信息过滤技术研究

毛颖　周源远 王继成

(南京大学计算机科学与技术系，南京 210093)

E-mail: maoy@graphics.nju.edu.cn

摘要： 随着 web 的普及，信息过滤技术得到越来越广泛的应用。本文介绍了基于内容和基于协作的两种不同的过滤方法；描述了信息过滤中的关键技术，包括用户兴趣文件的表示、建立、维护和相似度比较。

关键词： 信息过滤，基于内容的过滤，协作式过滤，用户兴趣文件，反馈

引言

我们生活在所谓的信息时代。尤其是随着互联网的出现，计算机用户越来越被庞大芜杂的信息淹没，变的无所适从。Nature 上的统计数据表明，Web 上约有超过 800,000,000 个页面，数据量达到 15TB，散布于大约 2,800,000 台 Web 服务器上。到 2002 年，仅 Google(www.google.com)就索引了 2,073,418,204 个页面，而 一般搜索引擎的覆盖率不会超过 34%。这样的现象即是所谓的信息过载。在互联网这一不断增长的数据流中，不是所有的信息都是用户感兴趣的，但似乎只有通览所有的信息用户才能找到真正感兴趣的东西。为了减轻用户的检索负担，信息过滤系统应运而生。

1、信息过滤简介

1．1 什么是信息过滤?

简言之，信息过滤系统监控信息源以找到满足用户需求的信息。一般来说，这样的系统都有一个用户的模型（称为用户兴趣文件 profile），这一模型被用来自动地过滤掉那些用户可能并不想看到的信息。

1．2 相关术语

(1) **信息过滤与信息检索**　信息过滤问题可以看作是信息检索的对偶问题。在信息检索中，信息用户有信息需求，这一需求被尽可能的表达为某种查询语言，再与信息对象的描述文件进行相关度的计算得到查询结果。在信息过滤中，信息对象有被分发的需求。故对信息对象的描述可以看作是对它的分配需求的描述。这一描述再与用户的兴趣文件（也即信息用户的描述文件）相比较。从某种意义上看，这两个问题也可看作是一个区间的两端。在信息检索中，用户的查询在某一时期内可能有很大的变化，而所要检索的信息集合相对而言是静止不变的；在信息过滤中，用户的兴趣是相对静止不变的，而信息源是动态的，如：可以是新出版的期刊杂志或是新一轮会议的学报，会刊等等。两者的相同之处在于都涉及到选取相关信息的技术。

从图 1 中可以清楚地看到两者的区别。

(2) **信息过滤与推荐系统**　推荐系统将知识发现技术运用于为个人推荐相关信息中去。和信息过滤

系统一样，也涉及到用户兴趣的学习，选取相关信息的技术。推荐和过滤看似两个相反的过程，一个是择取所需，一个是摒弃无关。但两者的目标和手段是相同的。在接下来的叙述中将不再区分这两个术语。

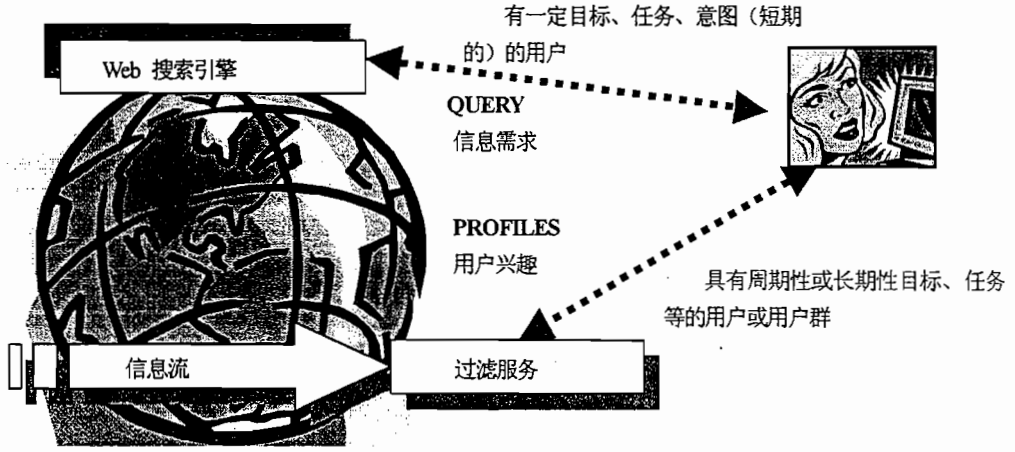

图1　信息过滤和信息检索的区别

(3) 信息过滤与分类、聚类　聚类和分类是信息过滤中的核心技术。一般而言，聚类用在用户兴趣文件的创建和修改中，而分类用在用户兴趣和信息对象的相关度比较中。值得一提的是，聚类和分类是通用技术，不仅用于信息过滤中，还广泛应用于其它信息处理中，如信息检索(retrieval)，信息摘要（summarization）等等。

2、相关工作

2．1 分类

总的来说，有两种信息过滤方式，分别为基于内容的信息过滤和协作过滤。

2．1．1 基于内容的信息过滤

基于内容的过滤源于信息检索，采用了与信息检索相似的技术。信息对象（如文本文档）的过滤是建立在其内容与用户兴趣文件（profile）相比较的基础上的。已有大量算法用来分析文本文档的内容，以作为过滤的基础。其中许多算法可以看作是分类学习器的具体化应用，其目的是要找到一种函数用以预测文档的类别（即是用户喜欢的还是不喜欢的），另一些算法的目标在于找到一种函数用以预测一个数值（即文档的评估值）。一般来说，会用到一些加权算法以给那些有辨别能力的单词很高的权重。一个纯粹的基于内容的过滤系统是仅仅基于用户的兴趣文件的，而这一兴趣的建立是通过分析用户以前所评估过的文档的内容。这种系统有：InfoFinder、NewsWeeder 和一些在 TREC 会议上为完成“文档路由（routing）”任务而开发的系统。

一个纯粹的基于内容的系统存在一些缺陷。首先，基于内容的技术在碰到相同主题的文档时，很难区分质量的高下。第二个问题是过分专业化。鉴于系统只能将同用户兴趣文件相比较得分高的文档推荐给用户，用户将局限于看到那些同已评估过的文档相似的文档。基因算法中的交叉和变异算法可以解决这一缺陷。

2．1．2 协作式过滤

与基于内容的过滤相比，基于协作的过滤有很大差别：不是推荐与用户以前喜欢文档相似的文档给用户，而是推荐相似用户喜欢的文档给用户。不是计算文档的相似度，而计算用户的相似度。典型的做法是每个用户都有一组近邻用户，近邻与该用户的评估纪录有很大的关联性。一个纯粹的协作式的过

滤系统并不分析文档——事实上用户所看到的文档只是一个唯一性的标记符而已。对用户来说过滤仅建立在同其它用户的相似度比较上。这种系统有：GroupLens，Bellcore 视频推荐系统，Tapestry， Jester 和 Ringo 。

纯粹的基于协作的过滤弥补了基于内容的过滤的所有不足。然而，该方法自身同样存在着一些缺陷。

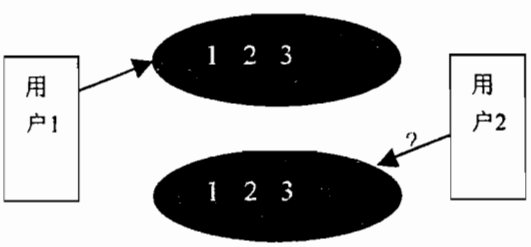

图2 Early Rater 问题

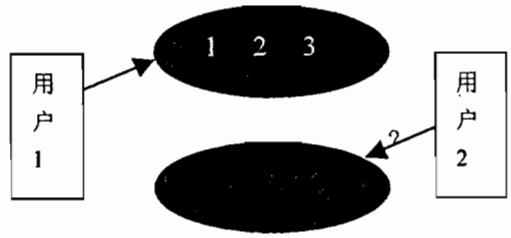

图3 稀疏问题

(1) Early-rater 的问题（如图2） 对于一篇新来的文档由于没有任何用户曾对其做过评估，故用纯粹的协作式过滤方法不能做出预测。类似的，即使对已经建立起来的过滤系统而言，一旦有新的用户加入，则对新用户做的预测也会不理想。

(2) 稀疏问题（如图3） 在许多信息领域，文档数量远远超过个人所能够浏览的数目，故而，若分别以用户和评估过的文档为纵横，则组成的矩阵非常稀疏。这样，用户的相似度计算非常困难，进而影响以此为基础的协作式过滤。

(3) “灰羊”问题（如图4） 在较小，甚至是中等大小的用户群中，存在着这样一种个体，找不到与其相似的用户，故而很难获得正确的预测。

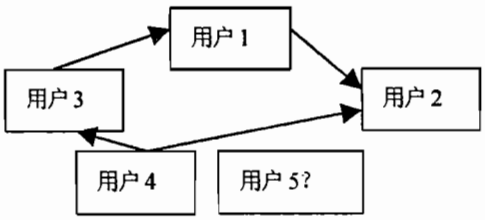

图4 “灰羊”问题

2．1．3 基于内容与基于协作相结合的过滤

实验证明，如果在协作式过滤方法中融入基于内容的技术，系统性能将有所提高。通过两者的结合，我们可以获得基于内容过滤的优点，包括能进行覆盖所有文档和用户的早期预测，同时也能获得协作式过滤的优点，即使用户评估过的文章数不断增加，系统仍能给出精确的预测。已有很多系统采取了两者相结合的方法。如 Fab 系统，ProfBuilder 系统等。

2．2 关键技术

2．2．1 用户兴趣（profile）的表示

(1) 基于内容的信息过滤中的文档及 profile 表示 根据“贝叶斯假设”，假定组成文档的字或词在确定文档类别的作用上相互独立，这样，可以就使用文档中出现的字或词的集合来代替文档，不言而喻，这将丢失大量关于文章内容的信息，但是这种假设可以使文档的表示和处理形式化，并且可以在信息过滤中取得较好的效果。目前，在信息处理方向上，文档的表示主要采用向量空间模型（Vector Space Model）。向量空间模型的基本思想是以向量来表示文档：（W1, W2, W3……Wn），其中 Wi 为第 i 个特征项的权重，根据实验结果，普遍认为选取词作为特征项要优于字和词组。最初的向量表示（即词条权重）完全是 0、1 形式，即，如果文档中出现了该词，那么文档向量的该维为 1，否则为 0。这种方法无法体现这个词

在文档中的作用程度，所以逐渐 0、1 被更精确的词频代替。词频分为绝对词频和相对词频，绝对词频，即使用词在文本中出现的频率表示文档，相对词频为归一化的词频，其计算方法主要运用 TF-IDF 公式，目前存在多种 TF-IDF 公式，如：

$$W(t,\vec{d})=\frac{(1+\log_2 tf(t,\vec{d}))\times\log_2(N/n_t)}{\sqrt{\sum_{t\in\vec{d}}\left[(1+\log_2 tf(t,\vec{d}))\times\log_2(N/n_t)\right]^2}}$$

其中，$W(t,\vec{d})$ 为词 t 在文档 $\vec{d}$ 中的权重，而 $tf(t,\vec{d})$ 为词 t 在文档 $\vec{d}$ 中的词频，N 为训练文本的总数，n_t 为训练文本集中出现 t 的文本数，分母为归一化因子。

用户兴趣文件（profile）同样用向量表示，任意 m 个词的兴趣文件用 $P(t_1,t_2...t_m)$ 表示。

(2) 协作过滤中 profile 的表示　基于协作的信息过滤是建立在用户的相似度比较的基础上的。故不存在文档内容的表示，用户关心的是其它用户对于该文档的评估值。相应的，用户兴趣文件（profile）即是由该用户对已知文档的评估值组成的。即 $P(r_1,r_2...r_m)$，其中，m 为用户所评估过的文档总数，r_i 为用户对文档 i 的评估值。

2．2．2 用户兴趣（profile）的建立

用户兴趣的建立可以看作是一个由文档矢量生成 profile 矢量的聚类的过程。因而很多聚类算法适用于建立用户的兴趣文档。聚类依赖于对观测间的接近程度（距离）或相似程度的理解，定义不同的距离量度和相似性量度就可以产生不同的聚类结果。有层次聚类、快速聚类、变量聚类等聚类过程。下面主要介绍层次聚类法。层次聚类是一种逐次合并类的方法，最后得到一个聚类的二叉树聚类图。其想法是，对于 n 个观测，先计算其两两的距离得到一个距离矩阵，然后把离得最近的两个观测合并为一类，于是我们现在只剩了 n-1 个类(每个单独的未合并的观测作为一个类)。计算这 n-1 个类两两之间的距离，找到离得最近的两个类将其合并，就只剩下了 n-2 个类……直到剩下两个类，把它们合并为一个类为止。聚类过程应该在某个类水平数（即未合并的类数）停下来，最终的类就取这些未合并的类。

对于协作式过滤而言，从协作式过滤的用户兴趣的表示中可以看到，用户兴趣的建立依赖于用户对文档的评估值。

2．2．3 相似度比较

判断信息对象是否是用户感兴趣的，可以看作是一个分类的过程，只不过这里的类别总体来讲只有两种，即用户感兴趣的，和用户不感兴趣的。而不论是基于内容的方法还是协作式的过滤方法，分类的过程中都需要进行相似度比较，即将信息对象与用户的兴趣文件相比较。一般而言，用户兴趣文件的表示方式决定了相应的相似度比较方法。下面分别介绍两种过滤方式下的相似度比较方法。

(1) **基于内容方式中的相似度比较**　对于采用向量空间模型来表示文档和用户兴趣文件的过滤系统，可以采取多种相似度比较方法。最简单的方法是仅考虑两个特征向量中所包含的词条重叠程度，即 $sim(d_k,c_i)=\frac{n_s(d_k,c_i)}{n_t(d_k,c_i)}$，其中，$c_i$ 为第 i 个类别，d_k 为第 k 篇文档，$sim(d_k,c_i)$ 为第 k 篇文档与第 i 个类别之间的相似度，$n_s(d_k,c_i)$ 是 d_k 和 c_i 具有的相同词条数目，$n_t(d_k,c_i)$ 是 d_k 和 c_i 具有的所有词条数目；最常用的方法是考虑两个特征向量之间的夹角余弦，即 $sim(d_k,c_i)=\frac{\overrightarrow{d_k}\bullet\overrightarrow{c_i}}{\left|\overrightarrow{d_k}\right|\left|\overrightarrow{c_i}\right|}$，其中，$\bullet$ 是点积，$\left|\overrightarrow{d_k}\right|=\sqrt{\overrightarrow{d_k}\bullet\overrightarrow{d_k}}$。

(2) **协作式过滤方式中的相似度比较**　通常，我们使用 Pearson 相关系数来计算用户间 profile 的相似

度。令，Ri,j 表示用户 i 对文档 j 的评估值。则用户 x 和用户 y 之间的相关度计算公式如下：

$$r(x,y)=\frac{\sum_{d=documents}(R_{x,d}-\overline{R}_x)(R_{y,d}-\overline{R}_y)}{\sqrt{\sum_{d=documents}(R_{x,d}-\overline{R}_x)^2\sum_{d=documents}(R_{y,d}-\overline{R}_y)^2}}$$

其中，$\overline{R}_x$ 是用户 x 总的评估值的平均值

用户 u 对文档的预测值计算通用公式如下：

$$prediction = \overline{u} + \frac{\sum_{i=1}^{n}(corr_i)*(rating_i-\overline{i})}{\sum_{i=1}^{n}(corr_i)}$$

其中$\overline{u}$ 是当前用户的平均评估值，$corr_i$ 是用户 i 同当前用户的 Pearson 相关系数，$rating_i$ 表示用户 i 对当前文档的评估反馈值，$\overline{i}$ 是用户 i 对所有已判断过的文档的平均反馈值，n 表示系统中与当前用户存在关联的且评估过当前文档的所有用户的总数。

2．2．4 用户兴趣（profile）的维护

在建立用户的兴趣文件时，用户有时难以准确表述自己的信息需求，但可以准确地判断返回的信息是否符合要求而采取针对性的处理，如：阅读、浏览、下载等，而通过用户的日常行为和对于内容的处理方式可以更好的学习用户的兴趣和偏好。另一方面，用户的信息需求是一个长期的过程，其兴趣和前后行为通常难以保持一致；而且随着时间的推移，Web 信息的内容也在不断地发生变化，引发新的需求，必须不断学习用户变化的兴趣。换言之，需要不断维护用户的兴趣文件，而维护的方式就是通过用户给出的反馈。比如，当一个页面被选中时，则该页面会展现给用户，同时由用户给出反馈。如果用户喜欢这个页面，则从这个页面抽取出的词的权重将会加到用户兴趣文件中相应词的权重上，这一过程被称为相关度反馈。当然这只是反馈的一种方式，有的过滤系统，采取成倍增减权重的方式，即如果用户喜欢这个页面，则用户兴趣文件中相应词的权重会成倍增加，反之，成倍减少。一种更常用的反馈方式，是采用 Rocchio 反馈模型。更有效的用户兴趣文件可由以下公式迭代产生：

$$P_{k+1}=P_k+\beta\sum_{k=1}^{n_1}\frac{R_k}{n_1}-\gamma\sum_{k=1}^{n_2}\frac{S_k}{n_2}$$

其中，P_{k+1} 是新的用户兴趣文件，P_k 是旧的用户兴趣文件，R_k 是用户反馈中认为感兴趣的（相关的）文档 k 的内容表示，S_k 是用户认为不感兴趣的（不相关）文档 k 的内容表示，n_1 是相关文档数，n_2 是不相关文档数，β,γ 值决定了正负反馈的相对作用。

3、结束语

随着 web 的普及，过滤技术正得到越来越广泛的应用。各式各样的过滤系统，推荐系统层出不穷，内容涉及 web 页面，Usenet 新闻，音乐，电影甚至笑话。

本文简要介绍了基于内容和基于协作的两种过滤方式，并着重描述了信息过滤中的关键技术，包括 profile 的表示、profile 的建立，相似度的比较和 profile 的维护。

参考文献

[1] Steve Lawrence, and C. Lee Giles. Searching the World Wide Web. Science, April., V.280: pp98-100, 1998

[2] Marko Balabanovic and Yoav Shoham. Content-Based, Collaborative Recommendation. Communications of the ACM,

40(3), March 1997.
[3] Michael J. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering. Department of Information and Computer Science University of California, Irvine Irvine, CA 92697 pazzani@ics.uci.edu
[4] James Allan. Incremental Relevance Feedback for Information Filtering. In Proceedings of SIGIR' 96, August 1996, Zurich Switzerland.
[5] 王继成、潘金贵、张福炎. Web文本挖掘技术研究
[6] 王继成、萧嵘、孙正兴、张福炎. Web信息检索研究进展. 计算机研究与发展 第38卷第2期 2001年2月
[7] 刘绍翰、武港山、张福炎. 相关反馈技术在互联网文本信息检索中的应用.

作者简介：毛颖（1980—），女，南京大学2002届本科毕业，将在南大继续攻读硕士学位，研究方向为信息系统，多媒体技术。

The Research of Information Filtering

Mao Ying, Zhou Yuan-Yuan and Wang Ji-Cheng

(Nanjing University, Nanjing 210093, China)

E-mail: maoy@graphics.nju.edu.cn

Abstract: As the popularity of web, information filtering techniques have been more and more widely used. This paper introduces two main filtering methods, named content-based and collaborative filtering; describes the key techniques in information filtering, including the representation, creation and rebuilding of user profile, and similarity computing.

Key words: information filtering, content-based filtering, collaborative filtering, profile, feedback