**信息过滤和信息检索：同一枚硬币的两面？**

信息过滤系统是为非结构化或半结构化数据而设计的，而数据库应用程序则使用非常结构化的数据。 这些系统还主要处理文本信息，但它们也可能包含图像、语音、视频或其他多媒体信息系统的数据类型。信息过滤系统还涉及大量数据和输入数据流，无论是从远程源进行广播还是由其他来源直接发送。过滤是基于描述个人或群体信息偏好或概况，通常代表长期利益。 过滤还意味着从传入流中删除数据，而不是在流中查找数据; 用户只能看到被提取出来的数据。信息检索和过滤的模型以及从检索研究吸取过滤的经验被提出。

信息过滤是用来描述涉及向需要它的人提供信息的各种过程的名称。 尽管这个术语在描述诸如电子邮件，多媒体分布式系统和电子办公文档等应用的流行和技术文章中经常出现，但是过滤和相关过程（如检索，路由，分类和提取）之间的区别往往不明确。 然而，只有通过这种区分，才能识别和解决与过滤相关的具体研究问题。

合理定义信息过滤的第一步是列出此过程的典型特征或功能。 以下功能是最常提到的：

* 信息过滤系统是为非结构化或半结构化数据设计的信息系统。 这与典型的涉及非常结构化数据的数据库应用程序（如员工记录）形成了鲜明对比。 这里使用的结构的概念不仅仅是数据符合格式，例如记录类型描述，而且记录的字段由具有明确定义的简单数据类型组成。 例如，可以为复杂文档定义数据库类型，例如期刊文章，但该类型的文本，图形和表格组件的含义远不如雇员记录类型的典型组件（如工资）定义得更好。电子邮件消息是半结构化数据的一个示例，因为它们具有明确定义的标题字段和非结构化文本正文。
* 信息过滤系统主要处理文本信息。 实际上，非结构化数据通常用作文本数据的同义词。 然而，它是更一般的，应该包括作为多媒体信息系统一部分的其他类型的数据，例如图像，语音和视频。 传统的数据库系统无法很好地处理这些数据类型，并且这些数据类型都具有难以表示的含义。
* 过滤系统涉及大量的数据。典型的应用程序将处理千兆字节的文本，或者更大量的其他媒体。
* 过滤应用程序通常涉及传入数据流，要么由远程源（例如新闻专线服务）广播，要么由其他来源（电子邮件）直接发送。 过滤也被用于描述从远程数据库访问和检索信息的过程，在这种情况下，传入数据是数据库搜索的结果。 生成“智能代理”的系统开发人员也使用此方案来搜索远程异构数据库。
* 过滤基于个人或群组信息偏好的描述，通常称为个人资料。 此类概况通常代表长期利益。
* 过滤通常意味着从传入流中删除数据，而不是在该流中查找数据。 在第一种情况下，系统的用户会看到删除数据后剩下的内容; 在后一种情况下，他们会看到提取的数据。 第一种方法的一个常见示例是用于删除“垃圾”邮件的电子邮件过滤器。 请注意，这意味着配置文件不仅可以表达人们想要的内容，还可以表达他们不想要的内容。

此功能列表表明，信息过滤是一个定义明确且独特的过程。然而，仔细研究一下，这些功能中的许多功能几乎与其他各种基于文本的信息系统中的功能相同。例如，文本路由（text routing）涉及将相关的传入数据发送给个人或群组。此过程与过滤基本相同。分类系统（Categorization systems）被设计为将一个或多个预定义的类别附加到传入对象（例如，通过新闻通讯服务来完成）。与profiles相比，过滤在这种情况下的主要不同是类别的静态性质。提取系统（Extraction systems）有些不同，因为它们强调从输入对象的文本中提取事实，确定哪些对象是次要问题。信息检索系统（Information retrieval systems）共享许多信息过滤功能。实际上，信息选择性传播（Selective Dissemination of Information , SDI），信息检索系统的原始功能之一，似乎对大多数信息过滤应用程序来说是相同的。

更深入地了解过滤和其他基于文本的过程之间的差异，以及所涉及的研究问题的定义，需要进行更详细的比较。本文将基于过去20年在该领域研究中开发的信息检索模型来进行这个比较。我们将开发一个类似的信息过滤模型，并通过比较这些模型来定义研究问题。通过阐明过滤和检索之间的相似点和不同点，过滤系统的开发者应该能够从相关检索实验中获得的结果中受益。

# 信息检索与过滤模型

## 信息检索与信息过滤的一般概念

信息检索（IR）可以用多种方式描述，从对其目标的描述到其相对抽象的组件和过程模型。虽然并非所有特征彼此一致，但它们都倾向于共享一些共性。通常，信息检索系统被认为具有“引导用户访问最能使他/她满足其信息需求的文档”的功能。更一般地说，“信息[检索]系统的目标是让用户从知识资源中获取信息，帮助她/他进行问题管理”。信息检索的这些功能或目标已在图1所示类型的模型中描述。该模型表示信息检索情况下的基本实体和过程。

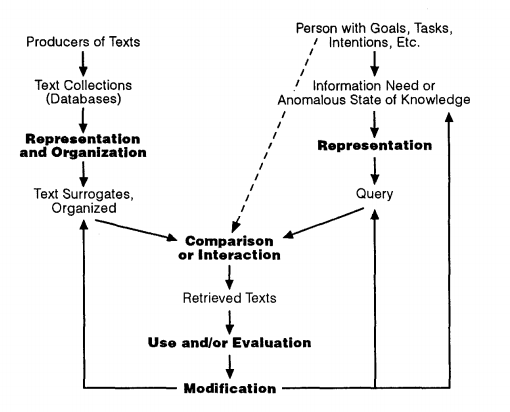


图1 信息检索的一般模型

在这个模型中，具有某些目标和意图（例如与工作任务有关）的人发现这些目标无法实现，因为他的资源或知识在一定程度上存在不足。这种“有问题的情况”的特征是异常的知识状态（ASK）或信息需求，其促使人参与主动的信息搜索行为，例如向信息检索系统提交查询。这个必须以系统理解的语言表达的查询，是信息需求的表示。这显示在图1的右侧。因为表示异常的知识状态的固有困难，信息检索系统中的查询总是被认为是近似的和不完美的。

在图1的另一侧，关注的焦点是信息检索系统的用户最终将访问的信息资源。这里，这个模型考虑文本的制作者或作者; 将文本分组为集合（例如，数据库）; 文本的代表性; 并且，将这些表示组织成文本代理的数据库。以更易于通过计算机处理的形式（有时称为索引）来表示文本含义的过程在信息检索中是至关重要的。典型的代理将由一组索引项或关键字组成。

查询和代理，或者在某些情况下，用户与文本或代理之间的直接交互（如在超文本系统中）的比较会导致可能相关的检索文本的选择。然后，将会对这些检索的文本进行评估或使用，之后用户会离开信息检索系统，或者根据评估对查询、信息需求或者更罕见的代理的进行一些修改。通过用户评估进行查询修改的过程称为信息检索中的相关性反馈。

信息检索的研究还没有考虑到图1所示的所有实体和过程。例如，几乎没有关于文本产生或者对它们的生产者的研究，并且对收集过程的研究几乎只在操作术语上进行。已有大量的实验研究集中于文本表示和组织、比较和查询修改的过程。该研究主要关注通过精度和召回来衡量系统性能的评估。另一项信息检索研究强调对信息检索系统参与者的研究，并研究了用户如何从目标或信息需求得到查询、基础查询的知识状态的表示、信息检索中的互动过程尤其是用户和人类中介之间的互动过程、 关于用户任务和目标的文本评估、互动系统的替代性能测量等问题。

基于图1中的信息检索的一般模型，以及先前对信息过滤特征的描述，图2给出了用于描述所涉及的主要实体和过程的信息过滤模型。在该模型中，信息过滤开始于具有相对稳定，长期或周期性目标或期望（例如，完成工作任务或娱乐）的人（过滤系统的用户）。团体，可以像个人一样以这样的目标为特征。这些人或团体将会导致定期的信息兴趣（例如，保持最新的主题），这些信息兴趣可能会随着时间的推移和条件、目标和知识的改变而慢慢地改变。这些信息兴趣导致人们参与相对被动的信息寻求行为，例如引起他们注意的文本。这是通过将信息兴趣的表示作为可配置到过滤系统的配置文件或查询来实现的。这样的配置文件一般被解释为信息兴趣的良好规范。

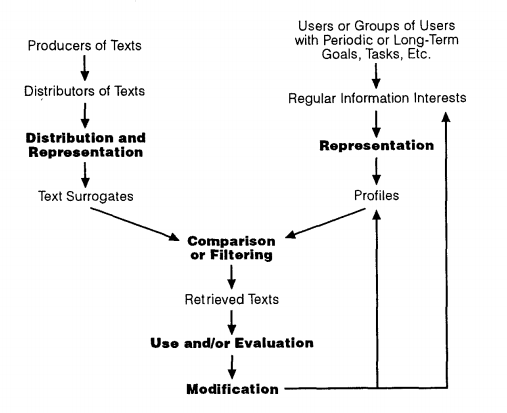


图2 信息过滤的一般模型

在图2的左侧，焦点是文本的制作者，他们通常是机构，如报纸，以及个人。这些机构或其他机构（如新闻组）承诺在文本生成时分发文本，以便引起用户的注意。为了实现这一点，文本被表示并与配置文件相比较。一些文本中的比较结果引起了用户的注意（被检索）。这些文本会根据它们对信息兴趣和激励目标的反应程度而被使用（或不使用）和评估。这个评估可能会导致配置文件和信息兴趣的修改。修改后的实体用于后续的比较过程。

在比较和讨论图1和图2的过程中，我们注意到，在这个抽象层次上，与信息过滤相关的实体和过程，和与信息检索相关的实体和过程几乎是相同的。主要区别似乎是：

* 信息检索通常涉及由具有一次性目标和一次性查询的人对系统的单一使用，信息过滤则涉及由具有长期目标或兴趣的人员对系统的重复使用。
* 信息检索认为查询的充分性是信息需求表示的固有问题，过滤则假定配置文件可以是信息兴趣的正确规范。
* 信息检索涉及文本的收集和组织，过滤涉及文本在群体或个人中的分配（推送）。
* 信息检索通常涉及从相对静态的数据库中选择文本，过滤主要涉及从动态数据流中选择或消除文本。
* 信息检索关注于用户与文本在单个信息寻求事件中的交互，过滤涉及一系列信息寻求事件的长期变化。

除了基于信息检索和过滤模型的这些区别之外，似乎还有一些其他的上下文差异也可能与研究兴趣相关。这是由于信息检索和过滤所涉及的在社会和/或实际情况中的差异。这些差异可以根据与文本、用户和每个人关注的一般环境的差异进行分类。

* 与文本有关的问题。对于信息过滤而言，文本的及时性往往具有重要的意义。对于信息检索，通常情况并非如此。
* 与用户相关的问题。总体而言，信息检索在明确定义的特定领域（主要是科学和技术领域）研究定义明确的用户群。这些用户几乎总是在积极主动地参与寻求信息的行为。然而，过滤通常涉及非常不确定的用户群体（例如在家中寻求娱乐的人）以及具有高度变化的领域。此外，过滤环境中的动机并不总能被假定。
* 环境问题。在这里，最显着的差异似乎是过滤在很多情况下都非常关注隐私问题; 信息检索由于各种原因，几乎没有关注这类问题。

## 信息检索的具体模型

讨论了信息检索和信息过滤在表示、比较和修改等过程中的强相似性，我们将通过简要介绍在信息检索中已经开发的更具体的模型来结束这一部分。这些模型主要关注比较过程。三个主要的选择是布尔，向量空间和概率检索模型。第一个基于所谓的“完全匹配”原则; 另外两个则基于“最佳匹配”的概念。详细综述见[2,22]。

布尔检索基于查询规范与一个或多个文本代理的精确匹配的概念。使用术语“布尔”是因为查询规范表示为单词或短语，使用布尔逻辑的标准运算符对其进行组合。在该检索模型中，包含查询中指定的单词或短语的组合的所有代理或更一般的文本会被检索出来，并且在任何检索的文档之间没有区别。因此，布尔检索中的比较操作的结果是将数据库划分为一组检索到的文档以及一组未检索到的文档。

布尔精确匹配检索模型是当前大规模操作信息检索系统的标准模型。该模型的主要问题是它不允许对检索到的文档集进行任何形式的相关性排序。也就是说，某些文本比其他文本更可能与信息需求相关（或更相关）是很明显的。以假定的相关顺序向用户呈现文档导致更有效和可用的系统。同样，排除与查询规范不完全匹配的文档会导致效率降低。

针对精确匹配检索的问题，最佳匹配检索模型被提出了。其中最广为人知的是向量空间模型。该模型将文本和查询视为多维空间中的向量，其维度是用于表示文本的单词。通过比较矢量（例如使用余弦相关相似性度量），来比较查询和文本。假设为表示文本的向量与表示查询的向量越相似，该文本与该查询相关的可能性越大。在此模型中，一个重要的改进是查询或文本表示的术语（或维度）可以加权，以考虑它们的重要性。这些权重是根据数据库中和术语中术语的统计分布计算的。

概率信息检索模型基于概率排序原则。这表明信息检索系统的功能是在给定所有可用证据的情况下，按照与查询相关的概率的顺序对文本中的文本进行排序。这一原则考虑到信息需求和文本的表达是不确定的，它们之间的关联关系也是不确定的。概率检索模型表明，可用于估计文本与查询的相关性的概率的证据的来源有很多种。这种证据最典型的来源是数据库中术语的统计分布，以及相关的和不相关的文本。下一节将详细讨论概率检索模型以及如何将其应用于过滤。

应该注意，这里提到的两个最佳匹配模型都可以使用布尔查询对文档进行排名。查询形式与基本检索模型之间的区别是一个重要的区别。

# 检索和过滤的概率模型

在本节中将讨论特定概率检索模型上下文中的过滤和该模型的实现。用于此目的的推理网络模型已被证明是通用的，因为它可以用来描述其他已知的检索方法，并且它有效，同时也因为在该模型中实现了相对于其他系统更高级别的查全率和查准率。推理网络模型还允许在构造查询和将查询概念与用于描述对象的概念联系起来的过程中具有大的灵活性。

## 检索模型

概率检索模型计算P（I！Object），P是给定特定对象满足用户信息需求的概率。对象一般被认为包含文本，尽管在复杂对象检索的情况下通常不是这样的。在本文中，我们关注的主要是文本，尽管我们将保留“对象”一词来表示模型更一般。我们将信息需求视为关于对象内容的复杂命题，其中可能的值为真或假。查询被视为信息需求的表示。推理网模型与其他概率模型的主要区别在于推理网络模型强调使用多个证据源来计算P（I！Object）。

推理网络模型基于贝叶斯推理网络。其是有向的非循环依赖图，其中节点表示命题变量或常量，边表示命题之间的依赖关系。如果由节点p“表示”的命题“引起”或暗示由节点q表示的命题，则我们绘制从p到q的有向边。节点q包含一个指定两个变量的所有可能值的P（q！p）的矩阵（链接矩阵）。当一个节点具有多个父节点时，该矩阵详细描述了该节点在父集合上的依赖性，并表现出该节点与代表其潜在原因的所有节点之间的依赖关系。给定网络根的一组先验概率，这些网络可用于计算与所有剩余节点相关联的概率或信任程度。

图3显示了本文中使用的基本推理网络。该网络由对象网络和查询网络组成。对象网络为集合构建一次，其结构在查询处理期间不会更改。查询网络由表示用户信息需求的单个节点和表示信息需求的一个或多个查询表达组成。针对每个信息需求建立查询网络，并根据交互查询公式或相关反馈来对查询网络进行修改。

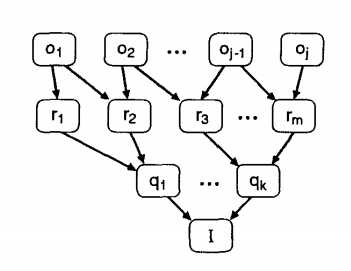


图3

对象网络由对象节点（[O.S.j]）和概念表示节点（[R.S.M]）组成。我们通过从每个表示已分配概念的对象的节点到表示节点的有向弧将一个特定的表示概念分配给一个对象。在给定其父对象节点集的情况下，表示节点包含与节点相关联的条件概率的规范。表示节点通过自动或手动索引生成。在典型的信息检索系统中，它们将与从文本中提取的单词相对应，尽管基于更复杂的语言分析的表示也是可能的。概率P([r.sub.m]/[o.sub.j])的估计是基于单个对象和大型对象集合中概念出现的频率。

查询网络包含对应于满足信息需求的事件的单个节点（I）和对应于表达信息需要的概念的多个根节点（[qk]）。可以使用一组中间查询节点来描述复杂的查询网络（例如用布尔表达式形成的那些网络）。

对于检索来说，通过与用户的交互来构建查询网络，并将其附加到对象网络。这允许我们计算满足任何特定对象的信息需求的概率，并由此产生对象的排序列表。有关此过程的更多详细信息，请参见[30]。

## 过滤模型

前一小节中对检索模型进行了描述，我们现在可以描述用于信息过滤的类似模型，该模型试图结合本文前面提到的特征。图4显示了该模型的结构。该模型与图3中的检索模型之间的差异反映了这样的事实：在过滤中，是将输入的对象流同时与多个配置文件进行比较，而不是将单个查询与大型相对静态的数据库进行比较。从概念上讲，这意味着，对于每个传入的对象[o.sub.j]，我们计算与所有配置文件节点[p.sub.1]到[p.sub.n]相关的概率。基于该计算，我们“过滤”对象，这可能意味着从给定配置文件的流中移除对象或为配置文件选择对象，具体取决于应用程序。然而，这种过滤模型提出了许多更详细的问题，必须解决这些问题才能构建过滤系统。

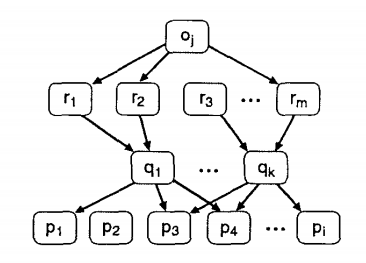


图4

通过考虑概率模型环境下的过滤定义，可以说明这些问题。在给定来自对象和一组配置文件传入流的特定对象的情况下，“过滤”该对象究竟意味着什么？从直观的角度来看，为对象选择最匹配的配置文件似乎是合理的。然而，这太简单了，无法作为一般模型。推理网络模型描述了如何计算传入对象中给定的配置文件（表示信息需要）为真的概率。在检索的情况下，该概率用于对对象进行排序以呈现给用户。然而，这种情况只会发生在过滤中，如果我们做出简化的假设，即传入的对象被批处理并相对于每个配置文件进行排序。在这种情况下，过滤成为检索的一个微小变化，它导致所有传入的对象(以不同的级别顺序)呈现给与每个配置文件关联的用户。尽管这对于某些应用程序可能是可行的，但是有许多应用程序无法对这些传入对象进行批处理。

如果我们不对批量传入的对象进行排名，而是必须确定每个对象的相关性，那么就有很多可能性。例如，我们可以将对象指向与排名靠前的配置文件集相关联的用户。这种方法的问题是，我们必须从排名的顶部选择一些固定的配置文件，而不考虑配置文件与对象匹配的程度。或者，我们可以尝试设置对象与配置文件的相似程度的阈值。这个阈值的更正式的定义来自于将推理网络模型解释为贝叶斯决策模型。这意味着，在假设决策错误的代价相等的情况下，如果P（[P.S.I]是真的[/O.S.j]）＞P（[P.S.I]是假的/ [O.S.j]），则假定一个对象[O.S.j]与一个配置文件[P.S.I]有关。然后设置阈值的问题成为获得准确概率估计的更一般的问题。

一般来说，过滤可以被定义为确定哪些配置文件有较高概率可以被来自传入流的特定对象所满足的过程。在特定配置文件中具有低概率的对象从指向与该配置文件相关联的用户的对象流中移除。如果适合应用程序，则可以使用概率对该流中的对象进行批处理并按排名顺序显示。

该模型可以直接处理“负面”配置文件。（如停用词表）这些配置文件描述了对象中不需要的功能，而不是所需的功能。不包含这些功能的对象具有满足配置文件的高概率，不会被删除。

基于该模型的过滤系统的实现涉及两个主要的概念问题和许多效率问题。第一个问题与索引或表示对象的内容有关。基于文本的过滤系统中的索引过程基本上与文本检索系统，尤其是处理异构数据库的系统中的索引过程相同。为了处理对象的许多不同格式以及这些对象中语言的动态特性，有必要使用相当简单的基于单词和短语的索引技术。然而，重要的是要意识到信息需求的表示不限于这些简单的特征。可以使用例如布尔运算符，短语识别技术和规则从这些简单的特征中构造出更复杂的特征。这些复杂的功能可以直接在推理网框架中建模。还可以使用更复杂的索引过程来识别这些特征。然而，在过滤的上下文中，当大量传入的文档流可能需要很快地被索引时，从改进索引获得的检索有效性的益处必须与索引效率的损失相平衡。

概率估计问题是任何检索系统中的主要问题（在某些系统中，概率是“权重”）。在过滤系统中，这个问题在某些方面更糟，在其他方面更好。问题更严重，因为对象是从流而不是静态数据库中获取。使用单个对象文本和对象数据库中的单词和短语频率来完成对索引概率（图3中的[R.S.M]，[O.S.j]）的估计。为了基于对象的“普遍”而获得对概率的准确估计，有必要将这些估计基于先前看到的大对象样本。甚至可能有必要为过滤系统的每个对象源维护这些样本概率。

由于相关信息需求的长期性，估计过滤系统中查询（或配置文件）网络中的概率比检索系统更容易。在这种情况下，可能会有更多满足配置文件的对象示例，也因此有更多机会了解正确的概率。在检索系统中使用的关联反馈技术通常能显著提高检索效率，在过滤系统中更是如此。

在效率方面，主要问题是检索系统通常使用文档代表的倒排文档来实现。在推理网络模型中，对象网络中的概率P([r.sub.m]/[o.sub.j])被预先计算并存储在倒排列表中，每个概念对应一个。当有许多对象要与单个查询进行比较时，这是一种非常有效的方法。但是，对于过滤系统，我们通常会将单个对象与大量（可能是数千个）配置文件进行比较，因此相同的实现不太可能。相反，每个传入的对象都可以被索引并具有在过滤时计算的相关概率。然后，可以使用这些概率来评估包含对象中存在的特征的配置文件网络。假设有大量的配置文件，为了确定哪些配置文件满足该约束，可以构造查询概念的倒排列表。

本节介绍的模型提出的过滤过程总结在图5中。我们相信，这个过程可以用来描述大多数已经提出的过滤应用。此外，这个过滤模型阐明了这些应用程序背后的假设和问题。

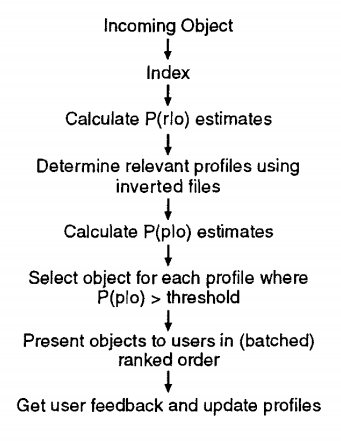


图5 基于推理网络的过滤过程

最后需要注意的是，与矢量空间模型等简单模型不同，对象和配置文件在推理网络模型中是不对称的。因此，我们不能简单地将推理网络“颠倒”，以使图4中的模型看起来更像图3所示的模型。我们不能这样做，因为我们并不真正理解概率P([o.sub.j]/[P .sub.i])的含义，也不知道如何计算它。信息需求从未被“观察到”，因为它在人们的脑海里。尽管这使得我们的过滤模型有些复杂，但我们相信概率方法可以使我们更好地理解关键问题和解决这些问题的新方法。

# 从检索研究中得到的过滤经验

鉴于基于文本的过滤系统的许多组件将与文本检索系统中的那些组件几乎相同，因此思考从文本检索系统的实验中学到了什么，以及这些结果如何应用于过滤系统是合理的。信息检索中的研究可以分为本文前面提到的三个主要类别，我们将基于它们对这一研究进行讨论。这些研究类别是文本表示，检索（比较）技术和信息需求的获取。

## 文本表示

文本表示，或索引，一直是信息检索研究的主要焦点之一。对过滤最重要的结果是，当与适当的检索模型结合时，简单的基于单词的表示会非常有效，而且实现起来效率高且直接。使用包括词汇识别、词法分析的方法对对象进行索引，以将不同的单词形式减少到常见的“词干”，并计算这些词干的出现次数。这个过程的简单性意味着即使有非常大量的传入对象，也可以使用概率方法进行过滤。此索引过程一个对某些应用程序非常有用的扩展是在扫描程序中包含专用识别器。可以通过这种方式识别的一些重要特征类型是公司名称、人们的姓名、日期和地点。

基于自然语言处理技术的更复杂的表示尚未显示出成本效益。这甚至包括如使用句法或随机解析来识别名词短语这样的简单的技术。尽管有一些证据表明在查询中使用这些技术来识别短语是有效的，但是通常可以使用简单的单词邻近度量来识别基于短语的概念在对象中的重要性。尽管在该领域难以取得进展，但最近对自然语言处理的大规模应用的兴趣的增加有望最终提高过滤系统的有效性。特别是在DARPA赞助的信息理解和评估会议下进行的文本提取研究表明，可以使用先进技术从文本中提取特定信息，并为文本对象的相关性提供更准确的证据。DARPA TIPSTER计划正在继续这项研究，并且还在进行首次大规模的过滤技术评估。

另一种在信息检索中被广泛研究的表示技术是聚类。文档聚类用于对具有相关表示的文档进行分组，术语聚类用于对相关的单词和短语进行分组。在文档集群的情况下，通常将集群的代表而不是原始文本表示用于与查询进行比较。因此，该技术可以被视为原始表示的一种转换。另一方面，术语聚类通常用于扩展（或转换）原始查询表示。虽然最近应用因子分析有一定的前景，但使用这些技术进行的实验尚未确定其有效性。

## 检索技术

本文前面已经讨论了使用检索模型作为检索技术的基础。该领域的信息检索文献中最重要的结果与不同检索技术和概率估计函数的相对有效性有关。

鉴于应该使用排名技术来实现良好的效果，一个基本问题是如何计算对象的“得分”。在概率检索模型中，这涉及估计概率。在向量空间模型中，术语权重可以被解释为概率估计，并且已经进行了大量的实验工作来评估替代形式。通常，这些被称为TF.IDF权重，因为它们包括基于对象文本中的词（或特征）的频率（术语频率分量或TF）的分量，和基于对象的“普遍”中的字的频率的分量（逆文档频率或IDF）。随着单词的频率减小，IDF权重增加（名称的由来）。基于推理网络模型的检索系统还使用TF.IDF权重的形式来估计P([r.sub.m]/[o.sub.j])的值。为了使过滤系统有效，使用类似的估计函数就很重要。

## 获取信息需求

在检索系统中获取准确的信息需求描述至关重要，在过滤系统中也同样重要。如前所述，过滤系统中的配置文件通常代表长期兴趣，并且可能有更多机会来提高配置文件的质量。与过滤这一方面相关的信息检索研究一直是查询公式和相关反馈。

查询公式的研究主要集中在查询语言和交互式辅助工具上。例如，已经证明布尔查询是非常难以生成的。还证明了布尔或结构化查询在与适当的检索模型一起使用时是非常有效的。布尔查询中的附加结构（与表示为术语集的查询相比）可以描述重要的语言特征，例如短语。这表明过滤模型应该能够处理结构化查询，并且应该设计一个可以支持结构化查询公式的接口。

用户输入的与初始查询中提到的概念相关的概念及其相对重要性可以显著提高检索有效性已经被证明了。与此相反，其他实验表明，通过让用户从系统建议的列表中选择其他概念来扩展查询通常是无效的。这些差异的原因尚不清楚，虽然似乎只使用系统建议过于严格，并没有充分利用用户的领域知识。因此，为过滤系统设计接口并不简单，主要组件应该鼓励用户尽可能地具体，但不限制用户从主题列表中进行选择。一种可能的方法是要求用户用自然语言描述兴趣，使用简单的自然语言处理技术分析这些描述以隔离概念，提示用户提供与初始陈述中的概念相关的概念并指出哪些概念是相关的。期望用户提供更复杂的信息需求描述的系统仅限于少量的应用，在这些应用中，这种期望是合理的。

相关性反馈的研究表明，通过使用简单的反馈技术可以获得显著的有效性改进效果。还有一些结果表明，在全文集和有大量相关文档可用于训练系统的应用程序中，从相关文档中选择新术语添加到查询的问题变得更加严重。在具有少量抽象长度文档的情况下有效地用于特征选择的技术在用于从数以千计可能的特征中选择时似乎没有足够的辨别力。这意味着虽然反馈是过滤系统的必要组成部分，但还需要更多的研究来确定最适合此任务的反馈技术。如果用户从相关文档的文本中而不是从相关文档自动选择的术语列表中选择特征，那么相关性反馈可以得到改进。

相关性反馈侧重于训练系统以响应特定的配置文件。也可以从许多profile-object比较的结果中学习概率估计函数(特别是用来估计P([r.sub.m]/[o.sub.j])的函数)。考虑到典型可用的大量训练数据（相关性判断），这对于过滤特别有意义。

## 评估

信息检索领域非常重视评估问题。检索系统的效率和有效性的区别是在早期就得到的，重点是测量有效性。已经发展了许多测量方法，其中最着名的是召回和精确度。精确度是实际相关的检索文档集的比例。召回是实际检索的所有相关文件的比例。这些数字通常表示为查询集的平均值。

在许多过滤应用中，召回和精确度足以评估有效性。然而，已经指出，评估过滤系统在响应输入文档中选择正确的配置文件时的性能可能需要标准测量的变化。其中一个不同之处在于，在过滤系统中，可能必须将每个传入文档分配给当前配置文件的子集，而在检索环境中，不必进行分配，因为所有文档都针对每个查询进行排序。对建立排名阈值以确定对配置文件的分配的关注至少导致了在评估中使用不同的平均技术。

信息检索社区也对在交互环境中标准召回和精确度量的价值和有效性表示关注。进行信息过滤实验的研究人员将能够从长期的信息检索评估经验中受益，但针对过滤系统评估的标准，测量和方法的发展是一个将对信息检索研究产生影响的重要问题。

# 结论

我们通过思考信息过滤和信息检索之间的关系开始本文。在研究了这些企业的基础之后，可以公平地说，两者之间在抽象层面上的差异相对较小。首先，他们的基本目标基本相同。也就是说，两者都关注向需要它的人提供信息，并且两者都关注或多或少的相同类型的上下文。此外，最初出现在信息过滤中的大多数问题都是信息检索问题的特殊化。对信息检索的概率推理网络方法及其在信息过滤中的应用的扩展讨论似乎相当具体地证明了这种关系。我们从中得出的结论是，许多信息检索的研究经验与过滤直接相关。

然而，很显然，信息检索研究忽略了信息检索和信息过滤本身所涉及的一般问题的一些方面，而这些正是与过滤的特定上下文环境特别相关的方面。下面是本文前面部分讨论过的具体问题的总结。

学习和适应是信息检索研究所关注的问题，主要是通过相关反馈的概念。然而，这种研究基于相对较少的训练集，并应用于相当小的数据库。信息过滤涉及更大的数据集，并且通常涉及在相对长的时间段内相对稳定的信息需求。

在信息检索中索引非文本数据的经验相对较少。在许多情况下，信息过滤与多媒体文本密切相关。虽然这个问题的兴趣正在两个领域趋同，但相对于信息检索，这在短期内似乎对信息过滤来说是一个更重要的研究问题。

数据的及时性是过滤特别关注的另一个领域。需要研究如何表示时间约束，如何理解文本何时可能对特定用户是及时的，以及在特定情况下的时效性意味着什么。

研究过滤的研究人员还需要对用户信息兴趣的维度进行大量研究：它们可能是什么，如何识别它们，如何表示它们以及如何修改它们。特别是因为过滤考虑了信息检索一般不会考虑的用户、用途和数据的新类别的情况而具有的相关结果。人们对文本的使用以及对这些用途显着的文本特征的研究将成为信息过滤背景下的主要关注点。特别是，诸如电视节目的娱乐使用之类的应用为过滤研究带来了特殊的问题和机会。

最后，信息过滤显然涉及许多与文本的生产和分发有关的经济和社会问题，而信息检索对这些问题相对不感兴趣。该领域的研究可能侧重于与隐私，版权和访问相关的问题。

因此，过滤确实存在着一个超出信息检索研究范围的研究事项。虽然这个事项与过滤可能发生的环境有关，但它的应用也基于它想要做的基础模型。这个模型虽然在许多方面等同于信息检索模型，但在一些有趣而重要的方面对它进行了扩展。这一扩展以及随之而来的研究议程似乎对信息检索和过滤具有重要意义，因为它解决了对信息检索来说非常重要但尚未解决的问题，主要是因为针对特定环境和用户的专门化。

我们得出信息检索和信息过滤确实是同一硬币的两个方面的结论。他们一起工作以帮助人们获得执行任务所需的信息。

我们在这里使用文本作为一般术语，也可以包括多媒体对象。