ISSN1000-9825 Journal of Software 软件学报 2000,11(4):545~550

利用 Bookmark 服务进行网络信息过滤*

卢增祥 关宏超 李衍达

(清华大学自动化系 北京 100084)

E-mail: luzxiang@public.bta.net.cn

摘要 信息过滤(information filtering)是解决网络信息查询困难的重要方法,其核心技术是用户信息需求的获取和信息匹配算法.该文从这两方面讨论了在 Internet 上进行信息过滤的问题.在用户信息获取方面,提出通过扩展浏览器上的 Bookmark 功能,跟踪用户信息需求,并直接利用用户评价文章来表达用户需求.在信息匹配方面,提出最大间距进行 ranking 的算法.实验表明,它比传统算法精度更高.作为对上述方法的实现,设计了 Bookmark 服务系统,给出其系统框图,并介绍了它的功能.

关键词 信息过滤,最大间距法,Bookmark 服务,Internet,WWW.

中图法分类号 TP393

进入 90 年代以来,由于互联网的飞速发展,Information Overload 的问题日趋严重.传统的信息查询技术所处理的信息格式与质量都有较好的保证.但是,在网络环境下的信息查询却不同:网上信息良莠不齐,形式各异,用户知识层次差别很大,信息需求也千奇百怪.目前人们利用的主要网络信息查询工具是搜索引擎(Internet search engine,利用关键词查询网络信息,如 Altavista,Infoseek,Excite 等)和网络目录(将网络信息利用人工分类方法,组织成一个树状的目录结构,如 Yahoo!,Looksmart 等).在使用搜索引擎时,用户需要将信息需求表达成由关键词组成的 Query.只要 Query 相同,搜索引擎给出的查询结果就相同,并不考虑用户的信息偏好.用户往往要尝试几次才能找到合适的 Query.另外,网络信息是动态的,用户时常关心这种变化,为了获得变化的信息,用户只有不断地在网络上查询同样的内容.网络目录,能让用户根据分类信息缩小搜索范围,但是,由于目录结构是固定的,而用户在理解分类时没有统一的标准,所以,利用时也会出现问题.另外,由于信息分类是由人工完成的,所以,信息的索引量受到很大的限制,运行成本也很高.在现有情况下,人们要想在网上找到自己所需要的信息,需耗费大量时间.传统的信息查询技术难以满足用户的信息需求,因此,对网络信息查询技术的研究日益受到重视.针对这个问题,1997 年 EI 上收录的文章比 1994 年增加了一倍.本文将从信息过滤和机器学习的角度,就此问题提出一些解决方法.

近年来,在信息查询领域中兴起了信息过滤(information filtering,简称 IF)技术,Croft 在文献[1]中指出了 IF 与信息查询(information retrieval)的不同,IF 关注用户的长线(long term)需求(长线需求是指在一段时间内,比较固定的信息需求).在 IF 中,用户的需求表示成 profile,profile 相当于 IR 中固定的"query",IF 系统根据 profile 对进入系统的文章流进行评价(ranking),同时从用户直接或间接地得到反馈信息,对 profile 进行修改(revising).由于反馈的存在,机器学习的方法在信息过滤中已得到广泛的重视,其中主要的方法有:Bayes 学习方法、神经网络方法、决策树、KNN(K nearest neighbor)、SVM(support vector machine)等[2].

把信息过滤技术用于互联网信息查询是非常重要的研究方向,它对于解决网络信息的个性化、动态化以及提高被查询信息对用户的可用度有很大作用.对信息过滤的研究和实验系统很多.例如,Syskill&Webert[3]通过关键词表达用户信息,利用 Bayes 方法进行过滤;WebWacher[4]根据用户依一定结构设计的目标,帮助用户寻找

---

* 作者卢增祥,1970 年生,博士生,主要研究领域为模式识别,智能系统.关宏超,1977 年生,硕士生,主要研究领域为模式识别,智能系统.李衍达,1936 年生,教授,博士生导师,中国科学院院士,主要研究领域为模式识别,智能系统.

本文通讯联系人:卢增祥,北京 100084,清华大学自动化系

本文 1998-10-27 收到原稿,1999-05-20 收到修改稿

信息;Letizia[5]可以监视用户的浏览路径,获取用户信息需求;WebHound[6]采用社会过滤的方法,将对同样的文章有同样的评价的用户联系在一起,认为他们有同样的信息需求,这样,他们之间就可以共享信息搜索的工作.

提供网络信息过滤服务的关键是获得用户的信息需求和信息过滤(推荐)算法.这种服务时刻监视用户的信息需求,同时监视网络上用户所关心的信息变化,利用智能技术进行信息匹配,及时主动地通知用户.信息过滤的主要技术集中在用户 Profile 的表示以及 Ranking 方法上.现有的信息过滤系统一般利用关键词、规则或分类信息来表达用户需求,对于不同的表示方法可以采用不同的信息匹配算法.例如,对于利用关键词表达的系统,可以利用 Boolean 模型、向量空间模型或概率模型等[5];对于利用分类信息表达的系统,可以利用自动分类的方法等.它仍然需要用户对自己的信息需求进行提炼和概括.对于一般 Internet 用户,如何生成合适的关键词、如何选择相关的类别,还是有一定困难的.这些信息过滤系统基本上采用信息检索的现有技术,在用户的信息表达和信息查准率方面没有多大改进.

在本文中,我们利用用户对文章的评价直接反映用户的信息需求,并通过浏览器的 Bookmark 功能作为用户界面,获取用户信息,同时,向用户进行信息推荐.为提高信息匹配的精度,结合向量空间法,我们提出了最大间距(maximize margin)方法.本文第 1 节介绍在网络上获得用户需求和搜集网络信息的方法.第 2 节讨论信息匹配的算法,介绍利用最大间距法进行过滤的算法,并给出它与传统算法比较的实验结果.第 3 节介绍实现上述方法的实用系统——Bookmark 服务.第 4 节得出结论,并提出进一步的研究方向.

1 获得用户的信息需求

在 Internet 网络上,信息过滤系统所使用的用户信息需求获取方法有很多,主要分成两种:(1) 由用户主动填写;(2) 监视用户的信息搜索与浏览过程.

用户主动设定信息需求可以通过设定关键词(如 informant)或设定主题(如 my-yahoo)来完成,这种方式需要用户事先总结自己的信息需求.由于语言表达的问题和分类的模糊性与多样性,用户往往不能通过这种方法将信息需求表达清楚.另外,因为它要求用户主动填写,所以系统不能主动跟踪用户的兴趣变化.

监视用户的信息查询过程的方法能自动获得用户的信息需求,其方法是在用户的终端上运行一个监视的信息代理(agent),信息代理将用户在 web 浏览时的相关信息不断传送给远端的服务器,服务器将信息进行整理、组织,并从中分析出用户的信息偏好,服务器根据用户的信息偏好进行新的信息的推荐(如 Alexa,Bullseye).由于用户的兴趣是时常变化的,用户的行为信息所反映的用户的信息需求往往是多条线索混合在一起,这给识别信息需求带来了很大的困难.一般需要在用户使用之前预先指定一个主题,但这就增加了用户的负担.

利用用户在浏览器上存储的 Bookmark 获得用户信息需求的方法属于"监视用户"类,它是一种有效的方法.在 Bookmark 中存储的信息往往是用户最关心的,需要记录下来以便以后再读.Bookmark 结构化的信息存储更能确切地表达用户需求.我们可以利用用户对文章的评价来表达用户的信息需求.对存入 Bookmark 的某一推荐目录的文章,可以认为是用户喜欢的文章,作为学习中的正例;在其他目录的文章,或是经过推荐没有被选中的文章将作为学习的反例.根据向量空间法将正反例表示成向量,就可以利用机器学习的方法对新文章进行推荐.

2 信息过滤(推荐)算法

根据用户信息需求表示的不同,所采用的信息推荐算法也有所不同.大多数现有的服务都利用关键词构成用户 profile,可以采用直接匹配方法推荐,也可以利用向量空间法,使用余弦准则推荐.我们利用 Bookmark 的信息进行推荐,将建立新的用户 profile——使用用户对文章的评价信息加以表示.由于一个关键词也可以看成是一篇文章,这种方法可以覆盖利用关键词表示的方法.

方法的实质是根据用户所感兴趣的和不感兴趣的文章来推荐新文章.这方面的研究可以追溯到 IR(information retrieval)中的 Relevance Feedback.在理论方面,关于 Relevance Feedback 的成果很多,如 Rocchio 方法[7]、Bayes 方法等.利用 Bookmark 信息,用户的反馈量可以很大,这为设计更准确的 Ranking 方法提供了条

件.本节在简要介绍已有的研究成果之后,提出最大间距方法.实验表明,最大间距方法有更好的性能.

2.1 传统 Relevance Feedback 方法

传统的 Relevance Feedback 方法有很多[8,9],实用方法集中在对 query 的调整上,其中主要的有两种,一种利用概率模型进行 query 中 term 的权值调整和增减 terms,另一种是利用向量空间法修正 query 方向.

为介绍反馈方法和下文讨论最大间距法,这里简要介绍向量空间法.在向量空间法中,文章要被表示成向量,首先需要进行预处理.对于西文,预处理的步骤为:去掉 stopword,如 the,that 等,而后进行 stemming,如将 played,playing 变为 play;对中文要增加切词的工作.预处理之后,文章变为一个词集(terms).词集中的每个词(term)都需要一个权值,通常是采用词(term)的 TFIDF(term-frequency inverse-document-frequency)加以计算的.在一个给定的文章集中,使用 TFIDF 方法,文章 i 中词 k 的权值由下式计算:

$$dw_{ik}=tf_{ik}*(\log_2(n)-\log_2(df_k+1)). \tag{1}$$

其中 tf_{ik} 为词 k 在文章 i 中的频率,df_k 为包含词 k 的文章数,n 为总文章数.

文章之间的相似度通常是利用向量的余弦来衡量的.设 w_j 为文章 j 的向量,文章 i,j 的相似度 R_{ij} 表示为

$$R_{ij}=\cos(w_i,w_j)=\frac{\sum_k w_{ik}w_{jk}}{|w_i||w_j|}. \tag{2}$$

把 query 依上述方法表示成向量 q,则 R_{iq} 可以用于 Ranking,在此基础上,传统的反馈方法是将对文章的评价信息转移到 query 上.最著名的是 Rocchio 中心向量方法[10],其公式为

$$q_1=q_0+\beta\sum_{k=1}^{n_1}\frac{R_k}{n_1}-\gamma\sum_{k=1}^{n_2}\frac{S_k}{n_2}. \tag{3}$$

其中 q_1 为新的查询命令,q_0 为原始查询命令,R_k 为第 k 篇相关文章的向量,S_k 为第 k 篇不相关文章的向量,n_1 为相关文章数,n_2 为不相关文章数,β 和 γ 为控制参数.实验表明,当 $\beta=0.75,\gamma=0.25$ 时,效果最好.

在概率模型下,给出样本的各项属性值后,可以决定文章 d_j 符合需求 Q 的概率.在独立假设下,可以使用 Bayes 方法,将文章表示成 terms,选择信息熵最大的 n 个 term(A_i),参见文献[3],得出

$$P(Q|d_j)=P(Q|A_1=T_{1j}\&\cdots\&A_n=T_{nj})=P(Q)\prod_k P(A_k=T_{kj}|Q). \tag{4}$$

其中 $P(A_k=T_{kj}|Q)$(在符合需求 Q 的文章中,term(A_k)发生的概率)和 $P(Q)$(在文章集中符合 Q 的文章的概率)可以从训练集合中得到.根据式(4)可以对文章进行 Ranking.

2.2 最大间距法

在传统的 Relevance Feedback 中,反馈信息通过修正 query 来影响 Ranking 的结果.Ranking 是一个影射的过程,将文章空间向用户需求符合度进行一维影射 $f(d)$.在 IF 中,$f(*)$ 由 profile 决定.上述方法都是力求找到一个最合理的影射方法,使相关与不相关的文章尽量分开.我们考虑线性影射

$$R_i=\boldsymbol{w}\cdot\boldsymbol{x}_i+b,\quad i=1,\Lambda,l, \tag{5}$$

其中 w 为影射方向,x 为文章向量,l 为训练样本数(文章数),b 为常数.这样就可以通过对 R_i 排序来实现 Ranking 了.

在训练样本线性可分的情况下,当向量被影射到一维时,两类可以被完全分开,则存在一个唯一的使距离最近的相关与不相关文章相距最远的影射 w^*.我们定义这种距离为两类间的最大间距,使用 w^* 进行 ranking,就形成了最大间距 Ranking 方法.在保证分类信息的情况下,取间距最大的问题可以转化为受限寻优的问题[10],具体算法如下:

目标函数: $\min\{\|\boldsymbol{w}\|^2/2\}$

限制条件: $y_i\cdot(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1,\quad i=1,\Lambda,l$

利用 Lagrange 乘数法得:

目标函数: $L_P\equiv\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{l}\alpha_i y_i(\boldsymbol{x}_i\cdot\boldsymbol{w}+b)+\sum_{i=1}^{l}\alpha_i$ 极小化

由 $\frac{\partial L_P}{\partial w}=0,\frac{\partial L_P}{\partial b}=0$,可以得出:

$$w=\sum_i \alpha_i y_i x_i, \tag{6}$$

$$\sum_i \alpha_i y_i=0. \tag{7}$$

代入后解对偶问题：$L_D=\sum_i \alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i x_j$ 极大化，通过数值解法解出 α_i，则根据式(6)得解. 对于文本分类的情况，由于空间维数很高，大部分为线性可分的情况[2].

对于线性不可分的情况可以通过增加松弛变量 ξ_i 来解决，即限制条件变为

$$y_i\cdot(w\cdot x_i+b)\geqslant 1-\xi_i,\quad i=1,\Lambda,l, \tag{8}$$

优化目标变为

$$\min\left\{\|w\|^2/2+C\left(\sum_i \xi_i\right)\right\}. \tag{9}$$

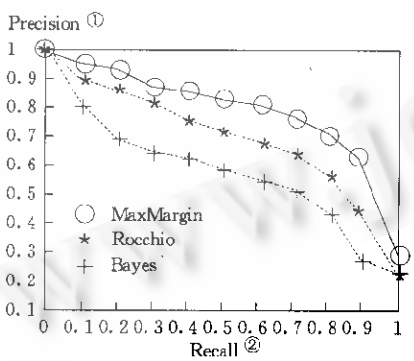
C 为惩罚因子，$C>\alpha_i>0$ 对应的点有效，其余的点删去，再由式(6)得解. 这种方法在使用上的一大好处是不必进行参数的调节. 它符合统计学习理论[11]的思想，因此具有很好的推广能力.

2.3 实验结果

实验使用了 Reuters-21578 数据集*. 有许多利用 Reuters-21578 数据集进行的实验，其中主要利用的切分方式之一为"ModApte"切分，利用其中的 9 603 篇作为测试集，3 299 篇作为反馈评价集. 在这种切分下，选择文章最多的 10 类进行试验. 去掉 stopword 和经过 stemming 后，共有 10 083 个 terms，对最大间距法(以后简称 MaxMargin 法)和 Rocchio 方法，使用全部 terms，利用 tfidf 进行权值调整后再进行归一化，形成了文章的向量. 对 Bayes 方法，选择信息熵最大的 900 个 terms.

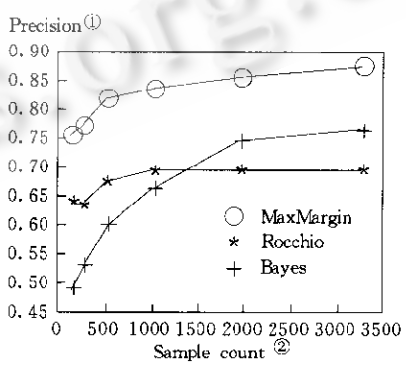
在信息查询中，对所用方法的两个重要的评价准则是查准率(precision)和查全率(recall). 一般情况下，两个指标成反比. 在对算法进行性能比较时，可以利用 precision=recall 的 breakeven 点[2]. 本文也采用此方法.

在分类器设计中，选择评价样本时，保证其中有至少 5 个"正例"和 5 个"反例"；在同等条件下对比 3 种 Ranking 方法. 图 1 是利用 100 个反馈样本的 11 点 P-R 图，数据获得方法为从训练集中随机取 100 个样本，重复 5 次，取平均，再根据各类的样本数取 10 类加权平均. 图 2 给出 3 种方法随反馈量变化的情况，取样本量为 100，200，500，1 000，2 000，3 299，对每种反馈量，从评价集中随机抽取 5 次，对结果取平均，再根据各类的样本数取 10 类平均，纵轴为 breakeven 值，横轴为反馈评价样本数. 实验在 SUN Ultra2 工作站上进行，算法计算时间均在 20 秒以内. 其中，惩罚因子 $C=1\ 000$.



①查准率，②查全率.

Fig. 1
图1



①查准率，②样本量.

Fig. 2
图2

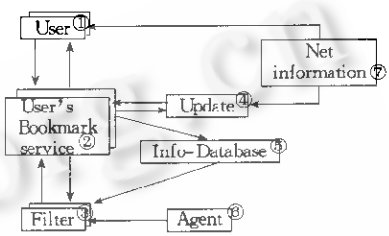
实验表明，最大间距 Ranking 方法与传统方法相比有更好的性能，在同样的条件下，精度可以增加 10%. 无论样本的反馈量大小如何，最大间距方法都能保持良好的性能.

* 数据集由 David Lewis 编辑，它来源于 1987 年路透社新闻. http://www.research.att.com/~lewis/reuters21578.html

3 Bookmark 服务

Bookmark 服务系统是一个网络信息过滤系统.它使用 Bookmark 服务与用户接口,获得用户的需求信息,对用户进行信息推荐;根据 Bookmark 信息,使用最大间距 Ranking 算法来设计信息过滤器;利用信息代理获取网络信息,通过信息过滤器推荐给用户;监视用户 Bookmark 中的网址,对信息进行更新;对用户从网络上直接获得的信息,可以通过 Bookmark 服务增加到信息资源数据库中,并且可以通过过滤器推荐给其他用户.系统框图如图 3 所示.

系统可以在 http://test.au.tsinghua.edu.cn/访问到.在建立服务时,系统中要求大量存储文章信息,我们将文章的文本部分经过处理后进行存储,只存储文章向量.在信息推荐中,虽然计算复杂,但由于不是实时对用户的推荐,所以速度不是主要问题.为了构造 Bookmark 服务,我们也将利用自己的智能信息代理(intelligent agent)从一些主要搜索引擎(Internet search engine)中获得信息,并根据特殊要求设计专门的信息代理(agent),独立地从网络中掘取信息,以此作为主动推荐的素材.



①用户,②用户书签服务,③过滤器,④更新,⑤信息资源数据库,⑥代理,⑦网络信息.

Fig. 3
图3

利用 Bookmark 进行信息过滤有很多优势.用户可以导入自己已有的 Bookmark 文件或直接在浏览网络时通过客户端软件加入的新书签,并且可以对其进行各种编辑操作.用户可以明确地对文章的内容与自己的需求是否一致作出评价,通过反馈信息将使信息推荐更为准确.用户可以灵活地设置推荐的范围,系统对用户信息的推荐信息直接并入 Bookmark 结构中,完全符合用户的习惯.另外,系统使浏览器的 Bookmark 功能实现网络化,即利用网络服务器存储用户的 Bookmark 信息,这样,可以使用户在任何地方、各种平台下,拥有一致的 Bookmark 信息.

4 结 论

通过 Bookmark 服务进行网络信息过滤是有效的方法.利用用户对文章的评价获取用户的信息需求,通过 Bookmark 服务向用户推荐信息是符合用户使用习惯的.它同时提供了一致的 Bookmark 服务,便于用户接受.在利用用户反馈信息进行过滤器的设计中,本文提出的最大间距 Ranking 法有很好的性能.它可以应用在非实时情况下,特别是在长线的信息过滤中.

本文主要考虑基于内容的过滤,在这方面有很多需要进一步研究的问题:(1) 与传统的关键词方法结合,对文章的评价可以推广到对关键词的评价;(2) 在对文章评价时,内容依赖于整篇文章,但是在很多情况下,文章中只有一部分与用户关心的内容有关,这需要系统增加选择部分文章进行推荐的能力;(3) 用户对文章的评价往往是多重的,只用"是""否"标准衡量将损失大量信息,这需要研究带多级评价的 Ranking 方法;(4) 网络信息是多语种的,在利用对文章的评价构造 profile 时,若所评价的文章集固定,对固定文章集进行翻译之后,就可以实现多语种过滤,这将是一个有效的多语种过滤方法.

上述系统所记录的用户的 Bookmark 信息为社会过滤的实现提供了很好的基础.社会过滤是利用用户之间的相似性进行信息推荐的方法.它可以与内容无关,而且能保证信息的质量.这将成为一种非常重要的网络信息过滤方法.

参考文献

1 Belkin N J, Croft W B. Information filtering and information retrieval: two sides of the same coin. Communication of ACM, 1992,35(12):29~38

2 Joachims T. Text categorization with support vector machine. Technical Report. LS VIII Number 23, German: University of Dortmund, 1997

3 Pazzani M, Billsus D. Learning and revising user profiles: the identification of interesting web sites. Machine Learning,

1997,27(3):313～331

4 Armstrong R, Freitag D, Joachims T *et al*. WebWatcher: a learning apprentice for the World Wide Web. In: American Association for Artificial Intelligence ed. Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, Heterogeneous Environments. Cambridge, MA: AAAI Press, 1995. 6～12

5 Lieberman H. Letizia: an agent that assists web browsing. In: AT&T Labs ed. Proceedings of the International Join: Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufman Publishers, 1995. 924～929

6 Lashkari Y. The WebHound personalized document filtering system. 1997, http://rg. media. mit. edu/projects/web-hound/

7 Rocchio J. Relevance feedback in information retrieval. In: Salton G ed. The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice-Hall Inc., 1997. 313～323

8 Harman D. Relevance feedback revisited. In: ACM ed. Proceedings of the SIGIR'92. Copenhagen: ACM Press, 1992

9 Allan J. Incremental relevance feedback for information filtering. In: ACM ed. Proceedings of the SIGIR'96. Copenhagen: ACM Press, 1996. 270～278

10 Cortes C, Vapnik V. Support-Vector networks. Machine Learning, 1995,20(3):273～297

11 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

Network Information Filtering Using Bookmark Service

LU Zeng-xiang GUAN Hong-chao LI Yan-da

(*Department of Automation Tsinghua University Beijing* 100084)

Abstract Information filtering is an important method to alleviate information overload. How to express user's information needs and how to rank them are two main problems in this area. This paper focuses on these two problems. Bookmark service, the solution that is an extension of bookmark function in browser, is used to catch user's information needs. The evaluation of the documents is used to represent user's information needs. Maxine Margin method, a novel ranking algorithm, is designed to increase precision. An experiment proves that it performs better than the traditional methods. A system called Bookmark Service System is designed for implementation. The system construction is given, and its main function is also introduced.

Key words Information filtering, maxine margin method, Bookmark service, Internet, WWW.