4) trec.nist.gov/pubs/trec-9/下的具体文件（结果 4）在关键资源代表的子站点内，但不是入口页面；

(5) www.trec.nist.gov/pubs/trec-9/（结果 5）则应该检查一下是不是 trec.nist.gov/pubs/trec-9/（结果 3）的别名；

(6) 最后一个结果也是一个关键资源。

有时候用户已经知道（或者猜测）有某一个网页（通常是某个机构，某个门户网站等等），但是不知道其具体的 url 地址是什么，因此用户关心的是找到该网页的入口地址。例如“中国国家图书馆”这样的查询。Web Track 的另外一个任务“查找网页”就是为这样的用户需求设计的。在该任务中，认为一个用户查询相对应的网页的 url 地址应该是唯一确定的，即使有多个网页，这多个网页之间也应该是内容相同的例如重定向链接或镜像网页等。

两个任务均要求检索全过程应自动完成，无人工干预。

2.3 测试数据

在“主题提炼”和“查找网页”两个任务中均使用.GOV[6]作为测试数据集。它包含有 1,247,753 个网页，是 2002 年 6 月从 Internet 上的.gov 域中抓取的真实数据，原始文档大约占 18G 的空间。这个.GOV 数据集与前几年的 TREC Web Track 使用的 WT10g 数据集（由 1997 年从 internet 上抓取的网页构成）相比，总的数据存储空间增加了（18G v.s 10G），但是文档数却有所减少(125 万篇 v.s. 169 万篇)，也就是说文档的平均长度大大增加(15K v.s. 7K)。这也反映了最近 5 年以来 Interent 的变化。“主题提炼”任务有 50 个用户查询，结果提交后共有 56650 个网页被进行手工评价，其中 1574 篇评价为关键资源，平均每个查询有 31 个关键资源。而“查找网页”任务则有 150 个用户查询，其中大多数都只有一个结果网页，但是有 2 个查询有三个结果，16 个查询有两个结果。表 1 中列出了 TREC2002 Web Track 使用的.GOV 数据集中不同类型文档分布信息[6]。

表 1 .GOV 数据集合中文档类型分布统计

文档类型	文档数	在数据集中所占比例
Application/pdf	131,333	10.53%
Text/plain	43,753	3.51%
Application/Msword	13,842	1.11%
Application/postscript	5,673	0.45%
Other (containing text)	42	0.003%
Non-html total	194,643	15.60%
Text/html	1,053,110	84.40%
Total	1,247,753	100.0%

2.4 评价方法

对于主题提炼任务，一共有 50 个用户查询。由于 TREC11(2002) Web Track 的主题提炼任务与前两届的查找相关信息的任务不同，是查找与用户查询相关的关键资源，因此评价标

准也有所不同，使用了前 10 篇文档的精确度作为评价指标。

对于查找网页任务中，一共有 150 个用户查询，检索返回排序最高的 100 篇文档，使用正确答案在结果中的平均返回位置倒数(*MRR*, mean reciprocal rank)作为评价指标。其中 MRR 的定义如下所以，其中 q_i 表示第 i 个查询，n 为查询的总个数：

$$MRR=\frac{\sum_{q_i}\frac{1}{\text{正确文档在结果中排名}}}{n}$$

2.5 THU 实验研究中的基本思想[7]

清华大学的实验研究中，主要思路是将基于文档内容的检索与基于链接的检索方法相结合，从两个不同的角度提高检索的性能。具体包括：

1．根据 HTML 文档结构特征，考察不同域(field)标记(tag)的文本内容在检索中的不同重要性。提出了主特征空间、主特征域和主特征词的概念。在 Internet 上，虽然网页作者不同，但是有一些词经常被大多数作者用来突出表示他们的网页，这些词通常属于某个特定的标记。它们可以分为两种：一种是功能性的或者问候式的语言，例如“版权所有”等；另一种是用来强调网页内容的，例如正文中被加粗的文字。第一种文本在信息检索中的很难带来更多的信息，而第二种突出表示的词，则反应了 HTML 文档中的更具有内容表现力的词，也是本文所关心的部分。

定义 1 在整个 Web 中，有一些词在被使用时，被不同的作者认为具有更丰富的主题或内容表现力，因而经常在一些特定的域中出现。这样的词被称作**主特征词**，相应的域被称作**主特征域**，整个网络中，所有网页的主特征域构成了**主特征空间**。

主特征域中的每个词都是主特征空间中的一维。可以看到这个空间的密度是不均匀的，也就是说空间中的每一维都带有一定的密度权值，可以用该词在整个特征空间中被不同作者使用的次数来决定。推荐该词的作者越多，则词的密度权值就越大。这个推荐的作者数，就等于该词在主特征域出现的文档频度 DF 值。这和传统模型中的 IDF 思想有着本质的区别。

考虑到网页文本主特征域所携带的信息，一个查询项与一篇文档的相似度的计算就改进为下面的公式 1，其中λ 是网页正文内容的影响因子。当λ =0 时，查询项的权重因子完全由该项的主特征信息权重因子决定，λ=0 时该公式退化为传统的概率模型相似度计算方法[8]。

$$\sum_{T\in Q}[\lambda w^{(1)}+(1-\lambda)w^{(2)}]\frac{(k_1+1)tf}{(K+tf)}\frac{(k_3+1)qtf}{(k_3+qtf)}$$

$$w^{(1)}=\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)},\quad w^{(2)}=\log(n_p+1) \tag{1}$$

其中 Q 是用户查询，T 为查询中的一项，tf 和 qtf 分别为该项在观察文档和查询中出现的次数，$K=k_1((1-b)+b\times dl/avdl)$，$dl$ 和 $avdl$ 分别为观察文档的长度和平均文档长度。$w^{(1)}$ 是查询项的 Robertson/Sparck Jones 权重因子[9]，N 是集合中的文档总数，n 是出现该项的文档数，R 是与该查询主题相关的文档数，r 是相关文档中含有该检索项的文档数。通常在第一

次检索时，R 和 r 取值为 0。n_p 是主特征域中含有该查询项的文档的个数，也是该查询项在主特征空间的密度。$w^{(1)}$ 反映了一个查询项作为正文信息的一部分所具有的权重。$w^{(2)}$则是该查询项作为主特征词所携带的权重。

研究表明粗体字(<bold>)、题目(<title>)以及链接文字(anchor text)对检索都有很大帮助。

2．对检索结果进行站点合并。这一方法的基本假设是：来自一个站点的网页中只有一个或者很少的几个网页能够成为关键资源，其他网页则因为站点内的链接关系而可以直接被关键资源网页访问到。因此，我们对一次检索的结果列表根据网页之间链接的关系进行了站点合并和结果重排序。研究表明，网页的出度比入度在站点合并和关键资源选择上更可信。

3．提出一种使用遗传算法进行检索系统参数无指导学习的方法，从而解决了传统上根据经验值设定的检索系统参数泛化能力弱的问题。

三．新信息抽取任务 Novelty Track

3.1 新信息抽取的背景和要求

由于 Internet 的飞速发展，现在互联网上可供检索的文档是海量的，而用户的需求往往只描述为两三个关键词，因而能够反应出来的用户需求非常模糊。在这两方面因素的作用下，往往对用户的一个查询所得到的结果个数上千。这些结果中大量的信息是重复的。这种重复有两种层次：1. 文档本身是多个内容相同的副本；2. 文档的内容不同，但是提供的信息相似。其中后者占了主要影响，也是相对前者来说更难解决的问题。同时已有的检索系统一般返回的是一个文档列表，这并不是理想的形式，因为一篇文档和查询相关的常常只是其中的几段或者几句话，大部分内容提供的是查询以外的信息。

从这个角度来讲，当前的信息检索还没有达到用户友好的要求。在用户查询检索结果的时候，首先要花大量的时间用于文档相关性的判断；其次在找到相关文档之后，还要自己阅读文档的全部内容，以从中找出直接满足自己需要的信息；最后，在找到相关信息之后，用户还有可能发现，这个信息是自己在前面的文档中已经找到了的信息。这个过程是非常繁杂而消耗时间的。

信息的重复性大大影响了检索结果的实际有效性，从而影响了用户对检索结果的满意程度。这就要求我们能够根据用户的需求，对检索出的不同来源不同描述特征的文档进行相关信息的抽取，并在整个文档集合范围内对信息的重复性进行判断及过滤。

3.2 任务描述

TREC2002 的 novelty track，分为两个任务：查找相关信息句(relevant)，查找新信息句(new)。给定用户查询，同时给出每个查询相关的文档集合。Novelty 项目的第一步是查找相关信息任务，找出文档中相关的句子是哪些；第二步是查找新信息任务，在第一个任务的结果基础上信息重复的句子(重复性按照句子在有序的文档中出现的先后顺序判断)。也就是说，第二个任务查找新信息句的结果集，一定是第一个任务抽取相关信息句的结果集的子集。

其中相关的句子应该是：

1．与给出的用户查询描述中内容相关的句子。

2．该相关信息句的相关性是与其周围的所有句子独立的。

而对于新信息句，则除了上面的 1 和 2 两个要求之外，必须满足：

3．该句子提供的信息，是在该句之前挑出来的所有句子中都找不到的。

4．不是一般性的介绍或者说明的句子。

5．如果有一组句子都表达了同样的一条简单信息，那么选择出的新信息句是其中描述得最详细的。

6．如果由于某种特殊的句法结构，或者是文档分句错误的原因，而造成两个相邻的句子合起来表示只一个简单信息，那么把这两句都选上。

检索中所有过程应自动完成，无人工干预。

3.3 测试数据

Novelty 项目的测试查询是从 TREC6，TREC7 和 TREC8 三年的 150 个用户查询(Topic300~Topic450)中挑选出来的 50 个查询，而提供的数据集合即每个查询对应的文档集，也是从前几届 TREC 中一些做的比较好的结果中选择出来的排在前面的一部分，一共有 1258 篇文档，平均每个查询有 25 篇相应的文档。所有文档认为已经是相关的了，并且已经有序。文档已经自动分成句子。

但是在 TREC2002 Novelty Track 的主席 Harman 的综述报告中指出该数据集合并不具有明显的重复性特征。在被标记为相关的信息句的重复度在 0%~50%之间，平均有 93%的相关信息句都是不重复的。事实上有 23 个用户查询的所有相关信息句都是新信息句（即不存在重复）[10]。这使得各种基于新信息查找的方法都无法表现出更明显的检索性能区别。

3.4 评价方法

对于每个查询，采用精度和召回率来衡量：

精度(P)＝结果中实际相关句（新信息句）个数 / 结果返回的句子总数；

召回率(R)＝结果中实际相关句（新信息句）个数 / 实际相关句（新信息句）总数。

然后再将两个标准结合起来评价系统性能。在结果提交后，最初大会曾经使用 P×R 作为统一衡量标准，后来在正式的官方公布的论文集中，改为由 F-measure 作为最终的系统性能评价方法。

3.5 THU 实验研究——基于覆盖和扩展的新信息抽取方法[11]

我们的研究主要分两个部分：信息的重复性判断方法，以及基于覆盖和扩展的新信息抽取方法，并使用了有效的扩展方法对查询和文档进行信息的补充和重构。

(1) 重复性判断

在一般的信息检索模型中，通常使用“相似性”来衡量两个文档之间的关系。最常见的

相似性评价标准有两个文档向量的内积等。这种传统的相似性是一个对于两个文档来说是对称的衡量标准。

但是，实际上在判断信息的重复性的时候，两个文档的关系并不是对称的。例如可能一个文档 P 所携带的信息被另外一个文档 Q 所完全包含。这样如果已经有了文档 Q，那么文档 P 就没有任何意义了。而在已知文档 P 的前提下，文档 Q 还有可能提供新的信息。因此，我们提出了基于“覆盖度”的思想。它表示信息 A 构成的集合 M 被信息 B 构成的集合 N 所覆盖的程度，它是非对称的，并与信息的出现顺序相关。

(2) 基于覆盖和扩展的相似性匹配

在现代信息检索中，文档中的词仍然是信息表达的载体。考虑到信息表达的多样性，在判断信息重复性的时候，我们不能仅仅使用文档本身所包含的词项来进行覆盖度计算，而应该把形式不同但意义相似的或者隐含的信息扩展进来。

如图 1 所示为基于词项扩展的覆盖度匹配示意图。其中文档 A 和文档 B 经过词项扩展后分别成为 A'和 B'。

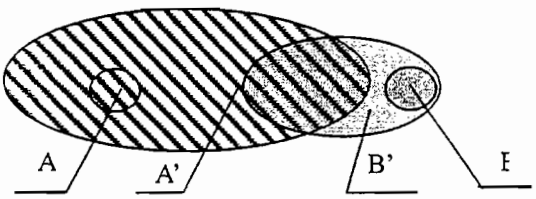

图 1 基于词项扩展的覆盖度匹配示意图

直观的基于扩展的覆盖度计算方法，可以用两个经过词项扩展之后的文档的覆盖度来表示，即如公式 2 所示：

$$Overlap'(A,B) = Overlap(A', B') \tag{2}$$

但是，注意到由于多义词等问题的存在，这种信息的扩展有可能引入较多的噪声。因此在进行相似度匹配的时候，我们提出如下公式 3 所示的覆盖度匹配方式（$\lambda_1,\lambda_2,\lambda_3$ 为覆盖系数）：

$$Overlap'(A,B) = \lambda_1 Overlap(A,B) + \lambda_2 Overlap(A',B) + \lambda_3\ Overlap(A, B') \tag{3}$$

也就是说，在两个文档的新扩展部分 A'和 B'之间是不比较覆盖度的，因为二者都可能含有大量的噪声而影响实际的覆盖度评价。

四．自适应过滤 Adaptive Filtering

4.1 项目背景

信息过滤技术主要用于文档流中主动获得与用户主题相关的文档信息。相对传统的批过滤技术，自适应过滤只需要很少量的初始相关文档，并且能够在过滤过程中不断对已过滤文档进行自主学习逐步提高过滤精度。由于其特性十分适合 Web 实际应用的特点，因而成为近年来信息过滤研究的热点。

4.2 任务描述

自适应过滤是TREC中文本过滤项目(filtering track)的一个任务。主要任务是根据给定的用户查询主题和相关训练文档，构造过滤模型，然后过滤有序的测试语料集合，输出与查询主题相关的新闻报道。每个查询主题只提供少量的初始训练文档和一个训练集。训练中不能使用任何测试集的统计信息。测试语料集为有序文档流，每个文档只被评价一次。如判为与主题相关则输出，否则被抛弃。自适应学习过程中，只能使用过滤得到的主题相关文档进行反馈，且测试与学习过程应自动完成，无人工干预。

4.3 测试数据

全部数据来自路透社新闻语料第1卷（Reuters News Corpus Volume 1），为XML格式。所有主题采用相同的训练集和测试集。训练集为该新闻语料的前83650篇文档，测试集为剩余的723141篇文档。从语料中共选择了100个测试新闻主题（R101-R200），每个主题提供3篇初始相关训练文档（来自训练集）和一个主题描述。

4.4 评价指标

100个查询主题被分为Assessor（R101-R150，前50个）和Intersection（R151-R200，后50个）两部分分别评测。对过滤结果的评价采用了多种指标，最重要的两个是T11F和T11SU。参赛机构可任选一种标准优化系统性能。对某个主题，设FD_i为过滤输出的文档集。$N=|FD_i|$，R为相关文档数，R_p和N_p分别为FD_i中实际的主题相关文档和不相关文档数。则有：

$$T11F=\begin{cases}0 & \text{if } R_p=0\\ (1.25*R_p)/(N+0.25*R) & \text{otherwise}\end{cases}$$

$$T11SU=\frac{\max(T11NU,MinNU)-MinNU}{1-MinNU}\quad MinNU=0.5\quad T11NU=\frac{2*R_p-N_p}{2*N}\quad N_p=N-R_p$$

4.5 THU 实验研究中的主要技术[12]

（1）采用了多种检索模型

检索模型分别采用向量空间模型（采用TF*IDF权重机制）和Language model [13]。后者作为新的检索模型，已在若干Web信息检索中取得良好的效果。此次我们在系统中引入该模型的相关检索机制，目的是测试其应用于自适应过滤中的实际性能。

（2）增量训练

由于初始相关训练文本非常有限，通常难以训练得到较好的初始主题过滤模型。我们使用了一种增量学习的机制，首先通过初始训练文本得到一个基本模型，然后基本模型在训练集中通过一种谨慎的选择机制，不断选择少量与主题最相关的新的伪正例文档，并通过伪正例文档的反馈提高模型的精度。同时，通过引入语义信息和改进反馈算法的权重评价方式，

也保证了引入的主题特征具有较高的主题相关性。

（3）查询扩展

对于训练得到的初始过滤模型，我们引入了查询扩展的技术，以进一步提高其定义的精度。对于过滤模型中的主题特征，部分具有较高权重信息的特征被扩展。扩展词源主要来自WordNet和一些相关词数据库。

（4）自适应学习机制

过滤过程中，过滤模型的自适应学习机制包括主题特征学习和阈值调节两方面。我们采用的方法如下：

- 主题特征学习：通过近期模型检索得到的主题相关文档进行反馈，以扩展模型中的主题相关特征。同时，考虑到多次反馈后模型将趋于完善，因此引入了衰减因子，以随着反馈学习的进行而逐步降低新特征的重要度。
- 阈值调节：采用了基于相关文档分布预期的调节方法。其基本思想是根据对测试文档集中主题相关文档分布的一个预期模型，调节阈值的升降，以使过滤结果符合预期模型。预期模型通过对训练集的统计建立（我们假设训练集和测试集具有相同的分布）。

五．TREC2002 实验测试结果

5.1 Web Track

在Web检索项目中，主题提炼任务（Topic Distillation）有23个参加机构提交了71组结果，查找网页任务（Named Page Finding）有19个参加机构提交了70组结果。其中清华大学在两个项目中各提交五组结果，经过测试得到在两个任务中均名列第一，且比第二名的结果有至少约5%的性能提高。在查找网页任务中的70组结果中包揽了前四位；主题提炼任务的71组结果中，排在前10位的有四组结果（分别排在第1、3、6、9位）。具体的官方评价结果如下：

表 2 TREC2002 Web Track 测试前三名的结果

寻找关键资源（平均精度）		
第一名	Tsinghua Uni.	0.2510
第二名	City Uni. of London	0.2408
第三名	Chinese Academy	0.2306

查找网页入口（MRR）		
第一名	Tsinghua Uni.	0.719
第二名	CMU	0.676
第三名	Yonsei Uni.	0.671

5.2 Novelty track

在新信息抽取项目的两个任务（抽取相关信息relevant，抽取新信息new）中各有42组结果提交。其中经过评测，清华大学在两个任务中也均名列第一，且包揽了两个任务提交结果中的前三位，比排在第二名的研究机构的最好结果在两个任务上分别领先11%和4%，具体的结果评价为（官方最终报告中使用的评价标准F度量值）：

表 3 TREC2002 Novelty Track 前三名评测结果（最终正式的评价标准 F-measure）

		抽取相关信息	抽取新信息
第一名	Tsinghua Uni.	0.235	0.217
第二名	UMass	0.211	0.209
第三名	Queens College	0.209	0.193

在新信息抽取项目最终的官方报告公布之前，最初曾使用 P×R 作为结果评价指标，并得到结果如下：

表 4 最初使用 P×R 作为评价标准，TREC2002 Novelty Track 的前两名评测结果

	抽取相关信息		抽取新信息	
第一名	Tsinghua Uni.	0.088	UMass	0.077
第二名	UMass	0.085	Tsinghua Uni.	0.075

5.3 adaptive filtering

按照比赛要求我们分别提交了基于向量空间模型（采用基于 Rocchio 的反馈）和 Language Model（采用 Mixture 反馈[14]）的过滤结果，均按照 T11F 标准优化。所有参赛机构共提交了近 40 组过滤结果。根据我们采用的优化标准，以 T11F（分布箱线图）评价，经过评测，清华大学位于第 4 名。表 5 和表 6 列出了前 4 名的研究机构与对应提交结果的平均评价值。

表 5 TREC2002 Adaptive Filtering 的前四名评测结果的平均值（使用 T11SU 作为评价标准）

名次	Assessor Topics（R101-R150）		Intersection Topics（R151-R200）	
	研究机构	平均值	研究机构	平均值
1	Chinese Academy of Sciences	0.475	Chinese Academy of Sciences	0.335
2	KerMIT Consortium	0.459	KerMIT Consortium	0.323
3	Carnegie Mellon University	0.433	John Hopkins University	0.322
4	Microsoft Research Cambridge	0.435	University of Iowa	0.294

表 6 TREC2002 Adaptive Filtering 的前四名评测结果的平均值（使用 T11F 作为评价标准）

名次	Assessor Topics（R101-R150）		Intersection Topics（R151-R200）	
	研究机构	平均值	研究机构	平均值
1	Chinese Academy of Sciences	0.428	Chinese Academy of Sciences	0.062
2	KerMIT Consortium	0.426	KerMIT Consortium	0.056
3	Carnegie Mellon University	0.401	Institut de Recherche en Informatique de Toulouse	0.054
4	Tsinghua University	0.422	Tsinghua University	0.052

清华大学参加 TREC2002 的成员介绍

清华大学参加 TREC 2002 的成员主要由智能技术与系统国家重点实验室的博士生与硕士生组成。主要参加人员包括张敏、马亮、宋睿华和林川。辅助开发人员有姜哲、金奕江、刘奕群和赵乐。指导教师为马少平老师和陈群秀老师。

参考文献

[1] D. Hawking. Overview of the TREC-9 Web Track. In Proc. of TREC9, pp87-102, 2000

[2] Anick, P.G. Adapting a Full-text Information Retrieval System to Computer the Troubleshooting Domain. Proc. of ACM SIGIR'94, 1994.

[3] Croft, W.B., Cook, R., and Wilder, D. Providing Government Information on the Internet: Experience with "THOMAS". U. of Mass. Technical Report 95-45, 1995

[4] Efthimiadis, E.N. A User-Centered Evaluation of Ranking Algorithms for Interactive Query Expansion. Proc. of ACM SIGIR'93, 1993.

[5] Bharat, K. and Henzinger, M.R. Improved Algorithms for Topic Distillation in a Hyperlinked Enviorenment. Proc. of ACM SIGIR'98, 1998.

[6] N. Craswell and D. Hawknig. Draft Overview of the TREC-2002 Web Track. In Proceeding of TREC-2002, 2002.

[7] Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, THU TREC2002 Web Track Experiments, the 11th Text REtrieval Conference, U.S.A, Nov., 2002.

[8] 张敏，马少平，Web 文本检索中信息的分布特性与检索策略研究，全国搜索引擎和网上信息挖掘技术研讨会，2003 年 3 月。

[9] S E Robertson and S Walker. Microsoft Cambridge at TREC-9: Filtering track. In Proceedings of TREC-9. 2000

[10] D. Harman. Overview of the TREC 2002 novelty track. in proceedings of TREC2002, 2002.

[11] Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma et al, Expansion-Based Technologies in Finding Relevant and New Information: THU TREC2002 Novelty Track Experiments, the 11th Text REtrieval Conference (TREC2002), U.S.A, Nov., 2002.

[12] Liang Ma, Qunxiu Chen, Shaoping Ma, Min Zhang, et al, Incremental Learning for Profile Training in Adaptive Document Filtering, the 11th Text REtrieval Conference (TREC2002), U.S.A, November, 2002.

[13] J. Lafferty, C. Zhai. Risk minimization and language modeling in information retrieval. In Proc. of 24th ACM SIGIR . 2001.

[14] Chengxiang Zhai, John Lafferty. Model-based Feedback in the Language Modeling Approach to Information Retrieval. The 10th International Conference on Information and Knowledge Management (CIKM), 2001